Quantifying Explainability of Counterfactual-Guided MRI Feature for Alzheimer's Disease Prediction

Kwanseok Oh[1], **Da-Woon Heo**[1], **Ahmad Wisnu Mulyadi**[2],
Wonsik Jung[2], **Eunsong Kang**[2], **Heung-Il Suk**[1,2,*]
[1]Department of Artificial Intelligence, Korea University
[2]Department of Brain and Cognitive Engineering, Korea University
{ksohh, daheo, wisnumulyadi, ssikjeong1, eunsong1210, hisuk}@korea.ac.kr

Abstract

The interpretability of deep learning (DL) for Alzheimer's disease (AD) prediction has provided supporting evidence for the timely intervention of disease progression. In particular, counterfactual reasoning is gradually being employed in the medical field, providing refined visual explanatory maps. However, most visual explanatory maps still rely on visual inspection without quantifying their validity, being a barrier for non-expert individuals. To this end, we propose a novel framework to analyze the counterfactual reasoning-based visual explanation by transforming them into quantitative features. Furthermore, we develop a simple shallow linear classifier to boost the effectiveness of quantitative features while promoting the model's interpretability and achieving superior predictive performance compared to the DL model. By doing so, our method further provides an ADness index that can be used to intuitively comprehend a patient's brain status with respect to AD.

1 Introduction

Alzheimer's disease (AD) is known as one of the most prevalent causes of dementia worldwide, which is an irreversible neurodegenerative disease accompanied by memory loss and impairment of cognitive functions [2]. However, as the currently available medication is merely to delay the AD progression, it is of paramount importance to early distinguish and manage the prodromal or preclinical stage, *i.e.*, mild cognitive impairment (MCI), that begins to decline cognitive function across the symptomatic spectrum from the cognitively normal (CN) [7].

In this light, structural MRI (sMRI)-based AD prediction via deep learning (DL) approaches has shown superior predictive performance against conventional machine learning (ML) techniques by automatically identifying AD-manifested patterns for each patient. Nevertheless, DL models' inherently opaque "black box" nature has the chronic drawback of making it difficult to describe the model's decision. To resolve this issue, explainable artificial intelligence [1] has been increasingly exploited in the field of medical research. In particular, counterfactual reasoning has recently emerged that can provide a qualitative explanation of the model's decision given hypothetical scenarios by generating refined visual explanatory maps—so-called counterfactual maps (CF maps) [9]. However, although those CF maps can visually exhibit brain regions affected by atrophic variations, visual explanations do not quantify clinical validity and still rely on subjective visual inspection. Furthermore, there are objectively interpretive limitations for clinicians to utilize as auxiliary diagnostic information.

In this study, we propose a novel framework for quantifying counterfactual-induced feature representation and interpretable brain regional-specific AD identification. To this end, we adopt the gray matter (GM) density map as the quantitative feature that precisely measures the volumetric changes in

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

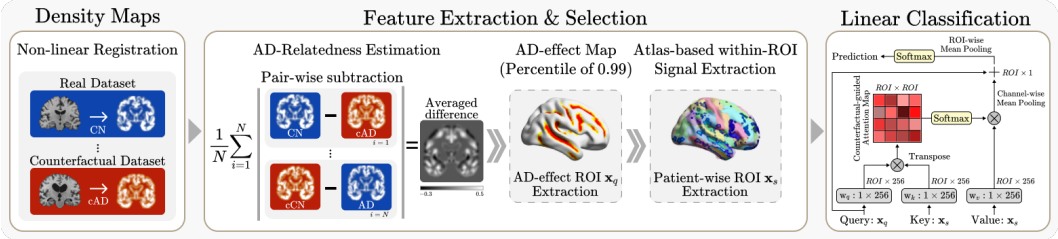

Figure 1: Illustration of GM density-based analysis to quantify the explainability of DL models.

