INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Interpretable Hierarchical Deep Learning Model for Noninvasive Alzheimer's Disease Diagnosis

Maryam Zokaeinikoo, Pooyan Kazemian, Prasenjit Mitra

To cite this article:
Maryam Zokaeinikoo, Pooyan Kazemian, Prasenjit Mitra (2023) Interpretable Hierarchical Deep Learning Model for Noninvasive Alzheimer's Disease Diagnosis. INFORMS Journal on Data Science 2(2):183-196. https://doi.org/10.1287/ijds.2020.0005

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 2, No. 2, October–December 2023, pp. 183–196
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Interpretable Hierarchical Deep Learning Model for Noninvasive Alzheimer's Disease Diagnosis

Maryam Zokaeinikoo,[a] Pooyan Kazemian,[a,*] Prasenjit Mitra[b]

[a] Department of Operations, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio 44106; [b] College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania 16802

*Corresponding author

Contact: mxz596@case.edu, https://orcid.org/0000-0002-9921-3844 (MZ); pxk409@case.edu, https://orcid.org/0000-0002-2846-3862 (PK); pum10@psu.edu, https://orcid.org/0000-0002-7530-9497 (PM)

Received: December 27, 2020
Revised: January 1, 2022; February 1, 2023; August 1, 2023
Accepted: October 4, 2023
Published Online in Articles in Advance: November 17, 2023

https://doi.org/10.1287/ijds.2020.0005

Abstract. Alzheimer's disease is one of the leading causes of death in the world. Alzheimer's is typically diagnosed through expensive imaging methods, such as positron emission tomography (PET) scan and magnetic resonance imaging (MRI), as well as invasive methods, such as cerebrospinal fluid analysis. In this study, we develop an interpretable hierarchical deep learning model to detect the presence of Alzheimer's disease from transcripts of interviews of individuals who were asked to describe a picture. Our deep recurrent neural network employs a novel three-level hierarchical attention over self-attention (AoS3) mechanism to model the temporal dependencies of longitudinal data. We demonstrate the interpretability of the model with the importance score of words, sentences, and transcripts extracted from our AoS3 model. Numerical results demonstrate that our deep learning model can detect Alzheimer's disease from the transcripts of patient interviews with 96% accuracy when tested on the DementiaBank data set. Our interpretable neural network model can help diagnose Alzheimer's disease in a noninvasive and affordable manner, improve patient outcomes, and result in cost containment.

History: Rema Padman served as the senior editor for this article.

Data Ethics & Reproducibility Note: The code capsule is available on Code Ocean at https://codeocean.com/capsule/2881658/tree/v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2020.0005). The study involves secondary use of already-collected data. None of the authors were part of the original study team. The authors had no interaction with living individuals and had no access to protected health information (PHI) or private identifiable information about living individuals.

Keywords: deep learning • Alzheimer's disease • natural language processing • attention over self-attention • interpretable

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder in which cognitive abilities (including memory and language) and executive function deteriorate gradually. AD has been known in the past century as the primary cause of dementia consequently leading to death (Katzman 1976). Alzheimer's disease is the sixth leading cause of death in the United States. AD is typically diagnosed through extensive tests and cognitive tools (Williams et al. 2013). To diagnose AD, imaging methods, such as positron emission tomography (PET) scan and magnetic resonance imaging (MRI), and invasive methods, such as cerebrospinal fluid analysis, are employed. The degeneration of brain cells can be reflected in a variety of ways in brain scans. However, a diagnosis of AD based only on these scans can lead to mistakes (both false positives and false negatives) because it is often not straightforward to distinguish normal age-related changes in the brain from the abnormal AD-related ones. Therefore, there is an immediate need for more accurate and reliable diagnostic methods to improve patient outcomes (Laske et al. 2015).

Attempts to develop neuropsychological tests using a series of cognitive tests containing a set of questions and images have been made to detect the early signs of AD with various accuracy levels (Mortamais et al. 2017). These screening tools, such as the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA), consist of questions and cognitive examinations to assess cognitive abilities of patients. However, the quality of the assessment is highly dependent on the physicians' experience and their ability to distinguish between different categories of the disease (Damian et al. 2011). Sometimes physicians need to combine MMSE with other cognitive tests, which makes it cumbersome and complicated to diagnose AD (Mitchell 2009). The National Institute on Aging at the National Institutes of Health (NIH) and the Alzheimer's Association have both called for better approaches to diagnose AD in a noninvasive way (Albert et al. 2011). Although

there is no cure currently available for AD, treatments are available (e.g., cholinesterase inhibitors) that can ameliorate some symptoms of the disease and lead to improvement in cognition, neuropsychiatric symptoms, and activities of daily living (Birks 2006). Also, early detection of Alzheimer's disease, even partially, is projected to result in a substantial $7 trillion cost savings compared with the status quo (Alzheimer's Association 2018).

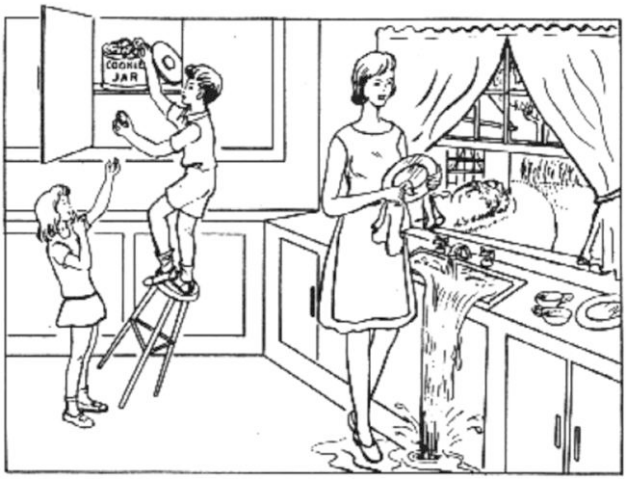
Artificial intelligence (AI) has shown promising potential, as a novel tool, for early diagnosis of Alzheimer's disease (Subasi 2020, El-Sappagh et al. 2021, Logan et al. 2021). Natural language processing (NLP) involves analyzing the transcript of an individual's speech to identify subtle clues that may indicate the early stages of AD (Ševčík and Rusko 2022, Yang et al. 2022). In addition, AI algorithms have been used to analyze brain images for changes indicative of AD, such as those seen with structural magnetic resonance imaging (Saleem et al. 2022). The use of AI in AD diagnosis has the potential to improve accuracy and efficiency, leading to earlier treatment and potentially improved patient outcomes (Fabrizio et al. 2021).

Hence, automatic mathematical tools and algorithms that can detect AD early and accurately are extremely valuable.

It is worth noting that *speech* is a valuable source of clinical information, which can be an indicator of cognitive status. The nerve cells that control cognitive ability and speech processing gradually deteriorate in individuals with AD. Thus, the linguistic deficit captured by verbal utterances can be an indicator of Alzheimer's disease (Fraser et al. 2016). Lately, AI methods have been proposed to detect AD by combining signal processing, machine learning, and NLP, which employ either recorded narrative speech (Lopez-de Ipiña et al. 2012) or recorded scene descriptions (Fraser et al. 2016). However, they lack interpretability, which makes them not suitable for integration into everyday practice.

In this study, we combine NLP and hierarchical deep learning to detect the presence of Alzheimer's disease based on longitudinal patient interview data, from patients who were asked to describe a Cookie-Theft picture (Figure 1). We propose a three-level hierarchical recurrent neural network (RNN) with a unique attention over self-attention (AoS3) mechanism to develop a powerful and interpretable model that can explain the most important lexical memory-related patterns without the need for any feature engineering. The interpretability of a "black-box" deep learning model is critical for several reasons (Olden and Jackson 2002). First, it can help build trust and confidence in clinicians who use the model in their daily practice, thus making model outputs more actionable. Second, interpretable outputs can be easily explained to patients, making physician-patient interactions easier. Third, in the case of detecting AD, interpretability of model outputs can shed light on subtle language deficits that may be associated with the presence of Alzheimer's but would otherwise go unnoticed. Fourth, interpretability of model outputs can assist test designers in improving cognitive tests based on the aspects of the test that are more useful in detecting Alzheimer's disease (Goodman and Flaxman 2017, Holzinger et al. 2017, Kim and Seo 2017).

