INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

EEG-based Short-time Auditory Attention Detection using Multi-task Deep Learning

Zhuo Zhang[1], *Gaoyan Zhang*[1], *Jianwu Dang*[1,2], *Shuang Wu*[1], *Di Zhou*[1,2], *Longbiao Wang*[1]

[1]College of Intelligence and Computing, Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]Japan Advanced Institute of Science and Technology, Japan

zhuozhang@tju.edu.cn, zhanggaoyan@tju.edu.cn, jdang@jaist.ac.jp, longbiao_wang@tju.edu.cn

Abstract

A healthy person can attend to one speech in a multi-speaker scenario, however, this ability is not available to some people suffering from hearing impairments. Therefore, research on auditory attention detection based on electroencephalography (EEG) is a possible way to help hearing-impaired listeners detect the focused speech. Many previous studies used linear models or deep learning to decode the attended speech, but the cross-subject decoding accuracy is low, especially within a short time duration. In this study, we propose a multi-task learning model based on convolutional neural networks (CNN) to simultaneously perform attention decoding and reconstruct the attended temporal amplitude envelopes (TAEs) in a 2s time condition. The experimental results show that, compared to the traditional linear method, both the subject-specific and cross-subject decoding performance showed great improvement. Particularly, the cross-subject decoding accuracy was improved from 56% to 82% in 2s condition in the dichotic listening experiment. Furthermore, it was found that the frontal and temporal regions of the brain were more important for the detection of auditory attention by analyzing the channel contribution map. In summary, the proposed method is promising for nerve-steered hearing aids which can help hearing-impaired listeners to make faster and accurate attention detection.

Index Terms: EEG, brain-computer interface (BCI), auditory attention detection, multi-task learning, the cocktail party problem

1. Introduction

The cocktail party problem has been a research hotspot since it was proposed in 1953 [1]. Understanding speech in a noisy environment requires not only the auditory system to separate simultaneous speech streams but also the ability to focus on a specific speech stream while suppressing irrelevant information. Although the normal-hearing person can easily analyze complex speech scenes, some people with hearing impairments (such as sensorineural hearing loss) encounter difficulties in understanding speech in a noisy environment, even if they wear auditory prostheses (such as hearing aids or cochlear implants) [2]. Therefore, some studies proposed to use the EEG signals to detect the speech stream that the listener wants to focus on in advance, which will promote the development of nerve-steered cochlear implants to help these hearing-impaired listeners [3, 4, 5, 6, 7].

When focusing on one of the mixed speeches of two speakers concurring at the same time, it is possible to detect which speaker the listener is following from the brain activities [8, 3], this detection is called auditory attention detection (AAD). Most of the previous studies proposed an idea that reconstructing the temporal amplitude envelopes (TAEs) of the speech from neural signals and compare the correlation of reconstructed TAEs with the presented attended and unattended speech to detect the auditory attention. They supposed that if one attended to the speech, the reconstructed correlation is higher in the attended side than the unattended side. This idea based on a hypothesis that the cortical signal can track the TAEs of the acoustic speech signal. Numerous electrophysiological studies have shown that the slower than 20 Hz time-domain fluctuations in speech signals are synchronized with low-frequency cortical activity in the delta (1-4 Hz) and theta (4-8 Hz) frequency bands [9, 10]. This idea has been conceptualized through theoretical models [11] and has been verified through extensive experimental research.

Common methods to record brain activity include EEG, Magnetoencephalography (MEG), and Functional magnetic resonance imaging (fMRI). Among all these physiological signals, EEG signals can directly reflect variation in human activities. At the same time, EEG is a non-invasive device with high temporal resolution that is ideal for capturing these fast, dynamic, and chronological cognitive events [12]. Therefore, many previous studies have tried to construct a model that can account for relationships between EEG signals and the TAEs of the attended speech and detect auditory attention. Many studies used linear models to decode the EEG signals in a two-speaker environment, that is, to reconstruct the TAEs of stimulus by the multivariate temporal response function (mTRF) method [3, 4, 7]. The accuracy of classification by comparing the correlation is about 83.9% when subject-specific in the 60s time condition [3, 4].

