Explorable CNN-attention Networks (C-Attention Network) for Automated Detection of Alzheimer’s Disease

Ning Wang
nwang7@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Mingxuan Chen
mchen20@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

K.P. Subbalakshmi
ksubbala@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

ABSTRACT

In this work we propose three explorable deep learning architectures to automatically detect patients with Alzheimer’s disease based on their brain abilities. The architectures use: (1) only the part-of-speech features; (2) only language embedding features and (3) both of these feature classes via a unified architecture. We use self-attention mechanisms and interpretable 1-dimensional Convolutional Neural Network (CNN) to generate two types of explanations of the model’s action: intra-class explanation and inter-class explanation. The inter-class explanation captures the relative importance of each of the different features in that class, while the inter-class explanation captures the relative importance between the classes. Note that although we have considered two classes of features in this paper, the architecture is easily expandable to more classes because of its modularity. Extensive experiments and comparison with several recent models show that our method outperforms these methods with an accuracy of 92.2% and F1 score of 0.952 on the DementiaBank dataset while being able to generate explanations. We show by examples, how to generate these explanations using attention values.

KEYWORDS

self-attention, 1-D CNN, interpretable AI and Alzheimer’s Disease

ACM Reference Format:

1 INTRODUCTION

In 2019, Americans spent $244B in caring for patients with Alzheimer’s Disease and Related Dementia (ADRD). The National Academy of Sciences, the National Plan to Address Alzheimer’s Disease, and the Affordable Care Act through the Medicare Annual Wellness, all identify earlier detection of ADRD as a core aim for improving the brain health for millions of Americans. The success of disease modification and preventive therapeutics for ADRD requires the identification of the disease in very early stages, at least a decade before onset. Approaches to early identification have included the use of brief cognitive screening tests and biological markers (usually neuroimaging or cerebrospinal fluid examination [34]). Neuroimaging modalities often include magnetic resonance imaging (MRI) [26] or the evaluation of positron emission tomography (PET) [15] targeting amyloid, Tau or both. The traditional biological marker methods tend to be invasive, expensive and create patient compliance problems. Hence, there is a strong motivation to consider early detection schemes using non-invasive markers of the disease.

In practice, cognitive assessment tools like Practitioner assessment of Cognition (GPCOG) [6], Cambridge Cognitive Examination (CAMCOG) [37], Mini-Cog [5], Mini-Mental State Examination (MMSE) [17] etc. are used to classify dementia and mild cognitive impairment (MCI). Even in primary care settings, cognitive impairment is unrecognized in 27%–81% of the affected patients [11].

As language functions play an important role in the detection of cognitive deficiency across different stages of ADRD, speech transcripts can assist in early detection of the disease. Hence techniques at the nexus of natural language processing and deep learning offer an inexpensive solution to this early detection problem [7, 12].

A Convolutional Neural Network (CNN) and Long Short Term Memory Network (LSTM) model (CNN-LSTM) was proposed in [14, 25], by using part-of-speech (PoS) tags to get accuracy up to 89.7% on classification for AD disease on DementiaBank cookie sub-corpus. Deep neural networks and deep language models were combined [31] to classifying the disease. On a sparse clinical language dataset, the model could predict MCI and AD type dementia with accuracy of 83%.

One of the key criticisms of deep learning (DL) methods is that the decision process of the DL model is intractable thereby making it difficult to understand the reasoning behind its decisions. For applications such as AD detection, it is imperative that some form of reasoning be provided, because of the human angle involved. Hence in this work, we develop explainable deep learning architectures using attention mechanism and 1-dimensional CNN (1-D CNN) to detect Alzheimer’s disease (AD) from transcripts of an individual’s speech.

The main contributions of our work are:

- novel explainable deep learning architectures for early detection of Alzheimer’s disease
- inter and intra-feature attentions that captures the relative importance between the different feature classes as well as the relative importance of features within a class. These can be used to provide explanation of the model’s conclusions
- modular architecture that can be extended to as many feature classes as desired
• extensive testing on the Dementiabank corpus, a popular Alzheimer’s disease dataset
• example explanations using the proposed models

The rest of the paper is organized as follows. Section 2, describes related work. Section 3, discusses the proposed explainable deep learning models. Section 4, details the datasets, the experimental set up, data bias compensation mechanism, example explanations and discussions of the results. Section 5 discusses the conclusions we draw from this work.

