Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Counterfactual-Guided MRI Feature Representation and Quantitatively Interpretable Alzheimer's Disease Prediction

Kwanseok Oh
Korea University https://orcid.org/0000-0002-7378-4569
Da-Woon Heo
Korea University
Ahmad Wisnu Mulyadi
Korea University
Wonsik Jung
Korea University
Eunsong Kang
Korea University
Kun Ho Lee
Chosun University https://orcid.org/0000-0003-4144-8162
Heung-Il Suk (hisuk@korea.ac.kr)
Korea University https://orcid.org/0000-0001-7019-8962

Article

Keywords: Alzheimer's Disease, Counterfactual Reasoning, Quantitative Feature-Based Analysis, Counterfactual-Guided Attention

Posted Date: August 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1985865/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Deep Counterfactual-Guided MRI Feature Representation and Quantitatively Interpretable Alzheimer's Disease Prediction

Kwanseok Oh[1], Da-Woon Heo[1], Ahmad Wisnu Mulyadi[2], Wonsik Jung[2], Eunsong Kang[2], Kun Ho Lee[3,4,5] and Heung-Il Suk[1,2*]

[1]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea.

[2]Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea.

[3]Gwangju Alzheimer's & Related Dementia Cohort Research Center, Chosun University, Gwangju 61452, Republic of Korea.

[4]Department of Biomedical Science, Chosun University, Gwangju 61452, Republic of Korea.

[5]Korea Brain Research Institute, Daegu 41062, Korea.

*Corresponding author(s). E-mail(s): hisuk@korea.ac.kr;
Contributing authors: ksohh@korea.ac.kr; dwheo@korea.ac.kr; wisnumulyadi@korea.ac.kr; ssikjeong1@korea.ac.kr; eunsong1210@korea.ac.kr; leekho@chosun.ac.kr;

Abstract

Deep learning for Alzheimer's disease (AD) prediction has provided timely prevention of disease progression yet still demands attentive interpretability. Recently, *counterfactual reasoning* has increasingly been exploited in medical research by providing refined visual explanatory maps. However, such visual explanatory maps alone are not self-sufficient unless we can intuitively demonstrate their validity via quantitative features. For this, we first synthesize the counterfactual-labeled structural MRI using our proposed framework. We further transform it as a gray matter density map to precisely measure its volumetric changes over the parcellated region of interests (ROIs). Furthermore, to boost the effectiveness of selected ROIs while promoting interpretability and achieving

comparable predictive performance to deep learning methods, we devise a novel lightweight counterfactual-guided attentive feature representation constructed to a linear classifier. Thus, our framework provides an AD-relatedness index for each ROI, offering an intuitive understanding of brain status for an individual subject and across subjects regarding AD.

Keywords: Alzheimer's Disease, Counterfactual Reasoning, Quantitative Feature-Based Analysis, Counterfactual-Guided Attention

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative brain disease accompanied by memory loss, difficulties in logical thinking, and problems with reading and writing [1]. Although many efforts have been made to enhance understanding and discover efficient treatment, the medication for AD is intended only to slow AD progression [2]. As numerous studies have revealed the advantages of early prevention [3], it is paramount to identify patients who have mild cognitive impairment (MCI)—a prodromal stage of AD that is likely to convert to AD [4]—to effectively delay cognitive decline. With advances in brain imaging techniques such as structural magnetic resonance imaging (sMRI), a predictive framework that integrates clinical examination and imaging techniques [5] has been developed to understand brain disease-related structural changes. However, clinical examinations may lead to different diagnoses, contingent on which diagnosis index is used, as each index includes various evaluation items. Moreover, the visual inspection of such brain imaging scans inevitably depends upon the qualitative evaluations and subjective decisions of radiologists and clinicians.

Originally derived from oncological studies [6–8], *radiomics* has recently been used to extract quantitative features that objective and reliable disease-related characteristics, reflecting anatomical landmarks from brain imaging data. These quantitative features (*e.g.*, density, volume, morphometry, and textures) contain phenotypic characteristics that are beneficial for disease analysis [9]. Concurrently, quantitative feature-based machine learning (ML) techniques have been proposed as an essential component of computer-aided diagnosis and detection systems [10]. Unlike statistical methods [11, 12], which are based on group-level analysis, ML methods allow individual-level precise predictions for subjects [13] while recognizing the intricate patterns of input features for downstream tasks. Although existing ML-based methods are helpful in identifying MCI/AD at an individual-level, most of the biomarker region selection is biased towards a small set of pre-defined regions for the sake of validation alone [14]. That is, undiscovered disease-related regions may get ignored, *i.e.*, regions that could feasibly be considered a potential biomarker.

Furthermore, obtaining sufficient generalizability and fidelity of region of interests (ROIs) used for assessment is challenging because the disease-relevant regions are manually predetermined for each study [12, 15].

On the other hand, deep learning (DL) techniques [16] have recently achieved impressive outcomes, and remarkable progress in 3D sMRI-based AD prediction over conventional ML methods [17–20]. The hierarchical structure of DL models make them proficient for discovering informative patterns automatically, which alleviates the need for an a priori handcrafted selection of disease-related regions by experts. Although DL models significantly enhance predictive performance, outstanding performance with interpretable final decisions remains an unattained goal [21]. Because of the unfavorable "black-box" nature of DL models, many researchers have been unable to explain and interpret how their proposed DL models make definitive decisions from a clinician's perspective [22]. To solve this issue, explainable artificial intelligence (XAI) [23] is increasingly utilized in medical image analysis to unravel the underlying model behavior in drawing final decisions *w.r.t.* the influences and contributions of the relevant features (*e.g.*, particular disease-related ROIs) as a piece of supporting evidence.

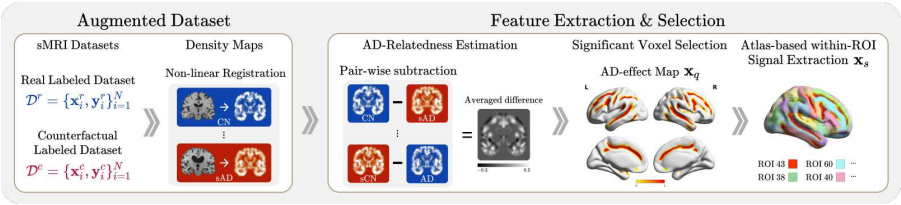
The XAI taxonomy can generally be grouped into two categories: intrinsic and post-hoc approaches. The intrinsic model is a self-explanatory model, as it is constructed with a particular mechanism to encourage interpretability by design. Intrinsic models, such as prototype-based models, have developed from comprehensible prototypes that can refer to representative parts of samples to determine the disease-related attribute that is prominently involved in distinguishing the input image [24]. However, the limitation of intrinsic models in practice is that their interpretability is restricted to a model-specific explanation and to significant individual variations in disease-related regions for each subject [21]. In contrast, a typical post-hoc model grants a model-agnostic explanation by seamlessly plugging into a well-trained DL model to describe the final prediction. With a post-hoc model, attribution-based approaches [25, 26] offer visual saliency maps that highlight the class-relevant attributes in given input images by analyzing the gradients or activations of the model. However, one general drawback of these attribution-based approaches is that they tend to produce similar saliency maps across the wrong class labels, owing to an "attribution vanishing" problem [27].

In response to these issues, we resort to a counterfactual reasoning approach as one of the post-hoc approaches. This approach possesses proficiency in providing refined visual explanatory maps—so-called counterfactual maps. Most preceding works that yielded counterfactual reasoning were built upon the generative adversarial network (GAN) [28] and its variants [29–31]. Such a counterfactual map reveals fundamental explanations regarding the models' decisions in hypothetical scenarios, similar to the human decision-making process [32]. Specifically, it enables an observer (*i.e.*, not necessarily an expert) to consider contrastive explanations and causal inference, such as why a particular decision was made instead of another, and to observe how an alteration of

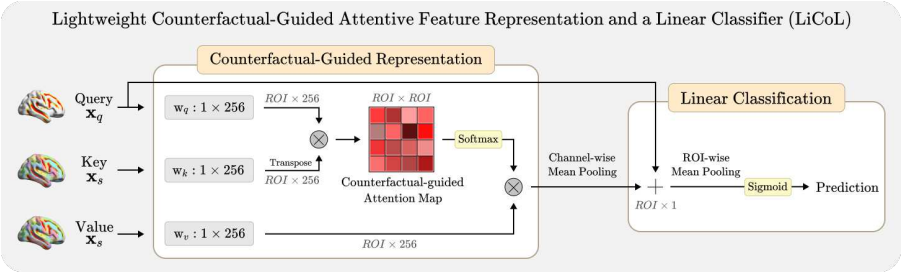
specific attributes will affect the model's output [33, 34]. In terms of MCI/AD prognosis, the counterfactual map serves as an explanatory map that includes anatomical changes caused by atrophies, assuming that the patient was diagnosed as a normal subject. However, such counterfactual maps still rely heavily on visual inspection according to the voxel-level morphometry owing to merely visualizing these counterfactual maps on the original images to understand the model's decision. Likewise, most medical imaging studies that have used applied XAI generally provided a visual interpretation of clinical performance without properly proving their clinical validity. This yields limited intuitive insight [35]. In practical application, exhibiting explanatory maps alone is not self-sufficient, nor do they fulfill an expectation to support MCI/AD prediction as clinically decisive auxiliary information. Accordingly, it is necessary to intuitively demonstrate the validity of the counterfactual map by analyzing the counterfactual reasoning via quantitative features [36].

For these premises, we propose a novel framework for counterfactual-induced feature representation and quantitatively interpretable and explainable brain disease identification. As an extension of our prior study, we exploit LEAR's counterfactual map generator (CMG) [31] to synthesize the counterfactual-labeled sMRI data (c-sMRI) by utilizing the generated counterfactual map via the given input (*i.e.*, real-labeled sMRI data, namely r-sMRI). The newly produced c-sMRIs reflect the target counterfactual attributes with high confidence, hence we could regard that constructed c-sMRIs as a disease-associated brain data augmentation set. *By leveraging the principal characteristics of those augmented data, we are able to unlock a way to investigate the brain that has not yet been converted to the disease-manifested brain or even a clinically improbable case of reversed disease progression by producing hypothetical scenarios.* We then transform both r-sMRIs and c-sMRIs into a gray matter (GM) density map (*i.e.*, rGM and cGM) that enables us to precisely measure the volumetric changes in GM [37], associated with brain atrophies caused by normal aging [38] and AD progression [39, 40].

