Received: 1 September 2023 | Revised: 22 April 2024 | Accepted: 25 April 2024

DOI: 10.1002/hbm.26707

RESEARCH ARTICLE

WILEY

Disease-driven domain generalization for neuroimaging-based assessment of Alzheimer's disease

Diala Lteif[1,2] | Sandeep Sreerama[2] | Sarah A. Bargal[3] | Bryan A. Plummer[1] | Rhoda Au[2,4,5,6,7,8] | Vijaya B. Kolachalama[1,2,8,9]

[1]Department of Computer Science, Boston University, Boston, Massachusetts, USA

[2]Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA

[3]Department of Computer Science, Georgetown University, Washington, DC, USA

[4]Department of Anatomy & Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA

[5]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA

[6]Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA

[7]The Framingham Heart Study, Boston, Massachusetts, USA

[8]Boston University Alzheimer's Disease Research Center, Boston, Massachusetts, USA

[9]Faculty of Computing & Data Sciences, Boston University, Boston, Massachusetts, USA

Correspondence
Vijaya B. Kolachalama, Department of Computer Science, Boston University, Boston, MA, USA.
Email: vkola@bu.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01-HL159620, R21-CA253498, R43-DK134273, RF1-AG062109; Karen Toffler Charitable Trust; American Heart Association, Grant/Award Number: 20SFRN35460031; National Institute on Aging's Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research, Grant/Award Number: P30-AG073104

Abstract

Development of deep learning models to evaluate structural brain changes caused by cognitive impairment in MRI scans holds significant translational value. The efficacy of these models often encounters challenges due to variabilities arising from different data generation protocols, imaging equipment, radiological artifacts, and shifts in demographic distributions. Domain generalization (DG) techniques show promise in addressing these challenges by enabling the model to learn from one or more source domains and apply this knowledge to new, unseen target domains. Here we present a framework that utilizes model interpretability to enhance the generalizability of classification models across various cohorts. We used MRI scans and clinical diagnoses from four independent cohorts: Alzheimer's Disease Neuroimaging Initiative (ADNI, $n = 1821$), the Framingham Heart Study (FHS, $n = 304$), the Australian Imaging Biomarkers & Lifestyle Study of Ageing (AIBL, $n = 661$), and the National Alzheimer's Coordinating Center (NACC, $n = 4647$). With this data, we trained a deep neural network to focus on areas of the brain identified as relevant to the disease for model training. Our approach involved training a classifier to differentiate between structural neurodegeneration in individuals with normal cognition (NC), mild cognitive impairment (MCI), and dementia due to Alzheimer's disease (AD). This was achieved by aligning class-wise attention with a unified visual saliency prior, which was computed offline for each class using all the training data. Our method not only competes with state-of-the-art approaches but also shows improved correlation with postmortem histology. This alignment with the gold standard evidence is a significant step towards validating the effectiveness of DG frameworks, paving the way for their broader application in the field.

KEYWORDS
Alzheimer's disease, cognitive impairment, domain generalization, magnetic resonance imaging

1 | INTRODUCTION

Dementia due to Alzheimer's disease (AD) is a progressive syndrome leading to loss of brain function that affects memory, thinking, language, judgment and behavior. The approach to dementia diagnosis involves careful consideration of the patient's demographics and symptoms, family, social and medical history, neurologic examination, cognitive, behavioral, and functional assessments along with neuroimaging (Hugo & Ganguli, 2014; McKhann et al., 2011). Magnetic resonance imaging (MRI) is typically recommended to evaluate the structural changes in the patient's brain that correspond to volume loss and atrophy patterns suggestive of AD and rule out other patterns indicative of non-AD dementias. Computational methods based on advanced machine learning techniques are increasingly considered to automatically process the MRI scans and classify persons with dementia due to AD from those with normal cognition (NC) and mild cognitive impairment (MCI) (Aderghal et al., 2017; Liu et al., 2015; Qiu et al., 2018; Qiu et al., 2020; Qiu et al., 2022). Some of recently reported frameworks have relied on training models using data collected from a single cohort followed by evaluation on independent test cohorts (Qiu et al., 2020). Such model development strategies can establish a proof-of-principle, but may lack generalizability because data collected from multiple cohorts contain variabilities stemming from independent scanning protocols, diversity of the study population and other sources. Furthermore, while recent advancements in public data sharing have made data more accessible, there is an increasing necessity to create models that yield findings which are both generalizable and consistent.

Recently, domain generalization (DG) approaches are being considered to train robust deep learning models that account for cohort-specific variabilities and work well across multiple datasets (Donini et al., 2018; Ghimire et al., 2020; Huang et al., 2020; Koh et al., 2021; Krueger et al., 2021; Li, Pan, et al., 2018; Li, Yang, et al., 2018; Zhang et al., 2020; Zhou et al., 2023). Most methods attempt to mitigate the distributional variance between domain-specific feature representations. We submit that additional aspects such as orienting the models to focus on disease-related information while performing model training can be a targeted approach to meet the objective of creating generalizable architectures for disease classification.

1.1 | Related work

DG frameworks are typically designed to learn a robust signal and a set of patterns possibly from single or multiple source domains with the aim of transferring them to unseen target domains. The expectation is that such frameworks lead to minimal performance degradation on the unseen target environment. In the setting of single-source DG, the model trained on this source learns robust representations that can generalize to out-of-distribution data. Single-source DG methods can also be applied to a multi-source setting, as training is done over pooled data across the different source domains (Zhou et al., 2023). Also, multiple source domains can be used for training domain-invariant feature representations that generalize well to unseen target data.

Most DG methods were originally designed to benchmark natural imaging datasets, with a limited number of frameworks focused on medical imaging data (Ghimire et al., 2020; Koh et al., 2021). A group of methods have been proposed to tackle DG via data manipulation, which could either be data augmentation or generation (Cubuk et al., 2020; Tobin et al., 2017; Volpi et al., 2018; Zhang et al., 2018; Zhou et al., 2020). One of those methods is Mixup (Zhang et al., 2018), a data-agnostic routine that constructs virtual training examples as convex combinations of pairs of examples and their labels sampled at random from the training distribution. Mixup is designed to regularize the neural network, encouraging it to adopt linear behavior between training examples (Zhang et al., 2018). Another group of methods belong to the use of representation learning to address domain shift, mainly by learning domain-invariant representations and feature disentanglement (Donini et al., 2018; Ganin et al., 2016; Huang et al., 2020; Krueger et al., 2021; Li, Pan, et al., 2018; Nguyen et al., 2021; Zellinger et al., 2017). Donini and co-workers proposed a multi-source algorithm that uses empirical risk minimization (ERM), which became the standard approach to the DG problem (Donini et al., 2018). ERM aims to minimize the training risk across all source domains. Recently, Kreuger and colleagues introduced risk extrapolation (REx) for out-of-distribution (OOD) generalization and proposed a penalty on the variance of training risks (V-REx) (Krueger et al., 2021). They showed that reducing differences in risks with V-REx can reduce a model's sensitivity to a wide range of extreme distributional shifts. Li et al., on the other hand, proposed using the maximum mean discrepancy (MMD) measure with autoencoders to align distributions across different domains via adversarial training (Li, Pan, et al., 2018). Another work introduced representation self-challenging (RSC) to force the model to discard dominant features activated on the training data and activate remaining features that correlate with ground-truth labels (Huang et al., 2020). Further, there exists a line of work that used meta-learning for DG. One of the proposed meta-learning methods was MLDG, meta-learning for domain generalization, which simulates domain shift during training by synthesizing virtual testing domains within each mini-batch (Li, Yang, et al., 2018).

Our approach represents a distinct advancement from prior research focused on learning domain-invariant feature representations. It uniquely contributes by employing interpretability techniques to extract disease-relevant information, which is then used for aligning features effectively. Related prior work used model explanations as means of disentangling domain-specific information from otherwise relevant features (Zunino et al., 2021). Contrastingly, our method utilizes the feature contributions leading to accurate predictions as a foundation of model-identified disease biology. This knowledge is then applied to direct the model's focus during training. We concentrate on the single-source DG setting, which is more practical in clinical environments where the model is trained on a single source

domain. The model's ability to generalize is subsequently assessed on external cohorts, which are considered the target domains.

