

DinoAtten3D: Slice-Level Attention Aggregation of DinoV2 for 3D Brain MRI Anomaly Classification

Fazle Rafsani
Arizona State University
frafsani@asu.edu

Jay Shah
Arizona State University
jgshah1@asu.edu

Catherine D. Chong
Mayo Clinic, Arizona
Chong.Catherine@mayo.edu

Todd J. Schwedt
Mayo Clinic, Arizona
schwedt.todd@mayo.edu

Teresa Wu
Arizona State University
teresa.wu@asu.edu

Abstract

Anomaly detection and classification in medical imaging are critical for early diagnosis but remain challenging due to limited annotated data, class imbalance, and the high cost of expert labeling. Emerging vision foundation models such as DINOv2, pretrained on extensive, unlabeled datasets, offer generalized representations that can potentially alleviate these limitations. In this study, we propose an attention-based global aggregation framework tailored specifically for 3D medical image anomaly classification. Leveraging the self-supervised DINOv2 model as a pretrained feature extractor, our method processes individual 2D axial slices of brain MRIs, assigning adaptive slice-level importance weights through a soft attention mechanism. To further address data scarcity, we employ a composite loss function combining supervised contrastive learning with class-variance regularization, enhancing inter-class separability and intra-class consistency. We validate our framework on the ADNI dataset and an institutional multi-class headache cohort, demonstrating strong anomaly classification performance despite limited data availability and significant class imbalance. Our results highlight the efficacy of utilizing pretrained 2D foundation models combined with attention-based slice aggregation for robust volumetric anomaly detection in medical imaging. Our implementation is publicly available at <https://github.com/Rafsani/DinoAtten3D.git>.

1. Introduction

Anomaly detection and classification in medical imaging pose significant challenges due to data scarcity and the high cost associated with obtaining expert annotations. Traditionally, supervised methods have been widely adopted, but these approaches

demand extensive labeled datasets and are susceptible to overfitting, especially in scenarios characterized by class imbalance and limited sample sizes. Consequently, unsupervised methods have emerged as alternative solutions, employing reconstruction-based frameworks utilizing generative adversarial networks (e.g., HealthyGAN [30], Brainomaly [34], f-AnoGAN [32]) or diffusion models (e.g., AnoDPM [39], AnoFPDM [6]). These unsupervised methods reconstruct healthy versions of potentially anomalous images and detect anomalies through deviations, yet they may also suffer from overfitting due to limited healthy training data, hindering generalization and performance [36, 41].

The recent rise of large-scale foundation models offers a potential remedy to these fundamental limitations. Models such as GPT [28], CLIP [29], and DINOv2 [26] achieve remarkable performance by extensively pretraining on large and heterogeneous datasets, thus effectively internalizing broad statistical regularities across language and vision modalities [37]. These foundation models exhibit emergent in-context learning abilities, enabling effective zero- or few-shot learning by leveraging generalized latent representations [2, 20].

In medical imaging, multimodal foundation models rely on paired image-text data. However, producing high-quality clinical captions is often impractical and laborious. This has motivated the development of vision-only models pretrained exclusively on medical imaging data. DINOv2 exemplifies this paradigm, demonstrating robust generalization as a feature extractor across diverse vision tasks, including medical imaging data [1].

Despite these advantages, foundation models like DINOv2 are inherently designed for 2D image processing and cannot natively handle volumetric medical images such as 3D MRI scans that are ubiquitous in clinical practice. Slice-based methods have

therefore been proposed, independently processing 2D slices from volumetric data. Additionally, multi-instance learning (MIL) approaches, traditionally applied to high-dimensional data like whole-slide images (WSIs), segment large images into smaller instances or patches [4, 14, 40]. While effective in identifying local discriminative regions, traditional MIL methods may overlook broader spatially distributed pathological features across multiple slices that are essential for accurate diagnosis [21, 40].