Given the rGM and cGM, we perform a series of GM density-based analyses and assessments, as depicted in Fig. 1. We first obtain the representative difference map by subtracting rGMs from their counterpart cGMs. We then take the first statistical moment (*i.e.*, mean of difference maps). We refer to such a map as an *AD-effect map*, which indicates the probable AD-induced anatomical variations in GM density, which are estimated in a data-driven manner. Subsequently, according to the aforementioned AD-effect map, we devise an ROI-based analysis mechanism over the derived significant ROIs (namely AD-effect ROIs). We further develop a shallow network of lightweight counterfactual-guided attentive feature representation and a linear classifier (LiCoL). It is noteworthy that the internal operations from the input to the output can be rewritten as a linear function, as in Supplementary H. This allows us to interpret the counterfactual-guided attentive representation as the regional status of a brain, which we call an "*ADness*" index, and to explain the output decision based on the regional ADness indices in a quantitative



(a) AD-effect map generation for the acquisition of the AD-effect ROIs



(b) A schematic overview of LiCoL framework

Fig. 1: Illustration of our GM density-based analysis workflow. We manipulate the given r-/c-sMRI data as a GM density map and calculate the representative difference map by subtracting rGMs from their counterpart cGMs, after which we take the first moment. Furthermore, by employing the percentile threshold to emphasize the meaningful voxels, we obtain the AD-effect map, which involves the most distinguishing features between clinical stages. According to its AD-effect map, we extract significant ROIs using the AAL3 atlas template, and finally, we utilize our proposed LiCoL to demonstrate the effectiveness of those selected ROIs as promising features for the predictive task. Here, the gray boxes in (b) denote the embedding layer. "+" and "⊗" denote the element-wise addition and element-wise multiplication, respectively. It should be noted that $\mathbf{x}_q$ for the query is fixed as extracted AD-effect ROIs, whereas $\mathbf{x}_s$ for the key and value is independently initialized to be subject-wise.

way [41]. To our best knowledge, this is the first sMRI-based AD prediction model that makes an input observation interpretable as a quantitative AD-relatedness index, and provides the AD-relatedness index comparable among the brain regions of a single subject and across subjects.

We validated our proposed method qualitatively and quantitatively via exhaustive experiments using 3D brain sMRIs from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [42] and private Korean 3D brain sMRIs of the Gwangju Alzheimer's and Related Dementia (GARD) cohort dataset [43].

Concretely, the main contributions of our method are summarized as follows:

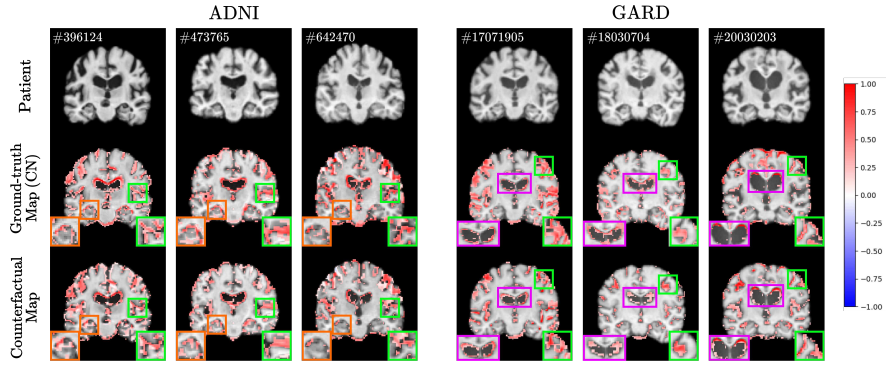


Fig. 2: Comparison of generated counterfactual maps with ground-truth maps on both ADNI and GARD datasets. In the first row, the #number at the top left of the respective image indicates the image ID. We also overlay purple, green, and orange boxes highlighting the ventricular, cortex, and hippocampal regions, respectively.

- We propose a novel methodology to derive fundamental scientific insights from a counterfactual reasoning-based explainable DL method. We demonstrate that our proposed method can interpret intuitively from the clinician's perspective by converting counterfactual-guided deep features to the quantitative volumetric feature domain, rather than using a direct inspection of DL-based visual attributions.
- We achieve similar or better performance than the DL model by designing a LiCoL with the AD-effect ROIs considered to be the distinctive AD-related landmarks via counterfactual-guided deep features.
- By exploiting the proposed LiCoL, we provide numerically interpretable ADness indices for each patient *w.r.t* anatomical variations caused by disease progression.
- We further investigate the discoveries of AD-manifested subregions based on common and potential regions by comparing the ADNI and GARD datasets.

2 Results

2.1 Counterfactual Reasoning

We conducted both qualitative and quantitative assessments to validate the reliability of counterfactual reasoning on ADNI and GARD datasets. Because there is no ground-truth to evaluate the generated counterfactual map, we constructed the "*pseudo*" ground-truth map (*i.e.*, an observed disease progression map), defined as the difference between longitudinal samples from two different clinical stages. To this end, we first gathered longitudinal MRIs converted from CN to MCI and AD at any given time. After that, the ground-truth map was created by subtracting the baseline image from the target class image. For

Table 1: Normalized cross-correlation (NCC) scores on the ADNI and GARD datasets. We differentiated NCC scores for each respective direction scenario of the generated counterfactual map. Here, we defined the ground-truth maps and counterfactual maps for various transformation scenarios, including CN←MCI, MCI←AD, and CN←AD as the "+" direction, and CN→MCI, MCI→AD, and CN→AD as the "-" direction, and calculated NCC(+) and NCC(-) for each direction, respectively.

Datasets	CN ↔ MCI		MCI ↔ AD		CN ↔ AD	
	NCC(+)	NCC(-)	NCC(+)	NCC(-)	NCC(+)	NCC(-)
ADNI	0.334±0.012	0.253±0.026	0.302±0.005	0.284±0.023	0.343±0.019	0.318±0.016
GARD	0.326±0.009	0.299±0.014	0.311±0.008	0.321±0.015	0.333±0.007	0.338±0.011

example, if the CN subject is selected as a baseline image, either the MCI or AD subject shall be defined as a counterpart of the target class image, depending on the scenario. This ground-truth map exhibited an excellent representation of disease localization, indicating which regions were changed according to the clinical label conversion. Eventually, the resulting ground-truth map and counterfactual map derived on the ADNI and GARD longitudinal datasets were overlaid onto a respective subject's sMRI, as shown in Fig. 2.

In Fig. 2, we illustrate the counterfactual map examples generated by our CMG [31] from the task of CN *vs.* AD classification. It can be seen that the generated counterfactual maps on both ADNI and GARD longitudinal samples exhibit remarkably good agreements with their corresponding pseudo ground-truth map. In fact, the hippocampal region (orange box) is known to be the most prominent AD region [44]; however, it is also known that morphological variations due to AD in many other regions are also involved. In terms of these disease characteristics in the brain, we observed that our CMG successfully captured reduced ventricle and hypertrophy in the hippocampus while providing a counterfactual map containing almost universal variations in all subtle regions of increased cortex.

As a numerical assessment metric of the generated counterfactual maps, we further calculated the normalized cross-correlation (NCC) score in Table 1, which measures the morphological and structural similarity between a generated counterfactual map and a pseudo ground-truth map as proposed in [31]. We leveraged the NCC score, as it has the primary properties of not being sensitive to the magnitude of the signals. Compared with LEAR's performance (which reported the performance on the ADNI dataset only), the NCC scores of GARD were comparable to those on the ADNI in all scenarios. Based on the superior NCC scores on both ADNI and GARD datasets, we hypothesized that the generated counterfactual maps on both datasets not only captured pathologically subtle changes associated with AD progression but also demonstrated the generalizability of the CMG over two independent and different racial data sources. Leveraged by those qualitatively and quantitatively promising

results, we can justify our proposed framework to exploit the counterfactually-generated samples and accept our findings from the experimental analysis described below.

2.2 AD-Effect Map and ROI Feature Representation

One of the key ideas in this study is to utilize the counterfactual map inferred from deep feature representations and a deep classifier for a lightweight network construction equipped with an interpretable feature representation and an explainable linear classifier. For this, we registered both r-sMRIs and c-sMRIs to a brain template and produced their respective GM density maps (refer to Section 4.5 for details).

We obtained quantitative measurements and discovered the morphological phenotype changes of AD effects by conducting an in-depth analysis of rGMs and cGMs. In particular, by utilizing voxel-based morphometry (VBM) [45] on rGMs and cGMs, we investigated the estimated volume-based GM variability that may have been caused by AD, which is suitable for discriminating between patients and normal controls based on the GM volume differences.

Using Eq. (11), we estimated an AD-effect map which would be exploited to identify cortical and subcortical brain regions where the GM density contrasts are considerably differentiated between clinical stages (the top two rows in Fig. 3). For comparison, we also computed a conventional group-level statistical map obtained only from rGMs (the bottom two rows in Fig. 3). Regarding the feature engineering, we considered these AD-effect and statistical maps to identify the class-discriminative areas and applied different thresholding strategies (*i.e.*, percentile *vs.* p-value) for ROI selection. The AD-effect map showed more fine-grained regional localization than the statistical map, as shown in Fig. 3. Whereas some areas are prominent in both the AD-effect map and the statistical map, other regions are revealed by only one map. Based on the classification performance in Supplementary E, we assume that the AD-effect map finds better-defined regions than does the statistical map. On the one hand, the statistically significant areas do not necessarily reflect class-discriminative information at an individual-level prediction. On the other hand, our proposed AD-effect map finds class-discriminative areas, which may not be statistically significant in a size-limited dataset, leveraging by the augmented cGMs.

