applied sciences

MDPI

Article

Alzheimer's Dementia Speech (Audio vs. Text): Multi-Modal Machine Learning at High vs. Low Resolution

Prachee Priyadarshinee *, Christopher Johann Clarke, Jan Melechovsky, Cindy Ming Ying Lin, Balamurali B. T. and Jer-Ming Chen [1]

Science, Mathematics and Technology, Singapore University of Technology and Design, Singapore 487372, Singapore; christopher_clarke@mymail.sutd.edu.sg (C.J.C.); jan_melechovsky@mymail.sutd.edu.sg (J.M.); cindylmy.93@gmail.com (C.M.Y.L.); balamurali_bt@sutd.edu.sg (B.B.T.); jerming_chen@sutd.edu.sg (J.-M.C.)

* Correspondence: prachee@sutd.edu.sg

Abstract: Automated techniques to detect Alzheimer's Dementia through the use of audio recordings of spontaneous speech are now available with varying degrees of reliability. Here, we present a systematic comparison across different modalities, granularities and machine learning models to guide in choosing the most effective tools. Specifically, we present a multi-modal approach (audio and text) for the automatic detection of Alzheimer's Dementia from recordings of spontaneous speech. Sixteen features, including four feature extraction methods (Energy–Time plots, Keg of Text Analytics, Keg of Text Analytics-Extended and Speech to Silence ratio) not previously applied in this context were tested to determine their relative performance. These features encompass two modalities (audio vs. text) at two resolution scales (frame-level vs. file-level). We compared the accuracy resulting from these features and found that text-based classification outperformed audio-based classification with the best performance attaining 88.7%, surpassing other reports to-date relying on the same dataset. For text-based classification in particular, the best file-level feature performed 9.8% better than the frame-level feature. However, when comparing audio-based classification, the best frame-level feature performed 1.4% better than the best file-level feature. This multi-modal multi-model comparison at high- and low-resolution offers insights into which approach is most efficacious, depending on the sampling context. Such a comparison of the accuracy of Alzheimer's Dementia classification using both frame-level and file-level granularities on audio and text modalities of different machine learning models on the same dataset has not been previously addressed. We also demonstrate that the subject's speech captured in short time frames and their dynamics may contain enough inherent information to indicate the presence of dementia. Overall, such a systematic analysis facilitates the identification of Alzheimer's Dementia quickly and non-invasively, potentially leading to more timely interventions and improved patient outcomes.

Keywords: Alzheimer's Dementia; deep learning; text/acoustic analysis; spontaneous speech

Citation: Priyadarshinee, P.; Clarke, C.J.; Melechovsky, J.; Lin, C.M.Y.; B. T., B.; Chen, J.-M. Alzheimer's Dementia Speech (Audio vs. Text): Multi-Modal Machine Learning at High vs. Low Resolution. *Appl. Sci.* **2023**, *13*, 4244. https://doi.org/10.3390/app13074244

Academic Editor: Michael Döllinger

Received: 23 February 2023
Revised: 14 March 2023
Accepted: 19 March 2023
Published: 27 March 2023

1. Introduction

Alzheimer's Dementia (AD) is a neuro-degenerative disease [1], where language and speech decline [2,3] can manifest years before other functions, such as behavior, memory, and sensory–motor skills, are compromised [2,4]. A wide range of speech and language impairments are well-documented and the focus of several review papers (e.g., [5–8], amongst others). Analysing spontaneous speech may afford timely intervention upon diagnosis of AD in its early stages.

With the rise of machine learning (ML) tools in the last decade, automatic speech analysis exploiting acoustic features for the screening, early detection and intervention of AD are becoming more robust and more practical to implement [9–11]. Furthermore, text feature extraction techniques due to breakthroughs in natural language processing (NLP)

tools [12], pretrained transformer-based language models [13,14] and automatic speech recognition (ASR) [15,16] methods have shown promising results in predicting AD.

Accordingly, speech-based approaches for AD prediction rely on acoustic features [16–19], text representations [20–22] or a fusion of both [23,24]. The authors in Ref. [17] compared conventional acoustic features (fundamental frequency, jitter, shimmer, etc.), pretrained acoustic embeddings using wav2vec 2.0 and a combination of both (by simply concatenating the representations from pretrained models and conventional acoustic features) on the ADReSSo-2021 (Alzheimer's Dementia Recognition through Spontaneous Speech only) dataset [25]. In addition, Ref. [17] investigated Logistic Regression (LR), Support Vector Machines (SVM), Neural Networks (NN), and Decision Trees (DT), reporting 67.6% accuracy on the combined representation of conventional acoustic features and pretrained acoustic embeddings (using SVM classifier). ADReSSo-2021 dataset [25] is also chosen for our investigation and comprises of 166 and 71 audio recordings in the training set and testing set, respectively. Using the same dataset, Ref. [16] achieved 78.9% accuracy using pretrained acoustic embeddings. Notably, Ref. [18] studied AD classification using only acoustic features on a DementiaBank dataset [26,27] (473 recordings: 233 speech samples from 97 cognitively normal (CN) and 240 samples from 167 AD patients) to report high accuracies of 94.7%, 92.3% and 90.9% using Bayesian Networks (BN), Meta-Bagging (MB) and Random Forest (RF) classifiers, respectively. On the other hand, Ref. [20] achieved an accuracy of 81.0% by utilizing only linguistic features from textual transcripts of the spontaneous speech (with pretrained transformer-based models) based on the smaller ADReSS-2020 [28] dataset (audio recordings in training set: 108; testing set: 48) [28]. Using only transcripts of this dataset, Ref. [22] achieved an accuracy of 83.3% by using FastText-based classifiers to which bigrams and trigrams were appended to the input transcription (which included discourse markers such as "um" and "uh").

Indeed, incidental comparisons between acoustic-based features and text-based features in dementia classification showed that text-based features performed better than acoustic-based features [24,29–32]. More specifically, Ref. [24] classified AD patients with 85.0% accuracy using language-only features as compared to 65.0% using acoustic-only features on the ADReSS-2020 test dataset. On the same dataset, Ref. [29] achieved 85.4% accuracy with BERT text embeddings (SVM classifier), whereas the acoustic feature accuracy using i-vectors (kNN classifier) was only 56.3%. Based on the ADReSSo-2021 dataset, Ref. [31] achieved the highest published accuracy of 84.5% (to date) using linguistic-only features, compared to 74.6% accuracy of acoustic-only features. On this dataset, however, Ref. [30] reported identical accuracies of 84.5% obtained for linguistic features as well as fusion of acoustic and linguistic features. Based on a subset of ADReSSo-2021 dataset (166 audio recordings), text-based features outperformed audio-based features in Ref. [23] (83.7% accuracy) and Ref. [32] (81.6% accuracy). Some recent studies indicate that a fusion of acoustic and linguistic features can improve AD prediction accuracy [23,33–35]. Ref [23] reported 83.7% accuracy by fusing acoustic features (IS10 paraling, fine-tuned wav2vec 2.0) and deep linguistic features (extracted using fine-tuned BERT) using an SVM classifier on the ADReSSo-2021 dataset. On the same dataset, Ref. [35] reported an accuracy of 81.6% by fusing linguistic and acoustic features. Ref. [36] utilized a C-Attention network trained on a combination of both linguistic features and acoustic embeddings to achieve an accuracy of 80.2%. However, similar to [30], Ref. [34] reported that both text features and fusion features achieved the same accuracy. Further, Ref. [34] reported an accuracy of 81.2% for both text features (Transformer-XL) and bimodal fusion features (ensembled output) on ADReSS-2020 [28] dataset. On this dataset, Ref. [33] achieved 75.0% AD classification accuracy by fusing acoustic-based models and transcript based models, compared to 72.9% and 58% accuracies of unimodal transcript-only and acoustic-only models, respectively. Furthermore, a fusion model investigated by [37] used a BiLSTM model with highway layers using linguistic (words, word probabilities), disfluency features, pause information and a variety of acoustic features to achieve an accuracy of 84.0% based on the ADReSSo-2021 dataset (the focus of this study).