1.2 | Contributions

Our work falls under the umbrella of medically interpretable machine learning, where we use feature contributions to adjust final predictions by emphasizing disease-relevant features. Through attention-based supervision, the model learns to focus on disease-correlated regions using pre-computed class-wise saliency map priors with voxel contributions. The main contributions of this paper are summarized as follows:

- We developed an interpretability-based computational framework to train deep neural networks that focus on model-identified disease regions of interest as a means to generalize across multiple cohorts.
- Using MRI scans and clinical data obtained from multiple cohorts, we developed a classifier that distinguishes between persons with NC, MCI and AD.
- We demonstrated that our method competes with state-of-the-art DG methods in the real-world single-source setting.
- Finally, we showed that our interpretable findings correlate strongly with postmortem histology, identifying disease presence in brain regions that are known to classically associate with disease.

2 | METHODS

2.1 | Study population

We obtained brain MRI scans and corresponding clinical and demographic data on participants from four different cohorts: the Alzheimer's Disease Neuroimaging Initiative (ADNI) $(n = 1821)$ (Petersen et al., 2010), National Alzheimer's Coordinating Center (NACC) $(n = 4647)$ (Beekly et al., 2004), the Australian Imaging Biomarkers & Lifestyle (AIBL) Study of Ageing $(n = 661)$ (Ellis et al., 2009), and the Framingham Heart Study (FHS) (Mahmood et al., 2014; Massaro et al., 2004) $(n = 304)$. There were 3697 cases with normal cognition (NC), 2323 cases with mild cognitive impairment (MCI), and 1413 cases with dementia due to Alzheimer's disease (AD) across all cohorts (Table 1). Statistical analysis of distributional variance, including variance in image quality and imaging equipment, across the four cohorts can be found in Figures S1–S3 of the supplement. Additionally, our study incorporated post-mortem histological evaluations of 23 participants from the ADNI and FHS cohorts, who deceased within 1 year after their last MRI scan. These assessments comprised pathology grades derived from three distinct stains, which quantified the extent

TABLE 1 Study population.

Dataset	Group [participants]	Age, years Mean [std]	Education, years Median [std]	Gender Male count (%)	MMSE Median [std]	APOE4 Positive count (%)
ADNI	NC [$n = 481$]	74.3 ± 6.0	16.3 ± 2.7	235 (48.9%)	29.1 ± 1.1	138 (29.6%)[a]
	MCI [$n = 971$]	72.8 ± 7.7	15.9 ± 2.8	572 (58.9%)	27.6 ± 1.8	438 (47.2%)[a]
	AD [$n = 369$]	74.9 ± 7.8	15.2 ± 3.0	203 (55.0%)	23.2 ± 2.1	229 (64.3%)[a]
	p-value	<.001	<.001	.001	<.001	<.001
NACC	NC [$n = 2524$]	69.8 ± 9.9	15.92 ± 3.0	871 (34.5%)	29.0 ± 1.3	599 (30.0%)[a]
	MCI [$n = 1175$]	74.0 ± 8.7	15.4 ± 3.4	555 (47.2%)	26.8 ± 2.5	322 (38.7%)[a]
	AD [$n = 948$]	75.0 ± 9.1	14.6 ± 3.6	431 (45.5%)	20.5 ± 5.7	346 (52.2%)[a]
	p-value	<.001	<.001	<.001	<.001	<.001
AIBL	NC [$n = 480$]	72.5 ± 6.2	N.A.	203 (42.3%)	28.7 ± 1.2	12 (2.5%)
	MCI [$n = 102$]	74.7 ± 7.1	N.A.	53 (52.0%)	27.1 ± 2.1	12 (11.8%)
	AD [$n = 79$]	73.3 ± 7.8	N.A.	33 (41.8%)	20.4 ± 5.5	14 (17.7%)
	p-value	.006	N.A.	.189	<.001	<.001
FHS	NC [$n = 212$]	73.4 ± 9.6	1.8 ± 0.9[b]	112 (52.8%)	28.1 ± 1.7	42 (20.2%)[a]
	MCI [$n = 75$]	76.2 ± 6.8	1.6 ± 1.0[b]	34 (45.3%)	27.2 ± 2.0	17 (23.6%)[a]
	AD [$n = 17$]	78.8 ± 7.2	1.8 ± 1.0[b]	4 (23.5%)	24.0 ± 2.1	7 (43.8%)[a]
	p-value	.007	.272	.049	<.001	.088

Note: MRI scans and corresponding clinical and demographic data were collected across four different cohorts: the Alzheimer's Disease Neuroimaging Initiative (ADNI), the National Alzheimer's Coordinating Center (NACC), the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL), and the Framingham Heart Study (FHS). The models were trained and tested to differentiate persons who have either normal cognition (NC), mild cognitive impairment (MCI) or dementia due to Alzheimer's disease (AD). Education information on the AIBL dataset was not available.

[a]Data were not available for some subjects.

[b]FHS education code: 0 = high school did not graduate, 1 = high school graduate, 2 = some college graduate, 3 = college graduate.

of disease in cortical and subcortical brain structures. In our approach, we adopted a single-source setting for DG. Here, the training, internal validation, and initial testing of our models were conducted using data from one source cohort. Subsequently, external validation and further testing were carried out on the target cohorts.

2.2 | Data selection criterion

To ensure uniformity and control for potential confounding factors, we uniformly applied a set of selection criteria across all cohorts, as detailed in Table 1. These criteria, derived from ADNI's baseline recruitment protocol (Petersen et al., 2010), were crucial in shaping our study's dataset. Our focus was on individuals aged 55 and above, a demographic choice reflective of AD characteristics, including the presence of brain atrophy observable in MRI scans. In our selection process, only subjects with MRI scans conducted within 6 months of their clinically confirmed diagnosis were included, prioritizing the scan closest to the diagnosis date when multiple MRIs were available. We excluded cases involving AD with mixed dementia, non-AD dementias, a history of severe traumatic brain injury, depression, stroke, brain tumors, or significant systemic illnesses. The MRI scans we analyzed adhered to a strict acquisition protocol, involving a T1-weighted sequence, 3D acquisition type (irrespective of the acquisition plane), and a field strength of either 1.5 or 3 Tesla. Additionally, most selected cases provided comprehensive demographic information, including gender, age, education level, and details about the scanner manufacturers or brands.

2.3 | MRI processing and quality assurance pipeline

The MRI scans, downloaded in NIFTI format, underwent a series of preparatory steps to ensure consistency and accuracy before skull-stripping using the FSL brain extraction tool (BET) (Smith, 2002), and subsequent alignment to the MNI152 template (Fonov et al., 2009). Initially, the scans were oriented to match the MNI template's axis order, utilizing the "fslorient2std" function within FSL. This step was crucial for standardizing the orientation across all scans. Following this, the "robustfov" function estimated the robust field of view, a process that efficiently removed extraneous areas such as the neck and lower head from the scan. This function provided bounding box 3D coordinates of the estimated field of view, crucial for the next processing step. Utilizing these coordinates, the "fslmaths" function cropped the scan to focus on the region of interest, which excluded voxels corresponding to white matter, cerebrospinal fluid, the brain stem, and the cerebellum. This precise cropping was imperative to isolate cerebral regions for in-depth analysis, ensuring the scans were optimally prepared for the subsequent steps in our study.