2.3 Predictive Performance Evaluation

We conducted exhaustive experiments on the rigorous predictive performance of our proposed LiCoL and the acquired AD-effect-map-selected-ROIs set $\mathcal{R}_{\text{eff}}$ by comparing the DL-baseline trained with 3D volumetric rGMs and the ML-baseline trained with the statistical-map-selected-ROI set $\mathcal{R}_{\text{stat}}$, as shown in Table 2 (*i.e.*, reported on each row w/o checkmark in cGM). Here, we employed six ML models, *i.e.*, random forest (RF), support vector machine (SVM), logistic regression (LR), decision tree (DTs), Lasso regression, and Ridge regression, as ML-baselines. In Table 2, we present the results of SVM,

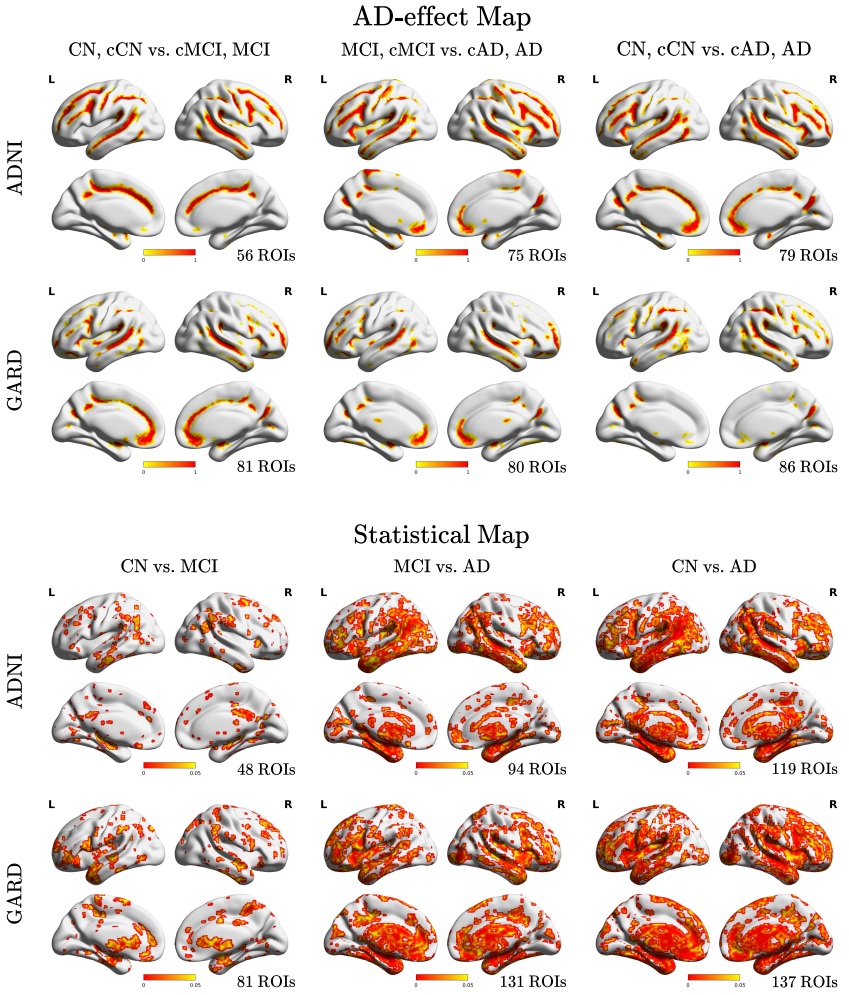


Fig. 3: Illustration of inferred AD-effect maps and statistical maps on ADNI and GARD datasets. Here, the cCN, cMCI, and cAD refer to the abbreviation of the CN-/MCI-/AD-labeled cGM, respectively. The colorbars over the AD-effect map and statistical map respectively indicate the range of normalized GM density and the p-value measured by the two-sample t-test. We also present the respective number of significant ROIs derived by each AD-effect map and statistical map below the visualized map in each scenario.

as it achieved the best performance among the comparative ML methods. The performance of the other ML methods with $\mathcal{R}_{\text{stat}}$ is in Supplementary Table S10. As for the DL-baseline, we employed ResNet18 [46], which was trained as a predictive model for counterfactual reasoning. The implementation details and hyperparameters for ML-/DL-baseline training are reported in Supplementary C.

Table 2: The classification performance results compared with the ML-/DL-based models in various classification scenarios on ADNI and GARD datasets with four evaluation criteria (*i.e.*, AUC, accuracy, sensitivity, and specificity). We divided the performance results on each dataset according to whether cGMs were used for augmentation. * and ** denote the statistical significance by the Wilcoxon signed-rank test at $p < 0.05$ and $p < 0.01$, respectively.

Scenarios	Datasets	cGM	Model	AUC	Accuracy	Sensitivity	Specificity
CN *vs.* MCI	ADNI		SVM	0.5278±0.036**	0.5481±0.031**	0.6361±0.046	0.5313±0.037
			ResNet18	0.6209±0.041**	0.6479±0.027**	**0.7487±0.063**	0.4931±0.112
		✓	SVM	0.6721±0.021**	0.6882±0.028**	0.6655±0.025	0.7385±0.011
		✓	ResNet18	0.7178±0.031*	0.7164±0.023*	0.7215±0.045	0.6976±0.036
		✓	**Ours**	**0.7678±0.013**	**0.7562±0.015**	0.7143±0.021	**0.8182±0.015**
	GARD		SVM	0.5667±0.054**	0.5819±0.042**	0.7091±0.048	0.5276±0.069
			ResNet18	0.6179±0.028**	0.6254±0.026**	0.6973±0.084	0.4871±0.103
		✓	SVM	0.7145±0.022**	0.7059±0.031**	0.7281±0.044	0.6798±0.031
		✓	ResNet18	0.7395±0.031*	0.7211±0.026*	0.7385±0.051	0.7661±0.036
		✓	**Ours**	**0.7778±0.020**	**0.7662±0.019**	**0.7485±0.024**	**0.8246±0.017**
MCI *vs.* AD	ADNI		SVM	0.6317±0.043**	0.6452±0.035**	0.5186±0.030	0.6869±0.046
			ResNet18	0.6946±0.062**	0.7121±0.053**	0.4819±0.196	0.8041±0.121
		✓	SVM	0.7519±0.013*	0.7329±0.021*	0.6657±0.030	0.7785±0.029
		✓	ResNet18	0.7705±0.013	0.7551±0.037*	0.6971±0.061	0.8495±0.044
		✓	**Ours**	**0.7727±0.018**	**0.7921±0.016**	**0.8014±0.022**	**0.8913±0.019**
	GARD		SVM	0.6418±0.032**	0.6039±0.031**	0.7984±0.089	0.5372±0.063
			ResNet18	0.7084±0.074**	0.7258±0.019**	0.5042±0.181	**0.8454±0.036**
		✓	SVM	0.7284±0.025*	0.7386±0.028*	0.8329±0.034	0.7391±0.036
		✓	ResNet18	0.7569±0.027*	0.7729±0.036	0.6619±0.043	0.8183±0.052
		✓	**Ours**	**0.7753±0.016**	**0.7814±0.014**	**0.8917±0.011**	0.7955±0.023
CN *vs.* AD	ADNI		SVM	0.8198±0.043**	0.8195±0.031**	0.6991±0.067	0.8712±0.055
			ResNet18	0.8685±0.043**	0.8714±0.042**	0.7609±0.107	0.9046±0.038
		✓	SVM	0.9167±0.009	0.9167±0.011	0.8889±0.023	0.9444±0.015
		✓	ResNet18	0.9265±0.033	0.9219±0.034	0.9078±0.032	0.9209±0.052
		✓	**Ours**	**0.9371±0.005**	**0.9285±0.011**	**0.9271±0.017**	**0.9591±0.015**
	GARD		SVM	0.7619±0.035**	0.7886±0.048**	0.6417±0.056	0.6985±0.051
			ResNet18	0.8343±0.045**	0.8503±0.029**	0.7455±0.099	0.9231±0.039
		✓	SVM	0.8774±0.021**	0.8533±0.024**	0.8912±0.014	0.8778±0.012
		✓	ResNet18	0.8930±0.031*	0.8913±0.032	0.8733±0.052	0.9093±0.034
		✓	**Ours**	**0.9304±0.008**	**0.9193±0.007**	**0.9441±0.010**	**0.9283±0.012**

Compared with ML-baseline, our proposed method considerably improved an area under the receiver operating characteristic curve (AUC) on both ADNI and GARD by +22.56%, +13.73%, and +14.29% on average for each of CN *vs.* MCI, MCI *vs.* AD, and CN *vs.* AD scenarios, respectively. Most remarkably, our method also achieved the greatest performance improvement in the CN *vs.* MCI scenario, despite the fact that it was difficult to identify disease-affected regions associated with brain atrophies owing to subtle anatomical variations. Having observed this predictive performance, we confirmed that our discovered $\mathcal{R}_{\text{eff}}$ plays a critical role in revealing a relatively substantial difference in GM density even for minor morphological changes. Thus, this piece of evidence encouraged us to exploit the produced $\mathcal{R}_{\text{eff}}$ because it exhibited promise for distinguishing most class-discriminative regions.

Table 3: Examination of precision and recall performance according to the number of samples for each age group in the MCI *vs*. AD scenario. These results can be certified to alleviate spurious correlations by additionally using the cGM. Here, we concisely define the set of AD-effect ROIs (Ours) and statistical ROIs as $\mathcal{R}_{\text{eff}}$ and $\mathcal{R}_{\text{stat}}$, and augmented cGM is indicated as w/ cGM.

Datasets	Age Range (years)	Samples (MCI)	Samples (AD)	$\mathcal{R}_{\text{stat}}$		$\mathcal{R}_{\text{stat}}$ w/ cGM		$\mathcal{R}_{\text{eff}}$ (Ours)	
				Precision	Recall	Precision	Recall	Precision	Recall
ADNI	60 ∼ 70	206	65	0.6897	0.8995	0.7871	0.9144	**0.8313**	0.9331
	70 ∼ 80	359	174	0.8201	0.9067	0.8641	0.9356	**0.8951**	0.9415
	80 ∼ 90	149	101	0.7273	0.9331	0.8318	0.9428	**0.8491**	0.9628
GARD	60 ∼ 70	130	19	0.6338	0.9339	0.7452	0.9773	**0.7780**	0.9840
	70 ∼ 80	181	53	0.7016	0.9211	0.7838	0.9276	**0.8093**	0.9777
	80 ∼ 90	50	34	0.5742	0.8942	0.7661	0.8997	**0.7957**	0.9146

In addition, our LiCoL surpassed the DL-baseline with statistical significance for AUC and accuracy. However, it should be noted that typical DL models are built on layers of networks with complex non-linearity, yielding high classification performance but inevitably sacrificing interpretability. In contrast, we thoughtfully designed the LiCoL with linearity so that the internal representations were interpretable, and the rationale behind any classification decision was straightforward to understand without additional tools or computations.

