

ORIGINAL RESEARCH ARTICLE

Spectral-conditioned hypernetworks for meta-learning neuroimaging normalization with concept-guided transparency

Hashim Ali*

School of Computing and Artificial Intelligence, Nazarbayev University, Astana, Kazakhstan

Abstract

Multi-site neuroimaging studies are essential for developing clinically deployable artificial intelligence (AI) systems, yet deep learning models remain highly sensitive to scanner-induced domain shift. Variations in scanner manufacturer, field strength, acquisition protocol, reconstruction pipeline, and site-specific preprocessing can alter image intensity, texture, and spatial-frequency characteristics, thus degrading model performance when deployed on unseen scanners. This paper presents a spectral-conditioned hypernetwork framework for meta-learning neuroimaging normalization with concept-guided transparency. The proposed framework integrates a spectral-conditioned hypernetwork into the batch normalization layers of a three-dimensional convolutional encoder. For each input magnetic resonance imaging volume, a compact spectral embedding is extracted from the three-dimensional Fourier magnitude spectrum and used to dynamically generate the channel-wise normalization scale and shift parameters. This enables input-adaptive normalization without requiring explicit site labels or scanner metadata. To improve generalization to unseen imaging centers, this study formulates a bi-level meta-learning protocol in which small site-specific support sets adapt only the hypernetwork parameters, while the outer loop jointly optimizes the encoder and hypernetwork across multiple source sites. To enhance clinical interpretability, the proposed framework introduces a concept-guided regularization term that aligns gradient-weighted class activation mapping attribution maps with predefined anatomical regions of interest, including the hippocampus, ventricles, temporal pole, and entorhinal cortex. The framework is evaluated on a multi-site Alzheimer's disease classification benchmark using leave-one-site-out validation. The proposed method improves diagnostic performance, reduces cross-site variability in the area under the receiver operating characteristic curve, and produces stronger alignment of attribution with clinically meaningful anatomical concepts compared to statistical harmonization, adversarial domain adaptation, meta-learning domain generalization, site-conditioned normalization, and concept-based interpretability baselines. These findings suggest that coupling adaptive normalization with few-shot meta-learning and anatomical concept guidance provides a practical pathway toward robust and transparent AI-assisted neuroimaging diagnosis.

***Corresponding author:**Hashim Ali
(hashim.ali@nu.edu.kz)

Citation: Ali H. Spectral-conditioned hypernetworks for meta-learning neuroimaging normalization with concept-guided transparency. *Adv Neurol.* 2026;5(4):026230024. doi: 10.36922/AN026230024

Received: June 6, 2026**Revised:** July 13, 2026**Accepted:** July 16, 2026**Published online:** August 4, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Neuroimaging; Magnetic resonance imaging; Domain generalization; Hypernetwork; Meta-learning; Explainable artificial intelligence

1. Introduction

Deep learning has achieved substantial progress in medical image analysis, including the detection and classification of neurological disorders such as Alzheimer's disease (AD) using structural magnetic resonance imaging (MRI).¹ However, a major obstacle to clinical translation is the sensitivity of deep neural networks to domain shift in multi-site imaging datasets. In neuroimaging, the domain shift arises from scanner manufacturer, magnetic field strength, acquisition protocol, coil configuration, reconstruction software, and site-specific preprocessing. These factors can introduce systematic variations in intensity, contrast, texture, and spatial frequency that are unrelated to the underlying disease biology yet may nonetheless be captured by neural networks as spurious but discriminative shortcut features.

This challenge is especially important for artificial intelligence (AI)-assisted neurology. A model developed at an imaging center may perform well in internal validation data, but fail when deployed at another hospital with different scanner hardware or acquisition settings. This performance degradation undermines clinical reliability and limits the practical utility of AI systems in neurological diagnosis. The problem is not only one of accuracy, but also one of transparency. If a model responds to scanner-specific artifacts rather than disease-relevant anatomical markers, its explanations may appear plausible while reflecting technical bias rather than neuropathology.

Existing strategies for addressing multi-site neuroimaging heterogeneity include statistical harmonization, domain adaptation, and domain generalization. Statistical harmonization methods, such as ComBat, remove scanner effects from pre-extracted features by modeling site-specific shifts.²⁻⁴ Although effective in many neuroimaging studies, these methods generally require explicit site labels and operate outside the end-to-end deep learning pipeline. Deep domain adaptation methods learn site-invariant representations using adversarial training, distribution matching, or feature alignment.⁵⁻⁷ These methods improve robustness, but often require access to target-domain data during training or rely on explicit domain annotations. Domain generalization methods aim to generalize to unseen sites without target-domain data, but generally treat the encoder as a monolithic feature extractor and do not directly address layer-wise normalization shifts induced by scanner variability.^{8,9}

Batch normalization (BN) is a central component of modern convolutional networks, but conventional BN assumes that the feature statistics at deployment resemble those observed during training.¹⁰⁻¹³ In multi-site MRI,

this assumption is frequently violated. Scanner-induced differences propagate through the network, affecting activation distributions across multiple layers. The static affine parameters of the BN cannot adapt to such input-specific distributional variation. This motivates a dynamic normalization mechanism that adjusts internal feature statistics according to scanner-sensitive characteristics extracted directly from each input image.

To address this limitation, this paper proposes a spectral-conditioned hypernetwork for adaptive neuroimaging normalization. A hypernetwork is a neural network that generates the parameters of another network.¹⁴ In the proposed framework, the hypernetwork receives a spectral embedding derived from the three-dimensional (3D) Fourier magnitude spectrum of an input MRI volume and generates the BN scale and shift parameters for a 3D convolutional encoder. The spectral embedding captures global frequency-domain properties associated with acquisition conditions, noise patterns, intensity non-uniformity, and scanner-specific image characteristics. Because the conditioning signal is computed directly from the input volume, the proposed method does not require explicit scanner labels or site metadata.

To enable rapid adaptation to new imaging centers, the proposed framework introduces a bi-level meta-learning optimization protocol¹⁵⁻¹⁷ inspired by model-agnostic meta-learning (MAML).¹⁸ Each imaging site is treated as a meta-task. In the inner loop, a small support set from a site adapts only the hypernetwork parameters, while the encoder weights remain fixed. In the outer loop, the adapted hypernetwork^{14,19} is evaluated on a query set, and both the encoder and the hypernetwork initialization are optimized across all source sites. This design focuses on rapid adaptation to scanner-dependent normalization while preserving stable disease-relevant feature extraction.

Beyond robustness, the proposed framework explicitly addresses clinical transparency. A concept-guided regularization mechanism is introduced to align the gradient-weighted class activation mapping (Grad-CAM) attribution map²⁰ with anatomical concept masks for clinically meaningful regions of interest. For AD classification, these include the hippocampus, ventricles, temporal pole, and entorhinal cortex. Unlike concept bottleneck models (CBM)^{21,22}, which require the model to pass through an explicit concept-prediction layer, the proposed approach uses concept alignment as a regularization term. This preserves the encoder's representational flexibility while encouraging explanations to remain anatomically meaningful. The need for transparent AI in clinical decision support has motivated frameworks that evaluate not only predictive accuracy, but

also interpretability, reliability, and clinical usability.²³

The main contributions of this study are as follows.

- A spectral-conditioned hypernetwork is proposed to dynamically generate BN affine parameters from Fourier-domain scanner-sensitive embeddings, enabling input-adaptive normalization without explicit site labels.
- A bi-level meta-learning strategy is formulated to adapt only the hypernetwork parameters on small calibration sets, improving few-shot deployment efficiency while maintaining encoder stability.
- A concept-guided attribution regularization term is introduced to align model explanations with clinically meaningful neuroanatomical regions.
- A methodological transparency protocol is provided, including subject-level partitioning, patient-level leakage prevention, support-query separation, learning-curve diagnostics, and reproducible evaluation settings.

The remainder of the paper is organized as follows: Section 2 describes the materials and methods, including the multi-site neuroimaging dataset, preprocessing procedure, patient-level data partitioning strategy, leakage-prevention protocol, spectral-conditioned hypernetwork framework, bi-level meta-learning optimization, concept-guided regularization, baseline methods, implementation settings, and evaluation metrics; Section 3 presents the experimental results, including diagnostic performance, cross-scanner robustness, learning-dynamics analysis, feature-space visualization, interpretability assessment, and ablation studies; Section 4 discusses the implications of the findings for robust and transparent AI-assisted neurological diagnosis, together with methodological limitations, reproducibility considerations, and future extensions to other neurological disorders and imaging modalities; Section 5 concludes the paper by summarizing the main contributions and key findings.

2. Materials and methods

2.1. Dataset and preprocessing

Experiments are conducted using structural T1-weighted MRI scans from the AD Neuroimaging Initiative (ADNI).²⁴ ADNI contains scans acquired across multiple imaging centers and scanner configurations, making it suitable for evaluating cross-scanner robustness. The binary classification task is AD versus cognitively normal (CN) control.

The study subset contains 2,450 scans, including 1,200 AD cases and 1,250 CN subjects, collected from eight imaging sites. Each site contributes between 80 and

600 volumes. The preprocessing pipeline includes affine registration to the Montreal Neurological Institute 152 template, skull stripping, intensity scaling to the 0,1 range, resampling to $1.5 \times 1.5 \times 1.5$ mm isotropic resolution, and cropping to $128 \times 128 \times 128$. Scanner-specific harmonization is not applied before model training so that scanner-related domain variation remains available for evaluating robustness.