Many deep learning (DL) models are now widely used in speech recognition research because they allow for the implementation of multiple layers that capture information at different granularity levels. Improved prediction accuracy scores have been reported for AD classification tasks using deep learning approaches. Based on the ADReSS-2021 dataset (the focus of our study), Ref. [16] used acoustic embeddings from pretrained models (trill, allosaurus, and wav2vec 2.0), achieving an accuracy of 78.9% using wave2vec 2.0. Using VGGish deep acoustic embeddings (ADReSS-2020 dataset) combined with other feature aggregation methods, such as Fisher Vector encodings (FVs) and Bag-of-Audio-Words (BoAW), Ref. [38] achieved an accuracy of 85.4%. Ref. [39] attained an accuracy of 83.3% using fine-tuned Bidirectional Encoder Representations from Transformers (BERT) based on the ADReSS-2020 dataset. For automatic AD identification from continuous speech, Ref. [40] utilized perplexity characteristics collected from N-gram language models and achieved 84.5% accuracy. Ref. [41] investigated both language features retrieved from the transcripts and encoded pauses and reported 89.6% accuracy by using pretrained language models along with pause encoding using the ADReSS-2020 dataset. Using this dataset, Ref. [42] achieved 83.3% accuracy by utilizing multi-layered perceptrons (MLP) and recurrent neural networks (RNN) while modelling several audio and linguistic characteristics. By combining the text embeddings (both word level and phoneme level) with audio features, Ref. [43] reported AD classification accuracy of 79.1%. They also used phoneme-level embeddings to train deep learning text systems on their small ADReSS-2020 dataset due to concerns of frequent overfitting. Ref. [44] implemented transformer-based models (ALBERT, XLNet, RoBERTa, BioBERT, BioClinical-BERT, ConvBERT, BERT) on the ADReSS-2020 dataset and achieved an accuracy of 87.5% using BERT text embeddings.

Previous studies, however, did not have the opportunity to compare these models on the same dataset. Therefore, to meaningfully assess the efficacy of the various approaches already deployed, here we report a survey comparing multi-modal feature spaces engaging multi-model and multi-feature cues in addressing Alzheimer's Dementia based solely on the same dataset: ADReSSo-2021 [25]. In this study, we offer a rich comparison of sixteen different methods to classify between AD and CN subjects from audio recordings of spontaneous speech. Such an assessment fairly compares the performance of these sixteen models and architectures on the same dataset, for both audio-based features and text-based features on the file-level and frame-level. These have not been reported before.

Accordingly, this study offers the following contributions:

- Provide a systematic comparison of the performance of sixteen features trained on the same dataset, offering insight towards facilitating the implementation of context-based tools.
- Notably, four of these feature extraction methods (Energy–Time plots, Keg of Text Analytics, Keg of Text Analytics-Extended and Speech to Silence ratio) are original approaches, contributing to existing methods.
- Specifically, we systematically compare the performance of both audio-based and text-based modalities at both frame-level and file-level resolution scales based on the same dataset. This rigorous approach allows comparison and insight on how audio-based and text-based modalities compare at different levels of granularity, facilitating accurate and effective AD classification based on the constraints of temporal and signal-to-noise ratio contexts towards applications in practical (clinical) field settings.

2. Experimental Methodology

In our investigation, we use a set of audio recordings in which participants are asked by the interviewer to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [45] provided as part of the ADReSSo-2021 dataset [25]. Participants include both CN subjects and people who have been diagnosed with AD. The Mini-Mental State Exam (MMSE) score was used to distinguish between CN and AD subjects.

2.1. Audio Pre-Processing and Choice of Features

To ensure the quality of our model's training, we aim to restrict the use of data solely to the relevant subject. Hence, all audio associated with the interviewer (including when it overlapped with the subject) was removed while retaining all other audio segments (such as silence and filler words), as these non-speech segments could still contain useful cues. To achieve this, segments corresponding to the interviewer were manually removed using Adobe Audition [46]. Various training features were then extracted in both audio and text domains, at both frame-level or file-level (detailed in Table 1). Text features were extracted from transcriptions of the recorded speech generated using Otter.ai [47], a commercial speech-to-text transcription service.

As speech cues can occur at different levels (individual phonated utterances or the entire string of utterances across the interview), we analyze the data on two levels: frame-level ('granular' or 'frame-by-frame' descriptors) and file-level (across the entire interview recording). Specifically, audio features at the frame-level were obtained by segmenting the audio signal into shorter frames (∼10–25 ms), whereas the audio features at the file-level were computed by utilizing the audio recording of the complete interview (minus the interviewer). Likewise, text features at the frame-level were extracted from words and short phrases (i.e., below the sentence level) of the transcribed speech (relying on a speech-to-text service), rather than employing the complete transcription of each subject's interview, where the latter was utilized in the generation of text features at the file-level. This approach facilitated insight into distinctive and efficacious features from both audio and text data at different levels of granularity.

Table 1. Training features used in our 16 models.

	File-Level	**Frame-Level**
Audio	eGeMAPS Emobase-Large Emobase Speech/Silence Energy-Time Plots	OpenSMILE (Prosody) VGG OpenL3
Text	Keg of Text Analytics Summed Word Embedding Keg of Text Analytics-Extended	Word Embedding BERT, RoBERTa, DistilBERT XLNet

2.2. Audio Feature Extraction

2.2.1. File-Level Audio Features

File-level features (or high-resolution features) were extracted from entire audio files provided (minus the interviewer speech). These include features extracted using various configuration files of the OpenSMILE library (Emobase, Emobase-Large, eGeMAPS) and our original features: Speech/Silence statistics and Energy–Time plots. OpenSMILE [48] (open-source speech and music interpretation by large-scale extraction) is a convenient toolkit which provides acoustic feature extraction and is capable of extracting low-level descriptors (LLD).

- **Emobase**: We extracted 988 acoustic features: 26 LLDs, along with their deltas and their 19 functionals [48,49]. The feature set contains the mel-frequency cepstral coefficients (MFCCs), voice quality, fundamental frequency (f_0), f_0 envelope, line spectral pairs (LSP), intensity features along with their first- and second-order derivatives and several statistical functions applied to these features, resulting in a total of 988 features for every speech segment.
- **Emobase-Large**: Using Emobase-Large [48,49], we extracted 6552 file-level features: 56 LLDs along with their 56 delta for their 56 delta–delta and 39 functionals. Local features or low-level descriptors (MFCCs, pitch, energy, quality of voice, etc.), their

first and second derivatives (i.e., delta and delta–delta) and statistical functionals (global features) were extracted.

- **eGeMAPS**: We extracted 88 file-level features (25 LLDs and their functionals) using the eGeMAPS [50] configuration in the OpenSMILE toolkit. The eGeMAPS configuration offers a reduced fundamental feature set to include 88 features, comprising of f_0 semitone, loudness, spectral flux, MFCCs, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and their most common statistical functionals [48] and can detect physiological changes in voice production.
- **Speech/Silence**: Pauses were identified in the preprocessed audio using a voice activity detection (VAD) algorithm [51]. Speech segments of less than 0.3 s and silence segments of less than 0.2 s are considered false detections and are converted into the opposite type. The feature set consists of nine features. Out of the nine features, five correspond to the number of pauses, which were organized into five "pause bins" (pb). Each pause bin is associated with a duration: <0.5 s (pb1), 0.5–1 s (pb2), 1–2 s (pb3), 2–4 s (pb4) and >4 s (pb5). The sixth feature is the "Centre of Mass of pause bins" (pbcm), where pbcm represents the number of pauses in a bin multiplied by the respective bin number, calculated as a weighted sum of the five bins, as shown in Equation (1).

$$pbcm = 1 * pb1 + 2 * pb2 + ... + 5 * pb5 \quad (1)$$

Furthermore, the remaining three features are "sprat" (speech chunk to pause chunk ratio), where the ratio of lengths of each pair of speech chunks and consequent pause chunks are calculated. This speech–silence subsequent segment pair duration ratio in the first, second (median) and third quartiles across the whole recording (spratq1, spratq2 and spratq3) followed by average duration of speech segments were then calculated. Figure 1 shows a typical distribution of normalized Speech/Silence features analyzed. As seen in this Figure 1, the pause bins of the healthy control (CN) group is always lower than the AD group. As expected, the opposite trend of higher speech chunk to pause chunk ratio was observed for spratq features in the CN group.