To address these limitations, we propose a DINOv2-based soft attention-driven global aggregation approach tailored specifically for 3D medical imaging (**DinoAtten3D**) that overcomes the inherent dimensionality constraints of 2D foundation models. It leverages DINOv2’s rich embeddings from axial slices of 3D MRI volumes and introduces a soft attention mechanism that adaptively weighs whole-slice (instead of patch) embeddings based on their diagnostic relevance. By emphasizing diagnostically significant slices, our approach efficiently captures both focal and distributed pathological patterns, bridging the gap between the inherent 2D capabilities of DINOv2 and the volumetric nature of medical imaging, without relying on computationally intensive 3D models that are often impractical for clinical deployment. In summary, our main contributions include:

- We propose a global attention-based aggregation framework for 3D medical imaging that adaptively fuses slice-level embeddings from all 2D slices via soft attention pooling. This highlights the most informative slices while retaining distributed pathological cues, enabling volumetric classification using lightweight 2D backbones such as DINOv2, without the computational overhead of fully 3D architectures.
- We validate our approach on two real-world clinical cohorts: Alzheimer’s Disease Neuroimaging Initiative (ADNI) MRI and a multi-class headache cohort, showing strong anomaly classification performance even under severe data scarcity and class imbalance.
- Beyond differentiating unhealthy from healthy samples, our framework also achieves promising results in distinguishing between different pathological subtypes, demonstrating its broader utility for downstream clinical analyses.

2. Related Works

Anomaly detection in medical imaging has garnered significant research attention, leading to a wide range of methodologies for identifying pathological deviations [35]. Early work focused primarily on supervised learning approaches, particularly Convolutional Neural Networks (CNNs), which are trained

using annotated datasets to learn discriminative representations of anomalies [10, 12, 15, 16, 18, 23, 27, 31, 33]. While CNN-based models have achieved success in both classification and segmentation tasks, their reliance on large, expert-labeled datasets poses significant challenges given the high cost and labor intensity of medical image annotation.

To alleviate the annotation bottleneck, unsupervised and self-supervised approaches have emerged. For example, Reconstruction-based anomaly detection methods leverage healthy images to learn normative distributions, flagging deviations at inference as potential anomalies. Generative adversarial networks (GANs), such as f-AnoGAN [32], Brainomaly [34], and HealthyGAN [30], generate healthy counterfactuals of test images to detect anomalies. More recently, diffusion-based methods [6, 11, 24, 39] have gained traction due to superior generative performance; however, these models still require substantial amounts of healthy training data, and efforts are still being made to address this issue with these models [9, 17, 38].

Medical imaging data are often high-dimensional and can possess complex three-dimensional structures, as seen in modalities such as MRI and CT. Whole Slide Images (WSIs), commonly used in cancer diagnosis, are particularly large and high-resolution. To efficiently process these data, patch-based training strategies and instance-based learning methods are frequently employed [7, 21]. In this context, multi-instance learning (MIL) has become a popular paradigm, especially in digital pathology and large-scale medical image analysis [4, 14, 21, 40]. Traditional MIL approaches divide images into patches or instances and use aggregation strategies such as max-pooling, mean-pooling, or attention pooling to generate bag-level predictions from instance-level features [5]. Attention-based MIL [14] is particularly effective, as it enables the model to learn which regions are most informative for downstream tasks. These methods, however, often focus on local or spatially sparse regions (e.g., tumor), which may not capture global or distributed pathological patterns, especially in volumetric imaging where relevant features may be distributed across multiple slices [8]. And, later, is particularly true to some neurodegenerative diseases such as Alzheimer’s and headaches.

To address class imbalance in anomaly detection and enhance representation learning in MIL frameworks, recent works such as SC-MIL (Supervised Contrastive Multiple Instance Learning) [19] have introduced supervised contrastive loss to promote more discriminative and robust bag-level embeddings. SC-MIL is particularly relevant as it addresses the challenge of imbalanced classification, a com-

mon issue in medical imaging, and forms a strong baseline for our work due to its close methodological proximity to our approach, differing mainly in its use of local instance aggregation compared to our proposed global attention-based aggregation strategy.