2 RELATED WORK

Automatically detecting Alzheimer’s disease using transcripts of conversations is not new. Some previous efforts [3, 16] used linguistic features, such as PoS tags and syntactic complexity; psycholinguistic features (e.g., lexical and morphological complexity, word information, and readability etc.) to detect Alzheimer’s disease. [16] used regression models to achieve an accuracy of 81% on classifying between AD patients and healthy controls. Some researchers have combined latent features obtained via language embeddings like word2vec [28] ‘GloVe’ [33] sentence2vector [27] along with hand-crafted features in a [14, 25, 29] hybrid approach using a CNN and RNN architecture to achieve accuracy around 89%.

While the above methods show varying degrees of accuracy, they do not provide adequate explainability in their models. Some researchers have introduced visualizations to help move the research in the direction of explainability. Activation clusters and first derivative saliency heatmap were used in [25]. K-means clustering was used [21] to find the most frequently used topic in the text data. Note that explainable AD models have been developed for MRI based AD detection methods, such as [30] which proposed a gradient-based visualization method to detect the most important biomarkers related to AD and progressive mild cognitive impairment (pMCI). However, these are not directly applicable to the language based detection problem that we are considering here.

Meanwhile, work on explainable AI has started to emerge in importance specifically to address the problem of trustability of AI systems. Multiple surveys have analyzed those works [9, 18, 43] which focus on different areas of explainable AI (XAI), [43] revisited visualization of CNNs and discussed trends in explainable artificial intelligence. [9] categorized prior work into model transparency and model functionality on multitude dimensions for model interpretability and analyzed the insufficiency of current work.

An interpretable convolutional neural network (ICNN) was developed in [43] for use on images and tested on three benchmark image datasets. The structure of LSTM was explored and contribution of variables to the prediction in multi-variable time series was captured in [20]. [35] developed local interpretable model-agnostic explanations (LIME) to explain both the predictions and models.

[38] presented a mechanism they call “deep learning important features(DeepLIFT)” to specify the contributions of all neurons in the network to every feature of the input. Meijer G-functions were used to disclose the functional forms learned by a model without much apriori assumptions [1]. However, none of these approaches provided precise explainable results and/or worked specifically in NLP domain.

In this work, we adapt the multi-head self-attention (MHA) proposed in [40] and use 1-D CNN [19] to define two types of explanations: intra and inter-feature explanations. These capture the relative importance between features within a set as well as between different classes of features respectively.

3 MODEL AND METHOD

We propose explainable AI models for detecting AD from speech transcripts of patients using attention mechanisms [40] and the 1-D CNN described in [19] to interpret the decisions that the AI models make. We use two types of features in the model, the self-attention mechanism and a 1-D CNN to understand the relative importance of the features in the final outcome. There is a debate on whether attention mechanisms are good for interpretation [23, 41]. However, this debate was settled in favor of using self-attention mechanisms as a viable method for interpretation for classification tasks [39]. Moreover, 1-D CNN can be used to interpret an AI’s decision as demonstrated in [19, 22] for NLP tasks.

3.1 Proposed Architectures: C-Attention Networks

We propose three architectures: one that uses only PoS features, one that uses only the latent features (language embeddings) and a unified architecture, which uses both features.

3.1.1 C-Attention-FT Network. The architecture of the model proposed for exploiting PoS features (C-Attention-FT Network) is depicted on the left hand side of Figure 1. This architecture comprises of a self-attention module that captures the intra-feature relationships; an attention layer together with a following 1-D CNN layer that can be used to generate feature level explanations followed by a softmax layer. The MHA module is the same as that proposed in [40] for the popular transformer architecture and is presented in Sec 3.1.4. Let $R = \{r_1, r_2, ..., r_n\}$ be the set of records, then $r_j$ is the $j^{th}$ record in the dataset. We compute PoS tags for each record using NLTK [16]. Let $P = \{p_1, p_2, ..., p_n\}$ be the set of PoS feature vectors and $p_i$ be the $i^{th}$ vector in the PoS matrix. We use $h$ Multi-Head-Attention (MHA) layers on $P$ to $(p_1, p_2, ..., p_n)$ to capture the relationship between the PoS features. The MHA transforms $P = \{p_1, p_2, ..., p_n\}$ to another matrix of n-dimensional vectors $A = \{a_1, a_2, ..., a_n\}$. The MHA module is followed by a 1-layer CNN and a softmax layer to get the final classification.