Recently, deep learning technologies have shown to use automatic feature extraction to achieve competitive accuracy for EEG decoding which has been widely used in the field of BCI [13, 14]. There was a study proposed to use a deep neural network model (DNN) to reconstruct the TAEs of interest speech. Compared with the linear model, this method improves the decoding accuracy to 97.6% in the 60s condition, but the accuracy of the 2s condition is about 67.8% [6]. In addition to stimulus reconstruction, many studies directly classify EEG signals. Some researchers used CNN to classify smaller periods (1-2s) to increase the accuracy to about 80% in the subject-specific condition [5]. These previous studies had problems that the classification accuracy is not so high, especially under cross-subject condition, which limited the model generalization in the practical application of the smart hearing aids.

The main purpose of this study is twofold: First, we proposed the multi-task learning model based on CNN to simultaneously detect auditory attention and reconstruct the attended

http://dx.doi.org/10.21437/Interspeech.2020-2013

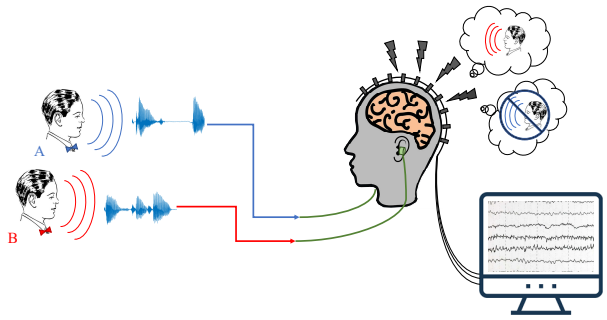

Figure 1: *The dichotic listening experiment.*

TAEs in a 2s time duration, because the CNN can extract the deep features of EEG signals and has good performance [5, 13, 14]. It is shown that the proposed multi-task learning frame can improve the decoding accuracies in both subject-specific and cross-subject conditions within a short time window of 2s. Secondly, through the Layer-Wise Relevance Propagation (LRP) method, we conclude that the temporal and frontal regions of the brain have made major contributions to auditory attention detection.

2. Materials

2.1. Experiment Setup

The dataset was collected by O'Sullivan et al. [3]. Forty subjects participated in the dichotic listening experiment. Participants reported no history of hearing impairment or neurological disease and were right-handed. In the experiment, the subject listened to two classic fictions that were played concurring, and different stories were played in the left and right ear respectively through the Sennheiser HD650 headphone. The story played in each trial is continuous with the previous trial. Another male speaker read each story. Participants were averagely divided into 2 groups, and each group listened to one of the stories in the left or right ear and ignored the other. After each trial, the subjects were required to answer 4 to 6 multiple-choice questions about the story to check the concentration. The amplitude of each speech stream was normalized to the same root mean square (RMS) intensity. All silence intervals of audio were shortened to 0.5 s to prevent the subject from distracting. Moreover, subjects should fix their eyes on the crosshairs on the screen during each trial and minimized blinks, head movements, and all other motor activities. During the experiment, the BioSemi Active Two system [3] was used to record the EEG signal.

2.2. Data Acquisition and Preprocessing

The experimental dataset published the EEG data and TAEs of 32 subjects, as shown in Table 1. Half of them attended to one story presented in the left ear. The other half subjects attend to the other story played in the right ear. There were 30 trials for each subject, and each trial recorded EEG signal for one minute, and the EEG signal was down-sampled to 128 Hz. The EEG data have 128 channels and two mastoid channels.

To remove line noise, the EEG data was filtered over the range of 1–45 Hz. Then, the Cleanline is used to eliminate the remaining line noise. Artifacts of the EEG channel were corrected by using Artifact Subspace Reconstruction (ASR). Moreover, bad channels were rejected and then interpolated. Finally, all EEG channels were re-referenced to the average of EEG data.

Table 1: *The format of the public dataset*

EEG data	Subject	32
	Channels	128
	Trial	30 (each trial 1 min)
	Sampling Rate	128 Hz
	Data format	$(128 \times 7680 \times 30)$ $(channel \times timepoint \times trial)$
Stimul data	Left Audio	'Twenty Thousand Leagues Under the Sea' (Subject 1-7, 9-17)
	Right Audio	'Journey to the Centre of the Earth' (Subject 18-33)
	Sampling Rate	128 Hz
	Data format	$2 \times 7680 \times 30$ $(audio \times timepoint \times trial)$

The TAEs of the speech signal were extracted by Hilbert transform and were kept to the same sampling rate of 128 Hz so that we can relate TAEs' dynamics to the dynamics of EEG. We used a 2s sliding window to intercept the corresponding EEG signals and speech stimuli and the window overlap rate is set as 50% so that this study can detect auditory attention in a shorter time. The training input sample format for each EEG is $128\,channel \times 256\,timepoints$ and the format of stimuli is $2\,direction\,(attended, unattended) \times 256\,timepoints$.