Following the initial preparation of the scans, we utilized the FSL brain extraction tool (BET) (Smith, 2002) for skull stripping. The BET function operates with two primary parameters: the fractional intensity threshold (f), which ranges between 0 and 1, and the vertical gradient in fractional intensity threshold (g), with values spanning from -1 to 1. To assess the quality of the processing, we conducted an inspection of the outputs. This involved generating and analyzing images from randomly selected slices across the axial, sagittal, and coronal planes of each scan. We visually assessed the extracted brain scans and moved the ones with issues such as residual skull fragments or unintended removal of gray matter by BET to a separate group. We then proceeded to reprocess the problematic cases in batches, adjusting the BET parameters to rectify the identified issues. This iterative approach allowed us to fine-tune the processing settings for improved outcomes. We discovered that setting the f value at 0.45 and the g value at 0 consistently produced the most accurate and reliable results in skull stripping, significantly enhancing the quality of the processed scans for our subsequent analyses. Finally, we applied intensity normalization and bias field correction to remove any intensity artifacts and increase data homogeneity, then we assessed the quality of the processed MRI scans. Results of the image quality assessment (IQA) can be found in Figure S2 of the supplement. Parcellation was performed on the processed scans of deceased persons from ADNI and FHS $(n = 23)$ with post-mortem histology who had their last MRI scan taken within 1 year of their death. This was done by applying a nonlinear warp of the Hammersmith Adult brain atlas, segmenting the brain into cortical and subcortical structures, allowing us to study region-based correlations between model-derived attention scores and post-mortem histology.

2.4 | Computational framework

Our framework is designed for the classification of 3D volumetric brain scans into three distinct cognitive states: Normal Cognition (NC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). The building blocks of our framework are a feature extractor, a class-wise attention module, and a classifier network (Figure 1). The training pipeline consists of two stages: the first is training a baseline model for the offline computation of class-wise priors, and the second is training a new independent model with the supervision of these priors.

2.4.1 | Feature extractor

We chose the UNet3D (Çiçek et al., 2016) architecture and started from a pretrained Models Genesis checkpoint on chest CT scans (Zhou et al., 2019; Zhou et al., 2021). Models Genesis are generic pretrained 3D models for 3D medical image analysis. They were trained in a self-supervised robust manner, and outperformed models trained from scratch (Zhou et al., 2021). To adapt the network to our classification task, we discarded the decoder module and kept the encoder of the UNet3D network as the feature extractor for our framework. Another feature extractor we tried was the transformer-

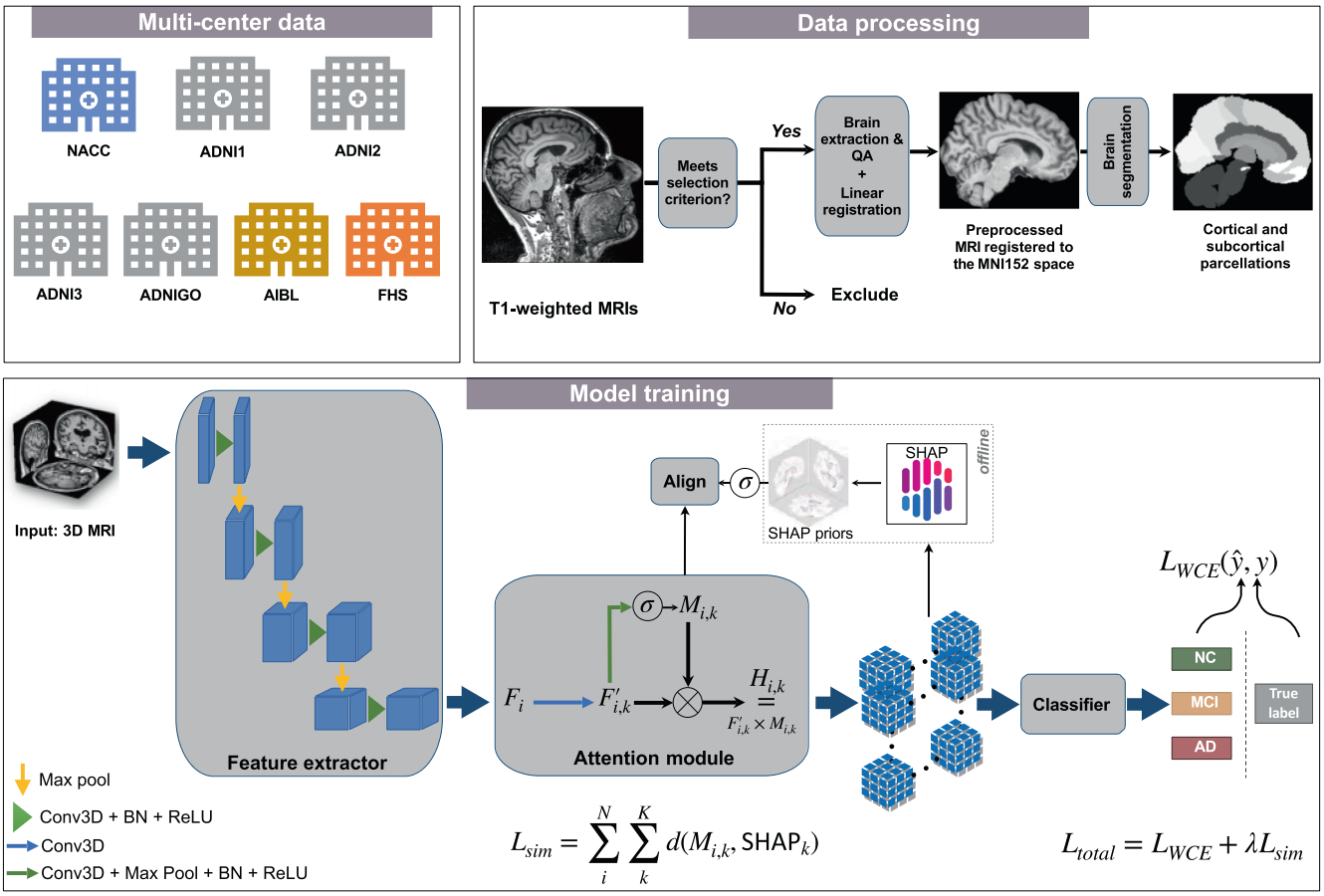


FIGURE 1 Schematic of the disease-informed domain generalization framework. MRI scans from various cohorts were processed via an image processing and quality assurance pipeline (see Section 2.3). Segmentation was applied to scans of deceased individuals from the ADNI and FHS cohorts $(n = 23)$ taken within 1 year of their death, with post-mortem histology available. Our approach takes 3D MRIs as input from the source domain and learns their feature representations in the latent space, and uses an attention module to learn class-specific saliency maps. These maps are then used to predict a class label (NC, MCI, or AD). We used SHAP offline to generate the averaged saliency maps, which we refer to as disease-informed prior knowledge, of NC, MCI, and AD classes over all samples of the source domain used for model training.

based Swin UNETR (Hatamizadeh et al., 2022) which employs a state-of-the-art window multi-head self-attention mechanism to learn embeddings in the latent space. We utilized pretrained weights yielded by the self-supervised pretraining of the Swin UNETR encoder on CT scans of the chest, abdomen, and head/neck. The Swin UNETR encoder was pretrained with multiple proxy tasks tailored for medical image representation (Hatamizadeh et al., 2022).

2.4.2 | Classifier module

We used a global average pooling (GAP) layer (Lin et al., 2013) followed by a softmax function as the classifier for the three-way classification of NC, MCI, and AD. Our choice of a GAP layer as opposed to a fully connected layer as the classifier encourages spatial awareness. The latter approach involves inputting a feature map that is pooled over the channel dimension and subsequently flattened into a one-dimensional vector. In contrast, the former approach processes a stack of 3D feature maps, where the channel dimension K corresponds to the number of classes. This method pools over the spatial dimensions, effectively preserving spatial information for each channel.

2.4.3 | Attention supervision

We added an attention module between the feature extractor and the classifier to learn class-wise attention over the source domain. It takes as input the feature maps F_k output by the feature extractor, and passes it through a 3D convolutional layer to get F'_k. The attention maps learned during this process are denoted by $M_k \in \mathbb{R}^{K \times D \times H \times W}$, where K is the number of classes, and D, H, and W are the depth, height, and width of the attention map, respectively. The final output of the attention module is then the element-wise multiplication of F'_k and M_k. The class-wise attention maps were later used in the second stage of training for alignment with visual saliency priors computed per class over the training data.