3.1.2 C-Attention-Embedding Network. The architecture of the proposed C-Attention-Embedding Network is shown on the right hand side of Figure 1. We propose this architecture as a means of capturing latent feature information implicit in language embeddings. Specifically we use the universal sentence embedding (USE) [8] to represent each sentence in a record. This architecture is similar to the proposed C-Attention (See 3.1.1) except for the addition of a positional encoding module. The positional encoding module is used to maintain the relative positions of the sentences and is the same as that used in the transformer [40] architecture.

Let $u_i$ be the USE vector corresponding to the $i^{th}$ sentence in the record. The positional encoding is applied to each vector and the resulting vectors, $u_i'$, are used to construct the matrix $U = \{u'_1, u'_2, ..., u'_N\}$. An h-layer MHA module is used to extract
the intra-feature relationships in this architecture. This is followed by an attention layer that captures interpretation at the embedding feature level. The output of the attention layer is fed to a 1-layer CNN and a softmax to get the final prediction.

3.1.3 Unified C-Attention Network. The third architecture we propose uses both the PoS and latent features of the sentences and is depicted in Figure 2. This architecture uses the proposed C-Attention-FT network and the C-Attention-Embedding network as two legs and combines them with another attention layer followed by a dense layer and the softmax layer. The dense linear layer is the same as that proposed in the transformer [40]. The attention layer captures the relative importance between the PoS and the USE features and helps in providing an interpretation at the feature class level.

3.1.4 Attention Mechanisms. Attention mechanisms have been proved to be efficient in capturing global dependencies without any restrictions on the distance between the input and output sequences [2, 42]. Vaswani et al. [40] use self-attention [32] mechanisms along with positional coding in the design of the transformer which has become very popular in language models like BERT [13]. In this paper, we use the attention mechanisms and MHA mechanism proposed in [40]. These use a scaled dot product attention, which is given by

\[
\text{Attention}(Q,K,V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]

where \(Q, K, V\) are the query, key and value matrices and \(d_k\) is dimension of the query and key vectors.

3.1.5 1-D Convolution Layer. We use a single layer 1-D CNN [10] as the penultimate layer, followed by a maxpooling layer which empowers the CNN to compress the information and extract global features. It was shown in [19] that 1-D CNNs filters essentially are n-gram detectors with each filter specializing in a closely-related family of n-grams. This feature can be used to interpret the action of the CNN layer in the architecture. We show how to trace back through these filters and the MHA layer to derive explanations for the classifier outcomes.

4 EXPERIMENT AND RESULT

We evaluate the proposed C-Attention Network architectures on the DementiaBank dataset and compare the performances of these architectures with each other as well as some recently published results [14, 25].
We implemented our proposed model by using Pytorch. The model is trained to minimize the cross-entropy loss function of predicting the class label of participants’ speech records in the training set. We randomly split the original data into 81% training, 9% validation and 10% testing.