Figure 1. The Boston Cookie-Theft Picture from the DementiaBank Data Set



Note. The DementiaBank data set is from Becker et al. (1994) and is designed to elicit language deficit that contributes to the diagnosis of AD.

1.1. Main Contributions

The contributions of this paper are fourfold:

1. We developed a new three-level hierarchical structure to capture the hierarchical dependencies between words, sentences, and longitudinal interview transcripts to detect the presence of Alzheimer's disease.

2. To the best of our knowledge, we are the first to develop an attention over self-attention mechanism, which is a powerful method to prevent information loss by considering the relation of words, sentences, and transcripts within the sequence. We further demonstrate its implementation to confront a pressing healthcare problem.

3. Our model is interpretable, which addresses a common shortcoming of black-box neural network models and provides valuable insight into the language deficit that heralds the presence of Alzheimer's disease. The attention mechanism at the word level helps identify the obscure or irrelevant words that can be indicators for memory loss due to Alzheimer's disease. At the sentence level, attention allows capturing the sentences with structural deterioration and semantic impairment that can help distinguish patients with AD from healthy individuals. Finally, attention at the transcript level provides insight into the particular interviews that contributed most to model prediction (AD or healthy) for each individual.

4. We provide proof of concept by conducting several numerical experiments on the DementiaBank data set. Our results indicate that the AoS3 deep learning model achieves a high accuracy (mean crossvalidation accuracy = 96%) and outperforms other models of similar nature (which do not need feature engineering) developed on the same data set.

Our interpretable deep neural network model can help diagnose Alzheimer's disease with a high accuracy, obviating the need for expensive imaging tools and invasive methods. Although proof of concept is given for Alzheimer's disease, this type of modeling framework can also be employed to detect other degenerative neurological diseases.

2. Related Works

2.1. Studies Relevant to Alzheimer's Disease Detection

Previous works to detect AD using language as their input data are mainly dependent on extracting linguistic features from transcripts (Orimaye et al. 2017). The main problem with feature-based methods is not only that the quality of predictions is highly dependent on the quality of features but also that some intricate features may not be recognizable by existing methods (LeCun et al. 2015). In addition, language evolution may affect linguistic features' extraction methods. Recently, deep learning models have outperformed other feature-based machine learning methods in speech recognition (Hinton et al. 2012, Mikolov et al. 2013, Sainath et al. 2013) and achieved promising results for various tasks in natural language processing including sentiment analysis (Bordes et al. 2014) and natural language understanding (Collobert et al. 2011).

Several studies used lexical features to detect AD from the DementiaBank data set (Zimmerer et al. 2016, Orimaye et al. 2017). Wankerl et al. (2017) proposed statistical approaches for detecting AD using n-gram models. They evaluated their approach on the DementiaBank data set and achieved an accuracy of 77.1%. Orimaye et al. (2018) proposed deep language models using decomposed higher-order n-gram N-dimensional vectors as discrete inputs on the DementiaBank data set. Their experimental results show that deep neural networks can learn linguistic markers with reasonable accuracy on this clinical data set; the area under the receiver operating characteristic (ROC) curve (AUC) for their best model was 83%. These models, however, are not interpretable, and do not provide insight into linguistic deficits that can indicate the presence of AD. It is important to note that models that lack transparency and interpretability are often not received well by clinicians and are rarely incorporated into the clinical practice (Goodman and Flaxman 2017, Fan et al. 2020).

Karlekar et al. (2018) applied three neural network models based on convolutional neural networks (CNNs), long short-term memory recurrent neural networks (LSTM-RNNs), and their combination to distinguish patients with AD from control patients based on documents in DementiaBank. Their best end-to-end model was the combined CNN-RNN model, which achieved an accuracy of 84.9%. More recently, Chen et al. (2019) proposed a network based on an attention mechanism by combining CNN and gated recurrent units (GRUs) to capture linguistic deficits of AD patients. They achieved a crossvalidation accuracy of 97% in distinguishing AD subjects from control subjects in the DementiaBank data set. Fritsch et al. (2019) improved the statistical approach proposed by Wankerl et al. (2017) by developing a neural language model based on LSTM cells and evaluating the perplexity of their model. They obtained an accuracy of 85.6% for the binary classification task of identifying subjects as either having AD or not. Chien et al. (2019) proposed a convolutional recurrent neural network (CRNN) model and obtained an AUC measure of 83.8%. Kong et al. (2019) developed a hierarchical attention network on the DementiaBank data set without feature engineering. Combining with demographic feature (age), they achieved an accuracy of 86.9%.

2.2. Studies Relevant to Interpretable Deep Neural Models

In recent years, extracting meaningful insights from black-box neural networks has attracted many researchers. For sequential input data, which we consider in this study, researchers have proposed interpretable RNN-based structures for disease diagnosis and prediction (Choi et al. 2016, Ma et al. 2017, Sha and Wang 2017). They achieved promising results by implementing an attention mechanism to discover where the model concentrates on (via attention weights) when making predictions. Ma and Hovy (2016) proposed a diagnostic model based on various attention mechanisms (the Dipole model). The attention layer of their model explains the importance of the RNN model's hidden states as their corresponding attention weights. Sha and Wang (2017) proposed a GRU RNN-based hierarchical attention (GRNN-HA) model. The GRNN-HA model is quite similar to the Dipole model, except that it has a hierarchical structure in which the bidirectional RNN (BRNN) and the attention mechanism in the lower layer encode every medical event within a visit. In the upper layer, the BRNN and the attention mechanism capture the dependencies among the sequence of medical visits. Yang et al. (2016) developed a bilevel structure and used an attention mechanism to demonstrate the interpretability of the model. Choi et al. (2016) proposed the Reverse Time Attention model (RETAIN), which processes the input in reverse-time order, unlike the BRNN structure in the Dipole and the GRNN-HA models.

Previous works on detecting AD using language as input data have primarily relied on human engineering of linguistic features from transcripts and applying

machine learning techniques. However, these methods can be limited by the quality of the features and the possibility of some intricate features being unrecognizable by existing methods. In addition, language evolution can impact the effectiveness of linguistic feature extraction methods. Some deep learning models have shown promising results in detecting AD, but they lack interpretability and transparency, making them less appealing to clinicians and less likely to be incorporated into clinical practice.

Modern natural language processing has introduced the ideas of word embeddings and learning representations. Additionally, the attention model has been proposed and very successfully used to force a machine learning model to pay attention to the most important parts of a sentence. These advances have resulted in (a) better accuracy and (b) more interpretable machine learning methods. The primary motivation of this work is to take advantage of these innovations and attempt to improve the accuracy as well as the interpretability of the methods.

Our study diverges from the existing works and develops a new methodology motivated by a pressing healthcare problem. In this study, we design a three-level hierarchical neural network that, unlike the bilevel model developed by Yang et al. (2016), captures the dependencies of longitudinal transcripts, sentences within each transcript, and words within each sentence. We extend the theory of interpretable deep neural network models by introducing a novel three-level hierarchical attention over self-attention (AoS3) mechanism that sheds light on the importance of each word, sentence, and transcript in making predictions. The main advantage of our AoS3 model is that it prevents loss of information and captures the relation of each subcomponent within the higher-level component (e.g., the relation of sentences within each transcript). This helps both model performance and interpretability. Numerical results demonstrate that our new AoS3 model outperforms the regular hierarchical attention-based neural network models and achieves a new benchmark accuracy for detecting the presence of Alzheimer's disease using longitudinal interview transcripts.

3. Methods

In this section, we first describe the components of our model. Then, we explain how we incorporated them in a three-level hierarchical structure to capture both the sequential dependencies and the importance of elements at each level. The input to our hierarchical neural network model is the transcript of individual interviews. In particular, each individual has one transcript per annual visit. Each transcript comprises a number of sentences, and each sentence includes several words. The words are embedded and then fed into the network as described in the next part.

3.1. Word Embedding Layer

A word embedding layer maps each word from the vocabulary set to a low-dimensional vector space using a pretrained word embedding model. This helps capture the context of words such that the words with similar or close semantic meanings have similar vector representations. In this study, we used Global Vectors for Word Representation (GloVe), a well-known pretrained word embedding model, to obtain the vector representation for each word (Pennington et al. 2014).

