# Alzheimer's Dementia Detection Using Acoustic & Linguistic Features and Pre-trained BERT

Akshay Valsaraj, Ithihas Madala, Nikhil Garg, Veeky Baths

**HAL Id: hal-03666188**
**https://hal.science/hal-03666188**

Submitted on 12 May 2022

# Alzheimer's Dementia Detection using Acoustic & Linguistic features and Pre-Trained BERT

Akshay Valsaraj
*Cognitive Neuroscience Lab*
*BITS Pilani, K.K. Birla Goa Campus*
Goa, India
f20180608@goa.bits-pilani.ac.in

Ithihas Madala
*Cognitive Neuroscience Lab*
*BITS Pilani, K.K. Birla Goa Campus*
Goa, India
f20180607@goa.bits-pilani.ac.in

Nikhil Garg
*IEMN*
*Université de Lille*
59650 Villeneuve d'Ascq, France
Nikhil.Garg@Usherbrooke.ca

Veeky Baths
*Cognitive Neuroscience Lab*
*BITS Pilani, K.K. Birla Goa Campus*
Goa, India
veeky@goa.bits-pilani.ac.in

*Abstract*—**Alzheimer's disease is a fatal progressive brain disorder that worsens with time. It is high time we have inexpensive and quick clinical diagnostic techniques for early detection and care. In previous studies, various Machine Learning techniques and Pre-trained Deep Learning models have been used in conjunction with the extraction of various acoustic and linguistic features. Our study focuses on three models for the classification task in the ADReSS (The Alzheimer's Dementia Recognition through Spontaneous Speech) 2021 Challenge. We use the well-balanced dataset provided by the ADReSS Challenge for training and validating our models. Model 1 uses various acoustic features from the eGeMAPs feature-set, Model 2 uses various linguistic features that we generated from auto-generated transcripts and Model 3 uses the auto-generated transcripts directly to extract features using a Pre-trained BERT and TF-IDF. These models are described in detail in the models section.**

*Index Terms*—**Speech Recognition, Human-Computer-Interaction, Computational Para-linguistics**

## I. Introduction

Alzheimer's Dementia (AD) is a progressive Neuro-degenerative disorder that affects major brain functions. The usual manifestations of AD include visible changes in memory and language and cognitive function. It also leads to personality changes along with agitation and losing the train of thought while speaking. Unlike other illnesses, the onset of visible symptoms is not observed in the initial stages. Early detection of this disease can prevent further degeneration, improve life quality and life expectancy.

Hence, AD is a major health concern in the struggle for increased life expectancy in the general population. Initially, it was considered to be a rare disease but soon it became closely associated with aging. There is a lot of stigma attached with this disease but with increased research and new treatments for patients with AD, these misconceptions are slowly fading away. Hence, it is the need of the hour to come up with inexpensive, quick and accurate diagnostic tools and early detection methods to provide utmost care to people suffering from early stages of the disease.

Current methods of diagnosis are human intensive, lab intensive and hence expensive and time consuming. While traditional Brain Imaging techniques are effective in detecting the onset of disease, these are expensive and tedious to repeatedly perform on a patient or a healthy individual. Microphones are present in most laptops and smartphones, and most people possess at least one of the two. Thus, speech data could be easily collected from any healthy individual without requiring any medical intervention. This spontaneous speech data could be used for early detection of AD when setup in the form of an interview taken by a virtual assistant and recording the responses.

The comparison of previously proposed models (ADReSS 2021 Challenge) for the classification task is summarized in Table I. This includes the features extracted, ML/DL models used and Accuracy achieved on the test dataset.

In our study, we discuss three different models based on the input features. Model 1: The Acoustic Model, uses various acoustic features from the eGeMAPs feature-set, Model 2: The Lingustic Model, uses various linguistic features that we generated from auto-generated transcripts and Model 3: Pre-trained BERT model + TF-IDF, uses the auto-generated transcripts directly to extract features using a Pre-trained BERT and TF-IDF.