3. Methods

3.1. Multivariate Temporal Response Function Model

The mTRF model assumes the human brain as a linear convolution system where the temporal response functions describing a mapping between speech stimulus and neural response in both encoding and decoding directions. [4, 15].

When using mTRF for backward decoding, the method can be described as follows: $r(t, n)$ expressed the EEG response at time $t = 1 \cdots T_r$ recorded by the channel n, $n = 1 \cdots N$. The stimulus is represented by $s(t)$, $t = 1 \cdots T_s$. At the corresponding time $t = 1 \cdots T$, a linear model is constructed as follow:

$$\hat{s}(t) = \sum_{n} \sum_{\tau} r(t - \tau, n) g(\tau, n), \tag{1}$$

where $\hat{s}(t)$ represents the reconstruction stimulus, and $g(\tau, n)$ represents the parameter that needs to be estimated by minimizing the distance between the original stimulus $s(t)$ and the reconstruction stimulus $\hat{s}(t)$. We use root mean square as a formula to measure the distance between $s(t)$ and $\hat{s}(t)$:

$$\min\, \varepsilon(t) = \sum_{t} [s(t) - \hat{s}(t)]^2. \tag{2}$$

In this experiment, the number of channels N is 128, and the original stimulus s is TAEs of attended speech. We used the Subject-Specific and Grand-average method proposed by [3] to train each sample of $r(t, n)$ and corresponding stimuli s in order to obtain the corresponding parameter g. Finally, all the g obtained from each sample were averaged as the final parameter $\bar{g}$ to test the testing set. The model can be regarded as a linear regression method to compare with our proposed model.

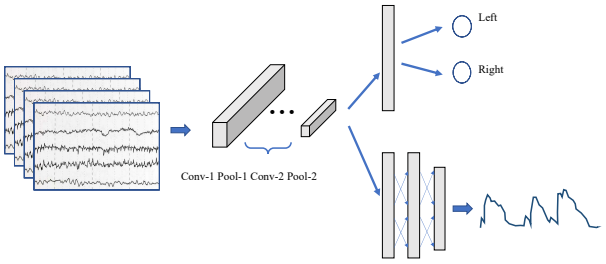

Figure 2: *Proposed CNN-based D + R model.*

3.2. Proposed CNN-based Multi-task Learning Model

CNN can automatically extract abundant features through layer-by-layer convolution. The lower layer extracts local and spatial features, such as boundaries. And higher layer can extract global features, such as complex shapes and complete objects. EEG spatial maps can be regarded as a two-dimensional image, one dimension is the temporal context, and the other is the electrode spatial dimension. Therefore, our study employed CNN to perform automatic feature extraction on EEG signals. The proposed multi-task learning model (D + R model) innovatively combined the task of attended TAEs reconstruction with the classification task for the left and right direction by sharing hidden convolutional layers. The model used the task of stimulus reconstruction to facilitate the learning of the classification task for direction. As shown in Figure 2, considering both decoding performance and efficiency, we used two convolutional layers after the EEG signal input layer and each of them connected with the max-pooling layer and the Relu function in turn. The convolution kernel size of the first convolution layer is 3×3, a total of 16 filters, the kernel size of second layer is 3×3 with a total of 32 filters. The window size of the max-pooling layer is 2×2, and the stride of the window in each dimension is 2. The max-pooling layer makes our convolutional features more robust, and the Relu function, which represents a non-linearly activated neuron, prevents gradient exploding and gradient vanishing problem. These layers were shared by two tasks as hidden layers. After them, the classification task was followed by a fully connected layer with a dropout of 0.5, and then a softmax layer, which outputs a 1×2 matrix represented the direction. The cost function of this task is the cross-entropy function. The TAEs reconstruction task is to connect three fully connected layers after the hidden layer and the numbers of neurons of which are 1024, 512, and 256, respectively. The dropout is set to 0.3 for each fully connected layer to prevent it from overfitting. The final fully connected layer outputted a 1×256 reconstruction TAEs matrix with the same dimensions as the original TAEs. The root mean square error (MSE) is set as the cost function:

$$C(\hat{e}_a, e_a) = \overline{[\hat{e}_a - e_a]^2}, \tag{3}$$

where $\hat{e}_a$ is the reconstructed TAEs and e_a is the original TAEs of the attended stimuli. The simple and efficient Adam optimizer was used in this model. The convolution layers use two-dimensional convolution kernels to automatically extract the features of the EEG signal on the temporal context and channel spaces respectively.