3.2. Contextual Embedding Layer

We used a GRU-based RNN layer (Cho et al. 2014) on top of the component embeddings from previous layers to capture the temporal dependencies within every component category (words, sentences, and transcripts) at each level. We placed a GRU in both forward and backward directions (i.e., bidirectional) to capture more information from both past and future utterances and then concatenated the outputs of the two GRUs. Therefore, if the input vector has d dimensions, the output of this layer will be $2d$-dimensional.

3.3. Three-Level Attention over Self-Attention Mechanism

Suppose we have a sentence (S) constructed from a sequence of n word embedding representation vectors (v_i).

$$V = (v_1, v_2, \ldots, v_n), \tag{1}$$

where (v_i) is a u-dimensional word embedding for the i-th word in the sentence. In the first layer, a bidirectional GRU is applied to encode the embedding representations as $E^1 = (e_1^1, e_2^1, .., e_n^1)$. Then, the self-attention structure is employed to extract different features of the sequence into a vector representation. The self-attention mechanism takes the vector of inputs (E), and outputs the vector of weights as α (Lin et al. 2017):

$$\alpha = \text{Softmax}(w_2 \tanh(W_1(E^1)^T)) \tag{2}$$

Here, W_1 is the weight matrix with shape $d - by - 2u$ (recall that u is the GRU dimension), which is going to be learned during the training process. The vector w_2 is a vector of parameters with size d (d is a hyperparameter).

In the second layer, each component is concatenated with the self-attentive representation in order to keep the relation information (Feng et al. 2018).

$$\hat{e}_t = \Sigma_{j=1}^{n} \alpha_j^t e_j^1 \tag{3}$$

$$e_t^2 = [v_t, \hat{e}_t] \tag{4}$$

Each e_t^2 captures the relationship between v_t and other words in the sentence.

We then apply an attention layer on top of the vector representation $E^2 = (e_1^2, e_2^2, \ldots, e_n^2)$ to capture the global

attention scores (β) of each component (word).

$$\beta = \text{Softmax}(w_3^T \tanh E^2) \tag{5}$$

Here, w_3 is the weight vector that is learned during the training process.

3.4. Three-Level Hierarchical Attention over Self-Attention Network

In this section, we first develop a three-level hierarchical attention mechanism that considers the dependencies between words, sentences, and documents for longitudinal interview transcripts of patients. Then, we propose a novel attention over self-attention (AoS3) mechanism to capture the importance of the components constructing the three levels (words, sentences, and documents) of our algorithm. The architecture of this algorithm is illustrated in Figure 2.

Recall that our goal is to detect the presence of Alzheimer's disease for each participant of the study using interview transcripts. Suppose that each participant has a sequence of at most N transcripts t_i such that each transcript contains L sentences. We denote each sentence as s_{ij}, which represents the j-th sentence in the i-th transcript of this patient. Each sentence s_{ij} contains T words such that w_{ijk} represents the k-th word in the j-th sentence of the i-th transcript. In the first level of our algorithm, we use GloVe to obtain a low-dimensional representation vector for each word, w_{ijk}, via

$$x_{ijk}^1 = W_{emb} w_{ijk} + b_{emb}, k \in [1, T], \tag{6}$$

where W_{emb} is the embedding matrix obtained through the pretrained GloVe embedding.

Then, we encode each word with GRU-BRNN as h_{ijk} and its hidden representation u_{ijk} as follows:

$$h_{ijk}^1 = [\overrightarrow{\text{GRU}}(x_{ijk}^1); \overleftarrow{\text{GRU}}(x_{ijk'}^1)], k \in [1, T], k' \in [T, 1], \tag{7}$$

$$u_{ijk}^1 = \tanh(W_w^1 h_{ijk}^1 + b_w^1), k \in [1, T]. \tag{8}$$

The self-attention weights are computed via

$$\alpha_{ijk}^1 = \text{Softmax}(u_{ijk}^{1\,T} u_w^1), k \in [1, T]. \tag{9}$$

The last step in this encoding level is to encode each sentence as a weighted sum of h_{ijk} with the attention scores,

$$s_{ij}^1 = \Sigma_k \alpha_{ijk}^1 h_{ijk}^1, k \in [1, T]. \tag{10}$$

Then, we concatenate the calculated sentence embedding s_{ij} with the word embedding in order to avoid information loss.

$$x_{ijk}^2 = [s_{ij}^1, x_{ijk}^1] \tag{11}$$

Because we aim to determine the contribution of each word within each transcript to the overall prediction, we applied a global attention layer on top of the concatenated vector representations (x_{ijk}^2) from the self-attention layer to obtain the attention scores α_{ijk}^2 for each word w_{ijk} in this level.

$$\alpha_{ijk}^2 = \text{Softmax}(w_w^T \tanh(x_{ijk}^2)), k \in [1, T] \tag{12}$$

Here, w_w is the trained weight vector, and its superscript, T, indicates the transpose of the vector. The attention helps the network to recognize which parts of the sequence play an important role (importance score) in text classification. We use those importance scores to interpret our classification model.

The last step in this level is to encode each sentence as a weighted sum of h_{ijk}^2 with the attention scores,

$$s_{ij}^2 = \Sigma_k \alpha_{ijk}^2 x_{ijk}^2, k \in [1, T]. \tag{13}$$

In the second level of this structure, we encode each sentence representation obtained from the first level s_{ij}^2, applying GRU-BRNN to incorporate both future and past information within a transcript. Then, we calculate the sentence-level attention α_{ij}^1 with the sentence-level context vector u_s^1 as follows:

$$h_{ij}^1 = [\overrightarrow{\text{GRU}}(s_{ij}^2); \overleftarrow{\text{GRU}}(s_{ij'}^2)], j \in [1, L], j' \in [L, 1], \tag{14}$$

$$u_{ij}^1 = \tanh(W_s^1 h_{ij}^1 + b_s^1), j \in [1, L], \tag{15}$$

$$\alpha_{ij}^1 = \text{Softmax}(u_{ij}^{1\,T} u_s^1), j \in [1, L], \tag{16}$$

$$t_i^1 = \Sigma_j \alpha_{ij}^1 h_{ij}^1, j \in [1, L]. \tag{17}$$

We again concatenate the calculated transcript encoding t_i with the sentence encoding obtained from the previous level (s_{ij}^2) to maintain the relation of each sentence within the transcript.

$$s_{ij}^3 = [t_i^1, s_{ij}^2], j \in [1, L] \tag{18}$$

Next, we compute the global attention score for each sentence within each document via the following equation:

$$\alpha_{ij}^2 = \text{Softmax}(w_s^T \tanh(s_{ij}^3)), j \in [1, L]. \tag{19}$$

As the last step in this level, we compute the transcript representation with the extracted attention scores as

$$t_i^2 = \Sigma_j \alpha_{ij}^2 s_{ij}^3, j \in [1, L]. \tag{20}$$

In the third level, we repeat the same process for the transcript representation t_i^2 computed from the second level to determine the attention score for each transcript in our longitudinal data set as

$$h_i^1 = [\overrightarrow{\text{GRU}}(t_i^2); \overleftarrow{\text{GRU}}(t_{i'}^2)], i \in [1, N], i' \in [N, 1], \tag{21}$$