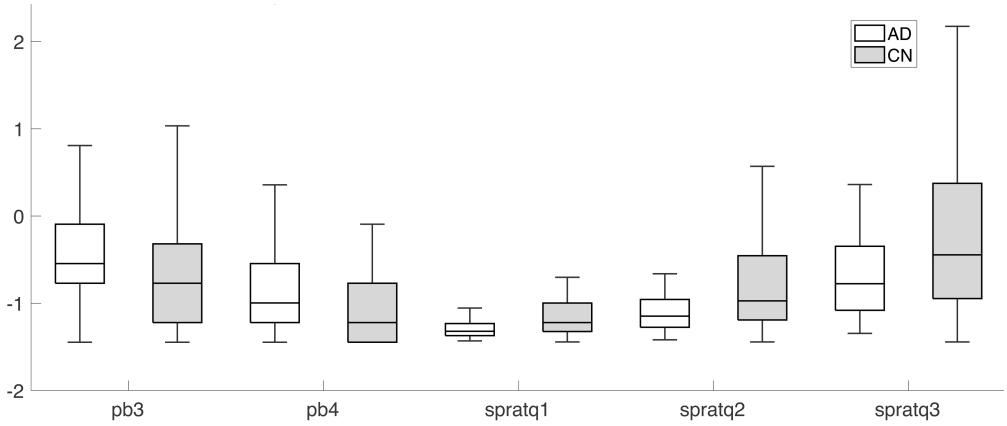

Figure 1. A typical distribution of normalized speech/silence features analyzed for Alzheimer's Dementia (AD) and Control Normal (CN). pb3: pause bin of 1–2 second duration; pb4: pause bin of 2–4 second duration; spratq1, spratq2 and spratq3 are the speech–silence segment ratio in the first, second (median) and third quartiles across the entire audio recording.

- **Energy–Time plots**: Two types of images were generated to represent the time amplitude signal of the segmented audio files: a plot with Cartesian coordinates (x = time, y = absolute value of amplitude) (Figure 2a,b) and a polar plot with time mapped to angle θ and absolute amplitude values mapped to the radii ρ (Figure 2c,d). All the values were normalised to a number between 0 and 1, except for θ which was normalised to an angle between 0 and 2π. The resulting images were then used

to train an image-based model. (A side benefit: because the tangential direction in polar representation represents the temporal domain, the noise floor is consequently diminished in visual weight, thereby making the polar representation somewhat less sensitive to the noise floor.)

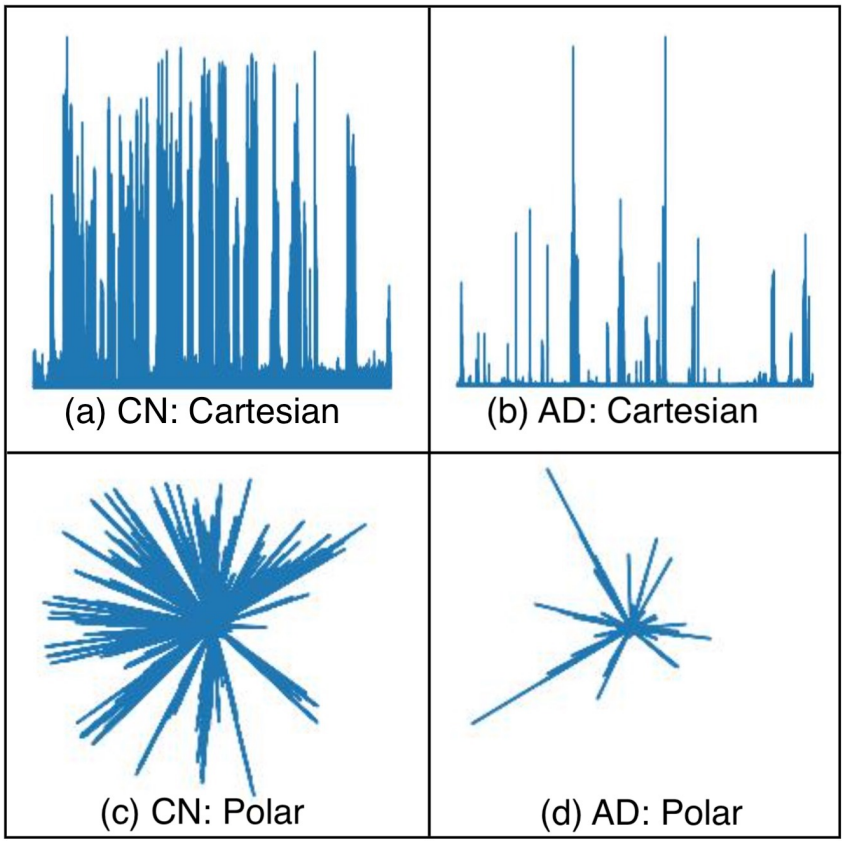

Figure 2. A typical Energy–Time plot is shown. Cartesian representation for (**a**) Control Normal (CN) and (**b**) Alzheimer's Dementia (AD) and polar representation for (**c**) Control Normal (CN) and (**d**) Alzheimer's Dementia (AD) shows rhythmic/metrical representations as different spatial features.

2.2.2. Frame-Level Audio Features

Frame-level features extracted from audio frames include VGG, OpenL3 feature embeddings and features extracted using another configuration file, Prosody, using the OpenSMILE toolkit [48]:

- **VGG**: Semantically meaningful feature embedding extracted from the audio using VGGish resulted in a 128-dimension embedding of the input audio feature (modified mel-spectrogram to a log scale) extracted from an audio frame [52,53].
- **OpenL3**: OpenL3 embeddings were extracted from the audio signal, resulting in a 512-dimension embedding of the input audio feature (mel-spectrogram) extracted from an audio frame [54].
- **Prosody**: Features include f_0, voice probability and PCM loudness, among others [48].

2.3. Text Feature Extraction

Automatic transcription was performed for the manually segmented audio signal using Dropbox integration with Otter.ai [47]. From the generated transcripts, text features were extracted. Unfortunately, the automated transcription excluded instances of hesitation such as uhm, errr, uh, etc., which may in fact contain useful cues for classification. We also note that the transcription accuracy may be compromised by the audio quality. Thus, the reliability of the transcription obtained from Otter.ai then had to be checked manually. Sim-

ilar to the audio feature extraction, both file-level and frame-level features were extracted from the transcript.

2.3.1. File-Level Text Features

Linguistics metrics were extracted from the automatically transcribed text. Features from the BERT family (BERT, DistilBERT, RoBERTa) and XLNet were extracted using the HuggingFace transformer [55] library. Keg of Text Analytics is an original feature.

- **Keg of Text Analytics:** The file-level text features extracted using Matlab's text analytics toolbox include: total number of words (W_t); total number of unique words W_u; unique words normalised (W_u/W_t) (looking out for repeated words or repetitive use of simple words); speech rate in words per seconds (W_t/t); number of words which are not 'stop words' ($W_t - W_s$) ('stop words' are words which can be omitted without losing meaning of the sentence, e.g., 'a', 'the', 'to', 'and'); ratio of number of words with $\geq$4, $\geq$5 and $\geq$6 letters to the number of unique words (although the correlation between longer words and complexity is not constant in English, it aids in filtering out short words); the number of nouns, pronouns, adjectives, verbs, adverbs, auxiliary verbs, ad-positions, coordinating conjunctions, interjections, subordinating conjunctions and determiners and lastly, binary representation of the presence of the word "cookie" (included because many dementia subjects struggle to use the word "cookie" in the "Cookie Theft" picture). Figure 3 shows a typical parts-of-speech comparison for AD and CN. Further, a word cloud of 45 randomly selected AD and CN subjects from training is shown in Figure 4 (words with fewer than 4 characters were ignored in these charts). The difference in word count between CN and AD is obvious (seen in both Figures 3 and 4): the CN group has a larger vocabulary bank than the AD group. Furthermore, CN subjects produce longer words frequently, indicating a possibly higher cognitive capacity. Two sets of file-level text features were investigated: Keg of Text Analytics and Keg of Text Analytics-Extended. In the former, a total of 12 features were used (W_t, W_u, ratio of number of words with more than $\geq$4, $\geq$5 and $\geq$6 letters to the number of unique words, $W_t - W_s$, W_t/t, number of pronouns, nouns, adverbs, adjectives and auxiliary verbs), and in the latter, a super-set with 18 features (containing the former 12 plus 6 additional features: number of ad-positions, coordinating conjunctions, interjections, subordinating conjunctions and determiners and binary representation of presence of "cookie") was used.