GM, associated with brain atrophies [6]. Furthermore, we devise a lightweight counterfactual-guided attentive feature representation and a linear classifier (LiCoL), which can derive outstanding performance compared to DL models and provide a quantitative interpretation of the model's decision with an *ADness index* that numerically indicates the subregional state of the brain via an attentive score.

2 Method

To construct the counterfactual data (c-sMRI), we exploit the counterfactual map generator proposed in [9] and real data (r-sMRI). Subsequently, acquired r-/c-sMRI are transformed into the GM density map (*i.e.*, r-GM and c-GM) by performing thorough preprocessing and non-linear registration steps.

Given the manipulated r-GM and c-GM, we perform a series of GM density-based in-depth analyses and assessments (Fig. 1). We first estimate the representative difference map by subtracting and averaging r-GM images $\mathbf{X}^r$ from their counterpart c-GM images $\mathbf{X}^c$. The percentile is then calculated to merely highlight the significant voxels. We thus obtain the AD-effect map via $\frac{1}{N}\sum_{i=1}^{N}|\mathbf{X}_i^r - \mathbf{X}_i^c|$, where N and $|\cdot|$ respectively denote the total number of samples and the absolute operation. Subsequently, the AAL3 atlas template [10] is overlaid on the AD-effect map to parcellate it into individual regions, so that we attain trimmed regions and their corresponding voxel indices $\mathcal{V}$. By averaging the voxel values of indices $\mathcal{V}$ within the adopted regions, we eventually define the region of interests (ROIs), so-called AD-effect ROI $\mathbf{x}_q \in \mathbb{R}^{R\times 1}$, which consists of R number of ROIs.

Our LiCoL is built on the self-attention mechanism [11], which is comprised of query, key, and value as inputs to identify the significantly contributed ROIs by itself, considering the global relationship among ROIs. We utilize the query as AD-effect ROIs $\mathbf{x}_q$ whereas the key and value are individually established from each training sample according to the voxel indices $\mathcal{V}$ as $\mathbf{x}_s \in \mathbb{R}^{R\times 1}$. By multiplying the learnable embedding layer $\mathbf{w} \in \mathbb{R}^{1\times D}$ to the query, key, and value, we obtain the embedded matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{R\times D}$, accordingly. Thereafter, we compute the counterfactual-guided attention matrix as $\mathbf{A} = g\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$, where $\top$ and d denote a transpose operation and a scaling factor, respectively, and g denotes the softmax function. Having applied the channel-wise mean pooling $\mathrm{MP}_{\rightarrow}$ to reshape the output of the counterfactual-guided attention $\mathbf{A}$ to match the size of the input query $\mathbf{x}_q$, we then employ a residual connection using an element-wise addition, followed by ROI-wise mean pooling $\mathrm{MP}_{\downarrow}$ to obtain the final prediction score $\hat{\mathbf{y}} = g(\mathrm{MP}_{\downarrow}(\mathrm{MP}_{\rightarrow}(\mathbf{A}) + \mathbf{x}_q))$. Finally, our LiCoL is trained by a cross-entropy (CE) loss against the ground truth $\mathbf{y}$, *i.e.*, $\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}[\mathrm{CE}(\hat{\mathbf{y}}, \mathbf{y})]$.

Table 1: Results of sorted top-10 ROIs that contributed the most to prediction among AD-effect ROIs according to averaged attentive scores. L/R indicate the left and right hemisphere, respectively.