$$u_i^1 = \tanh(W_t^1 h_i^1 + b_t^1), i \in [1, N], \tag{22}$$

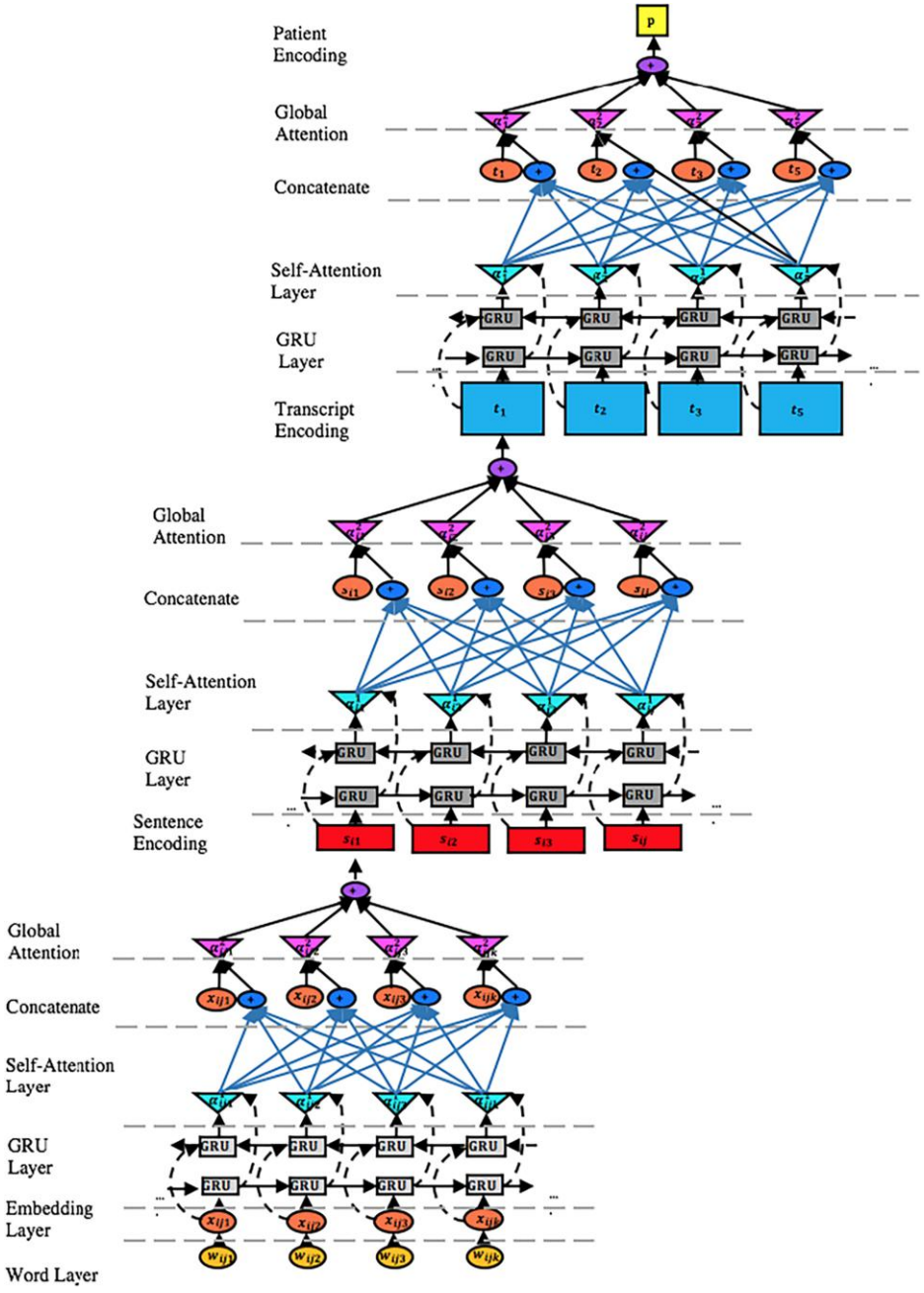
$$\alpha_i = \text{Softmax}(u_i^{1\,T} u_t^1), i \in [1, N], \tag{23}$$

$$p^1 = \Sigma_i \alpha_i^1 h_i^1. \tag{24}$$

The representation vector for each patient is obtained by

$$t_i^3 = [p^1, t_i^2], i \in [1, N]. \tag{25}$$

Figure 2. (Color online) Hierarchical Attention over Self-Attention (AoS3) Structure



Then, we apply the global attention mechanism on top of each transcript encoding as follows:

$$\alpha_i^2 = \text{Softmax}(w_t^T \tanh(t_i^3)), i \in [1, N]. \quad (26)$$

We obtain the patient representation via

$$p^2 = \Sigma_i \alpha_i^2 t_i^3, i \in [1, N]. \quad (27)$$

Finally, we use p^2 obtained from the last level to build a binary classifier as

$$g = \sigma(W_y p^2 + b_y). \quad (28)$$

4. Experiments

4.1. Data Set

We use the DementiaBank clinical data set (Becker et al. 1994) to train and validate our model. Data were collected longitudinally by the University of Pittsburgh School of Medicine for the study of communication in dementia. The data set contains transcripts of the participants' interviews. Participants were asked to describe everything happening in a Cookie-Theft picture (Figure 1). The descriptions were then used to detect language disorders due to AD.

Table 1. CHAT Disfluency Codes

Disfluency	Code
Whole word repetition	Follow word with [/]
Multiple whole word repetition	[x 'number of repetitions']
Phrase repetition	<> [/]
word revision	[//]
phrase revision	<> [//]
pause	(.) or (..) or (...)
filled pause	&-
unintelligible words	xxx

The audio files of participants' interviews were transcribed to Codes for the Human Analysis of Transcripts (CHAT) format. The CHAT transcription format is a tool that helps automatically transcribe audio files. Table 1 presents some CHAT disfluency transcription codes from the CHILDES manual that are used in our data set (MacWhinney 2000).

The DementiaBank data set contains 99 healthy subjects and 194 subjects with probable or definite Alzheimer's disease, with annual follow-up visits up to five years (i.e., maximum of five transcripts per individual). The summary statistics of the DementiaBank data set are presented in Table 2.

4.2. Model Configuration and Training

We developed a three-level hierarchical structure, including transcripts, sentences, and words, to detect the presence of AD. We also made the model interpretable such that it provides insight into how the black-box deep neural network model distinguished subjects with AD from healthy subjects. First, we broke down each transcript into sentences and tokens. Then, we set aside 1/10 of the training set for validation (10-fold crossvalidation) and obtained the 100-dimensional word embeddings by using the pretrained GloVe model (Pennington et al. 2014) on both training and validation sets.

The word, sentence, and transcript context vectors were set to have a dimension of 100 and were initialized at random. We set the GRU dimension to 50 for all word, sentence, and transcript levels; hence, the bidirectional GRU has 100 dimensions. For model training, we used a minibatch size of 10 subjects with the same number of transcripts of 5 per patient and the same number of 25 sentences per transcript (note that transcripts had no more than 25 sentences (Table 2)). For those subjects who did not have five visits, we filled the corresponding visit representation vectors with 'nan' so that the embedding matrix gives a big negative number to every element of these vectors.

We used the Adaptive Subgradient (AdaGrad) optimizer to train the hierarchical model with an initial momentum parameter of 0.1 (Sutskever et al. 2013, Neyshabur et al. 2014). We investigated various parameters using the standard grid search approach (Chollet 2021). We used a crossentropy loss function and evaluated our model using the 'accuracy' metric.

5. Results

To evaluate our model, we conducted three sets of validation experiments. In the first set, we used the balanced data set from 99 healthy subjects and the first 99 probable AD subjects to provide the same validation setting as Orimaye et al. (2017). We performed 10-fold stratified crossvalidation. In each fold, we divided the data set into 90% (training) and 10% (validation). Each subject had multiple interview transcripts. We used stratified crossvalidation to ensure that each set contains the same ratio of healthy to AD subjects. To evaluate the effect of the GloVe embedding dimension on our model's final performance, we also have tested different GloVe word embeddings using dimensions of 100, 200, and 300. Table 3 presents the mean and standard deviation (SD) of the model's performance in terms of accuracy, precision, recall, and AUC.

In our second set of experiments, we further investigated the performance of our proposed AoS3 model on the whole unbalanced DementiaBank data set, including 194 probable AD subjects and 99 healthy subjects. Table 3 presents the mean (SD) of several performance metrics across the 10-fold stratified crossvalidation outcomes using 100-, 200-, and 300-dimensional GloVe embeddings. For comprehensiveness, we report accuracy, precision, recall, and the AUC. We can see that the best mean accuracy result for the whole unbalanced DementiaBank data set is 0.96, which is obtained using 300-dimensional GloVe embedding. We also explored data augmentation, but that did not improve the model's performance.