We further examined the reliability and generalizability of cGMs by exploiting the augmented data to train the ML-/DL-baselines. Interestingly, we found that training the ML/DL models with the augmented data helped improve performance substantially relative to the baseline ML/DL models trained with only rGMs (*i.e.*, results w/ checkmark in cGM in Table 2 and Supplementary Table S11). Nevertheless, our method achieved the best performance in terms of AUC and accuracy across all classification scenarios over the ADNI and GARD datasets.

2.4 Counterfactual Images for Practical Example

We explored whether issues in practical examples that could be encountered in the medical field could be overcome using the cGM augmentation strategy. Indeed, accurate prevalence estimates of dementia by age, or other demographic information for different levels of severity could support effective medical treatment. In particular, age constitutes the greatest of the various risk factors *w.r.t* AD manifestation, with the percentage of the population who have AD increasing dramatically with age. However, although 3% of people aged 65 ∼ 74, 17% of people aged 75 ∼ 84, and 32% of people aged 85 or older have been diagnosed as AD patients [47], it is important to note that AD is not a normal part of aging [48], and older age alone is not considered sufficient to cause AD. Various studies in the medical field [49, 50] call this phenomenon spurious correlation, which occurs when two factors appear to be

correlated to each other, but in fact are not [51]. Furthermore, there are few cases of AD where the patient's age is younger than 65 years, and further data on the prevalence of brain disease are scarce. Hence, samples for a specific age or clinical stage are lacking, which leads to undesirable bias. We thus exploited the cGMs as augmented samples to alleviate these practical issues. For this purpose, we used statistical ROIs $\mathcal{R}_{\text{stat}}$ and AD-effect ROIs $\mathcal{R}_{\text{eff}}$ as classifier input to demonstrate the fidelity of cGM augmentation by showing that the performance shifts depending on whether augmented data are included in the training. Here, we trained our LiCoL for prediction, and we reported the experimental results in Table 3.

One can observe that the trained LiCoL that used statistical ROIs reveals lower precision in younger adults with AD (*i.e.*, the age range of $60 \sim 70$) and for MCI in older adults (*i.e.*, the age range of $80 \sim 90$) on both ADNI and GARD. Unlike in the ADNI, a lack of samples appeared in the $80 \sim 90$ age range in GARD, and this age range exhibited the lowest precision. However, by using the cGM to construct the statistical ROIs (*i.e.*, $\mathcal{R}_{\text{stat}}$ w/ cGM), the precision for the $80 \sim 90$ age range was dramatically increased by $+19.19\%$. We further observed that the performance with augmentation improved overall at all ages. To this end, we argue that the auxiliary cGMs of a patient at different clinical stages provided informative features, so that LiCoL was able to learn how to correctly distinguish the discriminative characteristics of each clinical stage, depending on the age range. As a result, we suggest that utilizing the cGMs is truly beneficial for mitigating the data-hungry problem in practical use cases. As such, data augmentation strategies may be a promising alternative for solving spurious correlation.

3 Discussion

As we devised our LiCoL by revoking the incomprehensible non-linearities in mind, we attentively embraced a linear function that intuitively interprets the final decision. Accordingly, we present the mathematical development of the LiCoL as a linear function.

Using the simplified function $\bar{y} = \sigma\left(\text{MP}_{\downarrow}\,\text{MP}_{\rightarrow}\left(\mathbf{A}\mathbf{x}_s\right)c_2 + c_1\right)$ in Supplementary H, we analyzed the input-dependent representation of $\mathbf{A}\mathbf{x}_s$ that reveals the contribution of each ROI to patient-specific prediction. We used the counterfactual-guided attention map $\mathbf{A}$ as the region-wise ADness index to support the LiCoL decision in any classification scenario. In Fig. 4 and Fig. 5, we illustrate the averaged importance score of the region-wise ADness index for all test samples in each scenario over ADNI and GARD, and the region-wise ADness index for each patient is summarized in Supplementary G.

Across all scenarios, we observed that the disparity of importance scores in CN *vs.* AD was apparent. In this particular scenario, as the morphological difference *w.r.t* brain atrophies was relatively immense, our LiCoL attended to the ROIs where GM density changes were prominent. Overall, the ROIs of L/R.HIP, L/R.STG, L/R.IFGoperc, and L.PreCG exhibited notably high



Fig. 4: Illustration of averaged importance scores (*i.e.*, $\mathbf{Ax}_s$) for each AD-effect ROI based on longitudinal samples of ADNI. Here, "importance" refers to the quantitative figures of the ADness index that contribute to prediction. It should be noted that the hatched bar indicates an ROI that contributed significantly to the prediction as the importance score was lower, whereas the other bar indicates an ROI that is more influential, as the importance score is higher.

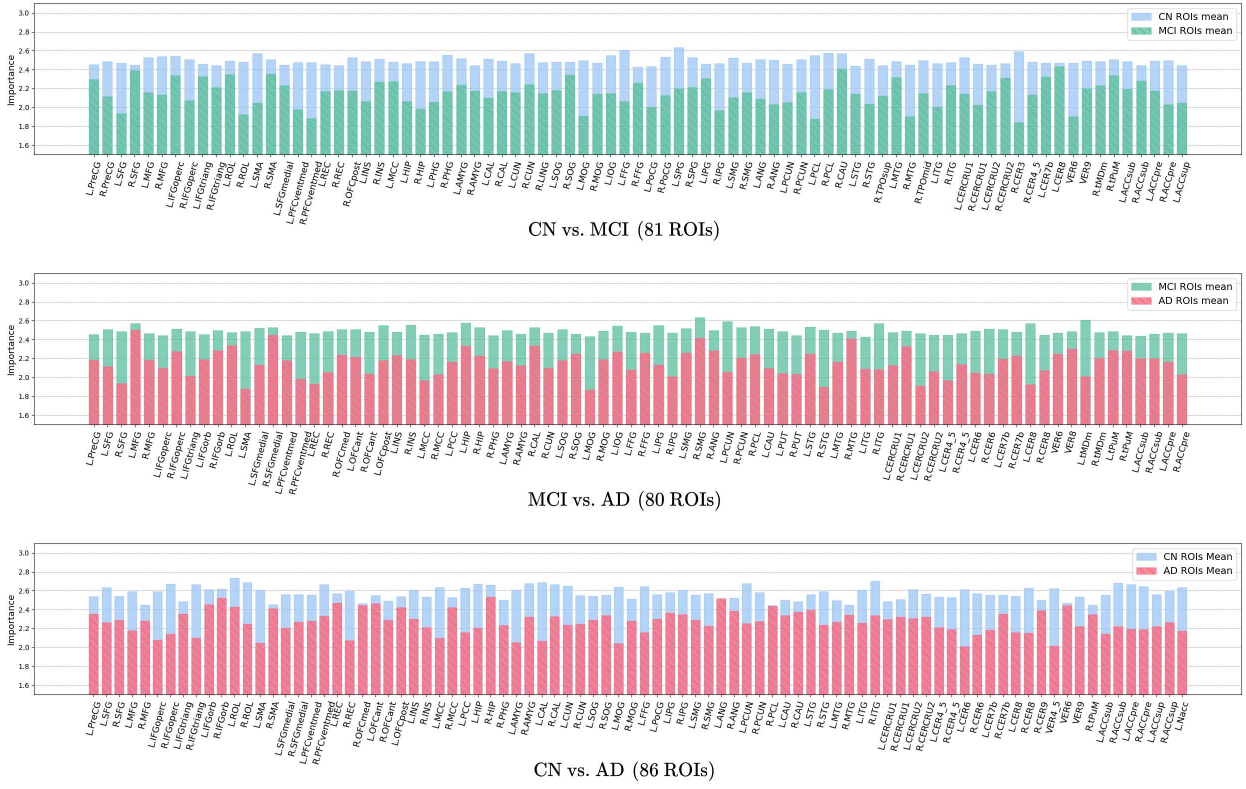


Fig. 5: Illustration of averaged importance scores (*i.e.*, $\mathbf{Ax}_s$) for each AD-effect ROI based on longitudinal samples of GARD. Here, "importance" refers to the quantitative figures of the ADness index that contribute to prediction. It should be noted that the hatched bar indicates an ROI that contributed significantly to the prediction as the importance score was lower, whereas the other bar indicates an ROI that is more influential, as the importance score is higher.

Table 4: Comparison of performance (AUC) between the AD-effect ROIs $\mathcal{R}_{\text{eff}}$ (Ours) and statistical ROIs $\mathcal{R}_{\text{stat}}$ using the LiCoL when exploiting the common or additional discrepant ROIs.

Datasets	ROIs	CN ↔ MCI		MCI ↔ AD		CN ↔ AD	
		Ours	$\mathcal{R}_{\text{stat}}$	**Ours**	$\mathcal{R}_{\text{stat}}$	**Ours**	$\mathcal{R}_{\text{stat}}$
ADNI	Baseline	**0.7678**±0.013	0.7222±0.016	**0.7727**±0.018	0.7514±0.021	**0.9371**±0.005	0.8912±0.014
	Intersection	0.7482±0.026	0.7284±0.011	0.7727±0.017	0.7538±0.014	0.9371±0.012	0.8933±0.009
	Union	0.7604±0.017	**0.7389**±0.027	0.7694±0.020	**0.7685**±0.031	0.9308±0.016	**0.9146**±0.010
GARD	Baseline	**0.7778**±0.020	0.7385±0.022	**0.7753**±0.016	0.7497±0.034	**0.9304**±0.008	0.9013±0.012
	Intersection	0.7592±0.017	0.7456±0.021	0.7693±0.013	0.7512±0.025	0.9304±0.006	0.9074±0.011
	Union	0.7745±0.026	**0.7479**±0.019	0.7737±0.016	**0.7581**±0.024	0.9277±0.014	**0.9211**±0.009

importance scores among AD-effect ROIs in all scenarios. In particular, in the CN *vs*. MCI and CN *vs*. AD scenarios, R.INS and L/R.PCUN did not exhibit as essential ROIs for CN prediction (*i.e.*, lower importance score), whereas those ROIs appear as crucial markers when they are used to diagnose MCI or AD subjects. Furthermore, one should also note that L.SMA was an ROI that markedly contributed to diagnosing MCI, and in AD, R.IPG was found to be a superior ROI with a high importance score. Consequently, with such interpretability, we ensure that our LiCoL can provide a quantitative and numerical interpretation of the anatomical variations in the brain resulting from AD progression in each patient, as well as in a group of patients, as a novel ADness index.