Order	**CN** *vs.* **MCI**	**MCI** *vs.* **AD**	**CN** *vs.* **AD**
1st	Superior frontal gyrus, dorsolateral (L)	Superior frontal gyrus, medial (R)	Hippocampus (L)
2nd	Precentral gyrus (R)	Insula (L)	Superior frontal gyrus, medial orbital (L)
3rd	Anterior cingulate cortex, supracallosal (R)	Inferior frontal gyrus, triangular part (L)	Superior frontal gyrus, medial orbital (R)
4th	Lobule IX of cerebellar hemisphere (R)	Fusiform gyrus (R)	Middle temporal gyrus (R)
5th	Precuneus (R)	Anterior cingulate cortex, subgenual (L)	Precuneus (R)
6th	Superior frontal gyrus, medial (L)	Supplementary motor area (L)	Cuneus (L)
7th	Fusiform gyrus (R)	Middle cingulate & paracingulate gyri (R)	Superior temporal gyrus (L)
8th	Insula (R)	Hippocampus (L)	Insula (R)
9th	Middle frontal gyrus (R)	Middle cingulate & paracingulate gyri (L)	Inferior frontal gyrus, opercular part (R)
10th	Postcentral gyrus (L)	Lobule VI of cerebellar hemisphere (R)	Cuneus (R)

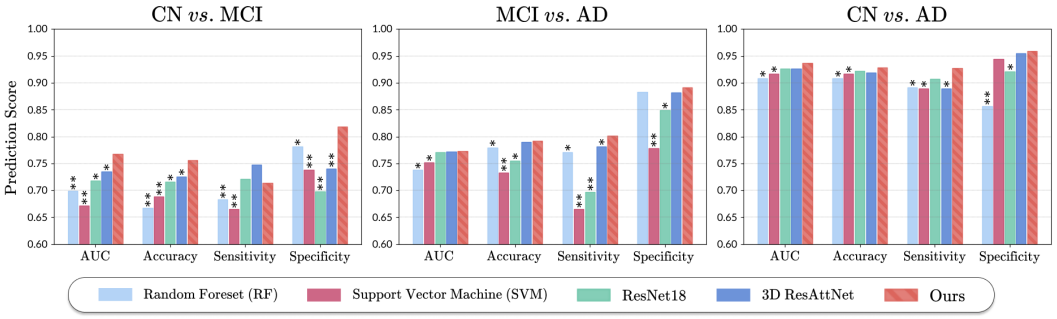

Figure 2: Classification performance results compared with ML-/DL-based models. $*$ and $**$ indicate the statistical significance by the Wilcoxon signed-rank test at $p < 0.05$ and $p < 0.01$, respectively.

3 Experimental Results

Dataset and Data Preprocessing We evaluated our method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [8], which consists of 1,540 subjects. By conducting a series of preprocessing [5, 4] and extra procedures (*i.e.*, down-scaling and subject-wise normalization) on raw 3D MRI scans, we attained the preprocessed MRI scans, each with a dimensionality of $96 \times 114 \times 96$.

AD-Effect Map To quantitatively investigate the anatomical variations in GM density, we estimated the AD-effect map, which involves disease-induced regional localization (Fig. 3). We observed that while some areas in which have related to the AD progression (*e.g.*, hippocampus and insula) were prominent in AD-effect maps of all scenarios, scenario-specific areas were evoked depending on the diverse levels of severity. As a result, the number of AD-effect ROIs in each scenario (a), (b), and (c) in Fig. 3 was composed of 56, 75, and 79 ROIs, respectively. Furthermore, based on the classification performance in Fig. 2, we were convinced that our AD-effect map reflected the most probable discriminative areas to classify the scenario-specific clinical stages.

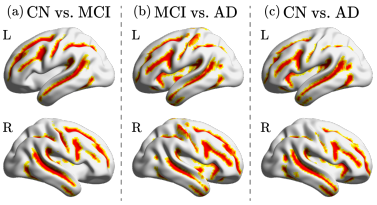

Figure 3: Visualization of AD-effect maps on various scenarios.

Performance Evaluation We evaluated the effectiveness of our LiCoL utilizing the AD-effect ROIs compared to ML-based models (*i.e.*, RF and SVM) and DL-based models (*i.e.*, ResNet18 [3] and self-attention-based 3D ResAttNet [12]) in terms of four evaluation metrics: a receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. We observed that our method achieved the highest performance over the AUC and accuracy in all classification scenarios as well as baselines (Fig. 2). Most remarkably, our LiCoL derived considerable performance improvement in the CN $vs.$ MCI scenario, despite that this scenario was challenging to distinguish AD-affected regions owing to subtle anatomical changes. Meanwhile, it should be noted that typical DL models are built on layers of complex networks with non-linearity, yielding superior predictive performance but inevitably sacrificing interpretability. In contrast, our LiCoL allowed outstanding classification performance while intuitively interpreting the rationale of the model's decision via a linear architecture (Fig. 1).