Amidst these developments, foundation models have demonstrated remarkable generalizability across diverse vision and language tasks by leveraging large-scale pretraining [42]. However, developing foundation models specifically for medical imaging remains challenging due to data scarcity and the cost of collecting paired image-text data. Vision-language foundation models such as BioMedCLIP [43] and MedSAM [22] leverage contrastive learning and prompts. Researchers are increasingly exploring vision-only models pretrained on natural images for medical imaging tasks [13]. Recent work has shown that DINOv2, a self-supervised vision transformer pretrained on large natural image datasets, provides robust and transferable feature representations, outperforming other pre-trained models in medical image classification [1, 26]. While DINOv2 is inherently 2D, its strong feature extraction capabilities make it an attractive choice for 3D medical imaging when paired with effective aggregation strategies.

In summary, our approach is motivated by (1) the proven effectiveness of attention pooling for large-scale and high-dimensional medical image analysis, (2) recent advances in supervised contrastive learning for imbalanced classification as exemplified by SC-MIL, and (3) the demonstrated generalizability of foundation models such as DINOv2 as feature extractors. Our approach contrasts with conventional local/instance-level MIL by proposing a global, attention-driven aggregation of slice-level DINOv2 embeddings that is specifically designed to address the different challenges of 3D medical images like brain MRI.

3. Method

In this section, we present the training procedure of the model and the architecture of the attention pooling-based 3D brain MRI classification task.

3.1. The Model Architecture: DinoAtten3D

The DinoAtten3D comprises a pre-trained feature extractor, an attention-pooled weighted aggregation block, and an MLP (Multi-Layer Perceptron) on top of the attention block. The full overview of the method is presented in Figure 1. The foundation model, DinoV2, trained through a self-supervised process, is used on 2D slices of the MRI volume to extract rich embeddings for each slice. Later, these 2D slice embeddings are given an attention score, further explained in section 3.1.2. The cu-

mulative weighted embeddings are then used to train a classifier. The training process employs a custom loss function that combines cross-entropy loss, contrastive loss, and a class variance loss (see section 3.2) to effectively learn from a relatively small dataset. While cross-entropy loss guides the model to correctly classify samples based on ground truth labels, it alone may not be sufficient to capture subtle intra-class variations or enhance feature separability, particularly in low-data regimes. To address this, contrastive loss is integrated to encourage the model to learn discriminative representations by pulling together embeddings of samples from the same class and pushing apart those from different classes in the feature space. Additionally, the class variance loss is introduced to minimize the intra-class variance, ensuring that embeddings of samples within the same class remain compact.

3.1.1. 2D Feature Extraction via Frozen DinoV2

Let a 3D MRI volume be represented by an ordered set of N axial slices, where \mathcal{S} represents the set of slices for each 3D MRI volume.

$$\mathcal{S} = \{S_j \in \mathbb{R}^{C \times H \times W} \mid j = 1, \dots, N\}.$$

Each slice S_j is passed through the pretrained, frozen DinoV2 backbone

$$f_{\text{Dino}} : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^d,$$

where $d = 384$, producing per-slice embeddings. In our work, we utilize the ViT-based backbone of DINOv2 (often denoted ViT-S/14), which tokenizes each input slice into a sequence of 14×14 patches and projects them into a 384-dimensional latent space. So

$$\mathbf{z}_j = f_{\text{Dino}}(S_j) \in \mathbb{R}^d, \quad j = 1, \dots, N.$$

3.1.2. Slice-Level Global Attention Aggregation

To aggregate the N slice embeddings into a single volume-level feature, we learn scalar attention scores with a two-layer MLP:

$$e_j = \mathbf{w}_2^\top \tanh(\mathbf{W}_1 \mathbf{z}_j)$$

with $\mathbf{W}_1 \in \mathbb{R}^{h \times d}$, $\mathbf{w}_2 \in \mathbb{R}^h$. Attention weights are then obtained via softmax:

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^N \exp(e_k)}$$

and the aggregated feature is:

$$\mathbf{z}_{\text{agg}} = \sum_{j=1}^N \alpha_j \mathbf{z}_j.$$

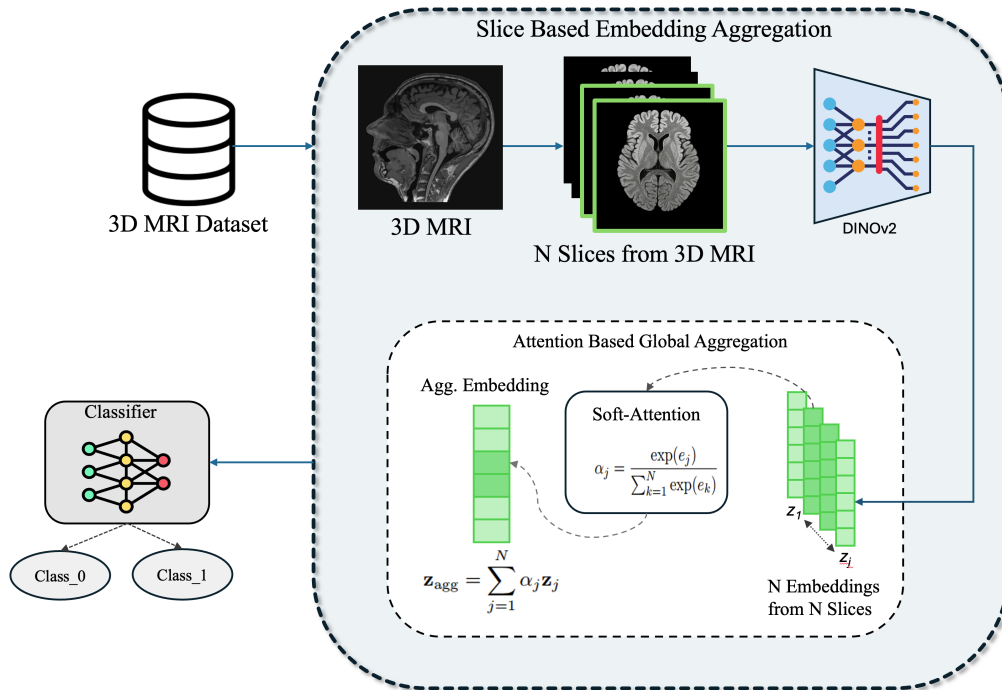


Figure 1. Overall architecture of the slice-based attention aggregation of 2D slice embeddings for 3D brain MRI

3.1.3. Embedding Head and Classification

The aggregated embedding \mathbf{z}_{agg} is mapped to a lower-dimensional embedding via a two-layer MLP. Finally, a linear classifier produces logits for the binary classification.

3.2. Training Objective

Let $\tilde{\mathbf{h}}^{(i)} = \mathbf{h}^{(i)} / \|\mathbf{h}^{(i)}\|_2$ be the normalized embedding for sample i in a batch of size B , and let $y^{(i)}$ be its class label. The overall loss combines three terms: a standard cross-entropy loss on classifier logits, a supervised contrastive loss on embeddings, and a within-class variance regularization.

Cross-Entropy Loss: Given predicted logits $\mathbf{o}^{(i)} \in \mathbb{R}^C$, the cross-entropy loss is:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{i=1}^B \log(\text{softmax}(\mathbf{o}^{(i)})_{y^{(i)}}).$$

Contrastive Loss: We compute pairwise similarities between normalized embeddings:

$$s_{ij} = \frac{1}{\tau} \tilde{\mathbf{h}}^{(i)} \top \tilde{\mathbf{h}}^{(j)},$$

where $\tau > 0$ is a temperature parameter. Let $\mathcal{P}(i)$ be the set of indices sharing the same label as sample i . The contrastive loss is:

$$\mathcal{L}_{\text{contra}} = \frac{1}{B} \sum_{i=1}^B \left[-\frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(s_{ij})}{\sum_{k \neq i} \exp(s_{ik})} \right].$$

Within-Class Variance Loss: To encourage compact clustering of same-class embeddings, we define for each class c the centroid:

$$\bar{\mathbf{h}}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \tilde{\mathbf{h}}^{(i)},$$

where \mathcal{I}_c contains the indices of samples with label c . The variance loss is:

$$\mathcal{L}_{\text{var}} = \frac{1}{C} \sum_{c=1}^C \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \left\| \tilde{\mathbf{h}}^{(i)} - \bar{\mathbf{h}}_c \right\|_2^2.$$

Total Loss: The total objective combines the three components:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{contra}} + \lambda \mathcal{L}_{\text{var}},$$

where we empirically set $\tau = 0.07$ and $\lambda = 0.1$ in our experiments.