2.4.4 | Training

In the first phase of training, we computed visual saliency maps over correct predictions by a baseline model trained with weighted cross-entropy over the training data. To achieve this task, we used SHapley Additive exPlanations (SHAP) to compute the feature contributions per class (Lundberg & Lee, 2017). For the purpose of smoothing out sample noise and variance, we used an averaged saliency map over samples of the same class as a representation of class-wise saliency. Figure 2 shows visualizations of the pre-computed SHAP priors specific to the AD class. For the purpose of visualization, Shapley values were scaled to the range of $[-1,1]$, which we chose to correctly represent negative and positive voxel contributions as in the original range. Once the SHAP priors were generated, we ran our explainability-based strategy to regularize the model through a combined weighted cross entropy (1) and similarity loss (2). We applied augmentation techniques to the training data using the Medical Open Network for AI (MONAI) framework (Cardoso et al., 2022), which included random contrast adjustment, random bias field, random spatial cropping, upsampling, and intensity scaling. We found that intensity scaling to the range $[0,1]$ worked best for data normalization of structural MRI scans.

$$L_{WCE}(\hat{y},y) = -\sum_{i}^{N} w_{y_i}.y_i \log(\hat{y}_i), \tag{1}$$

such that N spans the minibatch dimension, and w_{y_i} refers to the weight assigned to all samples belonging to the ground-truth class y_i. Class weights are computed by taking the inverse of the total count of samples belonging to each class, so that underrepresented classes have a higher weight.

After having the SHAP maps generated offline per class, we used a similarity loss defined in (2) to minimize the distance between each sample's extracted feature maps and the retrieved SHAP prior with respect to the same class as the ground truth label of that sample.

$$L_{sim} = \sum_{i}^{N} \sum_{k}^{K} d(M_{i,k}, \text{SHAP}_k), \tag{2}$$

with d being the distance metric of choice, which, in our case, is the L2 norm. We used the L2 norm loss to increase the semantic consistency between the attention maps $M_{i,k}$ and SHAP priors SHAP_k corresponding to class $k \in [1,K]$, thereby encouraging the model to focus its attention on disease-relevant regions that the pre-computed priors highlighted in the brain.

The final loss is then:

$$L = L_{WCE} + \lambda L_{sim}, \tag{3}$$

where λ is a hyper-parameter that can be optimized.

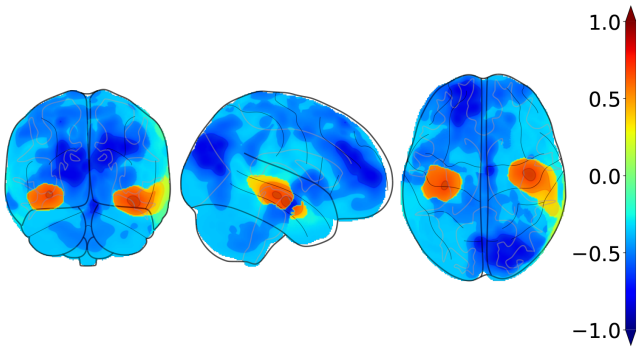


FIGURE 2 Orthogonal projections of the pre-computed AD-specific SHAP priors used in our computational framework. The above projections correspond to the averaged saliency maps with respect to correct predictions of AD over all samples of the source domain. We projected the resulting maps to 2D space onto the coronal, sagittal, and axial axes, respectively.

2.5 | Neuropathological validation

To validate model predictions with gold standard biological evidence, we correlated deep feature contributions with region-specific neuropathological scores obtained from autopsy on persons who had their last MRI within a year of their demise. We quantified regional disease presence based on the degree of amyloid β deposits, neurofibrillary tangles (NFT), and neuritic plaques (NP) on histology. These three pathologies are hallmarks of AD that increase in density and/or spread through the brain as the disease progresses, and they are associated with tissue/cellular damage and death (McKhann et al., 1984).

We obtained 23 participants from ADNI $(n = 13)$ and FHS $(n = 10)$ who had MRI scans taken within 1 year of death with available regional semi-quantitative histopathology scores. Presence and density of amyloid β deposits, neurofibrillary tangles, and neuritic plaques were assessed in the entorhinal, hippocampal, frontal, temporal, parietal, and occipital cortices. The regions were proposed based on the NIA-AA protocol for standardized neuropathological assessment of AD. Severity of the assessment was categorized into four score categories: 0 (None), 1 (Mild), 2 (Moderate), and 3 (Severe) (Hyman et al., 2012). We used the trained models to run inference on those cases and saved their corresponding class-wise attention maps for computation of region-level scores. Since postmortem histology grades assess for the presence of disease in the respective brain regions, we used the AD-specific attention map to compute region-level attention scores as model evidence for the prediction of AD. Using the MNI-152 template, we obtained a brain parcellation for each of the MRIs and aggregated voxel attention values per region, normalized by regional volume. Once model scores were computed, we ran the Spearman's rank correlation coefficient test with pathology grades of amyloid β, neurofibrillary tangles, and neuritic plaques in the various pre-identified brain regions. Following (Rothman, 1990; Saville, 1990), the resulting p-values were not adjusted for multiple comparisons.

3 | EXPERIMENTAL SETUP

We considered the NACC dataset as the source domain for training, validation and internal testing, and used ADNI, AIBL, and FHS as the target domains for external testing. All experiments were run with k-fold cross validation over the source domain with $k = 5$, and the average metrics over the five runs with their standard deviation were reported. Since the source domain we have access to suffers from class imbalance, wherein MCI and AD cases are significantly less than NC cases, we used stratified k-fold cross validation to ensure the target classes follow the same ratio in each fold as in the full dataset. We used a split ratio of 3 : 1 : 1, where 60% of the data were used for model training, 20% were used for internal validation, and the rest for internal testing. We trained our models for 60 epochs with 200 steps, that is, weight updates, per epoch. We also compared against two state-of-the-art methods in the single-source DG setting: RSC (Huang et al., 2020) and Mixup (Zhang et al., 2018). After hyperparameter tuning, we chose a $\lambda = 5 \times 10^{-5}$ for our training strategy and an $\alpha = 0.2$ for the Mixup method. Due to large size of the input image, that is, $(182 \times 218 \times 182)$ per MRI, we could only fit a batch size of 2 into GPU memory (48 GB) and had to resort to gradient accumulation over 8 steps to simulate a final batch size of 16, since the small batch size rendered weighted random sampling ineffective for mitigating class imbalance. We also modified the state-of-the-art DG methods to use weighted cross-entropy across all experiments, as their implementation was not designed to deal with heavy class imbalance.

3.1 | Performance metrics

Along with model accuracy, we reported the macro F1-score averaged over five folds as it better represents a balanced score between precision and recall through their harmonic mean. The macro F1-score in multi-class classification is the average of F1-scores over all classes (4). A higher macro F1 score represents lower false positives, that is, recall, and false negatives, that is, precision.

$$\text{Macro}\, F_1 = \sum_k^K \frac{2 \times \text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{4}$$

such that,

$$\text{Precision}_k = \frac{M_{kk}}{\sum_i M_{ik}} \tag{5}$$

$$\text{Recall}_k = \frac{M_{kk}}{\sum_i M_{ki}}. \tag{6}$$

We also reported Matthew's Correlation Coefficient (MCC), using Scikit-Learn's (Kramer, 2016) formula for multi-class classification (7). An advantage of having MCC as a single-value classification metric is that it is perfectly symmetric, unlike precision and recall that can be affected by swapping positive and negative classes. In addition, it quantifies how well the model is doing at predicting each class, regardless of class imbalance.

$$\text{MCC} = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{\left(s^2 - \sum_k^K p_k^2\right) \times \left(s^2 - \sum_k^K t_k^2\right)}} \tag{7}$$

such that,

$$t_k = \sum_i^K M_{ik} \tag{8}$$

$$p_k = \sum_i^K M_{ki} \tag{9}$$

$$c = \sum_k^K M_{kk} \tag{10}$$

$$s = \sum_i^K \sum_j^K M_{ij}, \tag{11}$$

where M refers to the confusion matrix, K the total number of classes, t_k the number of times class k truly occurred, p_k the number of times class k was predicted, c the total number of samples correctly predicted, and s the total number of samples.