ADness Index We used the counterfactual-guided attention matrix $\mathbf{A}$ (described in Section 2) as the quantitative region-wise ADness index to elucidate our LiCoL's decision in any classification scenario. As shown in Table 1, we illustrate the sorted top-10 AD-effect ROIs based on region-wise ADness index (*i.e.*, averaged attentive score) for all test samples in each scenario. According to this result, we found that the insula and superior frontal gyrus have contributed to the prediction in all scenarios. This implies the GM density of these regions was noticeably reduced by atrophic variations, suggesting that their ROIs are vital biomarkers across the AD spectrum. One can further observe that the hippocampus, fusiform gyrus, anterior cingulate cortex, and inferior frontal gyrus were influential regions contributing to disease prediction in scenarios involving MCI and AD patients. Consequently, based on these extended investigations, we argued that the ADness index enables us to interpret the counterfactual-guided attentive representation concerning the subregional status of each subject's brain and to explain the output decision via the subregional ADness indices in a quantitative way.

4 Conclusion

We introduced a novel quantitative feature-based in-depth analysis using counterfactual-guided deep feature representation with enhancing AD predictive performance compared to ML-/DL-based models. As a result, our method could be one of the milestones toward the comprehensible explanation of the DL models' decision from the clinician's perspective on neurodegenerative disease prediction.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) No. 2022-0-00959 ((Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making) and No. 2019-0-00079 (Department of Artificial Intelligence (Korea University)). This research was further supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2022R1A4A1033856).

Potential Negative Societal Impacts

In order to produce the counterfactual sample for the quantitative feature-based analysis in this study, we exploited the counterfactual reasoning generator, which is built upon the generative adversarial network (GAN). The GAN-based counterfactual reasoning model applied in the medical field has the capacity to generate a realistic synthesized sample without the need to use sensitive information over the patient while maintaining essential characteristics of the original sample. Even though the development of GAN-based approaches unlocks the way to mitigate the scarcity of labeled data in healthcare services as well as to investigate the status of patients before and after disease onset (*e.g.*, normal control or severe Alzheimer's), it may pose a social ethics issue that accompanies by potential risks. As an example, synthetic clinical data can be abused as a means to arbitrarily manipulate clinical outcomes to gain formal approval in receiving access or financial support for clinical-related scenarios. In other words, if such faulty medical applications are unexpectedly authorized and commercialized, it might suggest that improper medical treatment could exacerbate disease progression or have life-threatening fatal consequences. Therefore, it is necessary to take technical and institutional responses thoroughly that register the strict regulation or judge the results of a clinical trial.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[2] Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 15(3):321–387, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019.

[5] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *NeuroImage*, 62(2):782–790, 2012.

[6] GB Karas, Philip Scheltens, Serge ARB Rombouts, Pieter Jelle Visser, Ronald A van Schijndel, Nick C Fox, and Frederik Barkhof. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, 23(2):708–716, 2004.

[7] Shanshan Li, Ozioma Okonkwo, Marilyn Albert, and Mei-Cheng Wang. Variation in variables that predict progression from MCI to AD dementia over duration of follow-up. *American Journal of Alzheimer's Disease (Columbia, Mo.)*, 2(1):12, 2013.

[8] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

[9] Kwanseok Oh, Jee Seok Yoon, and Heung-Il Suk. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] Edmund T Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. *NeuroImage*, 206:116189, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Xin Zhang, Liangxiu Han, Wenyong Zhu, Liang Sun, and Daoqiang Zhang. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Journal of Biomedical and Health Informatics*, 2021.