5.1. Comparison with Simpler Models

5.1.1. Traditional Machine Learning Models. To compare the performance of our novel deep learning model

Table 2. Summary Statistics of Data Set

Visit	Year 1	Year 2	Year 3	Year 4	Year 5
No. of transcripts	293	151	69	28	11
AD	194	77	24	11	3
Healthy	99	74	45	17	8
No. of sentences per transcript, mean (SD)	13.7 (6.5)	13.4 (5.5)	14.1 (6.8)	13.1 (4.7)	13.3 (5.1)
No. of words per transcript, mean (SD)	121.7 (62.9)	122.4 (61.9)	133.8 (81.6)	120.8 (41.7)	132.2 (55.3)
No. of words per sentence, mean (SD)	9.0 (2.7)	9.4 (3.8)	9.6 (2.9)	9.6 (2.8)	10.2 (2.8)

Table 3. Comparison Between Different Experiment Settings of AoS3 Model Using 10-Fold Crossvalidation

			Mean (SD)			
Input data	Model	GloVe	Accuracy	Precision	Recall	AUC
Balanced (99-99)	AoS3	100	0.93 (0.08)	0.93 (0.07)	0.94 (0.11)	0.98 (0.04)
Balanced (99-99)	AoS3	200	0.93 (0.05)	0.95 (0.05)	0.91 (0.10)	0.98 (0.03)
Balanced (99-99)	AoS3	300	0.96 (0.06)	0.96 (0.07)	0.97 (0.07)	0.98 (0.03)
Unbalanced (194-99)	AoS3	100	0.93 (0.07)	0.96 (0.05)	0.93 (0.08)	0.96 (0.05)
Unbalanced (194-99)	AoS3	200	0.94 (0.07)	0.97 (0.03)	0.94 (0.11)	0.97 (0.05)
Unbalanced (194-99)	AoS3	300	0.96 (0.07)	0.97 (0.05)	0.96 (0.07)	0.99 (0.03)

(AoS3) with other alternative models, we have implemented two models of traditional machine learning algorithms, including support vector machine (SVM) and random forest (RF). We used the whole interview transcripts with all CHAT disfluency codes as model input. Then, we used the well-known term frequency-inverse document frequency (TF-IDF) vectorizer to extract numerical features for unigrams and bigrams (Salton and McGill 1983). TF-IDF measures the importance of each word to a transcript in a corpus for both unigrams and bigrams. Then, we fed those extracted features into the SVM and RF algorithms. We used transcripts from all patients' most recent visits to create settings similar to those used in recent studies (Orimaye et al. 2017, 2018) to assess the performance of their models.

5.1.2. Two-Level AoS Model (AoS2). We evaluated our model's performance in detecting AD using only a single transcript for each patient (rather than multiple transcripts from annual visits). To do so, we used the last transcript (i.e., the most recent one) for each patient. Thus, we removed the third level of our AoS3 model that captures the dependencies of the longitudinal transcripts. The resulting model, called the two-level AoS model (AoS2), predicts whether a given transcript indicates signs of AD. Figure 3 illustrates the structure of the AoS2 model. To assess the performance of this model, we used the last-visit interview transcript for each subject. Then, we designed the same data settings as the original AoS3 model (balanced, unbalanced). Table 4 presents the performance results of the AoS2 model.

5.1.3. Three-Level Hierarchical Bidirectional RNN (HBRNN) Model. In this model, we removed the attention over self-attention mechanism from our AoS3 model to assess the performance of a simpler model without the attention mechanism. In each level of this HBRNN model, we only applied the bidirectional RNN model to capture the dependencies between the components. We used the same data settings (balanced, unbalanced) of the DementiaBank data set as we used for the AoS3 model.

Table 4 demonstrates the results of 10-fold stratified crossvalidation for the Aos3, AoS2, and HBRNN models along with two other baseline models (SVM and RF). We can see that our three-level hierarchical deep neural network (AoS3) model achieved the best mean accuracy of 0.96 with an SD of 0.06 across the 10 folds using the balanced data set. Results were similar for the unbalanced data set.

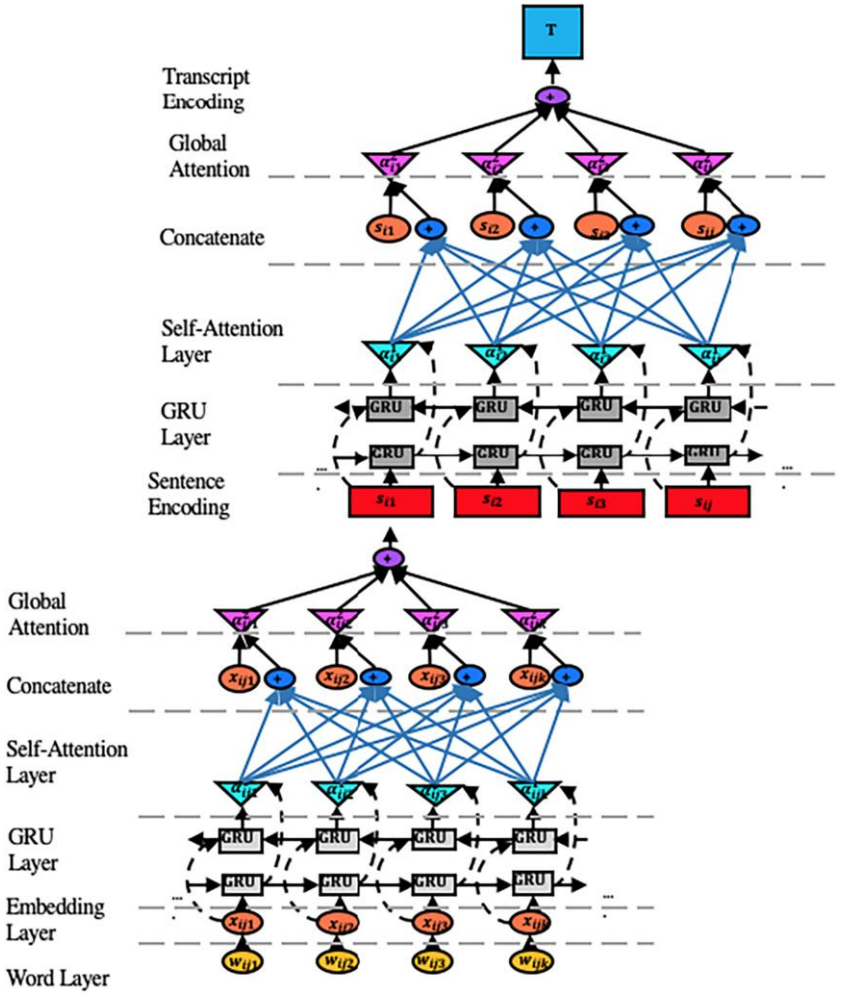
Both AoS2 and AoS3 performed substantially better than the HBRNN model. This is likely because of the integration of an effective attention over self-attention mechanism, which is the key difference between these models and the HBRNN model. Furthermore, the HBRNN model is not interpretable because it does not incorporate the attention mechanism. Moreover, the mean (SD) accuracy for SVM was 0.84 (0.08) and for RF was 0.86 (0.06) using the same balanced data set. Results were comparable in the unbalanced data set. Thus, it is clear that the AoS3 neural network model outperforms the classical machine learning models in detecting Alzheimer's disease.

AoS3 also had superior performance compared with AoS2. In the balanced set, AoS3 achieved an accuracy of 0.96 whereas AoS2 had an accuracy of 0.91. In the unbalanced set, the accuracies were 0.96 and 0.72 for AoS3 and AoS2, respectively. Thus, the hierarchical architecture contributes to improved performance. It also provides additional insights through interpretability.