3.2 | Computing infrastructure

We used PyTorch (v1.13.1) and a NVIDIA A6000 graphics card with 48 GB memory on a GPU workstation to implement the model. The training speed was about 2.25 s/iteration, and it took less than 24 h to reach convergence with a batch size of 16 after gradient accumulation. The inference speed was <0.2 s per MRI.

3.3 | Data and code availability

All the MRI scans and corresponding clinical and demographic data can be downloaded freely from ADNI, NACC and AIBL websites. FHS data is available upon request and subject to institutional approval. Python scripts and manuals are available on GitHub.[1]

4 | RESULTS

We compared the results of our computational framework against state-of-the art DG methods for the single-source setting in Table 2. We used a vanilla UNet3D model trained without DG on the NACC cohort as the baseline on which we ran three different experiments:

[1] https://github.com/vkola-lab/hbm2024.

TABLE 2 Model performance on the classification of NC, MCI, and AD.

Method	Class-wise attention	Pretrained		Source	ADNI	AIBL	FHS	Target mean
Baseline (Zhou et al., 2019)	✗	✗	Accuracy (%)	52.5 ± 5.4	38.9 ± 2.7	54.1 ± 11.5	38.6 ± 10.0	43.9 ± 6.9
Baseline (Zhou et al., 2019)	✓	✗		52.7 ± 2.5	42.9 ± 0.6	52.6 ± 4.8	42.7 ± 4.5	46.1 ± 2.5
Baseline (Zhou et al., 2019)	✓	✓		64.3 ± 4.0	42.7 ± 1.4	66.1 ± 3.7	48.1 ± 7.2	52.3 ± 2.9
RSC (Huang et al., 2020)	✓	✓		55.5 ± 3.4	43.8 ± 3.8	44.8 ± 9.0	35.9 ± 9.2	41.5 ± 3.0
Mixup (Zhang et al., 2018)	✓	✓		62.8 ± 1.5	43.5 ± 2.0	65.2 ± 3.6	34.3 ± 2.8	47.7 ± 2.1
Ours	✓	✗		51.6 ± 2.3	**44.0 ± 0.4**	47.3 ± 3.3	37.1 ± 3.6	42.8 ± 1.6
Ours	✓	✓		**66.5 ± 1.3**	42.9 ± 1.5	**73.4 ± 2.4**	**49.1 ± 6.5**	**55.1 ± 2.9**
Baseline (Zhou et al., 2019)	✗	✗	Macro F1 Score	0.50 ± 0.04	0.39 ± 0.03	0.44 ± 0.06	0.33 ± 0.07	0.39 ± 0.05
Baseline (Zhou et al., 2019)	✓	✗		0.50 ± 0.02	0.44 ± 0.01	0.45 ± 0.02	0.37 ± 0.03	0.42 ± 0.01
Baseline (Zhou et al., 2019)	✓	✓		0.58 ± 0.02	0.44 ± 0.02	0.54 ± 0.02	0.40 ± 0.05	0.46 ± 0.02
RSC (Huang et al., 2020)	✓	✓		0.52 ± 0.01	0.44 ± 0.03	0.42 ± 0.02	0.32 ± 0.07	0.39 ± 0.02
Mixup (Zhang et al., 2018)	✓	✓		0.58 ± 0.01	0.44 ± 0.02	0.54 ± 0.03	0.30 ± 0.02	0.43 ± 0.02
Ours	✓	✗		0.50 ± 0.02	**0.45 ± 0.00**	0.42 ± 0.02	0.34 ± 0.03	0.40 ± 0.01
Ours	✓	✓		**0.60 ± 0.02**	0.44 ± 0.02	**0.58 ± 0.02**	**0.41 ± 0.04**	**0.48 ± 0.02**
Baseline (Zhou et al., 2019)	✗	✗	MCC	0.27 ± 0.04	0.13 ± 0.03	0.21 ± 0.06	0.11 ± 0.06	0.15 ± 0.05
Baseline (Zhou et al., 2019)	✓	✗		0.26 ± 0.04	0.18 ± 0.02	0.21 ± 0.03	0.13 ± 0.03	0.17 ± 0.02
Baseline (Zhou et al., 2019)	✓	✓		0.40 ± 0.04	**0.21 ± 0.03**	0.34 ± 0.02	0.19 ± 0.37	0.25 ± 0.02
RSC (Huang et al., 2020)	✓	✓		0.31 ± 0.02	0.18 ± 0.02	0.23 ± 0.01	0.10 ± 0.03	0.17 ± 0.01
Mixup (Zhang et al., 2018)	✓	✓		0.39 ± 0.02	0.19 ± 0.01	0.33 ± 0.03	0.11 ± 0.02	0.21 ± 0.02
Ours	✓	✗		0.26 ± 0.04	0.18 ± 0.01	0.18 ± 0.02	0.11 ± 0.03	0.16 ± 0.02
Ours	✓	✓		**0.42 ± 0.02**	**0.21 ± 0.03**	**0.40 ± 0.02**	**0.20 ± 0.03**	**0.27 ± 0.03**

Note: We trained our model on the NACC cohort and used the ADNI, AIBL, and FHS cohorts as target domains. We reported accuracy on the test split of NACC, and on the entirety of the target datasets. Performance metrics including accuracy, macro F1-score and MCC are reported on each case. Note that model training was done via 5-fold cross validation on the NACC dataset, and testing was done on each of the models. Results are reported as mean ± standard deviation. The bold font is used to report the best model performance in each column.

one trained from scratch and not using attention (Row 1), another also trained from scratch but with our attention module (Row 2), and the third trained starting from the pretrained Models Genesis (Zhou et al., 2019; Zhou et al., 2021) checkpoint with our attention module (Row 3). First, the two methods we compared against, RSC (Huang et al., 2020) and Mixup (Zhang et al., 2018), did not show improvement over the baseline. In fact, performance was deteriorated going from Row 3 to 4 by 10.8% in terms of target mean accuracy, 0.07 (7%) in terms of target mean macro F1-score, and 0.08 (4%) in terms of target mean MCC. The same pattern of performance degradation was observed going from Rows 3 to 5, with a 4.6% lower target mean accuracy, a 0.03 (3%) lower target mean macro F1-score, and 0.04 (2%) lower target mean MCC. These findings suggest that while these methods have shown enhanced performance and resilience against distributional changes in natural and synthetic imaging benchmarks, their effectiveness may not extend to real-world clinical scenarios, specifically in the context of volumetric structural brain MRIs. On the other hand, training using our method improved performance, outperforming RSC, Mixup, and the baseline across the reported target mean metrics. We showed a 2.8% improvement over the baseline (Row

3 vs. Row 7) in terms of target mean accuracy. In fact, our method was able to achieve a 73.4% accuracy on the target cohort AIBL, a 7.3% improvement over the baseline (Row 7 vs. Row 3). This improvement is also reflected in the MCC value which increased by 0.07 (3%) from Rows 3 to 7. Receiver operating characteristic (ROC) and precision–recall (PR) curves supporting our findings were included in the supplement in Figures S8 and S9, respectively.