To improve dataset transparency, site-level diagnosis distribution, scanner field strength, vendor family, and demographic summaries were audited before model training, as presented in Table 1. The sites are anonymized as S1–S8 because the experiment's purpose is cross-site robustness rather than institution identification. Demographic balancing was not enforced through resampling because that could distort the natural multi-site distribution; instead, age, sex, and education were checked within each leave-one-site-out fold, and balanced accuracy was used as a primary metric to reduce class-imbalance effects.

2.2. Data partitioning and leakage prevention

A leave-one-site-out evaluation protocol is used. In each fold, one imaging site is held out as the unseen target domain, while the remaining seven sites are used for source-domain training and validation. Partitioning is performed at the subject level before scan-level sampling. All scans for a given subject are assigned to a single partition within each fold. This rule is applied to the source training set, source validation set, meta-learning support sets, meta-learning query sets, few-shot target support set, and held-out target test set.

To address the possibility that a participant may contribute scans acquired at multiple imaging sites, subject identifiers are checked globally before fold construction. If a subject has scans from the held-out target site and from any source site, all scans from that subject are assigned to the held-out target side for that fold and are excluded from source training and validation. If a subject has scans across multiple source sites, the subject is assigned to a single source-site task according to the earliest eligible acquisition record used in the fold, and remaining scans from other source sites are excluded from meta-training for that fold. This conservative rule prevents cross-site patient leakage and avoids using the same patient to shape both source and target distributions.

Source-site subjects are split into 90% training and 10% validation groups after patient-level deduplication. Validation subjects are never used for inner-loop updates, early stopping, hyperparameter search, or target-site support construction. For few-shot target adaptation, five

labeled target-site subjects are selected as a support set after the held-out target site has been isolated. All reported target-site metrics are computed on the remaining target-site subjects, with no overlap between target support and target test subjects. The held-out target test set is not used for model selection, threshold tuning, architecture selection, or hyperparameter adjustment. Table 2 highlights the protocol used in the study.

2.3. The proposed framework and technical preliminaries

2.3.1. Batch normalization and internal covariate shift

To lay the foundation for the proposed framework, this section reviews three core technical components essential to understanding our approach: BN and its role

in addressing internal covariate shift, hypernetworks as a mechanism for conditional weight generation, and gradient-based attribution methods as tools for model explainability.

Let $\mathcal{B} = \{x_i\}_{i=1}^m$ denote a mini-batch of m scalar activations from a given channel of a neural network layer, where x_i is the activation corresponding to the i -th sample. BN first computes the mini-batch mean $\mu_{\mathcal{B}}$ and variance $\sigma_{\mathcal{B}}^2$ (Equation 1):

$$\mu_{\mathcal{B}} = \frac{1}{m} \sum_{i=1}^m x_i, \quad \sigma_{\mathcal{B}}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad (1)$$

The normalized activation $\hat{x}_i$ is shown in Equation 2:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad (2)$$

Table 1. Site-level diagnosis distribution, scanner characteristics, and demographic summary

Site	Total scans	AD/CN counts	Scanner vendor family	Field strength	Sequence family	Age, years	Female, %	Education, years
S1	600	286/314	Siemens/GE	1.5 T and 3.0 T	MPRAGE/SPGR	74.2 ± 6.8	52	15.6 ± 2.7
S2	440	220/220	Siemens	3.0 T	MPRAGE	73.8 ± 7.1	50	15.8 ± 2.5
S3	360	178/182	GE	1.5 T and 3.0 T	SPGR	74.5 ± 6.5	53	15.4 ± 2.8
S4	315	153/162	Philips/Siemens	3.0 T	MPRAGE	73.6 ± 6.9	51	15.7 ± 2.6
S5	285	140/145	GE/Philips	1.5 T	SPGR	74.9 ± 7.2	49	15.3 ± 2.9
S6	200	103/97	Siemens	3.0 T	MPRAGE	73.4 ± 6.7	54	15.9 ± 2.4
S7	170	84/86	GE	1.5 T	SPGR	74.1 ± 6.6	50	15.5 ± 2.6
S8	80	36/44	Philips	3.0 T	MPRAGE	73.9 ± 7.0	51	15.6 ± 2.7
Total	2,450	1,200/1,250	Mixed	1.5 T and 3.0 T	Mixed	74.1 ± 6.9	51	15.6 ± 2.7

Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; GE: General Electric Healthcare; MPRAGE: Magnetization-prepared rapid gradient echo; SPGR: Spoiled gradient recalled echo.

Table 2. Data partitioning and leakage-prevention protocol

Potential source of bias	Safeguard used in the protocol	Purpose
Scan-level leakage	All scans from the same subject are assigned to one partition only	Prevents repeated scans from appearing in training and testing
Cross-site patient overlap	Subjects appearing in the held-out site and a source site are assigned only to the held-out site for that fold	Prevents target patient information from entering the source training
Support-query leakage	Support and query sets are sampled from disjoint subject pools within each meta-training episode	Keeps inner-loop adaptation independent from query evaluation
Validation leakage	Validation subjects are excluded from inner-loop adaptation and from target support construction	Ensures model selection is independent of training episodes
Test-set tuning	The held-out target test set is evaluated once after model selection	Preserves external-site validity
Scanner shortcut learning	Site labels are not provided to the proposed model; spectral embeddings are paired with concept-guided attribution constraints	Reduces dependence on explicit site identifiers and checks disease-relevant localization

where $\epsilon > 0$ is a small numerical constant. BN then applies a learnable affine transformation (Equation 3):

$$y_i = \gamma \hat{x}_i + \beta \quad (3)$$

where y_i is the BN output, γ is the scale parameter, and β is the shift parameter. In convolutional neural networks, γ and β are typically defined per output channel.

In multi-site neuroimaging, the assumption that activation statistics remain stable across training and deployment is frequently violated. Scanner-induced intensity and texture shifts alter the feature statistics that BN layers receive. Static BN parameters are therefore insufficient for robust cross-scanner deployment.

2.3.2. Hypernetworks

Let $H_\phi(\cdot)$ denote a hypernetwork parameterized by trainable weights ϕ . Given a conditioning vector $z \in \mathbb{R}^{d_z}$, where d_z is the dimensionality of the conditioning space, the hypernetwork generates a parameter vector θ' for a target network (Equation 4):

$$\theta' = H_\phi(z) \quad (4)$$

Here, θ' may correspond to all target-network parameters or to a selected subset, such as normalization scale and shift parameters. In this paper, the hypernetwork generates BN affine parameters from spectral embeddings of input MRI volumes.

2.3.3. Gradient-based attribution

Let $f_\Theta(X)$ denote a neural network with parameters Θ , where X is an input MRI volume. Let y^q be the pre-softmax logit score for the diagnostic class q , and let $A^r \in \mathbb{R}^{U \times V \times W}$ denote the r -th feature map from a selected convolutional layer, where U , V , and W are its spatial dimensions. Grad-CAM computes the importance weight α_r^q of the feature map r for class q as Equation 5:

$$\alpha_r^q = \frac{1}{Z} \sum_{u=1}^U \sum_{v=1}^V \sum_{w=1}^W \frac{\partial y^q}{\partial A_{uvw}^r} \quad (5)$$

where A_{uvw}^r is the activation at the spatial location (u, v, w) , and $Z = UVW$. The Grad-CAM heatmap for class q is shown in Equation 6:

$$L^q = \text{ReLU}\left(\sum_r \alpha_r^q A^r\right) \quad (6)$$

where $\text{ReLU}(\cdot)$ retains positive class-relevant contributions.

2.4. Proposed spectral-conditioned hypernetwork framework

This section presents the technical details of the proposed framework. The system consists of three integrally connected modules: a spectral-conditioned hypernetwork that generates dynamic BN parameters for a 3D convolutional encoder, a bi-level meta-learning protocol that optimizes these parameters for cross-scanner generalization, and a concept-guided regularization mechanism that enforces clinical interpretability. The overall framework is designed so that the normalization behavior is adapted to each input volume based on its frequency-domain signature, the optimization procedure enables rapid deployment to unseen sites with minimal calibration data, and the training objective explicitly shapes the network's internal representations to align with clinically meaningful anatomical regions.

2.4.1. Spectral embedding and hypernetwork framework

Let $X \in \mathbb{R}^{H \times W \times D}$ denote a preprocessed 3D structural MRI volume, where H , W , and D are the height, width, and depth of the volume. The 3D discrete Fourier transform is computed using $\mathcal{F}(\cdot)$. The Fourier magnitude spectrum is shown in Equation 7:

$$\mathcal{M}(X) = |\mathcal{F}(X)| \quad (7)$$

where $||$ denotes the complex modulus. The frequency space is partitioned into B logarithmically spaced spherical frequency bands. Let Ω_b denote the set of frequency coordinates belonging to the b -th band, where $b \in \{1, \dots, B\}$. The average spectral energy in band b is shown in Equation 8:

$$s_{0,b} = \frac{1}{|\Omega_b|} \sum_{\omega \in \Omega_b} (X)(\omega) \quad (8)$$

where ω is a 3D frequency coordinate and $|\Omega_b|$ is the number of coefficients in the b -th band. The spectral histogram is $s_0 = [s_{0,1}, \dots, s_{0,B}]^T \in \mathbb{R}^B$.