3.3. Visualization by Layer-wise Relevance Propagation

The Layer-wise Relevance Propagation (LRP) algorithm is proposed as a model visualization method [16]. In brief, the main idea of the LRP algorithm is to calculate the relevance of each input node layer by layer to the effect on the final output node [17]. The total relevance of each layer is the same, and the relevance is transmitted between the layers. Each node of the layer l which transmitted to the connected node j in the subsequent layer $l+1$ share part of relevance. In general, the LRP assumes that we have a relevance score R for each dimension of the vector on layer $l+1$. The idea is to find a relevance score R in each dimension of the next vector. Layer l closer to the input layer makes the following equation:

$$\sum_i R_{l,l+1}^{i \to j} = R_{l+1}^j. \tag{4}$$

In this study, we used the following basic LPR rules to traverse the network [18]:

$$R_{l,l+1}^{i \to j} = \sum_j \frac{z_{i,j}}{\sum_i z_{i,j}} R_{l+1}^j, \tag{5}$$

where i and j are two neurons in any continuous layer. The quantity $z_{i,j}$ models the extent to which neuron i has contributed to make neuron j relevant. We know the relevance of the output layer, so we will start there and iteratively use this formula to calculate R for each neuron in the previous layer.

3.4. Model Training and Testing

To compare with the proposed D + R model, in addition to training the mTRF model on the training and testing set, we also increased a CNN-based direction binary classification model (D model) to further test the performance of the proposed D + R multi-task learning model. We compared the subject-specific and cross-subject decoding performance in 2s time duration with the linear mTRF model, D model, and the proposed D + R model, respectively.

First, we divided the data of the same subject into 9: 1 as the training set and the testing set, and trained the data of 32 subjects respectively under the subject-specific conditions. Secondly, for the training of cross-subject, we randomly selected about 10% of the data from all 32 subjects as the test set (a total of 3 subjects). We shuffled the remaining 90% of the data in the temporal dimension and subject dimension (a total of 29 subjects) to use as a training set. EEG signals vary significantly between different subjects. Therefore, cross-subject training is essential for practical applications. There is no subject overlap between the samples in the testing set and the training set, which means that cross-subject training is achieved.

4. Results and Discussion

4.1. Experiment 1: Subject-Specific Training

The results of classification accuracy for different subjects in this experiment are shown in Figure 3. The mTRF linear model has a very poor classification effect under 2s conditions, and the average accuracy is only 56.41% as shown in Table 2. In contrast, the CNN based nonlinear models achieve great classification results under 2s duration. Among them, the classification accuracy of the D model is about 89.83%, and the average accuracy of the proposed D + R model reaches 90.68%. Since the nonlinear classification model can already obtain very high classification accuracy, in that case, the accuracy of the proposed D + R model dose not have great improvement in the subject-specific condition. However, for each subject, the classification accuracy of our proposed D + R model is improved compared to the D model.

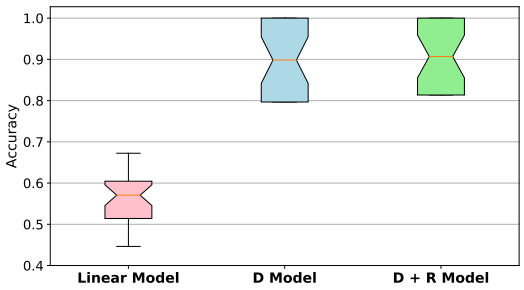

Figure 3: *Classification accuracy under the subject-specific condition.*

4.2. Experiment 2: Cross-Subject Training

The experiment results of cross-subject training are shown in Table 2. Due to the huge difference in EEG signals between subjects, the classification accuracy of these three models under the cross-subject conditions is lower than that of the subject-specific condition. The results have shown that the mTRF linear model had a poor effect of about 55.76%, however, the cross-subject decoding accuracy of the D + R model can reach 82.38% under 2s condition. Compared with the mTRF model and D model, the D + R model has higher classification accuracy in 2s condition when crossing the subject which also indicated that CNN had a good effect in extracting the EEG features.