The above quantitative results were reflected in Figure 3, where we used the t-distributed stochastic neighbor embedding (t-SNE) algorithm (Van der Maaten & Hinton, 2008) to visualize MRI embeddings learnt by the baseline model trained without DG (Row 3 in Table 2) and the model trained with our computational framework (Row 7 in Table 2). While the baseline t-SNE plot shows the MRI embeddings learned by the baseline model clustered by cohort, ours shows that our approach to aligning model attention with SHAP priors reduces cohort-specific clustering. In particular, the improvement in accuracy over the baseline on the AIBL cohort shows in the dispersion of MRI embeddings belonging to AIBL (orange) across the tSNE plot on the right (Ours) as opposed to a clear cluster highlighted in the plot to the left (Baseline). These results indicate that even though the SHAP priors used in training were derived only from the source domain, they effectively reduced the distributional variance across

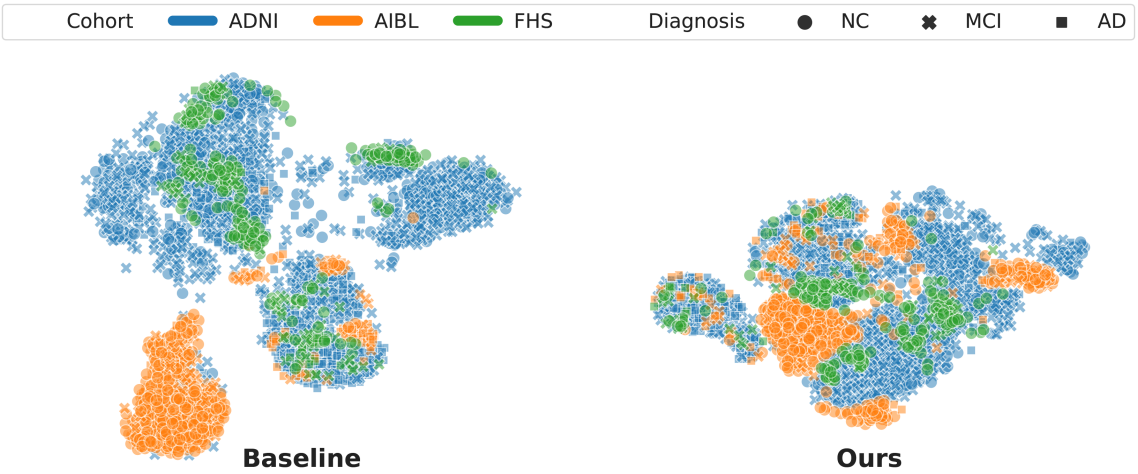


FIGURE 3 Visualization of MRI embeddings in the latent space. We generated MRI embeddings at the attention module level from two UNet3D models trained on the NACC cohort without domain generalization (Baseline, Row 3 in Table 2) and with our proposed DG framework (Ours, Row 7 in Table 2), and visualized them in a 2D space using t-SNE. For both models, data from the target cohorts (ADNI ($n = 1{,}821$), AIBL ($n = 661$) and FHS ($n = 304$)) were used. The data points were color-coded by diagnosis label and marked by cohort.

TABLE 3 Performance results of training without domain generalization (DG) using the Swin UNETR (Hatamizadeh et al., 2022) encoder as the feature extractor and different classifiers listed below.

Classifier module	Class-wise attention	Pretrained		Source	ADNI	AIBL	FHS	Target mean
Conv3D	✗	✓	Accuracy (%)	57.1 ± 5.4	35.2 ± 4.8	41.9 ± 9.2	56.6 ± 9.2	44.6 ± 6.4
Conv3D ($N = 3$)	✗	✓		57.2 ± 4.4	44.3 ± 4.4	41.0 ± 2.7	54.9 ± 7.2	46.7 ± 3.2
Conv3D	✓	✓		57.7 ± 1.1	40.7 ± 2.4	39.1 ± 2.0	58.2 ± 2.3	46.0 ± 1.4
GAP	✓	✓		59.8 ± 2.7	41.8 ± 2.9	44.6 ± 5.1	58.4 ± 2.7	48.3 ± 3.1
Conv3D	✗	✓	Macro F1-score	0.48 ± 0.05	0.34 ± 0.05	0.32 ± 0.05	0.38 ± 0.07	0.34 ± 0.05
Conv3D ($N = 3$)	✗	✓		0.55 ± 0.03	0.43 ± 0.03	0.36 ± 0.01	0.45 ± 0.05	0.41 ± 0.03
Conv3D	✓	✓		0.54 ± 0.01	0.40 ± 0.02	0.34 ± 0.02	0.46 ± 0.02	0.40 ± 0.02
GAP	✓	✓		0.55 ± 0.01	0.42 ± 0.03	0.37 ± 0.04	0.44 ± 0.03	0.41 ± 0.03
Conv3D	✗	✓	MCC	0.30 ± 0.04	0.15 ± 0.03	0.16 ± 0.04	0.17 ± 0.05	0.16 ± 0.03
Conv3D ($N = 3$)	✗	✓		0.35 ± 0.04	0.16 ± 0.02	0.17 ± 0.01	0.21 ± 0.05	0.18 ± 0.02
Conv3D	✓	✓		0.33 ± 0.01	0.13 ± 0.01	0.15 ± 0.01	0.20 ± 0.03	0.16 ± 0.02
GAP	✓	✓		0.35 ± 0.02	0.15 ± 0.02	0.18 ± 0.04	0.18 ± 0.05	0.17 ± 0.03

Note: The weights of the feature extractor were loaded from a pretrained checkpoint and fine-tuned while training on the classification of NC, MCI, and AD. The feature extractor has a window multi-head self-attention mechanism built in, and we ran training with and without the class-wise attention module before the classifier.

source and target domains. Moreover, we explored the effect of demographic variance on model performance and included a detailed comparison of our model (Row 7 in Table 2) against the baseline (Zhou et al., 2019) (Row 1 in Table 2) in the supplement (Figures S4–S7). Our model exhibited an overall improvement in performance over the baseline across different distributions of demographic groups.

For comparison with state-of-the-art, results of additional experiments were reported in Table 3 on the ternary classification task of NC, MCI, and AD with the transformer-based Swin UNETR (Hatamizadeh et al., 2022) encoder as the feature extractor. The model was trained with different classifiers, with and without the class-wise attention module described in Section 2.4. Table 3 shows a similar performance to the results with the UNet3D encoder in Table 2. Adding the class-wise attention module exhibited the same trend in performance as reported in Table 2 with the UNet3D (Çiçek et al., 2016) feature extractor. Remarkably, the results in Table 3 show that using a feature extractor with inherent, state-of-the-art self-attention did not provide an advantage over using class-wise attention supervision designed to focus on disease biology.

We further validated our method with gold standard evidence of disease pathology and compared it against the other methods, reporting the results in the form of a correlation heat map in Figure 4. We showed that not only did our method correlate more strongly with postmortem histology scores than other methods, but also, our results were more consistent across the three stains. Correlation of our method with pathology in the amygdala, hippocampus, parahippocampal and ambient gyri was positive for amyloid β, neurofibrillary tangles, and neuritic plaques. We then projected the computed correlation values onto their corresponding brain regions and displayed the projections (Figure 5). Figure 5a shows an improved correlation for our method with pathology grades of amyloid β in the hippocampal region and the middle frontal gyrus of the frontal lobe. Correlation in these brain regions is also consistent with pathology grades of neurofibrillary tangles and neuritic plaques (Figure 5b,c). As for the other evaluated methods, shown in the first three columns of each subfigure, the correlations were lower with pathology grades in the hippocampus of amyloid β, neurofibrillary tangles, and neuritic plaques, except for the baseline method in Figure 5c that had a positive—although lower than ours—correlation. In addition, our method showed the highest correlation in the parahippocampal and ambient gyri with pathology grades of neuritic plaques in Figure 5c. Our method demonstrated high correlations with specific brain regions, notably the hippocampal and parahippocampal areas, which were visually represented in the pre-computed AD-specific SHAP priors (Figure 2). These regions contributed positively to model predictions of AD, indicating the effectiveness of our technique in aligning model attention with established knowledge regarding disease indicators. Such observations indicating improved model correlation with regions that are well-known to be implicated with disease grounded our model predictions with biological evidence.