4. Datasets

We examined our method on two brain MRI datasets: the ADNI dataset and an in-house dataset of headache patients. Both datasets consist of 3D brain MRIs.

4.1. Alzheimer's Disease Neuroimaging Initiative Dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu) is a longitudinal, multi-center observational study launched in 2004 to

develop, standardize, and validate biomarkers for Alzheimer’s disease (AD) clinical trials. ADNI-1 initially enrolled 200 cognitively normal healthy controls (HC), 400 participants with mild cognitive impairment (MCI), and 200 with AD across 57 sites in the United States and Canada; later phases (ADNI-GO, ADNI-2 and ADNI-3) expanded the total cohort to over 1,000 subjects aged 55–90. Participants underwent serial neuroimaging (structural MRI T1-weighted MP-RAGE on 1.5T and 3T scanners; FDG-PET and amyloid PET tracers), biofluid collection (cerebrospinal fluid (CSF) and plasma biomarkers), genetic profiling (e.g., APOE genotyping), and comprehensive neuropsychological assessments (mini mental status exam (MMSE), Clinical Dementia Rating (CDR), Alzheimer’s Disease Assessment Scale – Cognitive Subscale (ADAS-Cog)) at baseline and follow-up visits (6, 12, 18, 24 months in ADNI-1; extended in subsequent phases). For our analyses, all T1-weighted MRI scans were preprocessed using non-linear registration to the MNI-152 template, N4 bias-field correction, skull-stripping, and intensity normalization via histogram matching. The dataset consists of 3 classes of images: healthy (HC), MCI, and AD. HC participants are participants with no subjective or objective memory complaints and normal performance on neuropsychological tests. Mild Cognitive Impairment (MCI) is the prodromal, intermediate stage between healthy aging and AD, characterized by subjective memory complaints, objective memory impairment (MMSE 24–30), CDR = 0.5, preserved activities of daily living, and absence of dementia. In our final processed dataset, we have 4769 T1-weighted brain MRI scans: 1831 HC, 1668 MCI, 1270 AD.

4.2. Institutional Headache Dataset

Headache data were collected via prospective research approved by the Institutional Review Board (IRB), and all participants provided written informed consent for their participation. At the time of enrollment, migraine participants were diagnosed with episodic or chronic migraine, with or without aura, based on the most recent edition of the International Classification of Headache Disorders (ICHD-3 beta or ICHD-3) [25]. Participants with acute post-traumatic headache (APTH) or persistent post-traumatic headache (PPTH) had PTH attributed to mild traumatic brain injury (mTBI) according to the latest ICHD criteria (ICHD-3 beta or ICHD-3). Individuals with a history of moderate or severe traumatic brain injury were excluded from the study. We collected MRIs of 96 individuals with migraine, 48 with APTH, 49 with PPTH, and 104 healthy controls from the institution. We extended our dataset by including MRIs of 428 healthy controls from the publicly available IXI dataset [3]. For our experiments,

we trained our model by first combining all headache types into one group and then investigated each subgroup’s performance separately. All 3D MRIs in this dataset were registered to the MNI152 1mm template and skull stripped.

5. Experiments and Results

We evaluated our slice-wise soft attention aggregation of DinoV2 features on two different tasks: Alzheimer’s disease detection using the ADNI dataset and headache detection on our private headache dataset. For the ADNI dataset, we performed three pairwise anomaly or disease detection tasks: HC vs. AD, HC vs. MCI, and MCI vs. AD. For each task, we split the data with corresponding classes in an 80:10:10 ratio for training, validation, and testing.

For the headache dataset, containing HC, migraine (MIG), APTH, and PPTH. We performed the experiments in 2 different settings. Firstly, we performed experiments with HC vs. different headache types to evaluate the performance of headache detection using the brain MRIs. We trained four separate models to evaluate the method with four different scenarios: HC vs. all headache (considering all headache types: migraine, APTH, PPTH as one class), HC vs. Mig, HC vs. APTH, and HC vs. PPTH. Secondly, we perform experiments with a view to differentiating between the subtypes of headache: Mig. vs. APTH. Mig. vs PPTH and APTH vs PPTH. For HC vs. all headaches, we split the dataset with an 80:10:10 ratio for training, validation, and testing. For the HC vs. subtype scenarios, we put 10 samples from each class in the validation set to avoid bias during evaluation and performed 5-fold cross-validation, reporting the mean values. The results are shown in section 5.2.

In each case, we report binary classification performance in terms of accuracy, area under the ROC curve (AUC), along with an F_1 score, False Negative Rate (FNR), and we visualize confusion matrices. We also compare the results with two baseline methods: SC-MIL [40] and 3D ResNet to demonstrate the effectiveness of our method over them.

Table 1. Binary classification performance on ADNI.

Task	Method	Accuracy (%)	AUC	F_1	FNR (%)
HC vs. AD	SC-MIL	59.16	0.502	0.031	97.42
	ResNet 3D	84.35	0.854	0.816	23.62
	DinoAtten3D	87.80	0.865	0.871	22.05
HC vs. MCI	SC-MIL	48.29	0.476	0.385	65.87
	ResNet 3D	62.29	0.668	0.420	71.40
	DinoAtten3D	70.50	0.702	0.700	23.95
MCI vs. AD	SC-MIL	56.46	0.514	0.218	85.75
	ResNet 3D	74.35	0.714	0.610	53.54
	DinoAtten3D	75.70	0.749	0.730	29.92

5.1. Alzheimer’s detection on ADNI dataset

Table 1 summarizes results on three pairwise anomaly or disease detection tasks in the ADNI cohort: HC vs. AD, HC vs. MCI, and MCI vs. AD. Figures 2a–2c display the corresponding confusion matrices.

For HC vs. AD, our model achieves 87.80% accuracy with an F_1 score of 0.871 in the test set after selecting the model with the lowest validation loss. As shown in Figure 2a, 174/184 HC and 99/127 AD scans are correctly classified with 0.865 AUC, indicating good convergence and limited overfitting in terms of Alzheimer’s disease detection. It also achieves better accuracy and AUC compared to the baselines.

In the HC vs. MCI task, we obtain 70.50% accuracy (AUC 0.702), with an F_1 score of 0.700. The confusion matrix in Figure 2b shows moderate misclassifications (133 true HC vs. 38 false negatives; 89 true MCI vs. 34 false positives), reflecting the subtlety of early-stage MCI detection.

For MCI vs. AD, accuracy reaches 75.70% (AUC 0.749), with an F_1 score of 0.730 for both classes. Figure 2c illustrates that 133 MCI and 89 AD volumes are correctly labeled.

	HC	AD		HC	MCI		MCI	AD
HC	174	10	HC	118	65	MCI	133	34
AD	28	99	MCI	40	127	AD	38	89
	(a)		(b)			(c)		

Figure 2. Confusion matrix for ADNI dataset experiments: (a) HC vs AD, (b) HC vs MCI, (c) MCI vs AD

As illustrated in Figures 3a–3c, the t-SNE projections of the aggregated patient embeddings for each ADNI classification task exhibit markedly different clustering behaviors. In the HC vs AD comparison (Figure 3a), the two cohorts form two well-demarcated clusters, indicating a high degree of separability in the learned representation space. By contrast, the MCI vs AD (Figure 3b) and HC vs MCI (Figure 3c) tasks display only moderate separation: while the embeddings tend to form cluster-like structures, there remains a non-negligible degree of overlap between classes. This suggests that, although the model captures discriminative features in all three scenarios, the boundary between mild cognitive impairment and either healthy controls or Alzheimer’s patients is less distinct than that between healthy and Alzheimer’s subjects.

5.2. Headache classification on private dataset

We further tested our method on a private headache MRI dataset for headache detection and inter-headache classification. For the HC vs all headache

Table 2. Binary classification performance on the headache dataset.

Task	Method	Accuracy (%)	AUC	F_1	FNR (%)
HC vs. Headache	SC-MIL	67.67	0.565	0.095	40.00
	ResNet3D	82.18	0.806	0.705	35.00
	DinoAtten3D	86.30	0.874	0.867	10.00
HC vs. MIG	SC-MIL	50.00	0.544	0.000	100.0
	ResNet3D	74.05	0.870	0.696	40.00
	DinoAtten3D	90.00	0.992	0.899	15.00
HC vs. APTH	SC-MIL	50.00	0.505	0.000	100.0
	ResNet3D	62.00	0.850	0.385	76.00
	DinoAtten3D	85.00	0.970	0.846	30.00
HC vs. PPTH	SC-MIL	50.00	0.570	0.000	100.0
	ResNet3D	65.00	0.828	0.565	54.10
	DinoAtten3D	90.00	0.980	0.899	20.00

scenario, the model achieved 86.30% accuracy in the blind test set with 0.874 AUC. In the case of migraine detection, the model achieved 90.00% accuracy with 0.993 AUC. For APTH and PPTH detection with respect to HC, the model also achieved 85.00% accuracy with 0.970 AUC and 90.00% accuracy with 0.980 AUC, respectively. Table 2 and Figures 4a–4d report the results and confusion matrix for headache detection.

To gauge our model’s ability to detect anomalies across different headache subtypes, we further trained it on three binary tasks: Migraine (Mig) vs. Acute Post-Traumatic Headache (APTH), Persistent Post-Traumatic Headache (PPTH) vs. APTH, and Mig vs. PPTH. Table 3 reports, for each task, the validation accuracy, F_1 -score, and AUC. For Mig. vs APTH and APTH vs PPTH, the model achieved high accuracy of 90% and 95% with almost perfect AUC scores. However, for the Mig vs PPTH classification, the model achieved 55% accuracy, close to the best-performing baseline method in terms of accuracy and AUC.

Table 3. Performance of DinoAtten for inter-headache classification.

Task	Method	Accuracy (%)	AUC	F_1
MIG vs. APTH	SC-MIL	50.00	0.388	0.667
	ResNet3D	78.01	0.920	0.798
	DinoAtten3D	90.00	0.980	0.890
APTH vs. PPTH	SC-MIL	50.00	0.540	0.667
	ResNet3D	84.00	0.989	0.802
	DinoAtten3D	95.00	0.991	0.949
MIG vs. PPTH	SC-MIL	50.00	0.520	0.000
	ResNet3D	58.00	0.562	0.698
	DinoAtten3D	55.00	0.680	0.436

Figures 5a–5d depict the t-SNE projections of the aggregated embeddings for test subjects in the headache dataset experiments. The results reveal well-defined clusters corresponding to healthy controls versus the various headache subtypes. In particular, the HC versus Migraine, HC versus PPTH, and HC versus all-headache comparisons exhibit pronounced separation in embedding space, with less inter-cluster overlap. While the HC versus APTH