We further investigated the variability in AUC performance when using common or additive $\mathcal{R}_{\text{eff}}$ discrepant ROIs (*i.e.*, $\mathcal{R}'_{\text{eff}}$) and $\mathcal{R}_{\text{stat}}$ discrepant ROIs (*i.e.*, $\mathcal{R}'_{\text{stat}}$). To evaluate the validity of those ROIs, we utilized LiCoL as a classifier and separated the set of ROIs into three categories: (i) baseline ($\mathcal{R}_{\text{eff}}$ and $\mathcal{R}_{\text{stat}}$); (ii) intersection ($\mathcal{R}_{\text{eff}} \cap \mathcal{R}_{\text{stat}}$); and (iii) union ($\mathcal{R}_{\text{eff}} \cup \mathcal{R}'_{\text{stat}}$ or $\mathcal{R}_{\text{stat}} \cup \mathcal{R}'_{\text{eff}}$), as depicted in Table 4. We observed that the best predictive performance was achieved by the model that utilized the baseline $\mathcal{R}_{\text{eff}}$. We also found noteworthy results of using intersection and union from the $\mathcal{R}_{\text{eff}}$ (Ours), which showed a slightly decreasing tendency compared with the baseline in several scenarios. That is, $\mathcal{R}'_{\text{stat}}$ could be regarded as ROIs that would not significantly contribute to the improvement of predictive performance despite showing statistical significance. Moreover, the union with $\mathcal{R}_{\text{stat}}$ exhibited an average performance improvement of 1.58% over baseline in all scenarios. In this respect, we concluded that $\mathcal{R}'_{\text{eff}}$ contains meaningful disease-related variations, which could be defined as potential biomarkers.

According to these comprehensive results, we investigated two instances over the set of significant ROIs, that is, the $\mathcal{R}'_{\text{eff}}$ and intersection ROIs that were revealed across all scenarios of both ADNI and GARD. These intersection ROIs were L.PreCG, L/R.SFG, L/R.MFG, L.IFGtriang, L.REC, R.INS, L/R.HIP, L.IPG, R.SMG, R.ANG, L/R.STG, L/R.MTG, L/R.ITG, and R.ACCpre. It can be observed that our $\mathcal{R}_{\text{eff}}$ was highlighted well and reflected prominent regions resulting from anatomical variations at each clinical stage,

Fig. 6: Illustration of selected common ROIs and discrepant ROIs $\mathcal{R}'_{\text{eff}}$ (*i.e.*, counterfactual-guided ROIs) on ADNI and GARD using 3D volume mapping. The first and second columns show the selected significant regions that correspond to AD and CN clinical stages in a longitudinal sample, and the cCN in the third column refers to the abbreviation of the CN-labeled cGM.

because those discovered regions are known as one of the most prevalent AD-related landmarks for AD progression [44, 52]. As $\mathcal{R}'_{\text{eff}}$, L.PFCventmed was observed across all scenarios for ADNI, and the CN *vs.* MCI scenario for GARD, and R.PFCventmed was discovered within the CN *vs.* MCI scenario for both ADNI and GARD. In particular, in ADNI, we observed that L.MCC, L.MOG, R.IPG, L.SMG, L/R.IFGoperc, and R.PCUN were shown as $\mathcal{R}'_{\text{eff}}$ in the CN *vs.* MCI scenario, whereas L.ACCpre was discovered in all scenarios. Moreover, L.ROL was found only in the MCI *vs.* AD scenario, whereas L.SMA appeared in both the CN *vs.* MCI and MCI *vs.* AD scenarios, and L.ACCpre was identified in all scenarios. On GARD, we discovered L.PCUN in the CN *vs.* MCI scenario. In Fig. 6, we adopt common and discrepant ROIs that appeared on ADNI and GARD as mentioned and visualize them via 3D volume mapping using longitudinal samples. We found that the density values of the selected region from the cGM progressively increased/decreased at specific region boundaries, similar to natural disease progression, and thoroughly mimicked specific regions with dense highlights in the CN.

Following the investigation, we believe that those $\mathcal{R}'_{\text{eff}}$ were considerably related to MCI and AD progression; hence $\mathcal{R}'_{\text{eff}}$ may cause predictive performance enhancement. In particular, the cognitive (*e.g.*, recognition, language, attention, *etc.*), sensory, or motor functions deteriorate in MCI and AD. Intriguingly, these $\mathcal{R}'_{\text{eff}}$ also possessed the ROIs related to cognitive and motor functions. For instance, R.INS and L/R.HIP is related to cognitive function [53]. Furthermore, information is exchanged between R.INS, L/R.ACCpre, and L.MCC [54]. L/R.SFG is connected with motor control regions, such as L.PreCG, and L/R.SFG further appeared as a region that was also related to motor function along with L.SMA [55]. L/R.STG, R.IPG, L/R.MTG, and R.PCUN were related to sensory functions, such as sound recognition and localization [56]. L/R.SMG and R.ANG were further associated with language and visual word recognition [57]. By doing this, we clinically demonstrated that the DL model well-generated the target properties of which regions in the brain to change the classifier's prediction to the target class as the brain disease progressed through quantitative analysis of counterfactual-guided deep features based on GM density. Consequently, we claim that our method could be a stepping stone toward a comprehensible explanation of the deep models' decisions from a clinician's perspective on neurodegenerative disease prediction.

4 Methods

4.1 Counterfactual Visual Explanation Model

The goal of the counterfactual visual explanation approach proposed in our LEAR [31] is to produce counterfactual reasoning for the decision of a predictive model $\mathcal{C}$ in diagnosing clinical stages in the AD spectrum (*i.e.*, CN/MCI/AD). It consists of several key components, namely the CMG, reasoning evaluator (RE), and discriminator (DC). Specifically, the CMG generates a counterfactual map conditioned on an arbitrary target label, whereas

the RE effectively guides the CMG towards comprehending the target label attributes in synthesizing realistic desired images. In addition, by exploiting the DC for carefully appraising the real and synthesized images (*i.e.*, r-sMRIs and c-sMRIs), the CMG will eventually ensure that the c-sMRIs are constrained as being realistic conversions. It should be noted that the RE is defined as the pre-trained predictive model $\mathcal{C}$, and the structure of the DC is identically imitated in the predictive model $\mathcal{C}$. Initially, we carried out pre-training for the predictive model $\mathcal{C}$ (see Supplementary B) by supervised learning with training samples $\mathbf{X}$ and one-hot encoded ground-truth labels $\mathbf{y}$:

$$\mathcal{L}_{\text{cls}}^{\mathcal{C}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\text{CE}\left(\mathcal{C}(\mathbf{X}), \mathbf{y}\right)\right], \tag{1}$$

where CE denotes a cross-entropy function. Henceforth, during the subsequent training phase to generate the counterfactual visual explanation, we fix the weights of the RE while jointly tuning the trainable parameters of DC $\mathcal{D}_{\psi}$ with the generator $\mathcal{G}_{\phi}$ in CMG.

4.2 CMG

The CMG is a variant of conditional GAN [29] devised to effectively synthesize a counterfactual map conditioned on a target label $\mathbf{t}$, where $\mathbf{t} \in [0,1]^{|\mathcal{Y}|}$ with $|\mathcal{Y}|$ denotes the size of the class space $\mathcal{Y}$. Suppose that the c-sMRI can be inferred through $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$ given an input r-sMRI $\mathbf{X}$, along with a target-specific counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$. The CMG shall be optimized in generating these maps $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ such that the c-sMRI $\tilde{\mathbf{X}}$ is diagnosed as the target label $\mathbf{t}$ with high confidence. Specifically, the architecture of the CMG is comprised of the encoder $\mathcal{E}_{\theta}$ and the generator $\mathcal{G}_{\phi}$, that is, a variant of U-Net [58] with a target label $\mathbf{t}$ concatenated to the skip connections. The subscripts θ and ϕ of the encoder $\mathcal{E}_{\theta}$ and generator $\mathcal{G}_{\phi}$ indicate the trainable parameters of the respective networks. Note that we replicate $\mathcal{E}_{\theta}$ from the pre-trained predictive model $\mathcal{C}$ in both its architecture and its weights (fixed). Thus, the encoder $\mathcal{E}_{\theta}$ is adequate for extracting a meaningful representation that reflects the class-relevant discriminative features from an input r-sMRI $\mathbf{X}$.

Because we require the CMG to take into account the target label $\mathbf{t}$ to produce the counterfactual map, we fuse the target-specific characteristics onto the feature maps obtained from the encoder $\mathcal{E}_{\theta}(\mathbf{X})$ via a concatenation operation. For this purpose, we tile the target label to match the shape corresponding to the respective feature maps of the l-th convolution layer. The size of the tiled target label is $w_l \times h_l \times d_l \times c_l$, where w_l, h_l, and d_l denote, respectively, the width, height, and depth of a feature map from the l-th convolution block, and c_l denotes the number of channels. To obtain the hierarchical discriminative representations *w.r.t* the target label, we devise an additional module that comprises a convolution operation (Conv3D) with a trainable $3 \times 3 \times 3$ kernel, a stride of one in each dimension, and zero padding, followed by a nonlinear

Leaky-ReLU activation function (LReLU):

$$\tau(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t}) = \text{LReLU}\left(\text{Conv3D}\left(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})} \oplus \text{Tile}(\mathbf{t})\right)\right), \tag{2}$$

where $\oplus$ denotes an operator of channel-wise concatenation, and $\{\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}\}_{l=1}^{L}$ denote the output feature maps of the L convolution layers from the encoder $\mathcal{E}_\theta(\mathbf{X})$. Subsequently, the target-fused feature maps $\tau(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t})$ are transferred to the generator $\mathcal{G}_\phi$ via skip connections. The generator $\mathcal{G}_\phi$ is then capable of seamlessly generating a counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ from the target label-informed feature maps:

$$\mathbf{M}_{\mathbf{X},\mathbf{t}} = \mathcal{G}_\phi\left(\mathcal{T}(\mathbf{X}, \mathbf{t})\right), \tag{3}$$

where $\mathcal{T}(\mathbf{X}, \mathbf{t}) = \{\tau(\mathbf{F}_1^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t}), ..., \tau(\mathbf{F}_L^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t})\}$. Lastly, we produce a c-sMRI by incorporating the counterfactual map with an input r-sMRI $\mathbf{X}$ using addition, *i.e.*, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$, which is supposed to be diagnosed as the target label $\mathbf{t}$.