### 4.2 Experiment Setup

We implemented our proposed model by using Pytorch. The model is trained to minimize the cross-entropy loss function of predicting the class label of participants’ speech records in the training set. We randomly split the original data into 81% training, 9% validation and 10% testing.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-LSTM</td>
<td>0.8384</td>
<td>0.8683</td>
<td>0.9497</td>
<td>0.9058</td>
<td>0.9057</td>
<td>6.3</td>
<td>15.6</td>
<td>5.3</td>
<td>102.6</td>
</tr>
<tr>
<td>C-LSTM-ATT</td>
<td>0.8333</td>
<td>0.8446</td>
<td>0.9778</td>
<td>0.9061</td>
<td>0.9126</td>
<td>2.6</td>
<td>19.3</td>
<td>2.3</td>
<td>105.6</td>
</tr>
<tr>
<td>C-LSTM-ATT-W</td>
<td>0.8512</td>
<td>0.9232</td>
<td>0.8949</td>
<td>0.9084</td>
<td>0.9139</td>
<td>14</td>
<td>8</td>
<td>11.3</td>
<td>96.6</td>
</tr>
<tr>
<td>C-BILSTM</td>
<td>0.8495</td>
<td>0.8508</td>
<td>0.9965</td>
<td>0.9178</td>
<td>0.9207</td>
<td>1</td>
<td>16.6</td>
<td>0.3</td>
<td>95</td>
</tr>
<tr>
<td>C-BILSTM-ATT</td>
<td>0.8466</td>
<td>0.8525</td>
<td>0.9895</td>
<td>0.9158</td>
<td>0.9503</td>
<td>1.3</td>
<td>16.3</td>
<td>1</td>
<td>94.3</td>
</tr>
<tr>
<td>C-BILSTM-ATT-W</td>
<td>0.882</td>
<td>0.9312</td>
<td>0.9298</td>
<td>0.9305</td>
<td>0.9498</td>
<td>11</td>
<td>6.6</td>
<td>6.6</td>
<td>88.6</td>
</tr>
<tr>
<td>C-BILSTM-ATT-W NO PSYCH.</td>
<td>0.879</td>
<td>0.887</td>
<td>0.9825</td>
<td>0.9319</td>
<td>0.9499</td>
<td>12</td>
<td>5.6</td>
<td>1.6</td>
<td>93.6</td>
</tr>
<tr>
<td>C-BILSTM-ATT-W NO SENT.</td>
<td>0.897</td>
<td>0.9239</td>
<td>0.9615</td>
<td>0.9321</td>
<td>0.9501</td>
<td>7.6</td>
<td>10</td>
<td>3.6</td>
<td>91.6</td>
</tr>
<tr>
<td>C-BILSTM-ATT-W NO DEMO.</td>
<td>0.8908</td>
<td>0.9005</td>
<td>0.9789</td>
<td>0.9308</td>
<td>0.9473</td>
<td>10.33</td>
<td>7.33</td>
<td>2</td>
<td>93.3</td>
</tr>
</tbody>
</table>

Table 1: The comparison of performance metric between our models and others. C-LSTM, C-LSTM-ATT and C-LSTM-ATT-W were referenced from [25], their standard model kernel is Convolutional-LSTM; C-BILSTM, C-BILSTM-ATT, C-BILSTM-ATT-W, C-BILSTM-ATT-NO PSYCH., C-BILSTM-ATT-W NO SENT. and C-BILSTM-ATT-W NO DEMO. were referenced from [14], the main kernel of these architectures is Convolutional-BILSTM. The rest six models are our models: Attention-FT is attention model with only PoS features; Attention-Embedding is the one with only universal sentence embedding; Attention-FT+Embedding is the combination of those two parts. C-Attention-FT, C-Attention-Embedding and C-Attention-FT+Embedding are the similar architectures but replaced the dense layer by a convolutional layer. All these six models are based on Attention Mechanism.

### 4.1 Data and Pre-processing

DementiaBank [4] is a database of multimedia interactions for the study of communications between people with dementia. Specifically the dataset contains transcripts of the description of a picture showing the theft of a cookie. This cookie theft corpus contains 1049 transcripts from 208 AD patients and 243 transcripts from 104 elderly control individuals for a total of 1229 transcripts.

### 4.3 The Benchmarks

We compare the performance of our architectures with recently published results in terms of accuracy, precision, recall, F1 score, area under the curve (AUC). These comparisons are shown in Table 1. We also include total number of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP) for completion.

- **Attention-FT**: The Attention-FT architecture is a slightly modified version of our proposed C-Attention-FT architecture. In this version, we replace the CNN with the dense linear network used in [40].
- **Attention-Embedding**: The Attention-Embedding architecture is a modification of the C-Attention-Embedding architecture. Just as in Attention-FT, we replace the CNN with a dense linear network.
- **Attention-FT+Embedding**: This is a variant of the proposed Unified C-Attention-FT+Embedding architecture, with the two “legs” of the original architecture replaced by the Attention-FT and Attention-Embedding architectures described above.
- **C-LSTM**: The C-LSTM architecture consists of a CNN followed by an LSTM layer. PoS and word embedding features are used [14].
Table 2: Samples of explanation for correct predictions of three speech records. We sort embedding sentences based on the attention value and show the top three sentences. The Label is the ground truth: 0 represents the healthy control; 1 represents the patient. Attention values means how much importance of these sentences are for prediction.

<table>
<thead>
<tr>
<th>Label</th>
<th>Speech Record</th>
<th>Important Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>okay, well the mother is drying the dishes, the sink is overflowing, um the little girl’s reaching for a cookie, and her brother’s taking cookies out of the cookie jar, and the stool is going to knock him on the floor laughs, he’s going to fall on the floor because the stool’s not uh what, with gravity, whatever, uh the uh curtains are blowing I think, that’s all I can see</td>
<td>um the little girl’s reaching for a cookie with gravity he’s going to fall on the floor because the stool’s not uh what I would like to have a lead pencil</td>
</tr>
<tr>
<td>1</td>
<td>I would like to have a lead pencil, the tree is blossoming, I hope my child doesn’t hafta go to the hospital, I hope my child doesn’t hafta go to the hospital, I shouldn’t say that because we have a daughter who’s pregnant, and I do want her to go to the hospital, okay then, this winter has been a very cold one, the doctor said I, I sat in the chair by a the doctor, brief, I’m not, I forgot to try make them brief, the bureau drawer stands open, , , ,</td>
<td>1 I shouldn’t say that because we have a daughter who’s pregnant and winter q exc and winter q exc last year was a cold winter</td>
</tr>
<tr>
<td>1</td>
<td>uh the pencil is on the desk, , leafing, meaning the leaves are opening, , , , , cold q exc, and winter q exc, last year was a cold winter</td>
<td>um the little girl’s reaching for a cookie with gravity he’s going to fall on the floor because the stool’s not uh what I would like to have a lead pencil</td>
</tr>
</tbody>
</table>

4.4 Performance Analysis

From the Table 1 we see that the best overall performance is achieved by both the proposed C-Attention-FT and unified C-Attention-FT+Embedding architecture. The C-Attention-FT architecture performs best in terms of accuracy and F1 scores, and the C-Attention-FT+Embedding performs best in terms of precision and AUC. Attention-Embedding does the worst in terms of accuracy, although its performance is not too bad in terms of precision recall and F1 scores. We also note that using attention only without a convolutional layer does not seem to result in the best performance as evidenced by the fact that all three modifications: Attention-FT, Attention-Embedding and Attention-FT+Embedding are not near the top in terms of most scores. Comparing only the Attention-FT and Attention-Embedding results, we note that the Attention-FT model performs better on all metrics, which seems to suggest that the PoS features (used in Attention-FT) may have more value than the latent features with it comes to using only attentions and no convolutional layers.

4.5 Explainability Analysis

The advantage of our model is that we can interpret the classification process of the model. Specifically, for each case, we can explain whether the model considers the PoS features or the latent feature as more important in its decision. Similarly, we can also use the self-attention weights of to determine the relative importance of utterances within a single picture description and the relative importance of the different PoS features in arriving at the final decision.

4.5.1 Explaining the Universal Sentence Embedding Features: Table 2 shows some sample sentences with their corresponding ground truth (labels). Label 0 indicates that the description was uttered
by a healthy individual (control) and Label 1 indicates that the description was uttered by a patient. In each of these examples the unified C-Attention-FT+Embedding network classified the utterance correctly. The column “Important Sentences” refers to the sentences that are captured by the attention and 1-D CNN layers, which indicates higher importance of those sentences. According to the analysis on entire testing dataset (129 speeches), we have following findings:

- the sentences that are considered most important by the attention layer (have highest attention values) is almost always captured by the filters in the 1-D CNN layer. We notice that 121 speeches out of 129 speeches show this pattern.
- We also note that the intra-feature attention value for a patient’s utterances seems to be more uniformly distributed compared to that of a healthy control which shows a definite higher value for some sentences compared to others. This might indicate that the AI is picking up on the “randomness” of the utterances of the patients compared to that of a healthy control.