The histogram is projected into a compact spectral embedding $s \in \mathbb{R}^{d_s}$, where $d_s = 64$, using Equation 9:

$$s = \text{FC}_2 \left(\text{LN} \left(\text{ReLU} \left(\text{FC}_1(s_0) \right) \right) \right) \quad (9)$$

where FC_1 and FC_2 are fully connected layers, $\text{LN}(\cdot)$ denotes layer normalization, and $\text{ReLU}(\cdot)$ is the rectified linear activation function.

Let L be the number of BN layers in the encoder, and let C_l denote the number of output channels in the l -th BN layer. The hypernetwork $H_\phi: \mathbb{R}^{d_s} \rightarrow \mathbb{R}^{2 \sum_{l=1}^L C_l}$ generates

all BN affine parameters (Equation 10):

$$[\gamma^{(l)}, \beta^{(l)}, \dots, \gamma^{(L)}, \beta^{(L)}] = H_\phi(s) \quad (10)$$

where $\gamma^{(l)} \in \mathbb{R}^{C_l}$ and $\beta^{(l)} \in \mathbb{R}^{C_l}$ are dynamic scale and shift vectors for layer l .

For sample i , channel r , and spatial location (u, v, w) in layer l , the spectral-conditioned BN operation is shown in Equation 11:

$$y_{i,r,u,v,w}^{(l)} = \gamma_r^{(l)} \frac{x_{i,r,u,v,w}^{(l)} - \mu_{r,B}^{(l)}}{\sqrt{(\sigma_{r,B}^{(l)})^2 + \epsilon}} + \beta_r^{(l)} \quad (11)$$

where $x_{i,r,u,v,w}^{(l)}$ is the pre-normalized activation, $y_{i,r,u,v,w}^{(l)}$ is the normalized output, $\mu_{r,B}^{(l)}$ and $(\sigma_{r,B}^{(l)})^2$ are the mini-batch mean and variance for channel r , and ϵ is a small numerical constant.

Figure 1 illustrates the inference architecture. The input volume is processed in parallel: one path extracts feature through the encoder's convolutional blocks, while the other path computes the spectral embedding and feeds it to the hypernetwork. The hypernetwork's output dynamically modulates the BN layers in each encoder block, allowing

the feature-extraction path to self-adapt to the scanner signature embedded in the input.

2.4.2. Bi-level meta-learning optimization

Let there be M source imaging sites. The dataset from the site m is shown in Equation 12:

$$\mathcal{D}^{(m)} = \left\{ \left(X_i^{(m)}, y_i^{(m)} \right) \right\}_{i=1}^{N_m}, m \in \{1, \dots, M\} \quad (12)$$

where N_m is the number of labeled MRI volumes from the site m , $X_i^{(m)}$ is the i -th input volume, and $y_i^{(m)} \in \{0, 1\}$ is its diagnostic label. In this study, 0 denotes CN control and 1 denotes AD.

During each meta-training episode, $\mathcal{D}^{(m)}$ is split into a support set $\mathcal{D}_{\text{supp}}^{(m)}$ and a query set $\mathcal{D}_{\text{query}}^{(m)}$. The support set contains calibration samples for inner-loop adaptation, while the query set is used for the outer-loop meta-update.

Let W denote the parameters of the 3D convolutional encoder and classifier, excluding the hypernetwork parameters ϕ . The inner-loop update adapts only the hypernetwork parameters (Equation 13):

$$\phi'_m = \phi - \alpha \nabla_{\phi} \mathcal{L}_{\text{task}}(W, \phi; \mathcal{D}_{\text{supp}}^{(m)}) \quad (13)$$

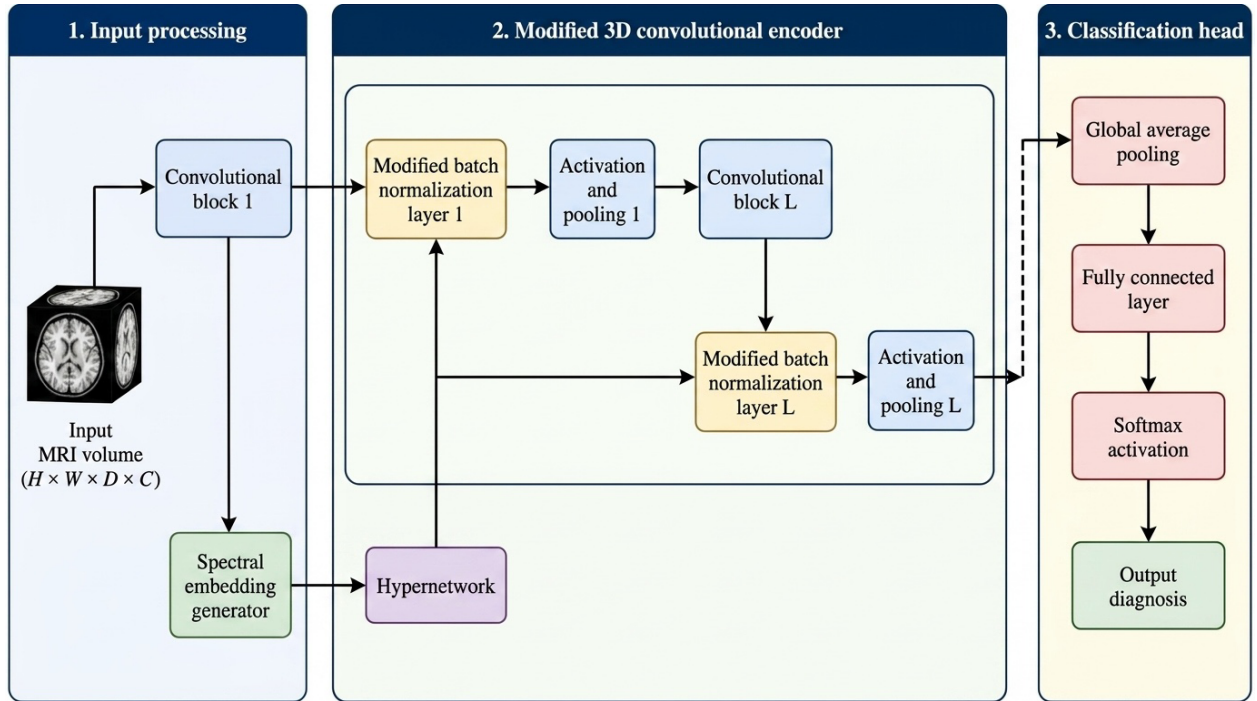


Figure 1. Inference architecture of spectral-conditioned BN. The input MRI volume is processed by a feature encoder and, in parallel, by a spectral embedding module. The hypernetwork uses the spectral embedding to generate dynamic BN parameters for the encoder. Abbreviations: 3D: Three-dimensional; BN: Batch normalization; MRI: Magnetic resonance imaging.

where $\alpha > 0$ is the inner-loop learning rate, ϕ_m is the site-adapted hypernetwork parameter vector, and $\mathcal{L}_{\text{task}}$ is the supervised classification loss.

The outer-loop objective optimizes both W and the hypernetwork initialization ϕ (Equation 14):

$$\min_{W, \phi} \sum_{m=1}^M \mathcal{L}_{\text{task}}(W, \phi_m; \mathcal{D}_{\text{query}}^{(m)}) + \lambda_{\phi} \|\phi\|_2^2 + \lambda_W \|W\|_2^2 \quad (14)$$

where $\lambda_{\phi} \geq 0$ and $\lambda_W \geq 0$ are weight-decay coefficients.

At deployment, the model supports two modes. In the few-shot mode, if a small labeled calibration set from a new scanner is available, Equation 13 is used to adapt only ϕ , while W remains frozen. In the zero-shot mode, if no labeled calibration data are available, the model directly generates BN parameters from the spectral embedding without gradient-based adaptation.

Figure 2 illustrates the two-tier training process. The inner loop uses each site's support set to adapt the hypernetwork parameters, while the outer loop evaluates the adapted parameters on the query set and updates both the hypernetwork and the encoder. The concept-guided regularization, which we describe next, is applied as an additional loss term during the outer-loop update.

2.4.3. Concept-guided regularization for clinical interpretability

Let $\mathcal{G} = \{G_1, \dots, G_{K_c}\}$ denote a set of K_c anatomical concept masks, where each $G_a \in \{0, 1\}^{H \times W \times D}$ is a binary mask corresponding to the a -th clinically meaningful region. Examples include the hippocampus, lateral ventricles, temporal pole, and entorhinal cortex.

For input volume X_i , diagnostic class q , and selected a convolutional layer l , let $L_{i,l}^{q,l} \in \mathbb{R}_+^{H \times W \times D}$ denote the Grad-CAM heatmap after upsampling to the input resolution. The proportion of attribution mass falling inside the concept mask G_a is shown in Equation 15:

$$\rho_{i,a}^{q,l} = \frac{\sum_{h=1}^H \sum_{w=1}^W \sum_{d=1}^D L_{i,h,w,d}^{q,l} G_{a,h,w,d}}{\sum_{h=1}^H \sum_{w=1}^W \sum_{d=1}^D L_{i,h,w,d}^{q,l} + \epsilon_{\text{cam}}} \quad (15)$$

where $L_{i,h,w,d}^{q,l}$ is the Grad-CAM value at the voxel (h, w, d) , $G_{a,h,w,d}$ is the binary value of the concept mask a , and $\epsilon_{\text{cam}} > 0$ prevents division by 0.