We also propose the implementation of models based on Vision Transformers (ViT) and sPCEN normalized spectrograms as future work.

## II. Dataset

We use the ADReSS Challenge Dataset which consists of 242 audio files. These were split into training and test sets with 70% for training and 30% for testing and validation. A set

| Input features | Models | Accuracy | Rf |
|---|---|---|---|
| Extracted voice features (acoustic) and transcripts | LDA,DT, INN,SVM, RF | 75% | [1] |
| Audio , TF-IDF and readability features, embeddings. | Xgboost, SVM, RF and LR | 77.8% | [2] |
| Fusion of Acoustic and silence features with transcripts | x-vectors , BERT | 75% | [3] |
| Acoustic and lexical information, including disfluency tagging | LSTM with Gating | 79% | [4] |
| Phonemes and Audio | FastText supervised classifier | 79.1% | [5] |
| Textual Transcriptions | PAR+INV / DistilBERT | 81% | [6] |
| Speaker Representation Vectors and Textual Feature Embeddings | pre-trained x-vectors SRE and LSTM-RNNs | 81.25% | [7] |
| Chat Transcipts , Acoustic and Semantic features | Pre-trained BERT and SVM | 83.3% | [8] |
| Audio Features and Transcripts | Bag-of-Audio-Words, Siamese and Attention Network | 85.2 % | [9] |
| Chat Transcripts | Pre-trained BERT and ERNIE models | 89.6% | [10] |
| Disfluency, acoustic and Interventions features | MLP ,RNN and ensemble models | 83% | [11] |
| Pre-trained language model and linguistic features. | (bi-LSTM) and CNN | 81.25% | [12] |
| Acoustic features and Speech Transcripts | SVM + Bert features | 85.42% | [13] |

of recordings of picture descriptions produced by cognitively normal subjects and patients with an AD diagnosis, who were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. The data set for the task was created from a study involving AD patients. The task involved classifying patients into 'ad' (suffering from Alzheimer's) and 'cn' (control subjects). From the data set provided we removed the voice of the interviewer and kept only the voice of the patient using the utterance segmentation files (diarization) provided with the dataset.

## III. TRANSCRIPTS

For our linguistic model, we require auto-generated transcripts of each audio file. We used the PyTorch based pre-trained enterprise-grade Silero STT (Speech-To-Text) English (en_v2) model for its ease of use & integration with our code and it is also comparable to the quality of Google's STT according to the benchmarks.

## IV. FEATURE EXTRACTION

We calculated various linguistic features such as Brunet's Index, Honore's Statistic, Standardised Entropy, Root Type-Token Ratio and Hyper Geometric Distribution Diversity for our linguistic model. We extracted the eGeMAPS acoustic feature-set for our Acoustic Model. These features have shown to be very effective in previous studies related to Alzheimer's Dementia.

### A. Linguistic features

Using the auto-generated transcripts, we decided to extract the following linguistic features based on extensive literature review.

*1) Brunet's Index:* Its a measure of lexical diversity that has been used in analysis of text and its often considered to be not dependent on text length. Its calculated by using the formula:

$$W = N^{V^{-a}} \qquad (1)$$

where $N$ is the text length, $V$ is the number of different words and $a$ is a scaling constant which is usually set at $0.172$ The richer the vocabulary the lower the values of $W$ and the values usually range between $10$ and $20$. Patients suffering from dementia tend to repeat their words which leads to a lower lexical diversity, hence an important feature to consider.

*2) Honore's Statistic:* This feature has been used in the stylometric analyses of texts. It is based on the fact that texts which have a higher proportion of words tend to have a richer vocabulary. Its calculated by using the formula:

$$R = \frac{100 \log (N)}{\left(1 - \frac{v_1}{V}\right)} \qquad (2)$$

where $v_1$ is the number of words used only one, $V$ is the number of different words and $N$ is the text length. Higher values of R correlate to a richer vocabulary and lower values correlate to poorer vocabulary. Patients suffering from dementia tend to have depleting vocabulary, hence an important feature to consider.