5.2. Interpretability

To demonstrate the interpretability of the model, we extracted attention scores on the three levels (transcripts, sentences, and words). On the transcript level, the model gave different attention score patterns to the transcripts of healthy subjects compared with the transcripts of individuals with Alzheimer's disease. Figure 4 illustrates the distribution and numerical value of the mean attention scores stratified by transcript number for AD and healthy subjects. For example, the mean attention score given to transcript 1 was 0.24 (AD subjects) and 0.37 (healthy subjects). This means that the first transcripts have contributed more to healthy predictions. In other words, of all attention the model has paid to the first transcripts, on average, 61% has been on transcripts of healthy individuals. On the other hand, the mean attention score for transcript 5 was 0.17 (AD subjects) and 0.10 (healthy subjects). This implies that

Figure 3. (Color online) Two-Level AoS Model (AoS2)

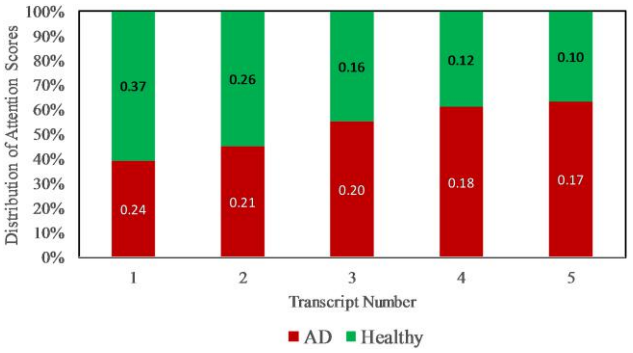


the fifth transcripts have contributed more to the prediction of AD. In other words, of all attention the model has paid to the fifth transcripts, on average, 63% has been on transcripts of individuals with AD. This is expected because as the AD progresses, the model should pay more attention to the later transcripts when predicting AD (e.g., transcripts 4 and 5 will be more determinative of AD than transcripts 1 and 2). Given that, by definition, attention scores for the transcripts of each individual should add to 1.0, this leaves fewer attention points for the initial transcripts. These findings should be interpreted with caution as our analysis is limited by the small size of the DementiaBank data set. In particular, there are only 11 patients with 5 transcripts, and noise

Table 4. Comparison Between Different Experiment Settings of Models Using 10-Fold Crossvalidation

		Mean (SD)			
Input data	Model	Accuracy	Precision	Recall	AUC
Balanced (99-99)	AoS3	0.96 (0.06)	0.96 (0.07)	0.97 (0.07)	0.98 (0.03)
Balanced (99-99)	AoS2	0.91 (0.04)	0.99 (0.01)	0.85 (0.06)	0.99 (0.03)
Balanced (99-99)	HBRNN	0.80 (0.06)	0.92 (0.04)	0.79 (0.08)	0.83 (0.08)
Balanced (99-99)	SVM	0.84 (0.08)	0.84 (0.08)	0.84 (0.08)	0.84 (0.08)
Balanced (99-99)	RF	0.86 (0.06)	0.86 (0.06)	00.86 (0.06)	0.86 (0.06)
Unbalanced (194-99)	AoS3	0.96 (0.07)	0.97 (0.05)	0.96 (0.07)	0.99 (0.03)
Unbalanced (194-99)	AoS2	0.72 (0.05)	0.71 (0.03)	0.82 (0.03)	0.73 (0.06)
Unbalanced (194-99)	HBRNN	0.64 (0.04)	0.66 (0.02)	0.75 (0.05)	0.70 (0.02)
Unbalanced (194-99)	SVM	0.81 (0.06)	0.81 (0.07)	0.81 (0.06)	0.76 (0.08)
Unbalanced (194-99)	RF	0.77 (0.05)	0.78 (0.06)	0.77 (0.05)	0.69 (0.07)

Figure 4. (Color online) Barplot Showing the Distribution and the Numeric Value of the Mean Attention Scores Given to Different Transcripts



from a couple of patients can substantially influence the results.

In the sentence level, the model automatically recognizes those sentences within a given transcript that can help distinguish subjects with AD from healthy individuals. Table 5, A and B, presents 10 sample sentences for healthy subjects and individuals with AD.

In the word level, we decided to provide example interview transcripts for a healthy subject and an individual with AD (Figure 5). In this figure, we have highlighted the words that received a high attention score (>0.08) in dark colors and those that received a medium attention score ($0.04 - 0.08$) in light colors. Words with a low attention score (<0.04) were not highlighted in Figure 5 (Yang and Zhang 2018). This type of analysis can help highlight the words that most significantly indicate the presence of AD and is of paramount importance to interpretability of the neural network model.

6. Discussion

In this study, we developed a new three-level hierarchical RNN model and combined it with NLP to detect the presence of AD based on longitudinal patient interview data. We provided proof of concept through several numerical experiments on the DementiaBank data set. The model can also be adopted to detect AD from similar verbal tests, such as the MoCA and the MMSE, which are commonly used nowadays to assess the presence and progression of dementia and memory loss. We made the black-box neural network model interpretable using a novel three-level attention over the self-attention (AoS3) mechanism. The primary advantage of our AoS3 model compared with the traditional attention models is that it can prevent information loss by considering the relation of words, sentences, and transcripts within the sequence.

We conducted several experiments using different combinations of GloVe embedding dimensions (100, 200, 300) and different input data structures (Table 3). The hierarchical AoS3 model using the DementiaBank data set obtained an overall accuracy of 96% (Table 6).

We also evaluated the attention scores our model gave to transcripts, sentences, and words to shed light on how the neural network model made predictions.

Of all attention scores the model gave to the first and second transcripts, a larger percentage was associated with the transcript of healthy individuals, whereas, of all attention scores given to the third, fourth, and fifth transcripts, the transcripts of individuals with AD had a larger share. This can be explained by the degree of language deficit in AD subjects. If a person suffers from AD, the disease will likely progress over the follow-up period. Thus, the later transcripts will be more determinative and pronounced in detecting AD.

These findings support previous research that looked at how language deteriorates in patients with Alzheimer's disease over time. Blair et al. (2007) showed the language declines in patients with AD through a longitudinal study. Fraser et al. (2016) demonstrated the relationships between language impairments and structural degeneration in AD using computational techniques and concluded the presence of significant signs of linguistic skills worsening, including semantic impairment, acoustic abnormality, syntactic impairment, and information impairment in patients' speech. In the early stages of Alzheimer's disease, patients have trouble finding words and lose their ability to speak fluently. In the moderate and severe stages of Alzheimer's disease, verbal fluency declines significantly, with numerous literal and semantic paraphrases (Feldman and Woodward 2005, Tang-Wai and Graham 2008).

At the sentence level, we provide 10 sample sentences of individuals with and without AD in Table 5, A and B, along with their corresponding sentence attention score (S_Score) and transcript attention score (T_Score) to illustrate the face validity of our AoS3 model. It can be seen that unlike the sentences in Table 5A, those in Table 5B seem to be affected by language deficit, which implies that those individuals may have suffered from memory loss and AD. Also, the sentences in Table 5B have more of the CHAT disfluency codes (such as those summarized in Table 1), which correspond to word repetition, word revision, phrase revision, filled pause, and pause disfluencies.

Finally, at the word level, we can see that the transcript sample of the individual with AD (Figure 5(b)) contains particular words that may suggest memory loss, such as "uh," "um," "oh," "maybe," and "xxx" (unintelligible words), as well as some words irrelevant to the picture (e.g., youngster) (Figure 6(b)).

An interpretable text-based AI model such as the one we developed in this study may have potential applications beyond predicting Alzheimer's disease. Several studies have demonstrated the potential of text-based AI

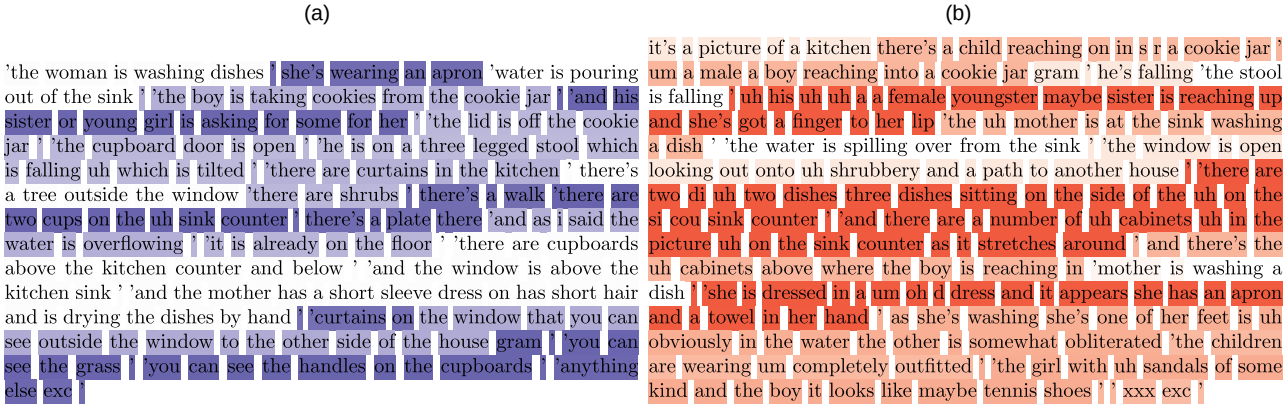
Table 5. Sample Sentences with High Attention Scores for Individuals with and Without AD