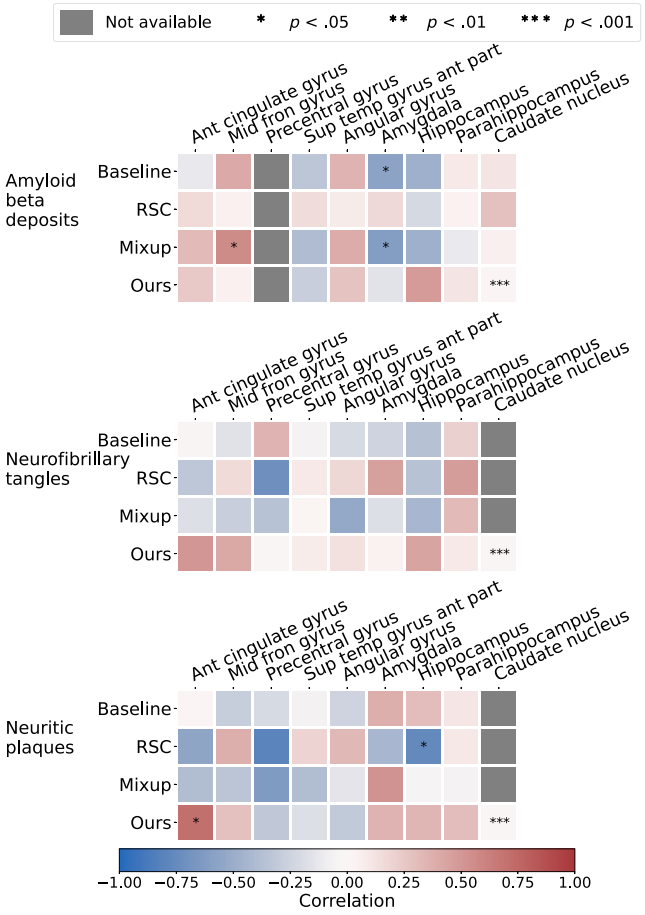
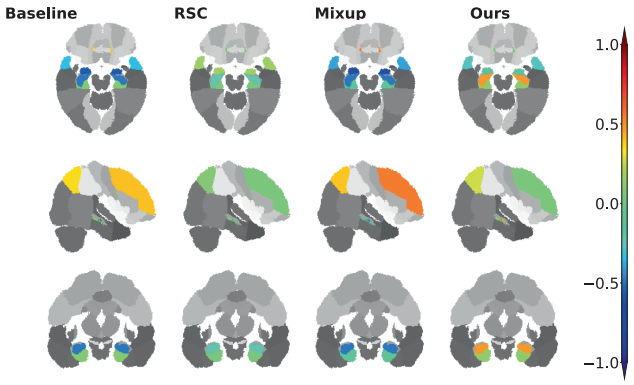


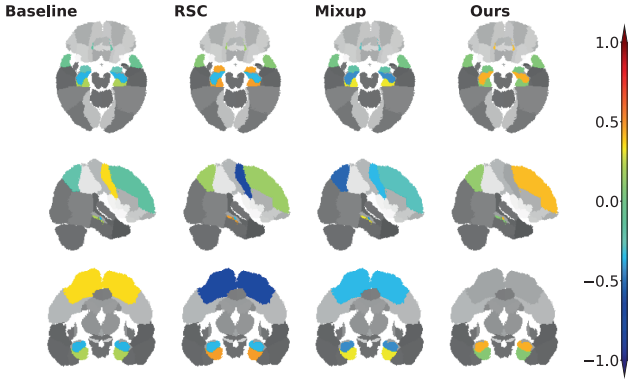
FIGURE 4 Correlation of model-generated attention scores with post-mortem histology. Pathology grades of amyloid β, neurofibrillary tangles and neuritic plaques in various brain regions on deceased ADNI and FHS participants were obtained $(n = 23)$. We compared model-identified importance in these brain regions with the degree of pathology severity, and compared them against predictions obtained using other well-known domain generalization methods.

5 | DISCUSSION

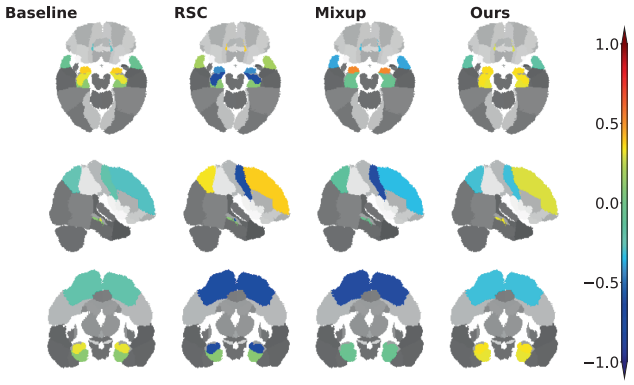
This work presents a computational framework for DG that adds disease-driven interpretability to deep learning models for AD prediction on volumetric MRI scans. While most of the existing methods focus on achieving high model performance on unseen data, they do not directly account for the underlying disease biology during model development. We achieved this goal by refining the model's attention to focus on brain regions that are most associated with disease based on pre-computed feature contributions. In such fashion, our method distinguishes itself by incorporating disease-driven interpretability into the training process. The added interpretability can provide a better understanding of the underlying disease mechanisms and aid in the clinical decision-making process. We compared the performance of our method with previously published DG frameworks, and showed that our approach shows competitive performance while incorporating disease relevance into the model training process. We confirmed



(a) Correlation of attention maps with the presence of amyloid β deposits.



(b) Correlation of attention maps with neurofibrillary tangles.



(c) Correlation of attention maps with neuritic plaques.

FIGURE 5 Visualization of correlations between model attention scores and post-mortem histology. We obtained region-specific pathology grades of amyloid β, neurofibrillary tangles and neuritic plaques on deceased ADNI and FHS participants $(n = 23)$. The pathology grades reflect the severity of the assessment which was categorized into four score categories: 0 (none), 1 (mild), 2 (moderate), 3 (severe). We obtained attention scores for each case from the model attention maps specific to AD, aggregated on a region level. We then computed Spearman's rank correlation coefficient between the model-derived attention scores and the region-specific pathology grades and projected them on the corresponding brain regions for visualization.

the degree to which our attention-based supervision strategy ultimately reflected disease biology by comparing model attention in predefined brain regions with postmortem neuropathology scores. Overall, our approach to creating a generalizable framework complements other published work in the literature.

We observed that our model achieved consistent, favorable performance on the test cohorts relative to other DG frameworks. While extensive testing is required to confirm any modeling framework's superiority in accurate prediction of disease, it is worth noting that model performance based on accuracy alone without downstream evidence of correlation with a reference standard may not be appealing in the context of medical machine learning. As such, classifying persons with NC from those who have MCI or AD solely on MRIs is a clinically challenging task, and often not part of routine clinical neurology work-up. Neurologists use a spectrum of patient data along with MRIs to assess a person's cognitive status. Nevertheless, our proposed framework has utility in the objective interpretation of brain MRIs, and broadly in the quantification of findings indicative of disease. Besides minimizing subjectivity, it also potentially fills gaps in healthcare settings where there is a lack of neuroradiology expertise.

Our study has a few limitations. Due to memory limitations, we resorted to offline computation of the saliency maps based on correct predictions by the trained baseline model. We also acknowledge that SHAP prior computation is solely dependent on the baseline model used, that is, the quality of prior knowledge and correctness of feature contributions extracted from the model are directly correlated with model performance. Also, it is possible that the offline computation and aggregation of class-specific SHAP maps may have reduced instance-to-instance variability and minimized radiologic artifacts, thereby facilitating model attention on disease pathology. In addition, it is possible that the model was able to capture the fine-grained nature of disease markers due to our choice of the voxel-wise L2 distance metric. We utilized this loss function to increase the semantic similarity between model attention and prior maps at the voxel level.