*3) Standardised Entropy:* This measure helps us to check the degradation in the variety of words and word combinations that a patient uses. It is calculated by dividing the word entropy by the log of the total word count. Formula:

$$\frac{-\sum P\left(xi\right) \times \log_2 P\left(xi\right)}{\log\left(N\right)} \qquad (3)$$

where $P(xi)$ is the probability of the word in the text and $N$ is the total number of words.

*4) Root Type-Token Ratio:* It has been noticed that in patients who suffer from Alzheimer's, there is a lot of repetition of words in their conversation as they tend to forget what they said and tend to repeat it again. Hence, to measure this we use Type-Token Ratio and to reduce its variance on text length Root Type-Token ratio is used. Its calculated by using the formula:

$$RTTR = \frac{t}{\sqrt{n}} \qquad (4)$$

where $t$ is the number of unique words and $n$ is the total number of words.

*5) Mean Segment Type-Token Ratio:* In this process the text to be analysed is divided into equal segments in terms of the number of words, in this case we used 16 words per segment.After this the TTR is calculated for each segment and the MSTTR is obtained after we take an arithmetic mean of all the TTR.

*6) Measure of Textual Lexical Diversity (MTLD):* This measure is similar to MSTTR but in this case the text is varaible and the number of words in each segment varies and the segment ends when the TTR reaches a threshold of 0.72.After that the following formula is obtained to obtain MLTD.

$$MLTD = \frac{L}{n} \qquad (5)$$

where $L$ is text length in number of words and $n$ is the number of segments.

*7) Hyper Geometric Distribution Diversity:* This index calculates the probability of finding any token for for each lexical type in a given text, in a random sample of 42 words taken from the same text. The sum of all the probabilities for all lexical types gives the final HDD score.

*8) POS-based features:* We used this marker as tagging the text with POS tags can indicate frequent adjectives, verbs and adverbs which indicate more descriptive and meaningful sentences, while frequent adverbs can indicate the ability to relate different utterances to each other. Hence, it can be a useful identifier of patients with Alzheimer's disease. The various POS tags we used are -

- Verb frequency: It is computed by dividing the total number of words with that verb tag by the total number of words spoken by the patient in the recording.
- Noun frequency: It is computed by dividing the total number of words with that verb tag by the total number of words spoken by the patient in the recording.

- Pronoun frequency: It is computed by dividing the total number of words with that Pronoun tag by the total number of words spoken by the patient in the recording.
- Adverb frequency: It is computed by dividing the total number of words with that Adverb tag by the total number of words spoken by the patient in the recording.
- Adjective frequency: It is computed by dividing the total number of words with that Adjective tag by the total number of words spoken by the patient in the recording.

### B. Acoustic Features

For the Acoustic Model, we extracted the eGeMAPS feature set on the entire audio file using the *opensmile* python package. The eGeMAPS feature set is a collection of 88 acoustic features that provide a baseline for evaluation in researches to avoid differences of feature set and implementations. We used a Recursive Feature Elimination (RFE) method to reduce the features which uses an external estimator that assigns weights to features and the RFE is used to select features by recursively considering smaller and smaller sets of feature. By using this technique We found that only 51 features contribute more towards the prediction, hence we decided to proceed with them for training our model.

## V. CLASSIFICATION

We used the following ML algorithms as final layers of our models -

### A. Support Vector Machine(SVM)

Support vector machines are one of the most prominent and widely used machine learning algorithms used for password authentication and accuracy prediction [18]–[20]. SVMs are based on the idea that the classes can be separated linearly using a hyper-plane to maximize the spaces between the most classes. In non-linear classification, a kernel function is used to map the non-linear classes to a linear problem by raising their dimension. The most widely used kernel functions are linear, polynomial, radial-basis function (RBF), and sigmoidal. In this study, the Polynomial kernel (degree=4) was found to have better performance than the other kernels.