No.	Sentence	S_Score	T_Score
	Panel A. Healthy individuals		
1	&uh the mother's drying dishes but the water is overflowing onto the floor.	0.13	0.41
2	the wind is blowing the curtains.	0.12	0.40
3	the mother is drying dishes and has the water turned on.	0.12	0.41
4	mother is drying dishes and the tap water is overflowing the sink and running on the floor.	0.12	0.41
5	&uh the mother is drying the dishes as the sink faucet has filled the sink bowl and is running over onto the floor.	0.11	0.54
6	mother [//] the reason the water's flowing out over the sink is because the water is running furiously &um and I'm looking out through the window.	0.11	0.42
7	at the point she's drying dishes ¡the water¿ [//] perhaps from the noise the water is spilling over the sink and onto the floor.	0.11	0.76
8	the mother's spilling the water and also drying the dishes.	0.11	0.77
9	'the boy is reaching for cookies and the stool is falling over.	0.11	0.37
10	the mother's drying the dishes, frowning but not turning off the faucet.	0.11	0.64
	Panel B. Individuals with AD		
1	and she's havin(g) problems because the sink's running over and she's standing in a puddle of water, some empty dishes on the counter.	0.10	0.31
2	she's drying [//] washing and drying dishes.	0.10	0.47
3	the tea cloth is drying the dishes.	0.10	0.58
4	well he's reachin(g) for the cookie &=laughs but he's handing the cookie to her", 'in the [/] the meantime the stool is falling over.	0.10	0.27
5	and mama's drying the dishes as usual for mamas (.)	0.10	0.25
6	&hm &hm (..) I don't see much more than that other than the kid's falling off of a stool.	0.10	0.27
7	and the [//] I guess it's the mother is drying dishes.	0.09	0.47
8	but the water is &flow still flowing.	0.09	0.40
9	<and the mother is> [//] well she's spilling her water which is not very good but she's doing [//] washing dishes and drying them.	0.08	0.21
10	the water is overflowing from the faucet into the sink onto the floor, ¡while she wipes¿ [//] &uh while she dries &uh a dish.	0.07	0.49

Note. S_Score, sentence attention score; T_Score, transcript attention score.

models in diagnosing a variety of diseases and conditions. These models can analyze speech patterns, choice of words, and sentiment in written text to help diagnose mental health conditions. For example, the use of NLP and deep learning has been shown to be useful in predicting depression, anxiety, and posttraumatic stress

Figure 5. (Color online) Sample Transcripts for a Healthy Individual and an Individual with Alzheimer's Disease



Notes. Words with a high attention score are highlighted in dark blue (healthy) and dark red (AD) and those that received a medium attention score are highlighted in lighter colors. (a) A healthy individual. (b) An individual with AD.

Table 6. Comparison of AD Detection Methods on the Interview Transcripts of DementiaBank

Model	Accuracy	Recall	AUC	Data split	Balanced data (Y/N)	Interpretability (Y/N)
Wankerl et al. (2017)	0.77	—	—	NA	N	N
Fritsch et al. (2019)	0.86	—	—	Leave-one-out	N	N
Chen et al. (2019)	0.97	—	—	NA	N	N
Pan et al. (2019)	—	0.85	—	80% training, 10% validation, 10% testing	N	N
Fraser et al. (2016)	0.82	—	—	90% training, 10% validation, no testing	N	Y
Kong et al. (2019)	0.87	0.90	—	90% training, 10% validation, no testing	NA	Y
Orimaye et al. (2018)	—	—	0.83	Leave-pair-out	Y	N
Orimaye et al. (2017)	—	—	0.93	Leave-pair-out	Y	N
AoS3 (ours)	**0.96**	**0.97**	**0.98**	**80% training, 10% validation, 10% testing**	**Y**	**Y**

Note. NA, not available.

disorder (PTSD) from social media posts, interviews, and clinical notes (Dinu and Moldovan 2021, Jiang et al. 2021, Zhang et al. 2022). Clinical notes stored in patients' electronic health records (EHR) are also a valuable source of information and have been used in machine learning and AI models to detect certain conditions, such as detecting persons living with human immunodeficiency virus (HIV) who may not be aware of their status (Ahlström et al. 2019, Marcus et al. 2019). Moreover, these models have demonstrated good performance in tasks such as predicting in-hospital mortality, unplanned readmission, and hospital length of stay based on clinical notes and EHR data (Rajkomar et al. 2018).

Although our numerical results are promising, they are limited by the size of our data set. Although DementiaBank is one of the best publicly available data sets of its kind, and it has been used in several prior studies to detect Alzheimer's disease, it is a relatively small data set. However, evaluating the performance of our AoS3 model on DementiaBank allowed us to compare our model to some of the previous benchmark models, all of which used the same data set. In general, deep learning methods train better on large amounts of data. However, it is common in healthcare to work on smaller data sets, given the challenges of obtaining patient-level data. Although we conducted crossvalidation to address this issue, further validation of our method on larger data sets is needed to evaluate the superior performance of our deep learning model. Furthermore, while this study focused on diagnosing Alzheimer's disease, future research could work on developing deep learning models to forecast the time to onset of AD, allowing clinicians to provide preventative measures.

In summary, we developed a unique three-level hierarchical attention over self-attention (AoS3) deep recurrent neural network model and presented an important application of it to detect the presence of Alzheimer's disease using longitudinal interview data. We demonstrated that by employing deep learning, we can unlock the patterns within the natural text that signal memory loss due to Alzheimer's disease. Furthermore, we have illustrated that a three-level AoS3 mechanism on words, sentences, and transcripts can shed light on how the model comes up with the predictions. This is extremely valuable, in particular for healthcare applications, as transparency of a predictive model is key to its adaptability to everyday clinical practice.

Figure 6. (Color online) Word Cloud with Attention Score > 0.03

Notes. (a) A healthy individual. (b) An individual with AD.

References

Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N (2019) Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine* 17:100203.

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, et al. (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7(3):270–279.

Alzheimer's Association (2018) 2018 Alzheimer's disease facts and figures. *Alzheimers Dement.* 14(3):367–429.

Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL (1994) The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51(6):585–594.

Birks JS (2006) Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database Syst. Rev.* 2006(1):CD005593.

Blair M, Marczinski CA, Davis-Faroque N, Kertesz A (2007) A longitudinal study of language decline in Alzheimer's disease and frontotemporal dementia. *J. Internat. Neuropsych. Soc.* 13(2):237–245.

Bordes A, Chopra S, Weston J (2014) Question answering with subgraph embeddings. Preprint, submitted June 14, https://arxiv.org/abs/1406.3676.

Chen J, Zhu J, Ye J (2019) An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech. *Proc. Interspeech 2019* (International Speech Communication Association, Baixas, France), 4085–4089.

Chien YW, Hong SY, Cheah WT, Yao LH, Chang YL, Fu LC (2019) An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Sci. Rep.* 9(1):1–10.

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN Encoder-Decoder for statistical machine translation. Preprint, submitted June 3, https://arxiv.org/abs/1406.1078.

Choi E, Bahadori MT, Kulas J, Schuetz A, Stewart W, Sun J (2016) RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems 29* (Neural Information Processing Systems Foundation, Inc., La Jolla, CA), 3512–3520.

Chollet F (2021) *Deep Learning with Python*, 2nd ed. (Manning Publications, Shelter Island, NY).

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. *J. Machine Learn. Res.* 12(Aug):2493–2537.