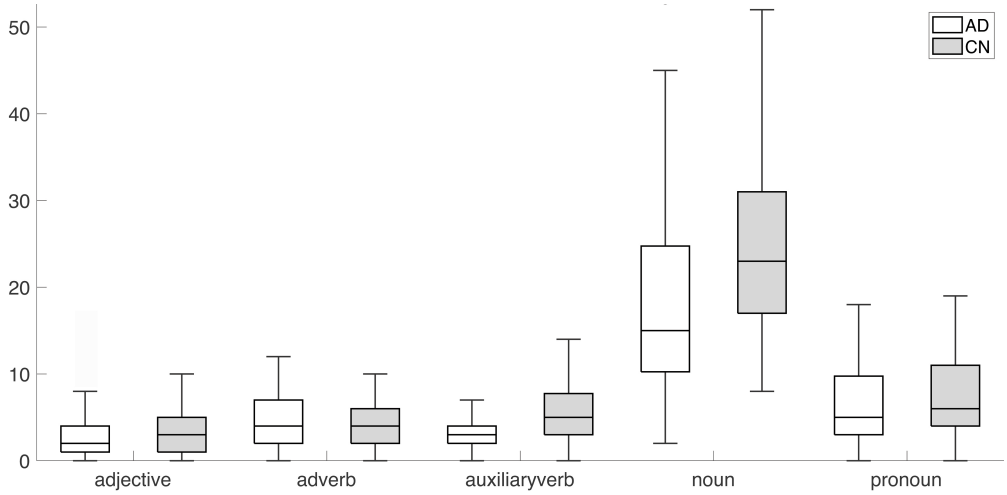

Figure 3. A typical parts-of-speech comparison extracted as part of Keg of Text Analytics. Differences between Alzheimer's Dementia (AD) and Cognitively Normal (CN) subjects can be observed, particularly for nouns.

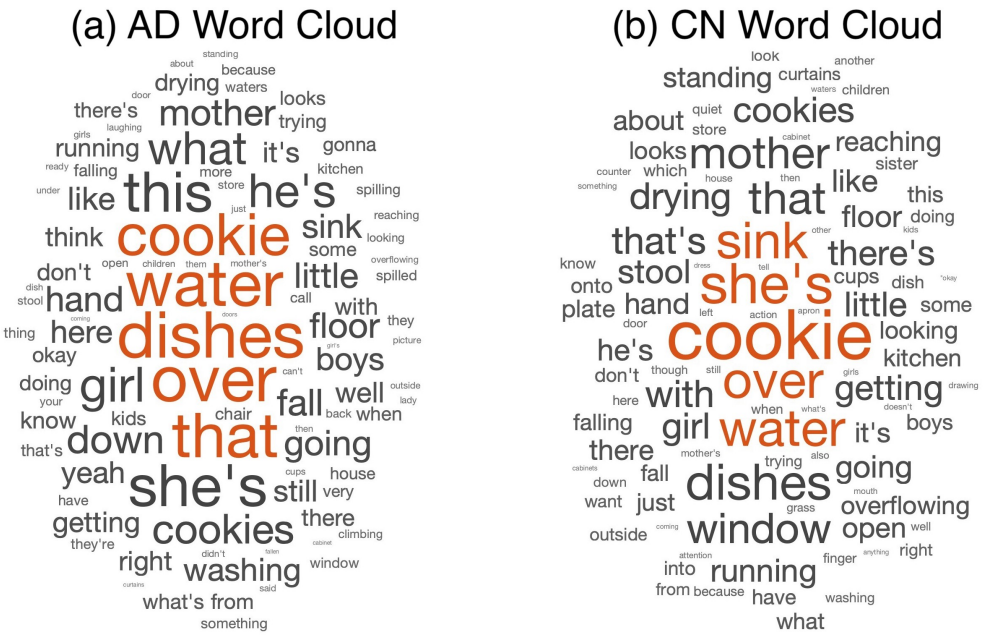

Figure 4. Word cloud of transcribed speech from (**a**) Alzheimer's Dementia (AD) and (**b**) Healthy Control (CN) groups.

- **Summed Word Embedding:** Words from the transcripts are embedded into a vector space model using FastText pretrained word embedding and sentence classification [56,57]. The hyperspace dimension for this embedding was 300, resulting in a vector of 300 elements for each word. The resulting vectors were then summed up to produce an overall representation of the transcription. Although the implication of this sum is not immediately intuitive, our results here suggest that, at least for this dataset, it seems helpful to provide cues for the overall transcript.
- **BERT:** The text transcriptions were used to generate feature embedding vectors from the pretrained BERT (Bidirectional Encoder Representations from Transformers) [14] model. BERT is used for pretraining a language representation over a large amount of unlabeled textual data and employs an attention mechanism (an encoder that reads the text and a decoder that predicts), resulting in deeper sense of language contextual relationships between words in a text.
- **RoBERTa:** RoBERTa (a Robustly optimized BERT approach) [13] uses dynamic masking (unlike BERT's static masking) which increases the data variability (by augmentation) and helps in yielding more robust features.
- **DistilBERT:** DistilBERT is a cheaper, smaller version of BERT [58], which is 60% faster and 40% smaller, while retaining 97% of its performance.
- **XLNet:** XLNet [59] uses an autoregressive pretraining method unlike BERT's autoencoding based pretraining.

2.3.2. Frame-Level Text Features

Word Embedding: This feature extraction is similar to the Summed Word Embedding extraction, except they are not summed; instead the 300 element feature representations of the individual words are analyzed.

2.4. Training Models

This study looked into a variety of machine learning and deep learning models. Table 2 shows the investigated models for the sixteen feature set. The choice of selected models was motivated by the success of these algorithms in supporting fast-prototyping when working with various frame-level or file-level features. Various combinations of classifiers and their hyperparameters were fine tuned for every feature set, and the resulting

training losses were compared. Those with the lowest training loss for a particular feature are reported here. Voting was also used to combine various ensemble models, including Bagging, Gradient Boost, Random Forest, and Adaboost classifiers because the resulting combination outperformed the individual classifiers. In the case of frame-level features, where long-term dependencies and dynamics are present in the sequences of data captured at the frame-level, GMM-UBM and BiLSTM were used. GMM-UBM uses a 512 mix of Gaussians, and the BiLSTM models were only one layer deep with 100 hidden units, except for OpenL3 that was two-layers deep. For deep neural networks used for modelling various text embeddings, a grid search was applied, and hyperparameter optimization was performed to decide on a dense (fully-connected) layer network with either ReLU or Softmax activations. A convolution neural network (CNN), a traditional approach for handling images, was chosen with four convolution layers of 3×3 with max pooling for modelling the Energy–Time plot feature set.

Table 2. Training features and models investigated in this work are shown. AD: Alzheimer's Dementia; CN: Cognitively Normal.