4.3 Objective Function for CMG

This subsection thoroughly presents each essential objective function for optimizing the overall counterfactual visual explanation framework to effectively generate a refined explanatory map by using an adversarial learning strategy in conjunction with several auxiliary losses.

To generate a realistic c-sMRI $\tilde{\mathbf{X}}$, we employ adversarial training. In particular, we adopt the least square GAN (LSGAN) [59] loss function, which enforces a stable optimization by penalizing samples that are far from the DC's decision boundary. Guided by this loss, the DC assists the CMG such that it will thoughtfully minimize a substantial distance between the real and generated distributions. Thus, the DC $\mathcal{D}_\psi$ and the generator $\mathcal{G}_\phi$ loss are defined as follows, respectively:

$$\mathcal{L}_{\text{adv}}^{\mathcal{D}_\psi} = \mathbb{E}_{\bar{\mathbf{X}}\sim P_{\mathbf{X}}}\left[(\mathcal{D}_\psi(\bar{\mathbf{X}}) - 1)^2\right] + \frac{1}{2}\left(\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\mathcal{D}_\psi(\tilde{\mathbf{X}})^2 + \mathcal{D}_\psi(\mathbf{X}')^2\right]\right), \tag{4}$$

$$\mathcal{L}_{\text{adv}}^{\mathcal{G}_\phi} = \frac{1}{2}\left(\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[(\mathcal{D}_\psi(\tilde{\mathbf{X}}) - 1)^2 + (\mathcal{D}_\psi(\mathbf{X}') - 1)^2\right]\right), \tag{5}$$

where $\tilde{\mathbf{X}}$ and $\mathbf{X}' = \tilde{\mathbf{X}} + \mathbf{M}_{\tilde{\mathbf{X}},\mathbf{t}'}$ denote the c-sMRI and its respective override c-sMRI (will also be utilized for cycle consistency), respectively, and $\bar{\mathbf{X}}$ denotes a r-sMRI sampled from $P_{\mathbf{X}}$. Note that $\mathbf{M}_{\tilde{\mathbf{X}},\mathbf{t}'}$ indicates the counterfactual map overriding the c-sMRI $\tilde{\mathbf{X}}$ to obtain $\mathbf{X}'$ which is expectedly to resemble the r-sMRI $\mathbf{X}$ conditioned by the posterior probability $\mathbf{t}' = \mathcal{C}(\mathbf{X})$.

Because the DC is employed merely to differentiate between the input r-sMRI $\mathbf{X}$ and the generated c-sMRI $\tilde{\mathbf{X}}$, it does not have the immediate capacity to explicitly guide the CMG to either preserve the appearance of the input or endorse the target attribution during the generative process. Therefore, we employ an auxiliary cycle consistency loss [60] based on the ℓ_1-norm to

produce enhanced target-dependent counterfactual maps. In this way, we force the CMG to consistently strive for various conditions without suffering from a mode collapse issue [61]:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}},\mathbf{t}\sim U(0,|\mathcal{Y}|)} \left\| \mathbf{X}' - \mathbf{X} \right\|_1 , \tag{6}$$

where $P_{\mathbf{X}}$ denotes the distribution of MRI samples, and $\mathbf{t} \sim U(0, |\mathcal{Y}|)$ is the one-hot encoded vector drawn randomly from a discrete uniform distribution.

We further utilize the total variation loss to reconcile the elaborate synthesis between the input r-sMRI $\mathbf{X}$ and its counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$. In particular, total variation loss ensures that the c-sMRI $\tilde{\mathbf{X}}$ imposes local spatial continuity and smoothness to mitigate unnatural and overly pixelated results, which encourages visual coherence in a real image. This regularizer is denoted as follows:

$$\mathcal{L}_{\text{tv}} = \sum_{i,j,k} \left| \tilde{\mathbf{X}}_{i+1,j,k} - \tilde{\mathbf{X}}_{i,j,k} \right| + \left| \tilde{\mathbf{X}}_{i,j+1,k} - \tilde{\mathbf{X}}_{i,j,k} \right| + \left| \tilde{\mathbf{X}}_{i,j,k+1} - \tilde{\mathbf{X}}_{i,j,k} \right| , \tag{7}$$

where $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$ and i, j, and k are the 3D coordinate of each index in volumetric images.

Moreover, elastic regularization is applied to the CMG to impose a constraint on the magnitude of the counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$. By doing this, the vast majority of fine-grained feature attributions for counterfactual reasoning are highlighted such that minimal modifications over the features induce a converted prediction from the posterior probability $\mathbf{t}' = \mathcal{C}(\mathbf{X})$ to the target label $\mathbf{t}$.

$$\mathcal{L}_{\text{map}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}} \left[\lambda_1 \left\| \mathbf{M}_{\mathbf{X},\mathbf{t}} \right\|_1 + \lambda_2 \left\| \mathbf{M}_{\mathbf{X},\mathbf{t}} \right\|_2 \right] , \tag{8}$$

where λ_1 and λ_2 are weighting constants as the hyperparameter.

As adequate assistance to the generator, we further utilize the classification loss function to produce counterfactual maps that transform an input r-sMRI $\mathbf{X}$ so that it is precisely classified as a target label $\mathbf{t}$:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}} \left[\text{CE} \left(\tilde{\mathbf{y}}, \mathbf{t} \right) \right] , \tag{9}$$

where $\tilde{\mathbf{y}} = \mathcal{C}(\tilde{\mathbf{X}})$ is a softmax activated probability over the c-sMRI $\tilde{\mathbf{X}}$.

Consequently, we define the composite objective function for the overall counterfactual map generation task as follows:

$$\mathcal{L}_{\text{CMG}} = \lambda_3 \mathcal{L}_{\text{adv}}^{\mathcal{G}_\phi} + \lambda_4 \mathcal{L}_{\text{adv}}^{\mathcal{D}_\psi} + \lambda_5 \mathcal{L}_{\text{cyc}} + \lambda_6 \mathcal{L}_{\text{cls}} + \lambda_7 \mathcal{L}_{\text{tv}} + \mathcal{L}_{\text{map}}, \tag{10}$$

where λ_* values are the hyperparameters of the model (including $\lambda_{1,2}$ in Eq. (8)). We empirically tune λ such that the magnitudes of the gradients for each loss term are roughly balanced.

4.4 Quantifying Explainability of Counterfactual-Guided Deep Features

Having sufficiently optimized the counterfactual map generation task, we extend our work [31] in quantifying the explainability of the acquired counterfactual-guided deep features in terms of AD diagnoses. For this purpose, we carefully devise steps, including i) establishment of image manipulation, ii) extraction of AD-related landmarks, and finally iii) development of a lightweight counterfactual-guided attentive feature representation and a linear classifier (LiCoL). The procedural steps of this in-depth analysis are depicted in Fig. 1.

4.5 Establishment of Image Manipulation

The goal of this image manipulation step is to establish a proportional approach to quantifying the explainability of deep features as a GM density map. The utilization of a GM density map enables us to carry out accurate numerical measurements and quantitative feature-based analysis for brain disease. To this end, we reverse the preprocessing procedures used to obtain the train-ready input (*i.e.*, r-sMRIs) $\mathbf{X}$ from a raw brain image $\mathbf{B}$ to obtain the GM density map, including counterfactual-guided deep features. Note that the preprocessing of raw brain sMRI scans is described in Supplementary A.

Specifically, the r-sMRI $\mathbf{X}$ and the c-sMRI $\tilde{\mathbf{X}}$ are fed into a series of reverse preprocessing steps consisting of reverse Gaussian normalization, reverse quantile normalization, and up-scaling steps. To calculate the reverse Gaussian normalization, each subject's mean and standard deviation values are stored during preprocessing and reused to obtain the original values. The subsequent step is the reverse quantile normalization, which is intractable by means of the conventional approach. To circumvent this issue, we resort to an alternative solution. In particular, rather than consider the original raw brain image $\mathbf{B}$, we instead adopt the input of the model $\mathbf{X}$ as the alternative ground-truth. A similar reverse preprocessing is performed over the c-sMRI $\tilde{\mathbf{X}}$. Thereafter, the final reversed raw image $\mathbf{B}'$ is acquired using an up-scaling operation to match the size of the raw brain image $\mathbf{B}$ (*i.e.*, $193 \times 229 \times 193$).

Finally, auxiliary steps are carried out to obtain the GM density map over a set of $\mathbf{B}'$. There are four kinds of steps: (i) brain-extracted images are segmented into GM, white matter (WM), and cerebrospinal fluid (CSF) volume probability maps using FMRIB's automated segmentation tool (FAST) [62]; (ii) a GM template is created by registering into MNI152 space (to avoid bias during the registration process) with FMRIB's linear affine registration tool (FLIRT). After nonlinear registration using FMRIB's nonlinear registration tool (FNIRT), the resulting images are averaged to create the template; (iii) in addition, all the GM images are nonlinearly registered to the template and modulated using the Jacobian of the warp field; (iv) these images are then smoothed with an isotropic Gaussian kernel with a σ of 2 mm to alleviate contrast and other irrelevant details in the image.

4.6 Extraction of AD-Related Landmarks

Given a set of manipulated GM density maps, we further investigate the validity of counterfactual-guided deep features by exploiting the GM density-based analysis. Let $\hat{\mathbf{X}}^r$ and $\hat{\mathbf{X}}^c$ denote the rGM and the cGM manipulated by an r-sMRI $\mathbf{X}$ and the c-sMRI $\tilde{\mathbf{X}}$, respectively. We first subtract rGMs $\hat{\mathbf{X}}^r$ from their corresponding cGMs $\hat{\mathbf{X}}^c$ in different clinical stages to acquire the stage-wise difference maps. For instance, for a CN-labeled rGM, the counterpart of its clinical label for the cGM shall be defined as the AD or MCI, depending on the given scenario. The favorable properties of difference maps have a discriminative capability to reflect inherent disease-associated regional localization, as well as the unique anatomical characteristics of the inter-subject variability. These difference maps are then averaged into a representative difference map, which we refer to as the AD-effect map $\mathbf{M}_{\text{eff}}$, by exploiting the first moment in statistics (*i.e.*, the mean of difference maps) using:

$$\mathbf{M}_{\text{eff}} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{\mathbf{X}}_i^r - \hat{\mathbf{X}}_i^c \right|, \tag{11}$$

where N and $|\cdot|$ denote the total number of samples and the absolute operation, respectively. Such an AD-effect map $\mathbf{M}_{\text{eff}}$ is finally masked by the percentile threshold that clearly reveals the most frequently highlighted regions among the set of fine-grained regions (illustrated in Fig. 1a).