### B. Logistic Regression

It is a supervised Machine Learning algorithm that often uses a Sigmoid function to model a binary dependent variable. Instead of fitting a straight line or hyper-plane in the case of other model,s the logistic regression model uses a logistic function to predict the output of a linear equation between 0 and 1.

### C. Random forests

Random Forest is a supervised Machine Learning algorithm [21] used for classification and regression tasks. It creates multiple decision trees by taking various combinations of classes and samples from the training set. Decision trees are similar to flow charts in terms of structure. A node represents a test on an attribute, each branch of the node holds an outcome of the test, and each leaf node is a class label. The root node

is at the topmost position. The impurity in the decision trees constructed is reduced by calculating the Gini index and using it in performing the appropriate splitting measures. The Gini index is calculated using the formula:

$$Gini = 1 - \sum_{i=1}^{n}(P_i)^2 \qquad (6)$$

where $P_i$ is the probability of a node being classified into a particular class.

## VI. MODELS

We have three models - Model 1: The Acoustic Model, uses various acoustic features from the eGeMAPs feature-set, Model 2: The Lingustic Model, uses various linguistic features that we generated from auto-generated transcripts and Model 3: Pre-trained BERT model + TF-IDF, uses the auto-generated transcripts directly to extract features using a Pre-trained BERT and TF-IDF.

### A. Model 1: The Acoustic model

In this model we used the eGeMAPS feature-set and obtained the final 51 features after feature reduction using Recursive Feature Elimination. We then classified them using various machine learning models such as Support Vector Machines (SVM), Linear Regression (LR) and Random Forest (RF).

### B. Model 2: The Linguistic model

In this model we used the 13 features obtained from the various linguistic features mentioned in the previous section and classified them using the same machine learning models.

### C. Model 3: Pre-trained BERT model + TF-IDF

Attention-based models [17] belong to a class of AI models, commonly referred to as sequence-to-sequence models. The goal of these models, as the name suggests, is to generate an output sequence from an input sequence that is generally of different lengths. The main technical innovation of bidirectional transformers (BERT) [15] is to apply the bidirectional training of a transformer, a popular attention model, to language modelling. Using this approach, we fine-tune the pre-trained BERT model by using the auto-generated transcripts of the audio files. We used the BERT model to obtain the pre-trained word embeddings for each sentence and obtained a total of 768 features and then we concatenated the BERT features with the TF-IDF features obtained from the text. TF-IDF is an information retrieval technique that weighs a term's frequency (TF) in the text and its inverse document frequency (IDF). A TF and an IDF value is obtained from each word and then they are multiplied. The formula to calculate TF-IDF for each word:

$$tfidf_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t} \qquad (7)$$

where $tf_{t,d}$ is the number of times the term $t$ occurs in the text $d$ and $df_t$ is the total number of documents containing the term t and $N$ is the total number of documents. After obtaining all the features we classified them using the same machine learning models - SVM, LR, RF.

## VII. RESULTS

The 5-Fold Cross Validation results of Models 1, 2 and 3 is summarized in Table 2 below. Both LR and SVM performed similarly considering the CV Accuracy achieved. But, considering the Precision, Recall, Specificity and F1 Score, LR came out on top with more consistency and hence we went with LR for all our models for the classification.

TABLE II
5-FOLD CROSS VALIDATION RESULTS OF MODELS 1, 2 AND 3

| Class | CV Accuracy | Precision | Recall | Specificity | F1 Score |
|---|---|---|---|---|---|
| **Model 1** | | | | | |
| LR | **0.640** | **0.645** | 0.741 | **0.522** | 0.690 |
| RF | 0.580 | 0.607 | 0.630 | **0.522** | 0.618 |
| SVM | **0.640** | 0.629 | **0.815** | 0.435 | **0.710** |
| **Model 2** | | | | | |
| LR | **0.660** | **0.867** | **0.464** | **0.909** | **0.605** |
| RF | 0.620 | 0.800 | 0.429 | 0.864 | 0.558 |
| SVM | 0.620 | 0.846 | 0.393 | **0.909** | 0.537 |
| **Model 3** | | | | | |
| LR | **0.700** | **0.731** | **0.704** | 0.696 | **0.717** |
| RF | 0.618 | 0.562 | 0.600 | 0.632 | 0.581 |
| SVM | **0.700** | 0.667 | 0.636 | **0.75** | 0.651 |

In Table 3, we have summarized the Test Set results. Clearly the Model 3 with the Pre-trained BERT + TF-IDF features is outperforming the Acoustic and Linguistic models. This is inline with the previous studies where Deep Learning based linguistic models outperform basic ML based models and acoustic models.