Resolution (Modality)	Feature	Models (Number of Estimators, If Applicable)
file-level (Audio)	eGeMAPS	Bagging Classifier (500)
	Emobase-Large	Hard Voting (Random Forest (1000) + Bagging Classifier (200))
	Emobase	Bagging Classifier (200)
	Speech/Silence *	Catboost (1000)
	Energy–Time plot	CNN
file-level (Text)	Keg of Text Analytics	(500)
	Keg of Text Analytics-Extended	Random Forest (500)
	Summed Word Embedding	Hard Voting (Random Forest (500) + Gradient Boosting (500) + Bagging Classifier (500))
	BERT *	DNN (4-layers)
	DistilBERT *	DNN (5-layers)
	RoBERTa *	DNN (3-layers)
	XLNet *	DNN (3-layers)
frame-level (Audio)	OpenSMILE (Prosody)	GMM-UBM
	VGG	BiLSTM
	OpenL3	BiLSTM
frame-level (Text)	Word Embedding	BiLSTM

* See Appendix A for details of model architecture.

The Summed Word Embedding feature and all the ensemble models, such as Random Forest, Adaboost, Gradient Boost and Bagging Classifiers, demonstrated similar losses, and hard-voting resulted in an even lower loss. For the Speech/Silence feature, SVM outperformed ensembles, whereas CNN appeared to be the best choice to analyze the Energy–Time plots, since the aim was to identify spatial patterns in those plots. The BiLSTM models chosen for frame-level features were only one layer deep with 100 hidden units.

2.5. Train–Test Dataset and Result Evaluation Metric

The ADReSSo-2021 dataset [25] has subjects from the healthy control group (CN cohort) and subjects with cognitive decline (assigned to the AD cohort). Of the 237 audio recordings made available, which were balanced for age and gender (to avoid biases), 166 were used for training, and the models were tested against a mutually exclusive 71 audio recordings. The number of audio recordings in the AD training and testing group were 87 and 35, respectively. The CN training and testing group contained 79 and 36 recordings, respectively.

The results for AD classification task are shown in terms of accuracy (see Equation (2)). We considered the AD class to be the positive one.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (2)$$

The ADReSSo-2021 dataset [25] is a subset of the English Pitt corpus of DementiaBank [26,27], where AD and CN subjects respond to the Cookie Theft stimulus from the Boston Diagnostic Aphasia Examination [45] consisting of the following: average age of AD patients is 69.3 ± 6.9 years; average age of CN subjects is 66.0 ± 6.3 years. A total of 237 audio recordings of the Cookie Theft picture description task is provided, constituting of 122 CN subjects (43 men, 79 women) and 115 patients with AD diagnosis (40 men, 75 women). These recordings also include the interviewer's instructions to the subjects or occasional short prompts. The average duration of interviews of AD and CN groups was 65.7 ± 38.6 and 61.6 ± 26.9 seconds, respectively. To minimize biases, not only was this dataset matched for age and gender, but also carefully matched to avoid commonly overlooked issues, such as repeated occurrences of speech from the same participant, variations in audio quality, etc. The audio was acoustically enhanced by removing stationary noise, and audio volume normalization to control for recording condition variations, such as microphone placement, was also performed [25].

3. Results

All our 16 models performed well above chance (see Table 3), with the poorest performance arising from Speech/Silence with 66.2% accuracy. On the other hand, the best performing model was RoBERTa at an accuracy of 88.7%, followed by DistilBERT at 85.9% and BERT at 84.5%. Looking at the confusion matrix (Figure 5) for RoBERTa, 29 out of 36 CN subjects and 34 out of 35 AD subjects were correctly classified. A comparison between file-level text features and file-level audio features reveals the file-level text features outperform file-level audio features: $78.8 \pm 7.6\%$ vs. $70.9 \pm 4.5\%$, respectively, on average. This observation is true even on a broader scale (overall text vs. overall audio): $78.8 \pm 7.0\%$ vs. $73.2 \pm 4.9\%$, respectively, on average. This agrees with previous observations [24,30,43] that text features contain more distinguishing cues than audio features for identifying AD.

Regarding automated transcription, whose accuracy depends on the overall intelligibility of utterances recorded, CN cases naturally transcribe richer (due to increased verbosity and word variety) and more faithfully (due to better speech clarity) than AD. Transcriptions that yield meaningful strings of words (or not) will result contrastingly in the text feature analysis (see Figure 3), facilitating classification between AD and CN. Accordingly, we indeed observed our best performances at 88.7%, 85.9% and 84.5% when using BERT-related features. While analyzing other text feature sets, the Keg of Text Analytics and Keg of Text Analytics-Extended both resulted in identical accuracies of 76.1%, a drop in performance compared to the BERT-related features (which was associated with more AD misclassifications). Finally, the Summed Word Embedding resulted in greater AD and CN misclassifications due to lower accuracy (73.2%) in comparison with Keg of Text features and BERT-related text features.

Now, frame-level audio features ($77.0 \pm 3.2\%$) performed better than file-level audio features ($70.9 \pm 4.5\%$), suggesting that extended audio recordings need not be aggregated, but rather brief speech audio samples representing intrinsic mechanics and dynamics of speaker idiosyncrasies may be preferred, being also easier to collect and execute. Among the various file-level audio features explored, however, Emobase and Emobase-Large yielded accuracies of 70.4% vs. 77.5%, respectively. In this case, Emobase-Large (consisting of a larger feature set) resulted in lower misclassification for both AD and CN. In contrast, eGeMAPS resulted in 67.6% accuracy, which is associated with greater AD and CN misclassification than Emobase (70.4%).

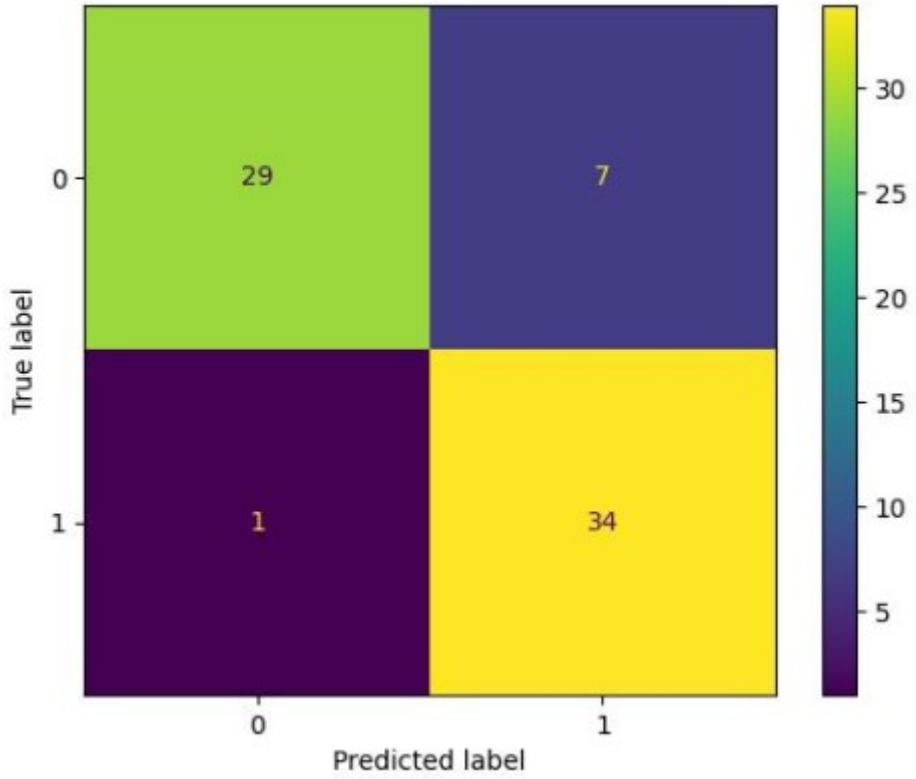

Figure 5. Confusion matrix (0- CN, 1- AD) for the best performing model: RoBERTa.

The best performing text-based feature at the file-level (RoBERTa) performed 9.8% better than the frame-level (Word Embedding). Conversely, the best audio-based feature at the frame-level (VGG) performed 1.4% better than the best file-level audio feature (Emobase-Large). Extracting deep features using VGG's deep neural networks at the frame-level could be a reason for its higher performance than the conventional acoustic features (low-level descriptors and functionals) of extractors, such as Emobase-Large. Mostly, VGG was able to find task-specific features leading to better classification accuracy.