4.5.2 Explaining the role of Part of Speech Features: PoS features are extracted at speech level. A total of 36 PoS tags are defined. Like the latent features we notice that the PoS features with the highest attention weight are 100% captured by the the filters in the following 1-D CNN layer. The top PoS features captured by attention layer and 1-D convolutional layers are shown in Figure 3. This result is obtained by analysing the entire testing dataset. It would indicate that the PoS features NNPS, NN, MD, MD, EX, PRP and PRP (personal pronoun) are the most important accounting for 80.9% of times for the PoS Features. Our findings are consistent with previous works [24] that AD patients tend to use more pronouns instead of nouns compared with healthy controls.

4.5.3 Explaining the relative importance of feature classes (PoS vs latent features): The final attention layer of the proposed unified C-Attention-FT+Embedding architecture (Figure 2), captures the relative importance between the PoS features and the latent features in a decision. Table 4 shows these attention values for three patients (with their record numbers indicated). Record number 1188 and 982 correspond to AD patients that were correctly identified as patients by the C-Attention-FT+Embedding model. Record 182 corresponds to a healthy control (indicated by a label value of 0) and also correctly identified by the model. For example, for speech record number 1188, the latent features played a bigger role in determining the overall decision with a weight of 0.703. By contrast, for record number 182, the PoS feature seems to have weighed slightly more (0.572) than the latent features (0.428). Figure 4 shows that the “Part of Speech” leg is assigned a higher attention value in 65.1% of the cases, while Universal Sentence Embedding is assigned a higher attention value in 34.9% of the cases, indicating that the PoS features seem to play a higher role in detecting AD. This fact is also demonstrated in Table 1.

Table 3: Example of PoS tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Americans</td>
</tr>
<tr>
<td>MD</td>
<td>modal could</td>
<td>will</td>
</tr>
<tr>
<td>EX$</td>
<td>existential there</td>
<td>there is</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>NNP$</td>
<td>proper noun, singular</td>
<td>Harrisons</td>
</tr>
</tbody>
</table>

Table 4: Attention Values for Part of Speech and Universal Sentence Embedding for three speech records numbered 1188, 182 and 982. Label 0 corresponds to a healthy control and Label 1 corresponds to an AD patient. In all cases, the C-Attention-FT+Embedding model identified the status of the individual correctly.

<table>
<thead>
<tr>
<th>Label</th>
<th>Record Number</th>
<th>PoS Weight</th>
<th>Embedding Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1188</td>
<td>0.297</td>
<td>0.703</td>
</tr>
<tr>
<td>0</td>
<td>182</td>
<td>0.572</td>
<td>0.428</td>
</tr>
<tr>
<td>1</td>
<td>982</td>
<td>0.157</td>
<td>0.843</td>
</tr>
</tbody>
</table>

5 CONCLUSION

We proposed three explainable architectures using CNN and attention to detect Alzheimer’s disease using two kinds of features: part-of-speech and language embeddings. One architecture uses only the PoS feature, one uses only the universal sentence embedding and the third is a unified architecture that uses both of these features. We propose the use of attention layers and 1-D CNN layer
to capture explanations at 2 levels: one each at the intra-feature level and inter-feature-class level. The intra-feature level attention weights and 1-D CNN filters capture the relative importance the model places on the individual features in the category, whereas the inter-feature level attention weights gives us an idea of the relative importance that the model placed between the two classes of features.

Extensive testing on the popular DementiaBank datasets and comparisons with several recently published models as well as minor modifications of our own models show that the C-Attention-FT architecture performs best in terms of accuracy and F1 scores, and the C-attention-FT+Embedding performs best in terms of precision and AUC while at the same time being able to generate explanations of the action of the AI. We also show by examples how to generate explanations for the actions of the models. Our results agree with some of the previous work who shows that AD patients tend to use more pronouns instead of nouns. Our work thus is an inexpensive, non-invasive, explainable AI model that can detect AD at good performance metric. Since it is based on only the spoken language, it can be potentially easily implemented in an app setting by giving the option of taking it at home. This in turn can have a positive impact on patient compliance and therefore early detection of AD.

REFERENCES