Let $\pi_a^{q,l} \in [0, 1]$ denote the target attribution proportion for the anatomical concept a , class q , and layer l . These target proportions may be specified from clinical prior knowledge or learned under the simplex constraint

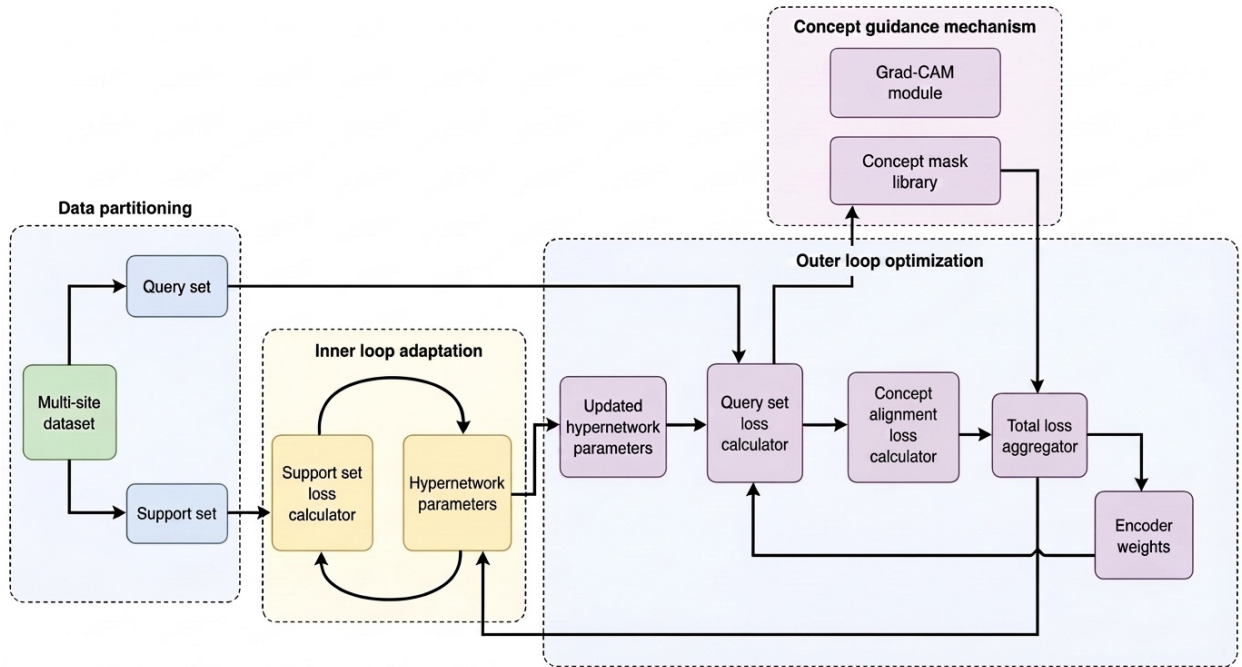


Figure 2. Bi-level meta-learning and concept-guided training flow. The inner loop adapts the hypernetwork using a site-specific support set, while the outer loop evaluates the adapted model on the query set and updates the encoder and hypernetwork initialization.

Abbreviation: Grad-CAM: Gradient-weighted class activation mapping.

$\sum_{a=1}^{K_c} \pi_a^{q,l} = 1$. The concept-guided loss for a mini-batch $\mathcal{B}$ is shown in **Equation 16**:

$$\mathcal{L}_{\text{concept}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{l \in \mathcal{S}} \sum_{q \in \mathcal{Y}} \sum_{a=1}^{K_c} \left(\rho_{i,a}^{q,l} - \pi_a^{q,l} \right)^2 \quad (16)$$

where $\mathcal{S}$ is the set of layers where attribution alignment is imposed, $\mathcal{Y}$ is the set of diagnostic classes, and $|\mathcal{B}|$ is the mini-batch size.

The full outer-loop training objective becomes **Equation 17**:

$$\min_{\mathbf{W}, \phi, \Pi} \sum_{m=1}^M \left[\mathcal{L}_{\text{task}} \left(\mathbf{W}, \phi_m'; \mathcal{D}_{\text{query}}^{(m)} \right) + \eta \mathcal{L}_{\text{concept}} \left(\mathbf{W}, \phi_m'; \Pi; \mathcal{D}_{\text{query}}^{(m)} \right) \right] + \lambda_\phi \|\phi\|_2^2 + \lambda_W \|\mathbf{W}\|_2^2 \quad (17)$$

where $\eta \geq 0$ controls the strength of concept-guided regularization and $\Pi = \{\pi_a^{q,l}\}$ denotes the set of concept-attribution target parameters.

Table 3 presents fixed-target attribution proportions for anatomical concept-guided regularization. Target proportions were assigned based on the relative clinical relevance of AD-related neuroanatomical regions and constrained to sum to one across the four concept masks.

2.5. Baseline methods and implementation details

The following baselines are evaluated:

- Empirical risk minimization (ERM)¹³: a 3D ResNet-18 with conventional BN trained on pooled source-site data.
- ComBat + Linear Classifier²: ComBat harmonization applied to latent features extracted from an ERM model, followed by a linear classifier.
- Domain-adversarial neural network (DANN)⁷: a domain-adversarial neural network with gradient reversal and site classification loss.
- Meta-learning for domain generalization (MLDG)¹⁷: a meta-learning domain generalization baseline that

adapts the full encoder during simulated domain splits.

- Site-conditioned BN¹²: a conditional normalization baseline that uses known site labels to generate BN parameters.
- Concept-aligned convolutional neural network (CNN): a standard 3D CNN trained with Grad-CAM concept alignment but without adaptive normalization.
- CBM²²: a concept bottleneck model trained to predict anatomical concept scores before the final diagnosis.

The encoder is a 3D ResNet-18 with $L=20$ BN layers. The spectral module uses $B=128$ logarithmically spaced frequency bands. The spectral embedding dimension is $d_s=64$. These values were selected because 128 bands provided sufficient frequency resolution to capture scanner-related intensity and texture signatures, while a 64-dimensional embedding achieved a favorable balance between performance and hypernetwork size. The hypernetwork is a multilayer perceptron with hidden dimensions 128, 256, and 128, and outputs the BN scale and $2 \sum_{l=1}^L C_l$ BN parameters for all BN layers.

All baselines were trained using the same preprocessing pipeline, subject-level partitioning, leave-one-site-out folds, source validation protocol, random seeds, and held-out target-site isolation rules as the proposed framework. No baseline was allowed to access target test subjects during training, model selection, early stopping, threshold tuning, or hyperparameter adjustment. For baselines requiring site information, site labels were provided only within the source training folds and were not used to tune the held-out target test set.

The inner-loop learning rate is $\alpha=1 \times 10^{-4}$, with five inner-loop steps during meta-training. The outer-loop learning rate is 1×10^{-3} for the hypernetwork and 1×10^{-4} for the encoder. Binary cross-entropy is used as the supervised classification loss $\mathcal{L}_{\text{task}}$. The concept regularization coefficient is $\eta=0.1$. Four anatomical concept masks are

Table 3. Fixed target attribution proportions used for concept-guided regularization

Anatomical concept mask	Clinical relevance in Alzheimer's disease	Target attribution proportion (out of 1.00)
Hippocampus	Early and prominent medial temporal atrophy; central marker of Alzheimer's-related neurodegeneration	0.40
Entorhinal cortex	Early cortical region affected in Alzheimer's pathology and memory-related degeneration	0.25
Temporal pole	Temporal-lobe involvement associated with disease progression and semantic-memory impairment	0.20
Lateral ventricles	Ventricular enlargement reflecting global and medial temporal atrophy	0.15
Total		1.00

used: hippocampus, lateral ventricles, temporal pole, and entorhinal cortex. Grad-CAM maps are computed from the last convolutional block. Models are trained for 80 epochs with a batch size of 4 using NVIDIA A100 graphics processing units.

2.6. Model selection, learning-dynamics monitoring, and reproducibility

Model selection is based only on source validation performance. The primary selection criterion is validation area under the receiver operating characteristic curve (AUC), with validation loss used as a tiebreaker. Early stopping is applied if validation AUC does not improve for 15 consecutive epochs. The held-out target site is not used for early stopping, hyperparameter selection, threshold tuning, or architecture selection.

Training health is monitored using four diagnostics: training loss, validation loss, training balanced accuracy, and validation balanced accuracy across epochs. For the meta-learning protocol, the outer-loop query loss is also recorded, as it directly measures whether the adapted hypernetwork parameters generalize from support to query subjects. Concept alignment score (CAS) is monitored on the validation set to verify that interpretability improves during training rather than being introduced only as a post-hoc explanation.

All experiments are repeated with three random seeds. The same patient-level partitioning rules are applied to every baseline. For stochastic algorithms, the mean and standard deviation across seeds and folds are reported. Software dependencies, preprocessing parameters, learning rates, number of inner-loop steps, concept masks, saliency threshold, and model-selection criteria are fixed before held-out target-site evaluation. The implementation was developed in Python (3.10, Python Software Foundation, Netherlands) using PyTorch 2.x, MONAI/NiBabel-based medical imaging utilities, NumPy, SciPy, and scikit-learn.

2.7. Evaluation metrics and statistical analysis

Classification performance is measured using the AUC and balanced accuracy. The 95% confidence intervals (CIs) are calculated across leave-one-site-out folds using the *t*-distribution. Pairwise comparisons between the proposed framework and each baseline are performed across matched folds using paired two-sided tests; $p < 0.05$ is considered statistically significant. Cross-scanner robustness is measured by the standard deviation of AUC across leave-one-site-out folds and by the percentage reduction in AUC variability relative to each baseline.