Our four original feature extraction methods proposed—Speech/Silence, Energy–Time plot, Keg of Text Analytics, and Keg of Text Analytics-Extended—achieved accuracies of 66.2%, 73.2%, 76.1% and 76.1%, respectively. Although the accuracies of both Keg of Text Analytics and Keg of Text Analytics-Extended features were 76.1%, the incorrectly classified subjects were not the same. These ab initio results may be currently lower than the well-established features, but with further exploration, we may expect overall improvements in performance.

The pretrained BERT, DistilBERT, RoBERTa and XLNet models used in this study achieved accuracies of 84.5%, 85.9%, 88.7% and 67.6%, respectively (Table 3). These pretrained models performed feature extraction by first processing the transcribed texts followed with a hyperparameter optimization for deep neural network (DNN) models. Appendix A details how the DNN models for BERT, DistilBERT, RoBERTa and XLNet were optimized.

Table 4 compares the best results of this study to state-of-the-art models (at the time of writing) for other AD classification studies based on the same ADReSSo-2021 dataset, surpassing them in accuracy by 4.2%. Consistent with earlier reports of successful pre-trained BERT models in AD classification, in this study, RoBERTa achieved the best accuracy of 88.7%.

Table 3. Training features and their associated performance are shown. Accuracy is presented as a percentage reflecting the total correct predictions out of total predictions. Accuracy = (True Positive+True Negative)/(True Positive+True Negative+False Positive+False Negative). AD class is positive. AD: Alzheimer's Dementia; CN: Normal Control.

Resolution (Modality)	Feature	Accuracy (%) (Correct CN Class/Total CN Class, Correct AD Class/Total AD Class)
file-level (Audio)	eGeMAPS	67.6 (26/36, 22/35)
	Emobase-Large	77.5 (30/36, 25/35)
	Emobase	70.4 (29/36, 21/35)
	Speech/Silence	66.2 (25/36, 22/35)
	Energy-Time plot	73.2 (25/36, 23/35)
file-level (Text)	Keg of Text Analytics	76.1 (30/36, 24/35)
	Keg of Text Analytics-Extended	76.1 (30/36, 24/35)
	Summed Word Embedding	73.2 (26/36, 26/35)
	BERT	84.5 (30/36, 30/35)
	DistilBERT	85.9 (29/36, 32/35)
	RoBERTa	88.7 (29/36, 34/35)
	XLNet	67.6 (25/36, 23/35)
frame-level (Audio)	OpenSMILE (Prosody)	78.9 (28/36, 28/35)
	VGG	78.9 (31/36, 25/35)
	OpenL3	73.2 (22/36, 30/35)
frame-level (Text)	Word Embedding	78.9 (28/36, 28/35)

Table 4. Comparison with other studies reporting from the ADReSSo-2021 dataset.

Reference	Best Performing Modality	Highest Accuracy % (Model)
[36]	Fusion (Audio + Text)	80.2 (C-Attention Network)
[30]	Fusion (Audio + Text)	84.5 (LR)
[16]	Audio	78.9 (DNN)
[37]	Fusion (Audio + Text)	84.0 (BiLSTM)
[31]	Text	84.5 ($BERT_{large}$)
[60]	Fusion (Audio + Text)	83.1 (Ensemble)
[61]	Fusion (Audio + Text)	83.1 (DNN)
[35]	Fusion (Audio + Text)	81.6 (Ensemble)
[25] (Baseline)	Fusion (Audio + Text)	78.8 (Ensemble)
This study	Text	88.7 (DNN)

4. Discussion

This study presents a systematic comparison of various approaches and methods, thus allowing us some insight into the speech information that should be considered depending on the speech sampling context. If the audio quality is amenable to faithful transcription (low background noise, good speaker clarity, monolingual and well-documented accents), our study shows that file-level text is most effective, with the BERT family performing rather satisfactorily at 86.2 $\pm$ 2.2%. However, care must be taken with file-level text as XLNet feature (67.6%) offered the second lowest accuracy among the 16 features studied.

In contrast, if the audio quality is good (high signal-to-noise ratio), but the speaker does not enunciate clearly or perhaps speaks in a way not compatible with automated text transcription (e.g., mixed languages used, poorly documented accent), the results guide us to suggest that frame-level analysis is then preferred. It can be expected that phonatory dynamics captured in short time frames may contain enough inherent information about the speaker's psycho-motor health. Further, such an approach relying on using audio feature sets alone has the added benefit of not being limited to a particular carrier language since it should reflect the speaker's cognitive/physiological state without the extra burden

and process of transcription pigeon-holed in a particular carrier language (which may introduce transcription errors compromising the efficacy of a text-based approach).

Frame-level features, such as sub-word or phoneme-level representations, are suitable for training on small datasets and have been shown to have better dementia classification accuracy than file-level features (such as word or sentence level) [43]. On the other hand, file-level features have the benefit of larger temporal samples and increased opportunities to rely on cues that arise randomly or infrequently [41].

Overall, text modality performs better on average (file-level + frame-level) than audio. This may suggest that with the onset of AD cognitive decline affects linguistic capacity more sensitively and may be apparent earlier than physiological decline. Given that factors, such as semantic distance, lexical resource and grammar structure, may become simpler as cognitive decline progresses, text-based methods could be more useful and provide greater insight into differentiating AD from CN, resulting in higher detection accuracy. However, text-based approaches require faithful and robust transcription processing as an additional step, complicating the analysis. Furthermore, many transcription services currently available automatically omit non-semantic utterances (such as *mmm*, *uh*, *um* and *ah*) which may in fact further enhance detection accuracy [41,62].

Most of the best performing AD classification approaches investigating the ADReSSo-2021 dataset [25] are trained on either text features alone or a fusion of features, such as text and audio [25,37], text and disfluency [60] and text and pauses [61], among others. Using solely text, Ref. [31] achieved an accuracy of 84.5%. On the other hand, several studies [25,30,35,37,60,61] also reported good accuracies by using fusion approaches, such as integrating audio and text data, with accuracies ranging from 78.8% to 84.5%. An audio-only study by [16] reported 78.9% accuracy using wave2vec 2.0 embeddings, which is comparable to our best-performing audio-only model accuracy of 78.9% using VGG and Prosody features. Our best performing approach using DNN was trained on RoBERTa text embeddings and offers comparable AD detection accuracy scores while still outperforming the existing state-of-the-art approaches on the ADReSSo-2021 dataset. While more approaches on other datasets exist in the literature, comparing them with this study will not offer a fair comparison. In our study, fine-tuned pretrained RoBERTa embeddings (using automatically transcribed text) achieved the highest accuracy (to date) score of 88.7%, offering a 4.2% improvement over Ref. [30] and Ref. [31] (both reported 84.5% accuracy). Thus, although on prima facie, our study may seem to suggest that text-based methods outperform audio-based methods, the successful implementation of text-based methods depends sensitively on the context in question (such as carrier language, choice and availability of reliable transcription, among others). Audio-based methods, on the other hand, will offer greater flexibility less sensitively to the context. As long as a good signal-to-noise ratio is achieved in recording the speaker's utterances, analysis can still proceed with some reliability, agnostic to the carrier language(s) involved. Lastly, depending again on the audio context and linguistic requirements, a simple voting- and/or fusion-model can then be implemented to reconcile the mutli-modal multi-model multi-feature approach to yield an optimal AD/CN classification outcome. However, this approach was not tested in our investigation due to the absence of a validation dataset. Thus, although text generally has better accuracy than audio, it is limited by the reliability of transcription services. Audio features, on the other hand, would not be constrained by the limitations of carrier/target language per se.

It should also be noted that the performance of the various models in this study is limited inherently by the reliability of the provided clinical labels and the quality and consistency of audio recordings provided in the dataset. In addition, in the absence of a validation set, we are unable to meaningfully fuse the results of this study further.