4.7 AD-Effect ROI Composition

The brain parcellation not only provides an understanding of the fundamental brain organization and function of closely interacting regions, but it also compresses information from hundreds of thousands of voxels or vertices into manageable sets. Because of these propitious characteristics, we adopt an ROI-based analysis that extracts and analyzes significant ROIs across voxel-level values that appear on the AD-effect map $\mathbf{M}_{\text{eff}}$. To this end, the automated anatomical labeling (AAL3) atlas [63] (see Supplementary I) is overlaid on the AD-effect map to further subdivide it into individual regions, so that each trimmed region is completely within the cortical and subcortical regions and has anatomical specificity. Subsequently, we discern parcellated regions and their voxel indices $\mathcal{V}$ with prominently highlighted voxels in the AD-effect map $\mathbf{M}_{\text{eff}}$. By averaging the voxel values of indices $\mathcal{V}$ within the adopted parcellated regions, we eventually define the set of representative ROIs, so-called AD-effect ROIs, that consist of R number of ROIs as $\mathbf{x}_q \in \mathbb{R}^{R \times 1}$.

4.8 LiCoL

To further verify the effectiveness of the acquired AD-effect ROIs $\mathbf{x}_q$, along with their respective voxel indices $\mathcal{V}$, we perform an objective classification task. Inspired by Transformer [64], we devise a LiCoL that takes into account

the global relationships among ROIs and effectively emphasizes the most important ROIs by itself (illustrated in Fig. 1b).

In a nutshell, our proposed LiCoL maps a set of ROIs (*i.e.*, the type of vectors) comprised of query, key, and value onto the input. We simply establish the query as AD-effect ROIs $\mathbf{x}_q$ that reflect the distinct discrimination subject to statistical significance to adequately guide the attention over the LiCoL. Unlike the $\mathbf{x}_q$ for the query that is fixed during the training step, the key and value are specifically devised to extract significant ROIs from each training sample independently according to voxel indices $\mathcal{V}$ as $\mathbf{x}_s \in \mathbb{R}^{R\times 1}$. Subsequently, by using the embedding layer $\mathbf{w} \in \mathbb{R}^{1\times D}$, the query, key, and value are transformed into the set of embedded matrices $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{R\times D}$, accordingly. We then compute the counterfactual-guided attention matrix of outputs as:

$$\text{Attention}(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = g\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{12}$$

where $\top$ and d respectively denote a transpose operation and the number of ROIs used from the key for scaling, while g denotes the softmax function. Note that the set $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V}$ represent the embedded matrices produced from a linear transformation, $\mathbf{Q} = \mathbf{x}_q\mathbf{w}_q$, $\mathbf{K} = \mathbf{x}_s\mathbf{w}_k$, and $\mathbf{V} = \mathbf{x}_s\mathbf{w}_v$, respectively, where $\mathbf{w}_q, \mathbf{w}_k$, and $\mathbf{w}_v$ are learnable weights.

Having applied the channel-wise mean pooling to reshape the output of the counterfactual-guided attention map to match the size of the input query $\mathbf{x}_q$, we then employ a residual connection using an element-wise addition, followed by an ROI-wise mean pooling to obtain the final predicted label $\bar{y}$ as:

$$\bar{y} = \sigma\left(\text{MP}_\downarrow\left(\text{MP}_\rightarrow\left(g\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}\right) + \mathbf{x}_q\right)\right) \tag{13}$$

where $\text{MP}_\downarrow$ and $\text{MP}_\rightarrow$ denote the ROI-wise mean pooling and the channel-wise mean pooling, respectively, and σ denotes the logistic-sigmoid function.

Eventually, we train the LiCoL by minimizing the classification loss via CE as:

$$\mathcal{L}_{\text{cls}}^{\text{LiCoL}} = \mathbb{E}_{\hat{\mathbf{X}}\sim P_{\hat{\mathbf{X}}}}\left[\text{CE}\left(\bar{\mathbf{y}}, \mathbf{y}\right)\right], \tag{14}$$

where $\hat{\mathbf{X}}$ and $\mathbf{y}$ denote a GM density map and its ground-truth label, respectively, and $\bar{\mathbf{y}}$ denotes one-hot embedded predicted label. It should be noted that during training, the LiCoL is learned by means of both rGMs and cGMs, but it utilizes only test samples acquired from the rGMs $\hat{\mathbf{X}}^r$ to evaluate the performance of the LiCoL.

4.9 Datasets

In order to evaluate the validity of the proposed framework, we conducted a series of comprehensive analyses on the ADNI and GARD datasets. Both datasets were comprised of 3D sMRIs acquired from various patient groups ranging from the CN to those having the MCI and AD. These three categories

were annotated based on the standard clinical criteria, including mini-mental state examination (MMSE) scores and clinical dementia ratings. Although some subjects had accumulated multiple MRIs through their follow-up in both datasets, we only selected and utilized their baseline (first-visit) MRIs. The sMRI preprocessing and details of the demographic information regarding the subjects are summarized in Supplementary A.

The ADNI dataset consisted of 1,540 sMRI scans that included CN subjects (n = 433), MCI subjects (n = 748), and AD subjects (n = 359) from the combined ADNI-1 and ADNI-2 studies. As a special case for ADNI, there existed two subgroups of MCI subjects; these were the progressive MCI (pMCI) (n = 251), including MCI subjects who have converted to AD within 36 months of screening, and the stable MCI (sMCI) (n = 497), including those who remained in the MCI group within 36 months of screening. The baseline of the ADNI-1 included 1.5T T1-weighted sMRI scans acquired from 822 subjects, and ADNI-2 included 3T T1-weighted sMRI scans acquired from 718 subjects. Meanwhile, the baseline GARD dataset contained 1.5T T1-weighted sMRI scans acquired from 745 subjects, including CN subjects (n = 261), MCI subjects (n = 375), and AD subjects (n = 109).

In addition, for the faithful measurement of intra-subject variability in brain atrophies related to clinical-stage conversion, we utilized auxiliary longitudinal samples of 12 subjects in ADNI-1/-2, and 9 subjects in the GARD dataset. Such selected longitudinal samples from both ADNI and GARD datasets were gathered using the same criteria: those who converted from CN (baseline) to MCI and AD at some point in their historical record. It should be noted that none of these selected longitudinal images were used in any of our model training procedures. Details of the subject ID with regard to each subject and image ID corresponding to CN, MCI, and AD are reported in Supplementary D.

4.10 Experimental Setup

Because the ADNI dataset consists of four categories (*i.e.*, CN, sMCI, pMCI, and AD), we merged the subjects of the sMCI and pMCI as MCI to maintain consistency with the GARD dataset. Henceforth, we evaluated both datasets for three scenarios: (i) CN *vs.* MCI, (ii) MCI *vs.* AD, and (iii) CN *vs.* AD. All experiments and evaluations were carried out using five-fold cross-validation.

4.11 Evaluation Metrics

To affirm the generalizability of the cGM for augmentation, we used precision and recall, which are preferred over other alternatives when the class distribution is significantly skewed [65]. All experiments for the predictive performance with GM density maps were quantitatively evaluated in terms of four criteria, that is, AUC, accuracy, sensitivity, and specificity. Note that these evaluation metrics were measured from test samples for both preprocessed types (*i.e.*, linear and non-linear registration).

4.12 Statistical Analysis

For the AD-effect map acquisition and augmentation of cGMs for the statistical map, we measured the statistical significance of categorizing with the same population using a two-sample t-test (p-value < 0.05) to group the same clinical stage of the rGMs and cGMs. This statistical test indicates that two independent samples have identical expected values; thus, it can be assumed that the cGMs are adequately mapped on the identical distribution of rGMs, respectively. In a nutshell, it was possible to consider the AD-labeled cGM and the AD-labeled rGM as the same AD category.

Meanwhile, we determined the value of the percentile threshold and a p-value based on the validation performance reported in Supplementary E. Specifically, we used a percentile threshold of 0.99 to obtain the fine-grained AD-effect map because this threshold achieved the highest validation performance in all scenarios. For comparison with the AD-effect map, we derived the conventional group-level statistical map by measuring the two-sample t-test (p-value < 0.01) between different stages of rGMs.

4.13 Criteria of ROI Selection

We set the minimum number of voxels to ten for a single-cluster composition for ROI selection. This was based on the statistical thresholding from the study by Lieberman *et al.* [66], which denoted a desirable balance of Type I and Type II error when using $p < 0.005$ and a ten-voxel threshold. Some imaging toolboxes follow a minimum ten-voxel threshold for a cluster setting [67].

Supplementary Information. We provide more details information regarding the proposed method architecture as well as exhaustive experimental results, which is note covered in the main manuscript.

Acknowledgments. 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)).

Author contributions. K.S.O and H.-I.S conceived the study and designed the methodology. K.H.L provided the GARD dataset comprised of raw 3D sMRI for the experiment. W.S.J and K.S.O preprocessed the data and constructed the design of figures and tables together with D.W.H. K.S.O wrote the implementation code and performed the related experiments. D.W.H and K.S.O provided the statistical analysis and interpretation of the experimental results. A.W.M supported K.S.O in drafting the manuscript under the supervision of H.-I.S. D.W.H, W.S.J, and E.S.K critically discussed the conceptualization of the experimental premise with K.S.O. All authors were involved in crucial revisions of the manuscript and have read and approved the final version.

Data availability. Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wpcontent/uploads/howtoapply/ADNIAcknowledgementList.pdf. In contrast, Gwangju Alzheimer's and Related Dementia (GARD) cohort dataset has not been published. The raw sMRI data are protected and are not made publicly available owing to data privacy laws, but are available from the corresponding author upon reasonable request. Demographic information over the source datasets is provided in this article.