Damian AM, Jacobson SA, Hentz JG, Belden CM, Shill HA, Sabbagh MN, Caviness JN, Adler CH (2011) The Montreal cognitive assessment and the mini-mental state examination as screening instruments for cognitive impairment: Item analyses and threshold scores. *Dement. Geriatr. Cogn. Disord.* 31(2):126–131.

Dinu A, Moldovan AC (2021) Automatic detection and classification of mental illnesses from general social media texts. Angelova G, Kunilovskaya M, Mitkov R, Nikolova-Koleva I, eds. *Proc. Internat. Conf. Recent Advances in Natural Language Processing (RANLP 2021)* (INCOMA Ltd., Shoumen, Bulgaria), 358–366.

El-Sappagh S, Alonso JM, Islam S, Sultan AM, Kwak KS (2021) A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.* 11(1):1–26.

Fabrizio C, Termine A, Caltagirone C, Sancesario G (2021) Artificial intelligence for Alzheimer's disease: Promise or challenge? *Diagnostics (Basel)* 11(8):1473.

Fan F, Xiong J, Wang G (2020) On interpretability of artificial neural networks. Preprint, submitted January 8, https://arxiv.org/abs/2001.02522.

Feldman H, Woodward M (2005) The staging and assessment of moderate to severe Alzheimer disease. *Neurology* 65(6 suppl 3):S10–S17.

Feng C, Cai F, Chen H, de Rijke M (2018) Attentive encoder-based extractive text summarization. *Proc. 27th ACM Internat. Conf. Inform. Knowledge Management* (Association for Computing Machinery, New York), 1499–1502.

Fraser KC, Meltzer JA, Rudzicz F (2016) Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49(2): 407–422.

Fritsch J, Wankerl S, Nöth E (2019) Automatic diagnosis of Alzheimer's disease using neural network language models. *Proc. ICASSP 2019-2019 IEEE Internat. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Piscataway, NJ), 5841–5845.

Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38(3):50–57.

Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, et al. (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* 29(6):82–97.

Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? Preprint, submitted December 28, https://arxiv.org/abs/1712.09923.

Jiang ZP, Zomick J, Levitan SI, Serper M, Hirschberg J (2021) Automatic detection and prediction of psychiatric hospitalizations from social media posts. Goharian N, Resnik P, Yates A, Ireland M, Niederhoffer K, Resnik R, eds. *Proc. Seventh Workshop Comput. Linguistics and Clinical Psychology: Improving Access* (Association for Computation Linguistics, Stroudsburg, PA), 116–121.

Karlekar S, Niu T, Bansal M (2018) Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. Preprint, submitted April 17, https://arxiv.org/abs/1804.06440.

Katzman R (1976) The prevalence and malignancy of Alzheimer disease: A major killer. *Arch. Neurol.* 33(4):217–218.

Kim J, Seo J (2017) Human understandable explanation extraction for black-box classification models based on matrix factorization. Preprint, submitted September 18, https://arxiv.org/abs/1709.06201.

Kong W, Jang H, Carenini G, Field T (2019) A neural model for predicting dementia from language. *Proc. Machine Learn. Res.* 106:270–286.

Laske C, Sohrabi HR, Frost SM, López-de Ipiña K, Garrard P, Buscema M, Dauwels J, et al. (2015) Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimers Dement.* 11(5):561–578.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Lin Z, Feng M, Nogueira dos Santos C, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. Preprint, submitted March 9, https://arxiv.org/abs/1703.03130.

Logan R, Zerbey SS, Miller SJ (2021) The future of artificial intelligence for Alzheimer's disease diagnostics. *Adv. Alzheimer Dis.* 10(4):53–59.

Lopez-de Ipiña K, Alonso JB, Solé-Casals J, Barroso N, Faundez-Zanuy M, Ecay-Torres M, Travieso CM, Ezeiza A, Estanga A (2012) Alzheimer disease diagnosis based on automatic spontaneous speech analysis. *IJCCI 2012: Proc. 4th Internat. Joint Conf. Comput. Intelligence* (SciTePress, Barcelona), 698–705.

Ma X, Hovy E (2016) End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. Preprint, submitted March 4, https://arxiv.org/abs/1603.01354.

Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J (2017) Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge*

Discovery and Data Mining (Association for Computing Machinery, New York), 1903–1911.

MacWhinney B (2000) *The CHILDES Project: Tools for Analyzing Talk, vol. II: The Database*, 3rd ed. (Psychology Press, New York).

Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE (2019) Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: A modelling study. *Lancet HIV* 6(10):e688–e695.

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Preprint, submitted January 16, https://arxiv.org/abs/1301.3781.

Mitchell AJ (2009) A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J. Psychiatr. Res.* 43(4):411–431.

Mortamais M, Ash JA, Harrison J, Kaye J, Kramer J, Randolph C, Pose C, et al. (2017) Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimers Dement.* 13(4):468–492.

Neyshabur B, Tomioka R, Srebro N (2014) In search of the real inductive bias: On the role of implicit regularization in deep learning. Preprint, submitted December 20, https://arxiv.org/abs/1412.6614.

Olden JD, Jackson DA (2002) Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* 154(1–2):135–150.

Orimaye SO, Wong JSM, Wong CP (2018) Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS One* 13(11):e0205636.

Orimaye SO, Wong JS, Golden KJ, Wong CP, Soyiri IN (2017) Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics* 18(1):34.

Pan Y, Mirheidari B, Reuber M, Venneri A, Blackburn D, Christensen H (2019) Automatic hierarchical attention neural network for detecting AD. *Proc. Interspeech 2019* (International Speech Communication Association, Baixas, France), 4105–4109.

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. Moschitti A, Pang B, Daelemans W, eds. *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA), 1532–1543.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, et al. (2018) Scalable and accurate deep learning with electronic health records. *NPJ Digit. Medicine* 1(1):18.

Sainath TN Mohamed AR, Kingsbury B, Ramabhadran B (2013) Deep convolutional neural networks for LVCSR. *Proc. 2013 IEEE Internat. Conf. Acoustics, Speech and Signal Processing* (IEEE, Piscataway, NJ), 8614–8618.

Saleem TJ, Zahra SR, Wu F, Alwakeel A, Alwakeel M, Jeribi F, Hijji M (2022) Deep learning-based diagnosis of Alzheimer's disease. *J. Personalized Medicine* 12(5):815.

Salton G, McGill MJ (1983) *Introduction to Modern Information Retrieval* (McGraw-Hill Book Co., New York).

Ševčík A, Rusko M (2022) A systematic review of Alzheimer's disease detection based on speech and natural language processing. *Proc. 2022 32nd Internat. Conf. Radioelektronika (RADIOELEKTRONIKA)* (IEEE, Piscataway, NJ), 01–05.

Sha Y, Wang MD (2017) Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. *Proc. 8th ACM Internat. Conf. Bioinformatics, Computational Biology, and Health Informatics* (Association for Computing Machinery, New York), 233–240.

Subasi A (2020) Use of artificial intelligence in Alzheimer's disease detection. Barh D, ed. *Artificial Intelligence in Precision Health* (Academic Press, New York), 257–278.

Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. *Proc. Machine Learn. Res.* 28(3):1139–1147.

Tang-Wai DF, Graham NL (2008) Assessment of language function in dementia. *Geriatrics Aging* 11(2):103–110.

Wankerl S, Nöth E, Evert S (2017) An N-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language. *Proc. Interspeech 2017* (International Speech Communication Association, Baixas, France), 3162–3166.

Williams JA, Weakley A, Cook DJ, Schmitter-Edgecombe M (2013) Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. *Proc. Workshops Twenty-Seventh AAAI Conf. Artificial Intelligence* (AAAI, Washington, DC), 71–76.

Yang J, Zhang Y (2018) Ncrf++: An open-source neural sequence labeling toolkit. *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA). https://aclanthology.org/P18-4013/.

Yang Q, Li X, Ding X, Xu F, Ling Z (2022) Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimers Res. Ther.* 14(1):1–16.

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. *Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Stroudsburg, PA), 1480–1489.

Zhang T, Schoene AM, Ji S, Ananiadou S (2022) Natural language processing applied to mental illness detection: A narrative review. *NPJ Digit. Medicine* 5(1):46.

Zimmerer VC, Wibrow M, Varley RA (2016) Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *J. Alzheimers Dis.* 53(3):1145–1160.