But it is worth noting that for the same network structure, the R model that only performs the reconstruction task, the classification accuracy is close to the chance level under both conditions of subject-specific and cross-subject. The main reason for this result of the R model may be that the length of time of a single sample is small and the output dimension is large. Perhaps the RNN model that are more sensitive to time series can be more effective for the reconstruction task. This requires further research in the future.

Table 2: *The decoding accuracy using different models*

Model	Accuracy	
	subject-specific	cross-subject
Liner model	0.5614	0.5576
D model	0.8983	0.7905
D + R model	0.9068	0.8238

4.3. Experiment 3: Channel Contribution

Figure 4 shows the LRP result of the D + R model in cross-subject condition. We averaged the relevance matrix corresponding to 5310 samples of the test set and found some channels have a relatively higher relevance. In order to show it more clearly, we normalize the matrix to (0, 1), as shown in Figure 4A. We averaged the matrix along the time axis and map to the positions of different channels on the scalp to obtain the scalp topology results shown in Figure 4B. As can be seen from this weight map, the relevance between the electrodes in the temporal lobe and frontal lobe is higher, which means that the temporal lobe and frontal lobe may contribute more to the task.

5. Conclusion

In this paper, we focus on achieving continuous EEG signal classification under a short time of 2s in cross-subject conditions to detect auditory attention. We propose a multi-task learning model based on CNN which combines the binary classification task for direction and the TAEs reconstruction task by sharing hidden layers to co-training model parameters to improve the classification accuracy. To that end, we compared three types of models, including the linear mTRF model, CNN-based D model, and the proposed D + R multi-task learning model. By comparing and analyzing our experimental results, it is proved that our proposed model can achieve better classification accuracy under both subject-specific and cross-subject conditions compared with the other two models in a short time of 2s. Compared with the traditional mTRF linear model, we increased the classification accuracy from 55.76% to 82.38% under the cross-subjects condition. Moreover, we obtained the contribution of all the channels through calculating the inter-layer relevance of the proposed model using the LRP algorithm and found that the temporal lobe and part of the frontal lobe channels contributed greatly to this AAD issue.

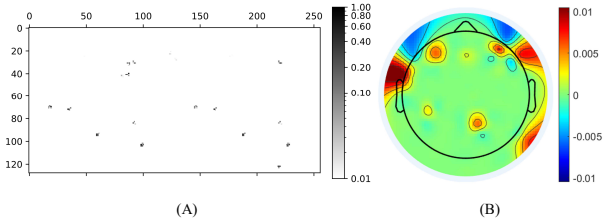

Figure 4: *Relevance matrix (A) and channel scalp contribution map (B).*

However, this study also has some shortcomings. First, the stimuli data of this public data set is TAEs that have been extracted by Hilbert transform but not the original audio so that we cannot make more improvements in the process of extracting the TAEs. Secondly, in order to be closer to the actual application, we did not specifically remove the blinking, eye movement and other motor artifacts in the preprocessing of the EEG signals, which led to the LRP results may be interfered by blinking and eye movement so that we cannot get more accurate channel contribution results. Finally, the experimental conditions we compared are insufficient. We will continue to discuss more experimental conditions in the future. Such as discussing classification accuracy under different time conditions, and analyzing whether our proposed model can behave good performance in different time conditions. Moreover, we will discuss the impact of the different frequency bands of the EEG signal on this experimental result, and detect whether our proposed model has robustness by using different data sets.

In conclusion, this study can achieve better accuracy in decoding EEG signals in a short-time when crossing the subjects which can be better applied to nerve-steered smart hearing aids or cochlear implants to help people with hearing impairments in the future. Auditory attention detection opens up new research possibilities in the fields of neuroscience, audiology, and signal processing, and it also plays an important role in the study of speech separation and in exploring the neural basis of speech separation in noisy environments.

6. Acknowledgements

This study was supported by National Natural Science Foundation of China (No.61876126) and JSPS KAKENHT Grant (No. 16K00297).

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] M. Armstrong, P. Pegg, C. James, and P. Blamey, "Speech perception in noise with implant and hearing aid." *The American journal of otology*, vol. 18, no. 6 Suppl, pp. S140–1, 1997.

[3] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[4] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[5] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "Eeg-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, 2018.

[6] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[7] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, vol. 13, p. 153, 2019.

[8] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[9] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, 2001.

[10] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.

[11] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, p. 511, 2012.

[12] M. X. Cohen, *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[13] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung, and C.-T. Lin, "Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *arXiv preprint arXiv:2001.11337*, 2020.

[14] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[15] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[16] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[17] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.

[18] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.