5. Conclusions

An approach for assessing audio recordings of spontaneous speech utterances related to Alzheimer's Dementia, utilizing a multi-model machine learning strategy, is presented.

The study analysed the effectiveness of various features extracted from audio recording and text derived from the audio, employing 16 different models to assess their relative performance. All models trained on various distinct feature sets demonstrated accuracy above chance levels. Overall, text-based features extracted from the transcribed audio outperformed the audio-based features. The best performing model was achieved by modelling the RoBERTa text embeddings, which attained an accuracy of 88.7%, achieving near-perfect classification for AD, identifying 34 out of 35 cases. This represents a 4.2% improvement surpassing other state-of-the-art models for AD classification trained on the same ADReSSo-2021 dataset.

When comparing the level of granularity (resolution) in the feature extraction process, frame-level audio features generally outperform their file-level counterparts. However, as the number of features increase, such as with Emobase-Large, the results become more comparable. This suggests that a sufficient number of low-level descriptors can effectively represent both AD and CN classes. In contrast, while file-level text embeddings have produced higher accuracy than frame-level word embeddings, this trend cannot be generalized since only one model was tested at the latter level.

The four original feature extraction methods we proposed, namely Speech/Silence, Energy–Time plot, Keg of Text Analytics and Keg of Text Analytics-Extended, demonstrated reasonable accuracy. The image-based Energy–Time plots may offer a new and promising avenue of dementia detection and invites further investigation. While their performance may not yet match off-the-shelf feature sets, additional exploration and fine-tuning of these four original feature extraction methods could lead to further improvements.

This investigation suggests that the transcribed textual data can produce meaningful word sequences that lead to contrasting results in text feature analysis. This is particularly helpful in classifying AD and CN patients, as factors such as semantic distance, vocabulary usage and grammar structures tend to be simpler for dementia patients. While transcribing English audio datasets is relatively straightforward, transcription resources for other languages may not be as easily available nor as mature and robust, making the text-based approach sensitive to the linguistic context of the target speaker. On the other hand, in an audio-based approach, data are not constrained by carrier language, making it more generalizable and useful for developing AD classification models that are more universal, as long as good signal-to-noise ratios of the speech utterances are captured. In practical terms, however, it is expected that the optimal solution is to implement a combination of both approaches to best reconcile differences arising from the context.

This study provides insight into the effectiveness of machine learning techniques trained on both textual and audio data for a given dataset, which is an improvement over previous studies that focused on only one (or one of the many) aspect(s) of the feature space or model selection. Insights from this study provide a fast and non-invasive means of reliable screening for Alzheimer's Dementia, assessing its severity and monitoring its progression. Profoundly, outcomes from the frame-level audio features hint at its potential to be generalizable across languages.

Author Contributions: Conceptualization, P.P., B.B.T., C.J.C. and J.-M.C.; methodology, P.P., B.B.T., C.J.C., J.-M.C., C.M.Y.L. and J.M.; software, P.P., B.B.T., C.J.C., C.M.Y.L. and J.M.; validation, P.P., B.B.T., C.J.C. and J.-M.C.; investigation, P.P., B.B.T., C.J.C., J.-M.C., C.M.Y.L. and J.M.; data curation, P.P., B.B.T., C.J.C. and J.-M.C.; writing—original draft preparation, P.P., B.B.T., C.J.C., J.-M.C. and J.M.; writing—review and editing, P.P., B.B.T., C.J.C. and J.-M.C.; visualization, P.P., B.B.T. and C.J.C.; supervision, J.-M.C.; funding acquisition, J.-M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SUTD Growth Plan (SGP) Grant to Healthcare Sector (PIE-SGP-HC-2019-01).

Data Availability Statement: The dataset is free and made available by Dementia Bank Pitt Corpus https://luzs.gitlab.io/adresso-2021/, accessed on 15 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Dementia
CN	Cognitively Normal
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
ASR	Automatic Speech Recognition
LR	Logistic Regression
MLP	Multi-Layered Perceptrons
BERT	Bidirectional Encoder Representations from Transformers
MMSE	Mini-Mental State Exam
LLD	Low Level Descriptors
GMM-UBM	Gaussian Mixture Model-Universal Background Model
NN	Neural Networks
DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
VAD	Voice Activity Detection
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Networks
BN	Bayesian Networks
RF	Random Forest
GB	Gradient Boosting
DT	Decision Trees
SVM	Support Vector Machines

Appendix A

BERT*: The DNN model used was four layers deep. Units used in the dense first, second, third and fourth layers are 103, 76, 65 and 105, respectively. The associated activation functions used were Softmax, ReLu, Softmax and ReLu, respectively. Adam optimizer with a learning rate of 0.01 was used.

RoBERTa*: The DNN model used was three layers deep. The units in the dense first, second and third layers were 65, 154 and 224, respectively. The activation functions used for these layers were Softmax, ReLu and Softmax, respectively. Adam optimizer was used with a learning rate of 0.023.

DistilBERT*: The DNN model used for DistilBERT was five layers deep. The units used in the dense first, second, third, fourth and fifth layers were 120, 29, 213, 98 and 152, respectively. The activation functions were ReLu, Softmax, Softmax, Softmax and ReLu, respectively. Adam optimizer with a learning rate of 0.075 was used.

XLNet*: Three layers were used in the DNN model. The units used in the dense first, second and third layers were 103, 137 and 73, respectively. The activation functions used for these three layers were Softmax, Relu and Relu, respectively. Adam optimizer with a learning rate of 0.05 was used.

Speech/Silence*: A 20-fold cross-validation and a grid search were conducted across learning rates and number of PCA components. The final model's learning rate was 0.05, and the PCA dimension was 6. The classifier used was Catboost.

References

1. Brookmeyer, R.; Johnson, E.; Ziegler-Grahamm, K.; Arrighi, H.M. O1–02–01: Forecasting the global prevalence and burden of Alzheimer's disease. *Alzheimer Dement.* **2007**, *3*, S168. [CrossRef]
2. Blair, M.; Marczinski, C.A.; Davis-Faroque, N.; Kertesz, A. A longitudinal study of language decline in Alzheimer's disease and frontotemporal dementia. *J. Int. Neuropsychol. Soc.* **2007**, *13*, 237–245. [CrossRef] [PubMed]
3. Meilán, J.J.; Martínez-Sánchez, F.; Carro, J.; Sánchez, J.A.; Pérez, E. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *Span. J. Psychol.* **2012**, *15*, 487. [CrossRef]