Code availability. The public toolkit of FSL and the AAL3 atlas template are available at: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki and https://www.gin.cnrs.fr/en/tools/aal/, respectively. Source code generated and used for this study is publicly available for download at https://github.com/ku-milab/LiCoL.

References

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

[2] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., *et al.*: The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception. Alzheimer's & Dementia **9**(5), 111–194 (2013)

[3] Cummings, J.L., Doody, R., Clark, C.: Disease-modifying therapies for Alzheimer disease: Challenges to early intervention. Neurology **69**(16), 1622–1634 (2007)

[4] Petersen, R.C., Roberts, R.O., Knopman, D.S., Boeve, B.F., Geda, Y.E., Ivnik, R.J., Smith, G.E., Jack, C.R.: Mild cognitive impairment: Ten years later. Archives of Neurology **66**(12), 1447–1455 (2009)

[5] Schneider, J.A., Arvanitakis, Z., Leurgans, S.E., Bennett, D.A.: The neuropathology of probable Alzheimer disease and mild cognitive impairment. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society **66**(2), 200–208 (2009)

[6] Wu, J., Li, C., Gensheimer, M., Padda, S., Kato, F., Shirato, H., Wei, Y., Schönlieb, C.-B., Price, S.J., Jaffray, D., Heymach, J., Neal, J.W., Jr, B.W.L., Wakelee, H., Diehn, M., Li, R.: Radiological tumor classification

across imaging modality and histology. Nature Machine Intelligence, 787–798 (2021)

[7] Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., *et al.*: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**(1), 1–9 (2014)

[8] Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M.: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communications **7**(1), 1–10 (2016)

[9] Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)

[10] López, M., Ramírez, J., Górriz, J., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R.: SVM-based CAD system for early detection of the Alzheimer's disease using kernel PCA and LDA. Neuroscience Letters **464**(3), 233–238 (2009)

[11] Zhou, H., Jiang, J., Lu, J., Wang, M., Zhang, H., Zuo, C., Alzheimer's Disease Neuroimaging Initiative and others: Dual-model radiomic biomarkers predict development of mild cognitive impairment progression to Alzheimer's disease. Frontiers in Neuroscience **12**, 1045 (2019)

[12] Lee, S., Lee, H., Kim, K.W.: Magnetic resonance imaging texture predicts progression to dementia due to Alzheimer disease earlier than hippocampal volume. Journal of Psychiatry and Neuroscience **45**(1), 7–14 (2020)

[13] Ij, H.: Statistics versus machine learning. Nature Methods **15**(4), 233 (2018)

[14] Li, Z., Duan, H., Zhao, K., Ding, Y.: Stability of MRI radiomics features of hippocampus: An integrated analysis of test-retest and inter-observer variability. IEEE Access **7**, 97106–97116 (2019)

[15] Zhao, K., Ding, Y., Han, Y., Fan, Y., Alexander-Bloch, A.F., Han, T., Jin, D., Liu, B., Lu, J., Song, C., *et al.*: Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: Diagnosis, longitudinal progress and biological basis. Science Bulletin **65**(13), 1103–1113 (2020)

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

[17] Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D.: Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific Reports **11**(1), 1–13 (2021)

[18] Zhang, X., Han, L., Zhu, W., Sun, L., Zhang, D.: 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)

[19] Yang, Z., Nasrallah, I.M., Shou, H., Wen, J., Doshi, J., Habes, M., Erus, G., Abdulkadir, A., Resnick, S.M., Albert, M.S., *et al.*: A deep learning framework identifies dimensional representations of Alzheimer's disease from brain structure. Nature Communications **12**(1), 1–15 (2021)

[20] Mulyadi, A.W., Jung, W., Oh, K., Yoon, J.S., Suk, H.-I.: Xadlime: explainable Alzheimer's disease likelihood map estimation via clinically-guided prototype learning. arXiv preprint arXiv:2207.13223 (2022)

[21] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

[22] Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), and Web **2**(2), 1 (2017)

[23] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)

[24] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: Deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8930–8941 (2019)

[25] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning, pp. 3319–3328 (2017)

[26] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, pp. 618–626 (2017)

[27] Wang, Y., Su, H., Zhang, B., Hu, X.: Learning reliable visual saliency for model explanations. IEEE Transactions on Multimedia **22**(7), 1796–1807 (2019)

[28] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in Neural Information Processing Systems **27** (2014)

[29] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[30] Dash, S., Balasubramanian, V., Sharma, A.: Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In: Winter Conference on Applications of Computer Vision, pp. 915–924 (2022)

[31] Oh, K., Yoon, J.S., Suk, H.-I.: Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–15 (2022). https://doi.org/10.1109/TPAMI.2022.3197845

[32] Byrne, R.M.: Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: International Joint Conference on Artificial Intelligence, pp. 6276–6282 (2019)

[33] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: 36th International Conference on Machine Learning, pp. 2376–2384 (2019)

[34] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)

[35] Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y., Park, S.H.: Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. Korean Journal of Radiology **20**(3), 405–410 (2019)

[36] Schwarz, C.G., Gunter, J.L., Wiste, H.J., Przybelski, S.A., Weigand, S.D., Ward, C.P., Senjem, M.L., Vemuri, P., Murray, M.E., Dickson, D.W., *et al.*: A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. NeuroImage: Clinical **11**, 802–812 (2016)

[37] Zhang, X., Mormino, E.C., Sun, N., Sperling, R.A., Sabuncu, M.R., Yeo, B.T., Initiative, A.D.N.: Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. Proceedings of the National Academy of Sciences **113**(42), 6535–6544 (2016)

[38] Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., Initiative, A.D.N., *et al.*: Learning to synthesise the ageing brain without longitudinal data. Medical Image Analysis **73**, 102169 (2021)

[39] Hirata, Y., Matsuda, H., Nemoto, K., Ohnishi, T., Hirao, K., Yamashita, F., Asada, T., Iwabuchi, S., Samejima, H.: Voxel-based morphometry to discriminate early Alzheimer's disease from controls. Neuroscience Letters **382**(3), 269–274 (2005)

[40] Karas, G., Scheltens, P., Rombouts, S.A., Visser, P.J., van Schijndel, R.A., Fox, N.C., Barkhof, F.: Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. NeuroImage **23**(2), 708–716 (2004)

[41] Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M.: Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: The SPARE-AD index. Brain **132**(8), 2026–2035 (2009)

[42] Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics of North America **15**(4), 869–877 (2005)

[43] Choi, K.Y., Lee, J.J., Gunasekaran, T.I., Kang, S., Lee, W., Jeong, J., Lim, H.J., Zhang, X., Zhu, C., Won, S.-Y., *et al.*: APOE promoter polymorphism-219T/G is an effect modifier of the influence of APOE $\varepsilon 4$ on Alzheimer's disease risk in a multiracial sample. Journal of Clinical Medicine **8**(8), 1236 (2019)

[44] Risacher, S.L., Saykin, A.J., Wes, J.D., Shen, L., Firpi, H.A., McDonald, B.C.: Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Current Alzheimer Research **6**(4), 347–361 (2009)

[45] Ashburner, J., Friston, K.J.: Voxel-based morphometry—the methods. NeuroImage **11**(6), 805–821 (2000)

[46] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[47] Hebert, L.E., Weuve, J., Scherr, P.A., Evans, D.A.: Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology **80**(19), 1778–1783 (2013)

[48] Nelson, P.T., Head, E., Schmitt, F.A., Davis, P.R., Neltner, J.H., Jicha, G.A., Abner, E.L., Smith, C.D., Van Eldik, L.J., Kryscio, R.J., *et al.*: Alzheimer's disease is not "brain aging": Neuropathological, genetic, and epidemiological human studies. Acta Neuropathologica **121**(5), 571–587 (2011)

[49] DeGrave, A.J., Janizek, J.D., Lee, S.-I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7),

610–619 (2021)

[50] Mahmood, U., Shrestha, R., Bates, D.D., Mannelli, L., Corrias, G., Erdi, Y.E., Kanan, C.: Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems. Frontiers in Digital Health **3**, 671015 (2021)

[51] Simon, H.A.: Spurious correlation: A causal interpretation. Journal of the American Statistical Association **49**(267), 467–479 (1954)

[52] Liu, Y., Paajanen, T., Zhang, Y., Westman, E., Wahlund, L.-O., Simmons, A., Tunnard, C., Sobow, T., Mecocci, P., Tsolaki, M., *et al.*: Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups—the addneuromed study. Neurobiology of Aging **32**(7), 1198–1206 (2011)

[53] Gasquoine, P.G.: Contributions of the insula to cognition and emotion. Neuropsychology Review **24**(2), 77–87 (2014)

[54] Taylor, K.S., Seminowicz, D.A., Davis, K.D.: Two systems of resting state connectivity between the insula and cingulate cortex. Human Brain Mapping **30**(9), 2731–2745 (2009)

[55] Li, W., Qin, W., Liu, H., Fan, L., Wang, J., Jiang, T., Yu, C.: Subregions of the human superior frontal gyrus and their connections. NeuroImage **78**, 46–58 (2013)

[56] Maeder, P.P., Meuli, R.A., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., Pittet, A., Clarke, S.: Distinct pathways involved in sound recognition and localization: A human fMRI study. NeuroImage **14**(4), 802–816 (2001)

[57] Stoeckel, C., Gough, P.M., Watkins, K.E., Devlin, J.T.: Supramarginal gyrus involvement in visual word recognition. Cortex **45**(9), 1091–1096 (2009)

[58] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)

[59] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)

[60] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image

translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

[61] Goodfellow, I.: NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)

[62] Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging **20**(1), 45–57 (2001)

[63] Rolls, E.T., Huang, C.-C., Lin, C.-P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. NeuroImage **206**, 116189 (2020)

[64] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 6000–6010 (2017)

[65] Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)

[66] Lieberman, M.D., Cunningham, W.A.: Type I and Type II error concerns in fMRI research: Re-balancing the scale. Social Cognitive and Affective Neuroscience **4**(4), 423–428 (2009)

[67] Roiser, J., Linden, D., Gorno-Tempinin, M., Moran, R., Dickerson, B., Grafton, S.: Minimum statistical standards for submissions to neuroimage: Clinical. NeuroImage: Clinical **12**, 1045 (2016)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf