2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Towards Explainable Diagnosis of Alzheimer's

Mariia Sidulova*, Chung Hyuk Park*
*Assistive Robotics and Tele-Medicine (ART-Med) Laboratory at The George Washington University

***Abstract*—Alzheimer's disease (AD) has become one of the most common neurodegenerative disorder, currently affecting nearly twenty-five million worldwide. There is no cure available at this moment, however early diagnosis, socially assistive robotics therapy showed promising results in slowing down the progression of the diseases. This paper discusses recent developments in early AD diagnosis and non-pharmacological treatment options.**

I. INTRODUCTION

Alzheimer's disease (AD) accounts for 60%-70% of all dementia cases, and its clinical diagnosis at the early stages is very difficult [1]. Even though several novel drugs endeavoring to mitigate disease progression and relieve symptoms are being developed, there is no effective cure yet. This paper proposes a framework that uses EEG signals for early diagnose AD and MCI. Moreover, this study attempts to find an explanation for the predicted diagnosis using the Explainable Artificial Intelligence algorithm (XAI).

II. RELATED WORKS

Alzheimer's disease is a progressive brain disorder. Unfortunately, there is no effective treatment or medication for such a disease, however studies have shown that robot-guided therapy sessions can slow down the progression and ease some of the symptoms. Alzheimer's disease can be diagnosed with complete certainty only after death, however, multiple medical assessment techniques can confirm the risk of acquiring Alzheimer's or dementia. Common clinical assessments include PET, CT, EEG, mental status and neuropsychological testing, blood, and genetic testing. Numerous studies have attempted to develop a user-friendly framework to assist in processing of these clinical assessments.

A. Machine Learning for Early Diagnosis of AD

Since the identification of pathological EEG sequences indicative of AD, there has been extensive investigation into linear and non-linear classification algorithms, or classifiers. Of the studies conducted, a significant portion relies on one-class classifiers. Methods such as Gaussian Naive Bayes, Fisher Linear Discrimination, and Gaussian Processes, as well as Support Vector Machines (SVM) with a linear kernel, maintained average accuracy of 60-65 [2],[3]. Several methods that combine techniques have been investigated in an attempt to streamline performance and improve accuracy. The study by Huagn et al has successfully implemented a new convolutional neural network (CNN) increasing the model training and validation accuracy to a range of 87-96% by adding more convolutional and pooling layers, thus deepening the network [4].

In the medical domain, there is a thriving demand for AI approaches, which are not only accurate but also easily understandable and easily interpretable. Another reason why the explainability of machine learning algorithms is in such great demand is that personalized medicine has grown a lot over the past couple of years. A couple of explainable AI (XAI) approaches have been applied in the medical domain. For example, in the paper by Benjamin Letham, a model called Bayesian Rule Lists, which generates a posterior distribution over possible decision trees, has shown promising results for a better stroke prediction[5]. XAI has also been applied in the field of digital pathology. For example, Chen et al, have proposed a novel framework for a quick and accurate method to detect mitosis. The study showed a novel deep cascaded convolutional neural network, which composed of CNNs and a discrimination model that is assembled by transferring deep and rich feature hierarchies trained by CNN [6].

B. Socially assistive robots therapy for AD

Social agents that strive to assist individuals in health-related tasks have been closely studied in the human-computer interaction (HCI) community. In the study by Francisco Martin, NAO robots were used for therapy for patients with dementia [7]. Therapy sessions were conducted twice a week for a month and preliminary medical assessments on real patients with moderate dementia shown promising results. Their neuropsychiatric symptoms have a tendency to improve compared with patients following classic therapy methods. The robot was able to capture the attention of the elderly due to its human shape, its human-like movements, and its music capabilities.

In the recent study by Juan Fasola, robots took the role of coaches for robot-guided sessions for the elderly population [8]. They were able to personalize the social interaction which resulted in higher engagement and motivation scores. Based on the post-treatment surveys, sessions with the robot physically present in the room resulted in an enjoyable and valuable experience.

III. METHODS

A. Dataset

The original dataset that was used in this study is the dataset from the paper by Cejnek [9]. The set contains EEG data collected from patients with probable Alzheimer's, Mild Cognitive Impairment (MCI), and healthy individuals of the age from 50 to 70. Every EEG recording is of the length 90s or less and was taken when subjects were at rest. The total numbers of recordings used for machine learning algorithms are 284 recordings for AD, 56 recordings for MCI, and 100

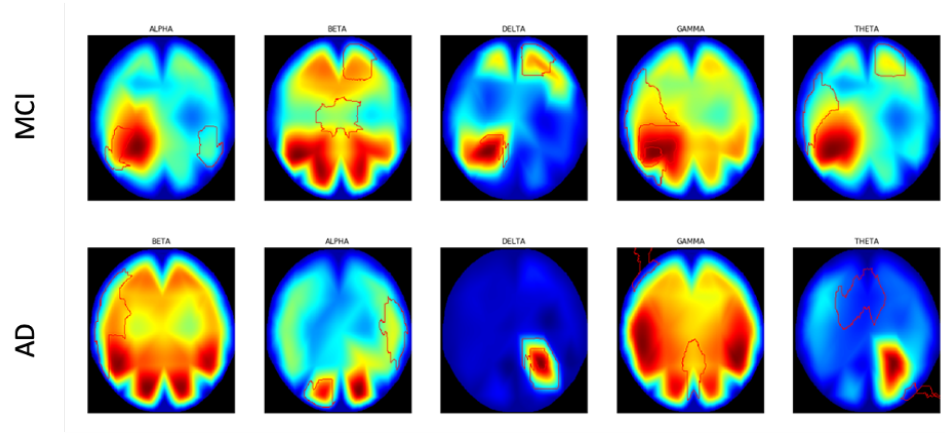

Fig. 1. This figure show sample LIME outputs for one AD and one MCI patient. The areas outlined in red are the areas of the brain that were identified by LIME as representative features for specific diagnosis

recordings for healthy. The EEG electrodes were positioned according to the 10-20 System. The recordings were conducted on a 21-channel digital EEG setup: Walter EEG PL-231, with a sampling frequency of 256 Hz, and TruScan 32, Alien Technik Ltd, with a frequency of 128 Hz.

B. Signal Processing and Visualization Method

A 10th-order lowpass Butterworth filter with a cut-off frequency at 50 Hz was applied to the original signal to remove muscle activity artifacts, which typically consist of frequencies of 300 and higher [10]. Each recording was divided into five frequency bands: Gamma band greater than 30 Hz, Beta band is between 13 Hz and 30 Hz, Alpha band is between 8 Hz and 12 Hz, Theta band is between 4 Hz and 8 Hz, and Delta band is less than 4 Hz [11]. Since EEG recordings were made in the resting state, the mean power of each of the band for each electrode was extracted as a feature. Extracted mean power for each electrode was used to create head-plot spectrograms. Head-plot spectrograms were constructed using the EEG mapping function in MATLAB [12]. This visualization method was used to produce a topographic map of the scalp by cubically interpolating values from each electrode.

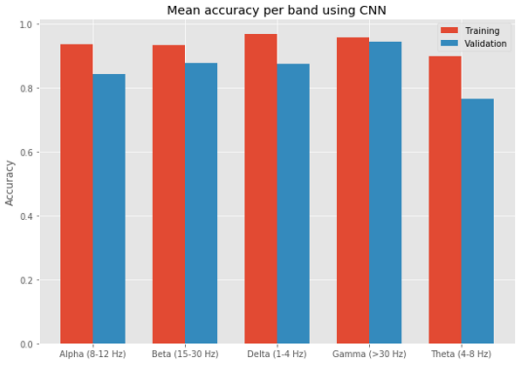

Fig. 2. Summary of all CNN models. Histogram is a summary of the training and validation accuracy for each of the frequency bands. Validation accuracy for all of the models fall into the range between 87 - 97%. Mean validation accuracy 0.89.

C. Machine Learning: CNN and LIME

The generated headplot spectrograms were divided into training and testing sets, and CNN models were trained using Keras packages [13]. Models for each of the frequency bands were trained and used for further Local Interpretable Model-Agnostic Explanations (LIME). To further understand the underlying trends of classification, a LIME network was trained to evaluate the performance of the proposed CNN model. LIME is an algorithm that attempts to explain the model by perturbing the input of data samples and understanding how the predictions change. For each of the explanation instances, LIME creates a new dataset, which consists of slightly altered samples and the corresponding predictions of the pre-trained model. On this newly created dataset, LIME trains an interpretable model, which consists of weighted similarity between sampled instances and the sample that has been inputted to the model. Keras library was used for the implementation of the LIME algorithm [14].

IV. RESULTS

The accuracy for CNN models for each of the spectral band are presented in Figure 2. The highest classification accuracy results were obtained using Gamma frequency band (¿30 Hz).

LIME algorithm was able to locally interpret each of the spectrograms. Sample LIME outputs for AD, MCI, and healthy individuals are presented in figure 1. This approach provides a qualitative insight into CNN classification. For testing the LIME algorithm, sample head-plot spectrograms were randomly chosen from the validation dataset and 400 perturbations are generated per each input. The identified region is the region that has a stronger association with the predicted diagnosis. The areas identified by the algorithm are mainly temporal, frontal lobes and motor cortex are affected in AD and MCI patients.

REFERENCES

1 J. C. Waymire, “Chapter 10: Cns aging and alzheimer’s disease,” *Neuroscience Online*.

2 C. F. B. S. A. B. P. B. C. D. C. Fiscon, Weitschel, "Combining eeg signal processing with supervised methods for alzheimer's patients classification.," *BMC Medical Informatics and Decision Making*.
3 N. K. Al-Qazzaz, S. H. B. Ali, S. A. Ahmad, K. Chellappan, M. Islam, J. Escudero, *et al.*, "Role of eeg as biomarker in the early detection and classification of dementia," *The Scientific World Journal*, vol. 2014, 2014.
4 Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, A. D. N. I. (ADNI, *et al.*, "Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network," *Frontiers in Neuroscience*, vol. 13, p. 509, 2019.
5 B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.
6 H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
7 F. Martín, C. Agüero, J. M. Cañas, G. Abella, R. Benítez, S. Rivero, M. Valenti, and P. Martínez-Martín, "Robots in therapy for dementia patients," *Journal of Physical Agents*, vol. 7, no. 1, pp. 48–55, 2013.
8 J. Fasola and M. J. Matarić, "A socially assistive robot exercise coach for the elderly," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 3–32, 2013.
9 M. Cejnek, I. Bukovsky, and O. Vysata, "Adaptive classification of eeg for dementia diagnosis," in *2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pp. 1–5, 2015.
10 P. Durongbhan, Y. Zhao, L. Chen, P. Zis, M. De Marco, Z. C. Unwin, A. Venneri, X. He, S. Li, Y. Zhao, *et al.*, "A dementia classification framework using frequency and time-frequency features based on eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 826–835, 2019.
11 J. J. Newson and T. C. Thiagarajan, "Eeg frequency bands in psychiatric disorders: a review of resting state studies," *Frontiers in human neuroscience*, vol. 12, p. 521, 2019.
12 I. Silva, "Eegplot." https://www.mathworks.com/matlabcentral/fileexchange/3279-eegplot, 2020.
13 F. Chollet *et al.*, "Keras," 2015.
14 M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.