TABLE III
TEST SET RESULTS

| Model | Class | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|---|
| Model 1 | non-AD | 0.595 | 0.5000 | 0.6207 | 0.5538 |
| | AD | | 0.6857 | 0.5714 | 0.6234 |
| Model 2 | non-AD | 0.5775 | 0.7222 | 0.5652 | 0.6341 |
| | AD | | 0.4286 | 0.6000 | 0.5000 |
| Model 3 | non-AD | 0.7606 | 0.7222 | 0.7879 | 0.7356 |
| | AD | | 0.8000 | 0.7368 | 0.7671 |

## VIII. FUTURE WORK

### A. Deep Acoustic Model: Vision Transformer and sPCEN

Vision Transformer (ViT) [14] is now the state-of-the-art architecture for image classification tasks. The first layer of the Vision Transformer proposed in the paper involves segmentation of images into small patches (16x16). These patches are then linearly embedded, followed by the addition of position embeddings and then fed into a standard Transformer Encoder. The final classification token was also tweaked according the number of classes required for the task. We worked on implementing ViT for AD classification by constructing mel-spectrograms and fine-tuning pre-trained ViT, but we were not able to achieve respectable results.

We propose to tweak the first layer such that we take a segment of dimension 128x16 which keeps the full height of the spectrogram (assuming the height of the spectrogram is 128 pixels and width is a multiple of 16) and train the model from scratch on the ADReSS dataset. This way of segmentation may improve the learnability of the model specifically for spectrograms, due to the way spectrograms are constructed. Previous studies have shown that breaking the audio files into 3-second clips and then constructing spectrograms separately for each clip improves accuracy.

We also propose to normalize spectrograms with sPCEN [22]. A typical spectrogram uses linear frequency scaling so that each frequency range is the same number of Hertz apart, shown to outperform the point-wise logarithm of mel-frequency spectrogram as an acoustic front-end. However, in the case of PCEN (Per-Channel Energy Normalization), it improves the robustness of the channel distortion. It has the ability to Gaussianize and whiten the background of acoustic recordings which is due to the operations of temporal integration, adaptive gain control, and dynamic range compression. Hence it is shown to outperform the point wise logarithm of Mel-frequency spectrogram.

### B. Ensemble model

We would like to explore further by combining the Deep Acoustic model with our Model 3 (Pre-trained BERT) to find out if it improves the prediction accuracy.

## REFERENCES

[1] S.Luz, F.Haider, S.dela Fuente, D.Fromm, and B.MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In Proceedings of INTERSPEECH 2020, Shanghai, China, 2020.

[2] Martinc M., Pollak S. (2020). Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia, in Proceedings of Interspeech 2020 (Shanghai: ), 2157–2161. 10.21437/Interspeech.2020-2202 - DOI

[3] Pappagari R., Cho J., Moro-Velázquez L., Dehak N. (2020). Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity, in Proceedings of Interspeech 2020 (Shanghai: ), 2177–2181. 10.21437/Interspeech.2020-2587 - DOI

[4] Rohanian M., Hough J., Purver M. (2020). Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech, in Proceedings of Interspeech 2020 (Shanghai: ), 2187–2191. 10.21437/Interspeech.2020-2721 - DOI

[5] Edwards E., Dognin C., Bollepalli B., Singh M. (2020). Multiscale system for Alzheimer's dementia recognition through spontaneous speech, in Proceedings of Interspeech 2020 (Shanghai: ), 2197–2201. 10.21437/Interspeech.2020-2781 - DOI