4. Murdoch, B.E.; Chenery, H.J.; Wilks, V.; Boyle, R.S. Language disorders in dementia of the Alzheimer type. *Brain Lang.* **1987**, *31*, 122–137. [CrossRef] [PubMed]
5. Klimova, B.; Kuca, K. Speech and language impairments in dementia. *J. Appl. Biomed.* **2016**, *14*, 97–103. [CrossRef]
6. Geraudie, A.; Battista, P.; García, A.M.; Allen, I.E.; Miller, Z.A.; Gorno-Tempini, M.L.; Montembeault, M. Speech and language impairments in behavioral variant frontotemporal dementia: A systematic review. *Neurosci. Biobehav. Rev.* **2021**, *131*, 1076–1095. [CrossRef] [PubMed]
7. Swan, K.; Hopper, M.; Wenke, R.; Jackson, C.; Till, T.; Conway, E. Speech-language pathologist interventions for communication in moderate–severe dementia: A systematic review. *Am. J. -Speech-Lang. Pathol.* **2018**, *27*, 836–852. [CrossRef]
8. Heuer, S.; Willer, R. How is quality of life assessed in people with dementia? A systematic literature review and a primer for speech-language pathologists. *Am. J. -Speech-Lang. Pathol.* **2020**, *29*, 1702–1715. [CrossRef]
9. Pulido, M.L.B.; Hernández, J.B.A.; Ballester, M.Á.F.; González, C.M.T.; Mekyska, J.; Smékal, Z. Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.* **2020**, *150*, 113213. [CrossRef]
10. Petti, U.; Baker, S.; Korhonen, A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1784–1797. [CrossRef]
11. Yang, Q.; Li, X.; Ding, X.; Xu, F.; Ling, Z. Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimers Res. Ther.* **2022**, *14*, 1–16. [CrossRef]
12. Amini, S.; Hao, B.; Zhang, L.; Song, M.; Gupta, A.; Karjadi, C.; Kolachalama, V.B.; Au, R.; Paschalidis, I.C. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. *Alzheimers Dement.* **2022**, *19* , 946–955. [CrossRef] [PubMed]
13. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805
15. Lopez-de Ipiña, K.; Alonso, J.B.; Solé-Casals, J.; Barroso, N.; Faundez-Zanuy, M.; Ecay-Torres, M.; Travieso, C.M.; Ezeiza, A.; Estanga, A. Alzheimer disease diagnosis based on automatic spontaneous speech analysis. In Proceedings of the 4th International Joint Conference on Computational Intelligence, Barcelona, Spain, 5–7 October 2012; pp. 698–705.
16. Gauder, L.; Pepino, L.; Ferrer, L.; Riera, P. Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3795–3799.
17. Balagopalan, A.; Novikova, J. Comparing acoustic-based approaches for alzheimer's disease detection. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3800–3804.
18. Al-Hameed, S.; Benaissa, M.; Christensen, H. Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), San Francisco, CA, USA, 13 September 2016; pp. 32–36.
19. Meghanani, A.; Anoop, C.; Ramakrishnan, A. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Virtual, 19–22 January 2021; pp. 670–677.
20. Searle, T.; Ibrahim, Z.; Dobson, R. Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2192–2196.
21. Syed, Z.S.; Syed, M.S.S.; Lech, M.; Pirogova, E. Automated recognition of Alzheimer's dementia using bag-of-deep-features and model ensembling. *IEEE Access* **2021**, *9*, 88377–88390. [CrossRef]
22. Meghanani, A.; Anoop, C.; Ramakrishnan, A.G. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fastText and cnn models. *Front. Comput. Sci.* **2021**, *3*, 624558. [CrossRef]
23. Ying, Y.; Yang, T.; Zhou, H. Multimodal fusion for alzheimer's disease recognition. *Appl. Intell.* **2022**. [CrossRef]
24. Shah, Z.; Sawalha, J.; Tasnim, M.; Qi, S.a.; Stroulia, E.; Greiner, R. Learning language and acoustic models for identifying Alzheimer's dementia from speech. *Front. Comput. Sci.* **2021**, *3*, 624659. [CrossRef]
25. Luz, S.; Haider, F.; de la Fuente, S.; Fromm, D.; MacWhinney, B. Detecting cognitive decline using speech only: The adresso challenge. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3780–3784.
26. DementiaBank English Pitt Corpus. Available online: https://dementia.talkbank.org/access/English/Pitt.html (accessed on 15 January 2023).
27. Becker, J.T.; Boiler, F.; Lopez, O.L.; Saxton, J.; McGonigle, K.L. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Arch. Neurol.* **1994**, *51*, 585–594. [CrossRef]
28. Luz, S.; Haider, F.; de la Fuente, S.; Fromm, D.; MacWhinney, B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2172–2176.

29. Haulcy, R.; Glass, J. Classifying Alzheimer's disease using audio and text-based representations of speech. *Front. Psychol.* **2021**, *11*, 624137. [CrossRef]
30. Pappagari, R.; Cho, J.; Joshi, S.; Moro-Velázquez, L.; Zelasko, P.; Villalba, J.; Dehak, N. Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3825–3829.
31. Pan, Y.; Mirheidari, B.; Harris, J.M.; Thompson, J.C.; Jones, M.; Snowden, J.S.; Blackburn, D.; Christensen, H. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3810–3814.
32. Clarke, C.J.; Melechovsky, J.; Lin, C.M.Y.; Priyadarshinee, P.; Balamurali, B.; Chen, J.M.; Kapoor, S.; Aharonov, O. Addressing multi-modal multi-model multi-feature cues in Alzheimer's Dementia: The ADReSSo Challenge. In Proceedings of International Congress on Sound & Vibration (ICSV28) 2022, Singapore, 25–27 July 2022. Available online: https://www.researchgate.net/publication/365683202_Addressing_multi-modal_multi-model_multi-feature_cues_in_Alzheimer%27s_Dementia_the_ADReSSo_Challenge (accessed on 10 March 2023).
33. Pappagari, R.; Cho, J.; Moro-Velazquez, L.; Dehak, N. Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2177–2181.
34. Koo, J.; Lee, J.H.; Pyo, J.; Jo, Y.; Lee, K. Exploiting Multi-Modal Features From Pre-trained Networks for Alzheimer's Dementia Recognition. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2217–2221.
35. Chen, J.; Ye, J.; Tang, F.; Zhou, J. Automatic detection of alzheimer's disease using spontaneous speech only. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3830–3834.
36. Wang, N.; Cao, Y.; Hao, S.; Shao, Z.; Subbalakshmi, K. Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3835–3839.
37. Rohanian, M.; Hough, J.; Purver, M. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3820–3824.
38. Syed, M.S.S.; Syed, Z.S.; Lech, M.; Pirogova, E. Automated Screening for Alzheimer's Dementia through Spontaneous Speech. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2222–2226.
39. Balagopalan, A.; Eyre, B.; Rudzicz, F.; Novikova, J. To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2167–2171.
40. Guo, Z.; Ling, Z.; Li, Y. Detecting Alzheimer's disease from continuous speech using language models. *J. Alzheimers Dis.* **2019**, *70*, 1163–1174. [CrossRef] [PubMed]
41. Yuan, J.; Bian, Y.; Cai, X.; Huang, J.; Ye, Z.; Church, K. Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2162–2166.
42. Sarawgi, U.; Zulfikar, W.; Soliman, N.; Maes, P. Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2212–2216.
43. Edwards, E.; Dognin, C.; Bollepalli, B.; Singh, M.K.; Analytics, V. Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Shanghai, China, 25–29 October 2020; pp. 2197–2201.
44. Ilias, L.; Askounis, D. Explainable identification of dementia from transcripts using transformer networks. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4153–4164. [CrossRef]
45. Goodglass, H.; Kaplan, E.; Weintraub, S. *BDAE: The Boston Diagnostic Aphasia Examination*; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2001.
46. Adobe Audition-version 23.0. Available online: https://www.adobe.com/products/audition.html. (accessed on 30 December 2022).
47. Otter.ai. Available online: https://otter.ai/login (accessed on 21 March 2021).
48. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
49. Parlak, C.; Diri, B. Emotion recognition from the human voice. In Proceedings of the 2013 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, Turkey, 24–26 April 2013; pp. 1–4.

50. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
51. Brookes, M. Voicebox: Speech Processing Toolbox for Matlab. Software 1997. Available online: www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (accessed on 30 January 2023).
52. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
53. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
54. Cramer, J.; Wu, H.H.; Salamon, J.; Bello, J.P. Look, listen, and learn more: Design choices for deep audio embeddings. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3852–3856.
55. Transformers — Transformers 3.3.0 Documentation–Hugging Face. Available online: https://huggingface.co/transformers/v3.3.0/index.html (accessed on 30 December 2022).
56. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning word vectors for 157 languages. *arXiv* **2018**, arXiv:1802.06893.
57. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759.
58. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
59. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLnet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32* .
60. Qiao, Y.; Yin, X.; Wiechmann, D.; Kerz, E. Alzheimer's Disease Detection from Spontaneous Speech through Combining Linguistic Complexity and (Dis) Fluency Features with Pretrained Language Models. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3805–3809.
61. Zhu, Y.; Obyat, A.; Liang, X.; Batsis, J.A.; Roth, R.M. WavBERT: Exploiting Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia Detection. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czech Republic, 30 August–3 September 2021; pp. 3790–3794.
62. Davis, B.H.; Maclagan, M.A. UH as a pragmatic marker in dementia discourse. *J. Pragmat.* **2020**, *156*, 83–99. [CrossRef]