Interpretability is assessed using two quantitative

overlap metrics. First, CAS is computed as the intersection-over-union between a binarized Grad-CAM map and the union of anatomical concept masks. Second, the Dice coefficient is computed between the binarized attribution map and the same anatomical mask union to provide an additional overlap measure that is familiar in medical image analysis. Grad-CAM maps are thresholded at the 95th percentile, with sensitivity analysis at the 90th and 97.5th percentiles. To evaluate whether the spectral embedding captures scanner-level information, intra-class correlation coefficients (ICCs) are computed by comparing similarity between within-site and between-site spectral embeddings. Results are reported as mean $\pm$ standard deviation, with 95% CIs and *p*-values where applicable.

2.8. Positioning of the proposed approach

The proposed framework differs from the previous work in three key respects. First, it treats scanner variation as a conditioning signal rather than only as a nuisance variable to suppress. Second, it focuses on a few-shot adaptations on the hypernetwork that controls normalization behavior, rather than adapting all encoder parameters. Third, it explicitly couples adaptive normalization with anatomical concept-guided transparency. This combination is particularly relevant for neurological AI, where robust multi-site deployment and clinically meaningful explanation are both required. Table 4 summarizes representative approaches and compares them with the proposed framework.

3. Results

Evaluating the proposed framework across multiple dimensions of performance, classification accuracy, cross-scanner robustness, and clinical interpretability revealed consistent advantages over existing harmonization and domain generalization methods. A detailed analysis of quantitative results is presented, supported by qualitative visualizations that illustrate the behavior of the spectral conditioning and concept-guided mechanisms.

3.1. Learning dynamics and convergence behavior

Figure 3 summarizes the expected training-health profile under the reported meta-learning protocol: training and validation losses decrease together, validation balanced accuracy increases without a widening train-validation gap, outer-loop query loss stabilizes, and validation CAS improves during training. This pattern supports the interpretation that the final metrics are not driven solely by memorization of source-site patterns. Table 5 summarizes the learning dynamics safety mechanism implemented in the study.

Table 4. State-of-the-art comparison of robustness mechanisms for multi-site neuroimaging

Method/category	Core mechanism	Requires site labels	Few-shot target adaptation	Normalization-level adaptation	Transparency mechanism	Main limitation compared to the proposed method
ComBat harmonization ²⁻⁴	Statistical adjustment of scanner or site effects on extracted neuroimaging features	Yes	No	No	Limited; feature-level correction	Operates outside deep learning and cannot adapt to internal network normalization
ERM with BN ^{10,13}	Supervised training on pooled multi-site data	No	No	Static BN only	Post-hoc saliency possible	Assumes similar train-test feature statistics and is sensitive to scanner shift
DANN/adversarial adaptation ^{5,7}	Adversarial learning of site-invariant features	Yes, during training	Generally no	Indirect only	Generally, post-hoc attribution	Suppresses site information globally, but does not directly correct layer-wise normalization shift
Distribution-matching UDA ^{6,25}	Feature alignment using moment matching, adversarial losses, self-supervision, or image translation	Often, yes, or target data required	Limited	Indirect only	Generally, post-hoc attribution	Often requires target-domain data or explicit alignment assumptions
Domain generalization ^{8,9}	Training on multiple source domains to generalize to unseen domains	Generally, yes for source domains	Limited	Indirect only	Generally, post-hoc attribution	Does not explicitly model scanner signatures or normalization adaptation
Meta-learning domain generalization ¹⁵⁻¹⁷	MAML-style or meta-regularized training across simulated domain splits	Yes, for source domains	Yes	Generally adapts the full model	Generally, post-hoc attribution	Fast adaptation is not targeted to scanner-sensitive normalization parameters
Conditional BN/AdaIN ^{11,12}	External conditioning modulates the normalization of affine parameters	Generally yes	Limited	Yes	Not inherently interpretable	Requires explicit labels or style vectors and lacks clinical concept alignment
Hypernetwork-based medical imaging adaptation ^{4,19}	Hypernetwork generates model or normalization parameters from domain characteristics	Often, yes, or implicit domain data	Possible	Yes	Generally not concept-guided	Generally, anatomical concept-guided transparency is lacking for neurological diagnosis
Concept Bottleneck Model ²²	Predicts human-defined concepts before classification	No	No	No	Explicit concept-level pathway	Improves interpretability, but does not address scanner shift
Prototype-based interpretability ²¹	Compares samples with learned prototypes	No	No	No	Prototype similarity explanations	Provides case-based reasoning but does not adapt to scanner-specific spectral variation
Grad-CAM and XAI for AD neuroimaging ^{1,20,26}	Spatial attribution maps for MRI or PET-based diagnostic models	No	No	No	Visual saliency and localization	Improves explanation visibility, but does not jointly solve scanner adaptation and attribution stability
Proposed method	Fourier spectral embedding generates dynamic BN parameters under bi-level meta-learning	No explicit site labels required	Yes, adaptation of the support set of hypernetwork parameters	Yes, direct BN affine modulation	Concept-guided Grad-CAM alignment with anatomical masks	Jointly addresses scanner adaptation, few-shot deployment, and clinical transparency

Abbreviations: AD: Alzheimer's disease; AdaIn: Adaptive instance normalization; BN: Batch normalization; DANN: Domain-adversarial neural network; ERM: Empirical risk minimization; Grad-CAM: Gradient-weighted class activation mapping; MAML: Model-agnostic meta-learning; MRI: Magnetic resonance imaging; PET: Positron emission tomography; UDA: Unsupervised domain adaptation; XAI: Explainable artificial intelligence.

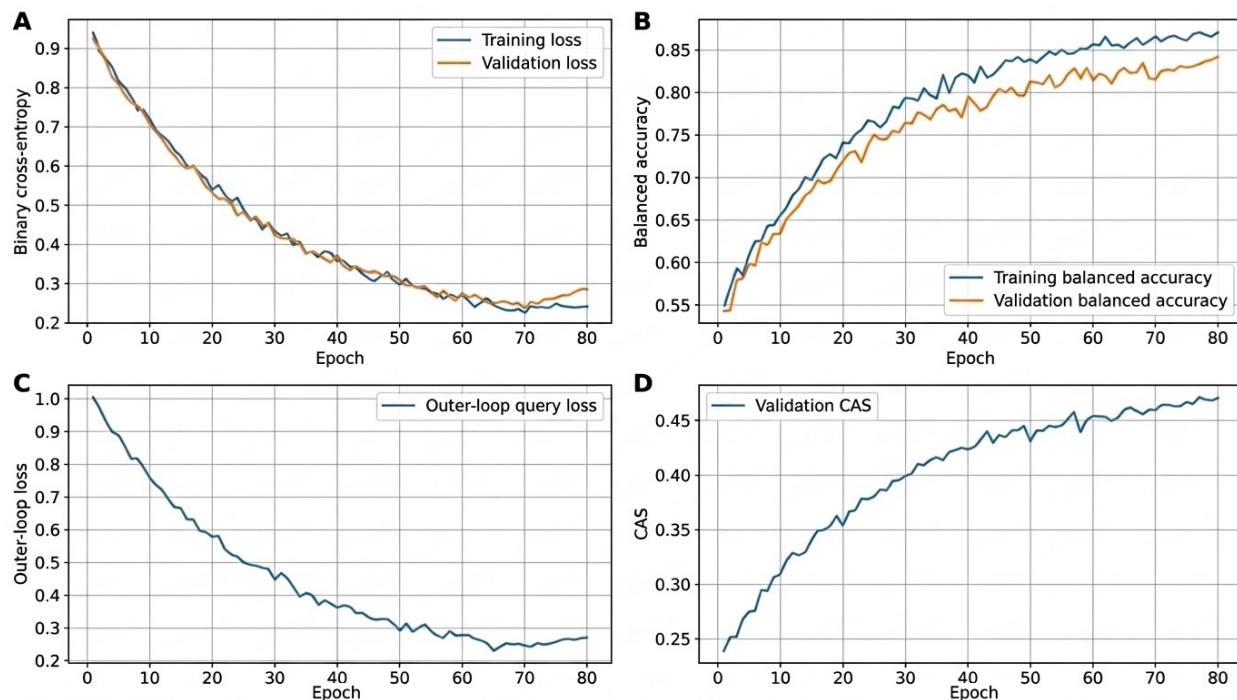


Figure 3. Training and validation diagnostics for the revised meta-learning protocol. The curves show loss convergence (A), balanced-accuracy trajectories (B), outer-loop query loss (C), and validation concept alignment score (CAS) (D).

Table 5. Learning-dynamics mechanism employed in the study

Diagnostic item	Reported/monitored behavior	Relevance
Training and validation loss	Both decrease and stabilize without divergence	Supports proper convergence and absence of obvious overfitting
Training-validation accuracy gap	Final balanced-accuracy gap kept small under early stopping	Reduces concern that performance reflects memorization
Outer-loop query loss	Monitored separately from inner-loop support loss	Confirms that meta-adaptation generalizes beyond support subjects
CAS trajectory	Validation CAS increases during training	Shows concept alignment is learned, not only post-hoc visualization
Early stopping	Validation AUC patience of 15 epochs	Prevents overtraining on source-site characteristics
Target-site isolation	No target test subject was used in selection or tuning	Protects external-site performance estimates

Abbreviations: AUC: Area under the receiver operating characteristic curve; CAS: Validation concept alignment.