In conclusion, our work contributes to the growing field of interpretable deep learning in medical imaging, paving the way for more accurate and personalized diagnoses of cognitive disorders. By highlighting the specific brain regions that contribute most significantly to disease, our approach can provide valuable insight into disease mechanisms and aid in developing targeted interventions. Furthermore, the disease-driven interpretability of our framework can help build trust and understanding between clinicians and patients, which is crucial for effective healthcare delivery.

AUTHOR CONTRIBUTIONS

DL, SAB, BAP, and VBK: study conception and design. DL and SS: data collection and processing. DL: implementation. DL and SS: analysis. DL, SS, SAB, BAP, RA, and VBK: data interpretation and manuscript write-up. VBK: study direction. All authors reviewed the results and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

This project was supported by grants from the Karen Toffler Charitable Trust, the American Heart Association (20SFRN35460031), the National Institutes of Health (RF1-AG062109, R01-HL159620, R21-CA253498, and R43-DK134273), and a pilot award from the

National Institute on Aging's Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research program.

CONFLICT OF INTEREST STATEMENT

V.B.K. is on the scientific advisory board for Altoida Inc., and serves as a consultant to AstraZeneca. R.A. is a scientific advisor to Signant Health and NovoNordisk. She also serves as a consultant to Davos Alzheimer's Collaborative. The remaining authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Vijaya B. Kolachalama https://orcid.org/0000-0002-5312-8644

REFERENCES

Aderghal, K., Benois-Pineau, J., Afdel, K., & Gwenaëlle, C. (2017). FuseMe: Classification of SMRI images by fusion of deep CNNs in 2D + ε projections. In *Proceedings of the 15th international workshop on content-based multimedia indexing CBMI'17*, New York, NY. Association for Computing Machinery. https://doi.org/10.1145/3095713.3095749

Beekly, D. L., Ramos, E. M., van Belle, G., Deitrich, W., Clark, A. D., Jacka, M. E., & Kukull, W. A. (2004). The national Alzheimer's coordinating center (NACC) database: An Alzheimer disease database. *Alzheimer Disease & Associated Disorders*, *18*(4), 270–277.

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., & Yang, I. (2022). *MONAI: An open-source framework for deep learning in healthcare*. arXiv preprint arXiv:221102701.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical image computing and computer-assisted intervention—MICCAI 2016: 19th international conference*, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19 (pp. 424–432). Springer.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops* (pp. 702–703). IEEE.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems (NeurIPS)*, *31*, 2791–2801.

Ellis, K. A., Bush, A. I., Darby, D., de Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., … the AIBL Research Group. (2009). The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*(59), 1–35.

Ghimire, S., Kashyap, S., Wu, J. T., Karargyris, A., & Moradi, M. (2020). Learning invariant feature representation to improve generalization across chest x-ray datasets. In M. Liu, P. Yan, C. Lian, & X. Cao (Eds.), *Machine learning in medical imaging* (pp. 644–653). Springer International Publishing.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In A. Crimi & S. Bakas (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 272–284). Springer International Publishing.

Huang, Z., Wang, H., Xing, E. P., & Huang, D. (2020). Self-challenging improves cross-domain generalization. In *Computer vision—ECCV 2020: 16th European conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (pp. 124–140). Springer.

Hugo, J., & Ganguli, M. (2014). Dementia and cognitive impairment: Epidemiology, diagnosis, and treatment. *Clinics in Geriatric Medicine*, *30*(3), 421–442.

Hyman, B. T., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Carrilllo, M. C., Dickson, D. W., Duyckaerts, C., Frosch, M. P., Masliah, E., Mirra, S. S., Nelson, P. T., Schneider, J. A., Thal, D. R., Thies, B., Trojanowski, J. Q., Vinters, H. V, & Montine, T. J. (2012). National Institute on Aging—Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimer's & Dementia*, *8*(1), 1–13.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Levine, S., Finn, C., & Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning PMLR* (pp. 5637–5664). PMLR.

Kramer, O. (2016). *Scikit-Learn* (pp. 45–53). Springer International Publishing. https://doi.org/10.1007/978-3-319-33383-0_5

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., & Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (REx). In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning, vol. 139 of proceedings of machine learning research PMLR* (pp. 5815–5826). PMLR. https://proceedings.mlr.press/v139/krueger21a.html

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32). AAAI.

Li, H., Pan, S. J., Wang, S., & Kot, A. C. (2018). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE.

Lin, M., Chen, Q., & Yan, S. (2013). *Network in network*. arXiv preprint arXiv:13124400.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M. J., & ADNI. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*, 4765–4774.

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective. *The Lancet*, *383*(9921), 999–1008.

Massaro, J. M., Srem, R. B. D'A., Sullivan, L. M., Beiser, A., DeCarli, C., Au, R., Elias, M. F., & Wolf, P. A. (2004). Managing and analysing data from a large-scale study on Framingham offspring relating brain structure to cognitive function. *Statistics in Medicine*, *23*(2), 351–367. https://doi.org/10.1002/sim.1743

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease. *Neurology*, *34*(7), 939. https://n.neurology.org/content/34/7/939

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

Nguyen, A. T., Tran, T., Gal, Y., & Baydin, A. G. (2021). Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems (NeurIPS)*, *34*, 5264–5275.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jr., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209.

Qiu, S., Chang, G. H., Panagia, M., Gopal, D. M., Au, R., & Kolachalama, V. B. (2018). Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*, 737–749. https://www.sciencedirect.com/science/article/pii/S2352872918300654

Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., Chang, G. H., Joshi, A. S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y. J., Swaminathan, A., Kedar, S., Saint-Hilaire, M. H., Auerbach, S. H., Yuan, J., Sartor, E. A., … Kolachalama, V. B. (2020). Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*, *143*(6), 1920–1933. https://doi.org/10.1093/brain/awaa137

Qiu, S., Miller, M. I., Joshi, P. S., Lee, J. C., Xue, C., Ni, Y., Wang, Y., de Anda-Duran, I., Hwang, P. H., Cramer, J. A., Dwyer, B. C., Hao, H., Kaku, M. C., Kedar, S., Lee, P. H., Mian, A. Z., Murman, D. L., O'Shea, S., Paul, A. B., … Kolachalama, V. B. (2022). Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, *13*(1), 3404.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43–46.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician*, *44*(2), 174–180.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 23–30). IEEE.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., & Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems (NeurIPS)* (Vol. 31). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2018/file/1d94108e907bb8311d8802b48fd54b4a-Paper.pdf

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., & Saminger-Platz, S. (2017). *Central moment discrepancy (CMD) for domain-invariant representation learning*. arXiv preprint arXiv:170208811.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *Mixup: Beyond empirical risk minimization*. arXiv preprint arXiv:171009412.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., & Xu, Z. (2020). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, *39*(7), 2531–2540.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2023). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415.

Zhou, K., Yang, Y., Hospedales, T., & Xiang, T. (2020). Learning to generate novel domains for domain generalization. In *Computer vision—ECCV 2020: 16th European conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16 (pp. 561–578). Springer.

Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., & Liang, J. (2021). Models genesis. *Medical Image Analysis*, *67*, 101840. https://www.sciencedirect.com/science/article/pii/S1361841520302048

Zhou, Z., Sodha, V., Rahman Siddiquee, M. M., Feng, R., Tajbakhsh, N., Gotway, M. B., & Liang, J. (2019). Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference*, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22 (pp. 384–393). Springer.

Zunino, A., Bargal, S. A., Volpi, R., Sameki, M., Zhang, J., Sclaroff, S., Murino, V., & Saenko, K. (2021). Explainable deep classification models for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3233–3242). IEEE.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lteif, D., Sreerama, S., Bargal, S. A., Plummer, B. A., Au, R., & Kolachalama, V. B. (2024). Disease-driven domain generalization for neuroimaging-based assessment of Alzheimer's disease. *Human Brain Mapping*, *45*(8), e26707. https://doi.org/10.1002/hbm.26707