[6] Searle T., Ibrahim Z., Dobson R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech, in Proceedings of Interspeech 2020 (Shanghai: ), 2192–2196. 10.21437/Interspeech.2020-2729 - DOI

[7] Pompili A., Rolland T., Abad A. (2020). The INESC-ID multi-modal system for the ADReSS 2020 challenge, in Proceedings of Interspeech 2020 (Shanghai: ), 2202–2206. 10.21437/Interspeech.2020-2833 - DOI

[8] Balagopalan A., Eyre B., Rudzicz F., Novikova J. (2020). To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection, in Proceedings of Interspeech 2020 (Shanghai: ), 2167–2171. 10.21437/Interspeech.2020-2557 - DOI

[9] Cummins N., Pan Y., Ren Z., Fritsch J., Nallanthighal V. S., Christensen H., et al. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition, in Proceedings of Interspeech 2020 (Shanghai: ), 2182–2186. 10.21437/Interspeech.2020-2635 - DOI

[10] Yuan J., Bian Y., Cai X., Huang J., Ye Z., Church K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease, in Proceedings of Interspeech 2020 (Shanghai: ), 2162–2166. 10.21437/Interspeech.2020-2516 - DOI

[11] Sarawgi U., Zulfikar W., Soliman N., Maes P. (2020). Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity, in Proceedings of Interspeech 2020 (Shanghai: ), 2212–2216. 10.21437/Interspeech.2020-3137 - DOI

[12] Koo J., Lee J. H., Pyo J., Jo Y., Lee K. (2020). Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition, in Proceedings of Interspeech 2020 (Shanghai: ), 2217–2221. 10.21437/Interspeech.2020-3153 - DOI

[13] Syed M. S. S., Syed Z. S., Lech M., Pirogova E. (2020). Automated screening for Alzheimer's dementia through spontaneous speech, in Proceedings of Interspeech 2020 (Shanghai: ), 2222–2226. 10.21437/Interspeech.2020-3158 - DOI

[14] Alexey Dosovitskiy & Lucas Beyer & Alexander Kolesnikov & Dirk Weissenborn & Xiaohua Zhai & Thomas Unterthiner & Mostafa Dehghani & Matthias Minderer & Georg Heigold & Sylvain Gelly & Jakob Uszkoreit & Neil Houlsby (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

[15] Devlin, J., Ming-Wei Chang, Kenton Lee and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT (2019).

[16] Zeghidour, Neil & Teboul, Olivier & Quitry, Félix & Tagliasacchi, Marco. (2021). LEAF: A Learnable Frontend for Audio Classification.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. ArXiv, abs/1706.03762.

[18] O. Nieves and V. Manian, "Automatic person authentication using fewer channel EEG motor imagery," 2016 World Automation Congress (WAC), Rio Grande, 2016, pp. 1-6, doi: 10.1109/WAC.2016.7582945.

[19] Y. Ma, X. Ding, Q. She, Z. Luo, T. Potter and Y. Zhang, "Classification of Motor Imagery EEG Signals with Support Vector Machines and Particle Swarm Optimization", Computational and Mathematical Methods in Medicine, vol. 2016, pp. 1-8, 2016. doi: 10.1155/2016/4941235.

[20] M. E. A. Abdel-Hadi, R. A. El-Khoribi, M. I. Shoman and M. M. Refaey, "Classification of motor imagery tasks with LS-SVM in EEG-based self-paced BCI," 2015 Fifth International Conference on Digital Information Processing and Communications

[21] Breiman, L. "Random Forests". Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.(ICDIPC), Sierre, 2015, pp. 244-249, doi: 10.1109/ICDIPC.2015.7323036.

[22] V. Lostanlen et al., "Per-Channel Energy Normalization: Why and How," in IEEE Signal Processing Letters, vol. 26, no. 1, pp. 39-43, Jan. 2019, doi: 10.1109/LSP.2018.2878620.