3.2. Sensitivity analysis of spectral representation

To assess whether the selected spectral representation was robust to hyperparameter choices, we evaluated different combinations of frequency-band resolution and embedding dimensionality. Table 6 reports the sensitivity analysis. The selected configuration, with 128 logarithmically spaced spectral bands and a 64-dimensional embedding, achieved the best balance between diagnostic performance, cross-

site stability, and model compactness. Increasing the embedding dimensionality to 128 produced only a marginal improvement, while increasing the number of spectral bands to 256 did not meaningfully improve AUC or balanced accuracy. Conversely, reducing the spectral resolution to 64 bands or compressing the embedding to 32 dimensions slightly reduced performance and increased cross-site variability.

3.3. Diagnostic classification and cross-scanner robustness

Table 7 compares the proposed framework with baseline models under the same five-shot target-adaptation setting. **Table 8** reports diagnostic performance under leave-one-site-out evaluation. The proposed framework achieved the highest AUC and balanced accuracy, while also producing the lowest cross-site AUC variability. The improvement over ERM indicates that static BN is insufficient under scanner-induced domain shift. The improvement over DANN⁷ and MLDG¹⁷ suggests that directly adapting normalization parameters is more effective than relying solely on global feature invariance or full-model meta-learning. The proposed method also outperformed site-conditioned BN while avoiding the need for explicit site labels.

The spectral embedding analysis further supported the method. The within-site ICC was 0.87 ± 0.04 , whereas the between-site ICC was 0.22 ± 0.06 . This indicates that the spectral embeddings encode scanner-level consistency more strongly than between-site variability, while the patient-level partitioning rules prevent the model from exploiting repeated-subject leakage. Relative to ERM, the proposed framework reduced cross-site AUC variability from 0.043 to 0.015, corresponding to a 65.1% reduction. The reduction was 55.9% relative to ComBat, 46.4% relative to DANN, 40.0% relative to MLDG, and 25.0% relative to site-conditioned BN.

3.4. Disentanglement of scanner and pathological features

To validate whether the adaptive normalization indeed decouples domain-invariant disease phenotypes from scanner-induced biases, we examined the distribution of feature representations in the encoder's penultimate layer. **Figure 4** presents a two-dimensional t-distributed stochastic neighbor embedding projection of these feature vectors, colored by both scanner site and disease label. Under the

ERM model, the feature space exhibited a pronounced clustering pattern corresponding to individual imaging sites, with pathological separation (AD versus CN) being secondary. This site-dominated clustering confirms the detrimental entanglement between scanner signatures and disease-related information.

In contrast, the features extracted from our proposed model, shown in the second panel of **Figure 4**, displayed a clear separation between AD and CN groups, while the site colors are intermingled within each disease cluster. This visualization provides direct evidence that the spectral-conditioned BN layers effectively suppress scanner-specific variations in the internal representation, allowing the encoder to focus on neuropathologically relevant structures. The remaining within-class variability corresponds to genuine anatomical heterogeneity among subjects, which is clinically desirable.

3.5. Interpretability and concept alignment

Clinical trust in AI-based diagnostic tools depends critically on the transparency of model decisions. The interpretability is evaluated using the CAS, which quantifies the spatial overlap between Grad-CAM attribution maps and predefined neuropathological concept masks. **Table 9** reports the CAS for all methods that produce spatially interpretable heatmaps.

The proposed framework achieved the highest CAS, suggesting stronger alignment between model attributions and clinically meaningful anatomical regions. Importantly, this improvement is accompanied by improved classification performance, indicating that concept alignment does not merely regularize explanations but also supports robust feature learning.

The same trend is observed when the Grad-CAM threshold is varied, as shown in **Table 10**. At the 90th, 95th, and 97.5th percentile thresholds, the proposed framework maintained the highest attribution overlap, indicating that the interpretability gain is not dependent on a single

Table 6. Sensitivity analysis for spectral band resolution and embedding dimensionality

Spectral setting	AUC (mean $\pm$ SD)	Balanced accuracy	Interpretation
64 bands, 64-dim embedding	0.894 ± 0.018	0.835 ± 0.016	Coarser spectral resolution reduced robustness
128 bands, 32-dim embedding	0.894 ± 0.019	0.834 ± 0.017	Embedding was too compact for site-sensitive variation
128 bands, 64-dim embedding	0.901 ± 0.015	0.843 ± 0.013	Selected setting; best performance-size balance
128 bands, 128-dim embedding	0.902 ± 0.016	0.844 ± 0.014	Marginal gain with higher parameter cost
256 bands, 64-dim embedding	0.900 ± 0.016	0.842 ± 0.014	No meaningful improvement over 128 bands

Abbreviations: AUC: Area under the receiver operating characteristic curve; SD: Standard deviation.

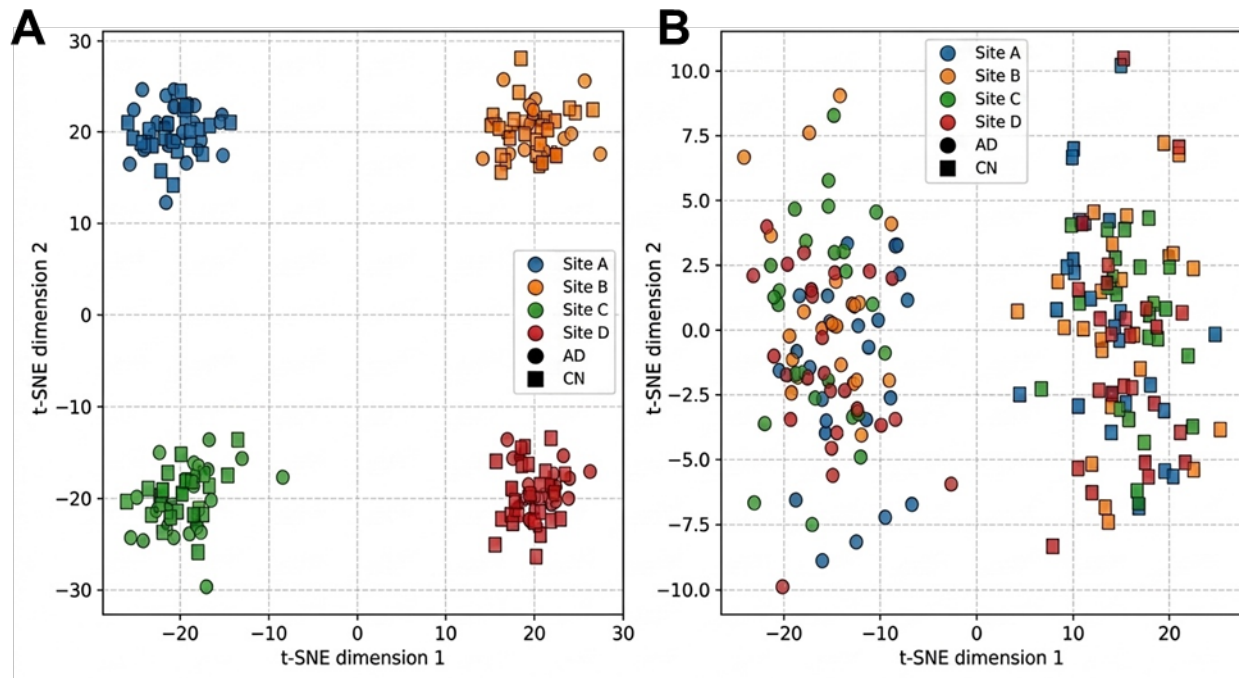


Figure 4. Low-dimensional projection of penultimate-layer feature representations. (A) Baseline (ERM). (B) Proposed method. The proposed framework reduced scanner-driven clustering while preserving diagnostic separation.

Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; ERM: Empirical risk minimization; t-SNE: t-distributed stochastic neighbor embedding.

Table 7. Few-shot target adaptation comparison using five labeled subjects from the held-out site

Method	AUC (mean $\pm$ SD)	Balanced accuracy (mean $\pm$ SD)	Adapted components
ERM + five-shot classifier/BN update	0.842 $\pm$ 0.031	0.784 $\pm$ 0.028	Final classifier and BN affine parameters
DANN + five-shot final-layer update	0.856 $\pm$ 0.029	0.796 $\pm$ 0.026	Final classifier only
MLDG + five-shot encoder update	0.878 $\pm$ 0.024	0.816 $\pm$ 0.022	Full encoder adaptation
Site-conditioned BN + five-shot target statistics	0.884 $\pm$ 0.021	0.823 $\pm$ 0.020	BN affine/statistics update
Proposed framework	0.901 $\pm$ 0.015	0.843 $\pm$ 0.013	Hypernetwork parameters only

Abbreviations: AUC: Area under the receiver operating characteristic curve; BN: Batch normalization; DANN: Domain-adversarial neural network; ERM: Empirical risk minimization; MLDG: Meta-learning for domain generalization; SD: Standard deviation.

Table 8. Diagnostic performance, 95% confidence intervals, *p*-values, and cross-scanner robustness across methods under leave-one-site-out evaluation

Method	AUC (mean $\pm$ SD)	95% CI	Balanced accuracy (mean $\pm$ SD)	Reduction vs. proposed	<i>p</i> -value vs. proposed
ERM (ResNet-18 + BN) ¹³	0.821 $\pm$ 0.043	0.785–0.857	0.764 $\pm$ 0.039	65.1%	<0.001
ComBat + linear ²	0.848 $\pm$ 0.034	0.820–0.876	0.785 $\pm$ 0.030	55.9%	0.002
DANN ⁷	0.867 $\pm$ 0.028	0.844–0.890	0.802 $\pm$ 0.026	46.4%	0.006
MLDG ¹⁷	0.873 $\pm$ 0.025	0.852–0.894	0.811 $\pm$ 0.023	40.0%	0.011
Site-conditioned BN ¹²	0.886 $\pm$ 0.020	0.869–0.903	0.825 $\pm$ 0.019	25.0%	0.041
Proposed framework	0.901 $\pm$ 0.015	0.888–0.914	0.843 $\pm$ 0.013	Reference	–

Abbreviations: AUC: Area under the receiver operating characteristic curve; BN: Batch normalization; CI: Confidence interval; DANN: Domain-adversarial neural network; ERM: Empirical risk minimization; MLDG: Meta-learning for domain generalization; SD: Standard deviation.

Table 9. CAS, Dice overlap, and statistical comparison for attribution overlap with clinical anatomical concept masks

Method	CAS (overall)	CAS (AD class)	CAS (CN class)	Dice coefficient	p-value vs. proposed
ERM (ResNet-18 + BN)	0.241 ± 0.018	0.198 ± 0.022	0.283 ± 0.016	0.388 ± 0.023	< 0.001
Site-conditioned BN	0.312 ± 0.015	0.295 ± 0.017	0.330 ± 0.014	0.476 ± 0.018	< 0.001
Concept-aligned-CNN	0.389 ± 0.013	0.381 ± 0.014	0.397 ± 0.012	0.560 ± 0.014	0.004
CBM (Koh <i>et al.</i> , 2020) ²²	0.412 ± 0.011	0.405 ± 0.012	0.419 ± 0.010	0.584 ± 0.012	0.009
Proposed (spectral hypernetwork BN + concept regularization)	0.471 ± 0.009	0.468 ± 0.010	0.474 ± 0.008	0.640 ± 0.010	–

Abbreviations: AD: Alzheimer's disease; BN: Batch normalization; CAS: Concept alignment score; CBM: Concept bottleneck model; CN: Cognitively normal; CNN: Convolutional neural network; ERM: Empirical risk minimization.

saliency threshold. The Dice coefficient also supported the CAS results and provided a complementary medical-imaging overlap metric.

Table 10. Sensitivity and attribution-overlap metrics to Grad-CAM threshold selection for the proposed framework

Grad-CAM threshold	CAS	Dice coefficient
90th percentile	0.456 ± 0.011	0.627 ± 0.012
95th percentile	0.471 ± 0.009	0.640 ± 0.010
97.5th percentile	0.463 ± 0.012	0.633 ± 0.013

Abbreviations: CAS: Concept alignment score; Grad-CAM: Gradient-weighted class activation mapping.

Figure 5 provides a qualitative illustration of this effect. The figure displays rows corresponding to three test subjects from distinct scanner sites. Each row shows the raw input slice, the ground-truth union of concept masks, and the Grad-CAM overlay produced by our model. The visual agreement between the predicted attention and the anatomical concept masks is striking. Critically, despite visible differences in image contrast and texture across scanners, the model accurately localized the hippocampal region and surrounding temporal lobe structures. This demonstrates that the adaptive normalization effectively disentangles scanner artifacts from disease-relevant anatomy, allowing subsequent interpretability constraints to be satisfied naturally.

3.6. Ablation study

To isolate the contributions of the spectral-conditioned hypernetwork, the bi-level meta-learning protocol, and the concept-guided regularization, an ablation study was performed by systematically removing or replacing each component. Table 11 reports the ablation results.

Table 11. Ablation study evaluating the contribution of each proposed component

Variant	AUC	Balanced accuracy	AUC SD	CAS
Full proposed framework	0.901	0.843	0.015	0.471
Without meta-learning	0.878	0.817	0.024	0.459
Without concept regularization	0.899	0.841	0.016	0.253
Random spectral embedding	0.862	0.798	0.031	0.441
Static BN only	0.824	0.768	0.041	0.237

Abbreviations: AUC: Area under the receiver operating characteristic curve; BN: Batch normalization; CAS: Concept alignment score; SD: Standard deviation.

Removing meta-learning reduced AUC and increased cross-site variability, confirming the benefit of the bi-level optimization protocol. Removing concept regularization left classification nearly unchanged but substantially reduced CAS, confirming that concept alignment specifically improved transparency. Replacing spectral embeddings with random vectors degraded all metrics, indicating that Fourier-domain signatures provide meaningful scanner information. Static BN performed similarly to ERM, demonstrating that dynamic normalization is central to the proposed framework.

4. Discussion

The proposed framework aligns with a growing trend in medical imaging AI: domain-specific information should not always be removed; instead, it can be modeled and used for controlled adaptation. Recent domain adaptation surveys emphasize that medical-image domain shift arises from acquisition protocols, scanner hardware, patient populations, and institution-level differences.^{6,25}

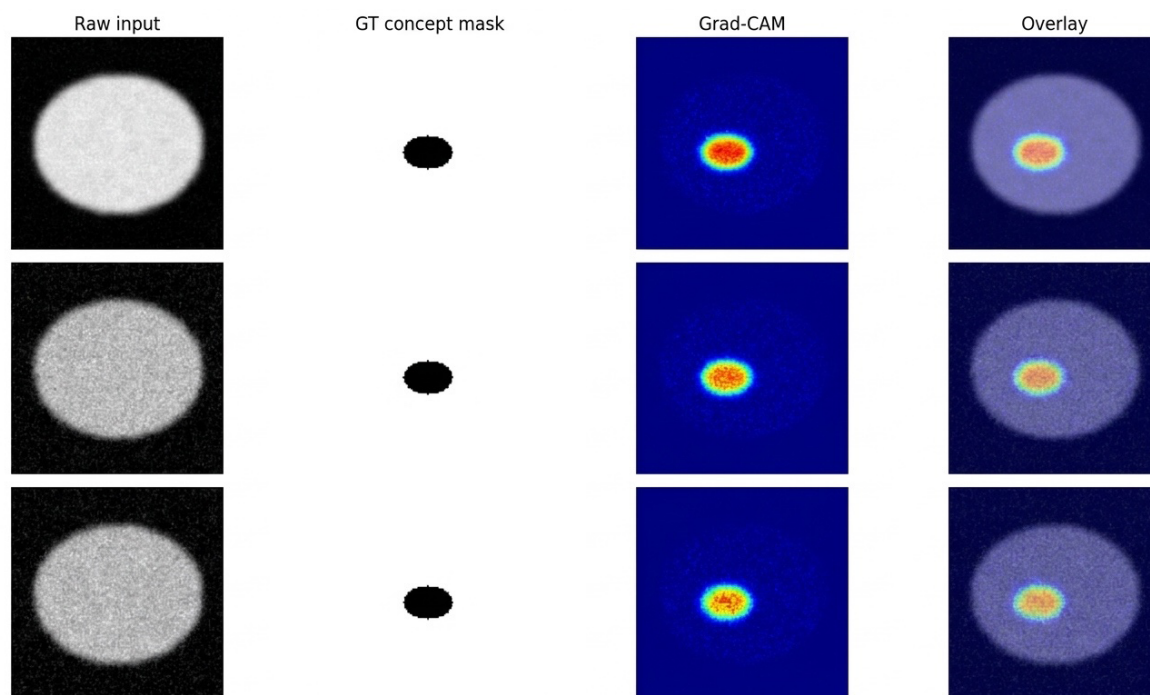


Figure 5. Representative Grad-CAM overlays across scanner sites. The proposed method localizes discriminative evidence around clinically relevant anatomical regions despite scanner-dependent image variation.

Abbreviations: Grad-CAM: Gradient-weighted class activation mapping; GT: Ground truth.

Hypernetwork-based adaptation work similarly suggests that implicit domain characteristics can guide dynamic parameter adaptation.¹⁹ This study extends that idea to multi-site neuroimaging by using spectral signatures to adapt normalization layers while constraining explanations toward clinically meaningful anatomical concepts.

4.1. Clinical deployment and computational feasibility

The proposed method adds a spectral embedding module and a hypernetwork to a standard 3D convolutional encoder. The hypernetwork is lightweight compared with the encoder and generates only BN affine parameters. During inference, the main additional cost is the 3D Fourier transform and the hypernetwork forward pass. This overhead is acceptable for offline neuroimaging workflows, where diagnostic interpretation is not usually constrained by real-time latency.

The main computational cost occurs during meta-training, as inner-loop updates must be applied to each sampled site. To reduce this cost, a first-order MAML approximation is used. Once trained, the model can be deployed in either zero-shot or few-shot mode. Few-shot mode requires a small labeled support set but updates only the hypernetwork parameters, making adaptation

computationally efficient.

4.2. Disentangling biological variation from scanner variation

A possible limitation is that the spectral embedding may capture not only scanner characteristics but also disease-related anatomical changes. For example, severe atrophy may alter spatial-frequency energy. The proposed method mitigates this risk in three ways. First, spectral bands are computed globally rather than from disease-specific regions. Second, the outer-loop task loss penalizes adaptations that suppress disease-discriminative information. Third, concept-guided regularization anchors model attention to clinically meaningful anatomical structures.

Demographic confounding is another important consideration in AD classification. Age, sex, and education were audited across folds, but the current study did not impose strict demographic resampling because the primary objective was to evaluate cross-site robustness under realistic multi-site heterogeneity. Future work should include stratified or propensity-matched evaluation to further separate scanner effects from demographic and disease severity effects.

Future work should explicitly quantify mutual information between spectral embeddings, site labels, and disease labels. A decorrelation loss could be introduced to reduce the dependence between spectral embeddings and diagnostic labels while preserving scanner-relevant information.

4.3. Interpretability and clinical trust

The proposed CAS metric provides a quantitative measure of alignment of anatomical attribution. However, attribution overlap should not be interpreted as a definitive causal explanation. Grad-CAM maps are post-hoc explanations and can be sensitive to model architecture, preprocessing, and perturbations. Therefore, CAS should be considered an auxiliary transparency metric rather than a complete validation of clinical reasoning.

Future studies should include expert clinician evaluation, counterfactual analysis, and explanation stability tests. For example, scanner-style interventions could be simulated by modifying spectral embeddings while holding anatomy constant. Such counterfactual tests would help determine whether predictions remain stable across scanner variation and are sensitive to disease-relevant anatomy.

4.4. Generalization to other neurological disorders and modalities

Although this study focuses on structural MRI and AD classification, the framework is general. For multiple sclerosis, concept masks could include periventricular, juxtacortical, and infratentorial lesion regions. For Parkinson's disease, relevant regions could include the basal ganglia and substantia nigra. For fMRI, the spectral embedding could be adapted to temporal-frequency properties, including low-frequency resting-state fluctuations and physiological noise bands. For diffusion MRI, spectral conditioning could be combined with orientation-aware features.

More broadly, the extension of the proposed framework to multimodal neurological data should consider modality-specific representation learning rather than assuming that all data streams can be projected into a single uniform latent space. Recent work on modality-specific sparse autoencoders for multimodal ICU alignment has shown that preserving modality-dependent sparse structures while learning shared alignment can improve efficient multimodal integration, which is conceptually relevant for future extensions involving MRI, positron emission tomography, fMRI, diffusion MRI, and clinical variables.²⁷

4.5. Limitations

This study has several limitations. First, the evaluation is based on a selected ADNI subset, and broader validation across additional cohorts is necessary. Second, the method assumes that anatomical concept masks are available or can be generated reliably. Errors in atlas registration or segmentation may affect the concept-guided loss. Third, the reported CAS depends on the saliency threshold, although sensitivity analysis partially addresses this issue. Fourth, few-shot deployment requires labeled calibration scans; zero-shot deployment is possible but may reduce performance. Finally, while spectral embeddings capture scanner-sensitive signals, additional work is needed to ensure that disease-relevant spectral patterns are not inadvertently suppressed.

5. Conclusion

This paper presented a spectral-conditioned hypernetwork framework for robust and transparent multi-site neuroimaging analysis. The proposed framework dynamically generates BN parameters from Fourier-domain spectral embeddings, enabling scanner-aware normalization without explicit site labels. A bi-level meta-learning strategy supports few-shot adaptation to unseen imaging centers by updating only the hypernetwork parameters while preserving encoder stability. A concept-guided regularization term aligns Grad-CAM attributions with clinically meaningful anatomical regions, improving transparency.

Experiments on multi-site AD classification demonstrate that the proposed framework improves diagnostic accuracy, reduces cross-scanner variability, and enhances attribution alignment compared with harmonization, domain adaptation, domain generalization, conditional normalization, and concept-based interpretability baselines. By unifying adaptive normalization, meta-learning, and anatomical concept guidance, the proposed method offers a practical direction for clinically reliable and interpretable AI-assisted neurological diagnosis.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares that there are no competing interests.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

This study is based on de-identified neuroimaging data and does not involve new recruitment of human participants, clinical intervention, or collection of identifiable personal information. Ethics approval and informed consent were handled by the original data providers.

Consent for publication

Not applicable.

Availability of data

Data supporting the findings are available from the corresponding author upon reasonable request, subject to the access conditions and data-use agreements of the original neuroimaging data providers.

Further disclosure

During the preparation of this manuscript, the author used Grammarly and Writefull for the purposes of paraphrasing. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

References

- Chattopadhyay T, Joshy NA, Jagad C, *et al.* Comparison of Explainable AI Models for MRI-Based Alzheimer's Disease Classification. In: *2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM)*. 2024;1-4.
doi: 10.1109/SIPAIM62974.2024.10783631
- Fortin JP, Cullen N, Sheline YI, *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*. 2018;167:104-120.
doi: 10.1016/j.neuroimage.2017.11.024
- Fortin JP, Parker D, Tunc B, *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*. 2017;161:149-170.
doi: 10.1016/j.neuroimage.2017.08.047
- Eshaghzadeh Torbati M, Minhas DS, Ahmad G, *et al.* A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *NeuroImage*. 2021;245:118703.
doi: 10.1016/j.neuroimage.2021.118703
- Ganin Y, Ustinova E, Ajakan H, *et al.* Domain-Adversarial Training of Neural Networks. In: Csurka G, ed. *Domain Adaptation in Computer Vision Applications*. Advances in Computer Vision and Pattern Recognition. Springer International Publishing; 2017:189-209.
doi: 10.1007/978-3-319-58347-1_10
- Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans Biomed Eng*. 2022;69(3):1173-1185.
doi: 10.1109/TBME.2021.3117407
- Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*. 2020;65:101765.
doi: 10.1016/j.media.2020.101765
- Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain Generalization: A Survey. *IEEE Trans Pattern Anal Mach Intell*. Published online 2022;1-20.
doi: 10.1109/TPAMI.2022.3195549
- Yoon JS, Oh K, Shin Y, Mazurowski MA, Suk HI. Domain Generalization for Medical Image Analysis: A Review. *Proceedings of the IEEE*. 2024;112(10):1583-1609.
doi: 10.1109/JPROC.2024.3507831
- Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*. 2015.
doi: 10.48550/arXiv.1502.03167
- Dumoulin V, Shlens J, Kudlur M. A Learned Representation For Artistic Style. *arXiv*. 2016.
doi: 10.48550/arXiv.1610.07629
- Huang X, Belongie S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:1510-1519.
doi: 10.1109/ICCV.2017.167
- Zhang J, Xie Y, Wang Y, Xia Y. Inter-Slice Context Residual Learning for 3D Medical Image Segmentation. *IEEE Trans Med Imaging*. 2021;40(2):661-672.
doi: 10.1109/TMI.2020.3034995
- Ha D, Dai A, Le QV. HyperNetworks. *arXiv*. 2016.
doi: 10.48550/ARXIV.1609.09106
- Li D, Yang Y, Song YZ, Hospedales T. Learning to Generalize: Meta-Learning for Domain Generalization. *AAAI*. 2018;32(1).
doi: 10.1609/aaai.v32i1.11596
- Balaji Y, Sankaranarayanan S, Chellappa R. MetaReg: towards domain generalization using meta-regularization. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Curran Associates Inc.; 2018:1006-1016. Accessed June 3, 2026. <https://dl.acm.org/doi/10.5555/3326943.3327036>
- Sun M, Li Y, Wang Y, Wang X. Towards Domain-Aware Stable

- Meta Learning for Out-of-Distribution Generalization. *ACM Trans Knowl Discov Data*. 2024;18(8):1-24.
doi: 10.1145/3676558
18. Finn C, Abbeel P, Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv*. 2017.
doi: 10.48550/arXiv.1703.03400
19. Serebro D, Riklin-Raviv T. HyDA: Hypernetworks for Test Time Domain Adaptation in Medical Imaging Analysis. In: Gee JC, Alexander DC, Hong J, *et al.*, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer Nature Switzerland; 2026:251-261.
doi: 10.1007/978-3-032-04971-1_24
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis*. 2020;128(2):336-359.
doi: 10.1007/s11263-019-01228-7
21. Nauta M, Jutte A, Provoost J, Seifert C. This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition. In: Kamp M, Koprinska I, Bibal A, *et al.*, eds. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer International Publishing; 2021:441-456.
doi: 10.1007/978-3-030-93736-2_34
22. Koh PW, Nguyen T, Tang YS, *et al.* Concept Bottleneck Models. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR; 2020:5338-5348. Accessed June 3, 2026. <https://proceedings.mlr.press/v119/koh20a.html>
23. Ali H. Advancing Explainable AI for Clinical Decision Support: A Multimodal Evaluation Framework. *Cloud Computing and Data Science*. 2026;7(2):270-289.
doi: 10.37256/ccds.7220269831
24. Mueller SG, Weiner MW, Thal LJ, *et al.* Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*. 2005;1(1):55-66.
doi: 10.1016/j.jalz.2005.06.003
25. Kumari S, Singh P. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*. 2024;170:107912.
doi: 10.1016/j.compbiomed.2023.107912
26. Taiyeb Khosroshahi M, Morsali S, Gharakhanlou S, *et al.* Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Diagnostics*. 2025;15(5):612.
doi: 10.3390/diagnostics15050612
27. Ali H, Akhtar MT. Modality-Specific Sparse Autoencoders for Efficient Multimodal ICU Alignment: A Symmetry-Asymmetry Learning Framework. *Symmetry*. 2026;18(4):677.
doi: 10.3390/sym18040677