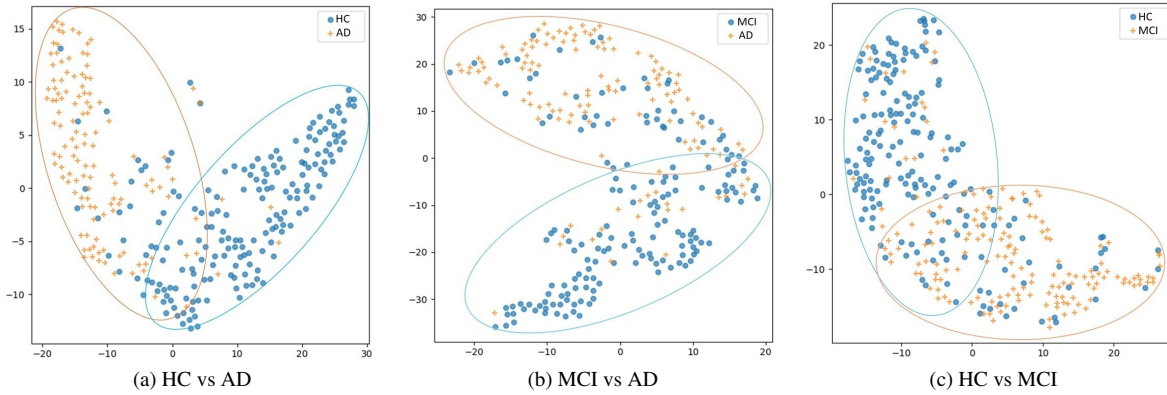


Figure 3. t-SNE plots of Aggregated embeddings for ADNI dataset test subjects

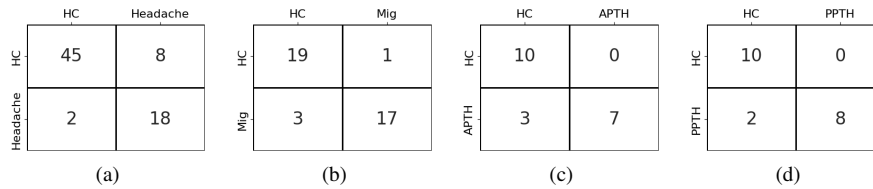


Figure 4. Confusion matrices for the four headache detection tasks. (a) HC vs All Headache, (b) HC vs Mig, (c) HC vs APTH, (d) HC vs PPTH

plot shows a modest degree of intermingling, the two groups remain largely distinguishable. These observations corroborate the efficacy of incorporating supervised contrastive learning alongside class-variance regularization during model training, as they yield representations that enhance class discriminability across headache detection tasks.

6. Discussion

Across the binary tasks, our attention-weighted DinoV2 embeddings exhibit consistently strong anomaly detection when the classes are clearly distinct (i.e., AD vs. HC, all headache types vs HC, Mig vs APTH, APTH vs PPTH), but struggle when they share similar features (e.g., PPTH vs Mig). In the Mig vs APTH experiment, the model achieved 90% accuracy, a 0.89 F_1 -score, and a 0.98 AUC, demonstrating robust separation of acute post-traumatic headache as the anomalous class. Performance improved slightly in the PPTH vs APTH comparison (95% accuracy, 0.95 F_1 -score, 0.99 AUC), indicating almost perfect discrimination of persistent post-traumatic headache against acute cases. By contrast, the Migraine vs PPTH task yielded only 55% accuracy, a 0.44 F_1 -score, and a 0.68 AUC, reflecting the model’s difficulty in distinguishing post-traumatic profiles with Migraine. This finding reflects the clinical observation that symptoms of Mig and PPTH are typically very similar, although the symptoms of PPTH are triggered by a brain injury, whereas those of migraine are not. These results demonstrate that our aggregation strategy excels at detecting anoma-

lies when the target condition is well-defined, yet also highlight the need for further feature refinement in cases of subtly differing health conditions. In the future, we plan to comprehensively assess our approach in more challenging scenarios where data scarcity and class imbalance are more prevalent.

7. Conclusion

In this study, we present a slice-level attention aggregation framework built upon DinoV2, a self-supervised vision transformer pretrained predominantly on natural images. Despite the domain shift, our method demonstrates strong performance across both neuroimaging and headache classification tasks. Notably, it achieves high accuracy in distinguishing healthy from pathological cases and shows competitive performance even in challenging inter-subtype classifications. The results on the ADNI dataset underscore the model’s capacity to reliably detect Alzheimer’s disease, while the promising outcomes in differentiating among headache subtypes (e.g., migraine vs. post-traumatic headache) highlight the discriminative power of the learned representations. Our findings suggest that the rich, generalized features extracted by DinoV2 are transferable to medical imaging contexts, enabling robust anomaly classification even in data-scarce scenarios. The effectiveness of our slice-level soft attention mechanism further validates the importance of localized features in volumetric medical data. Future work will explore leveraging these features for more granular multi-class disease sub-typing and investigating domain-

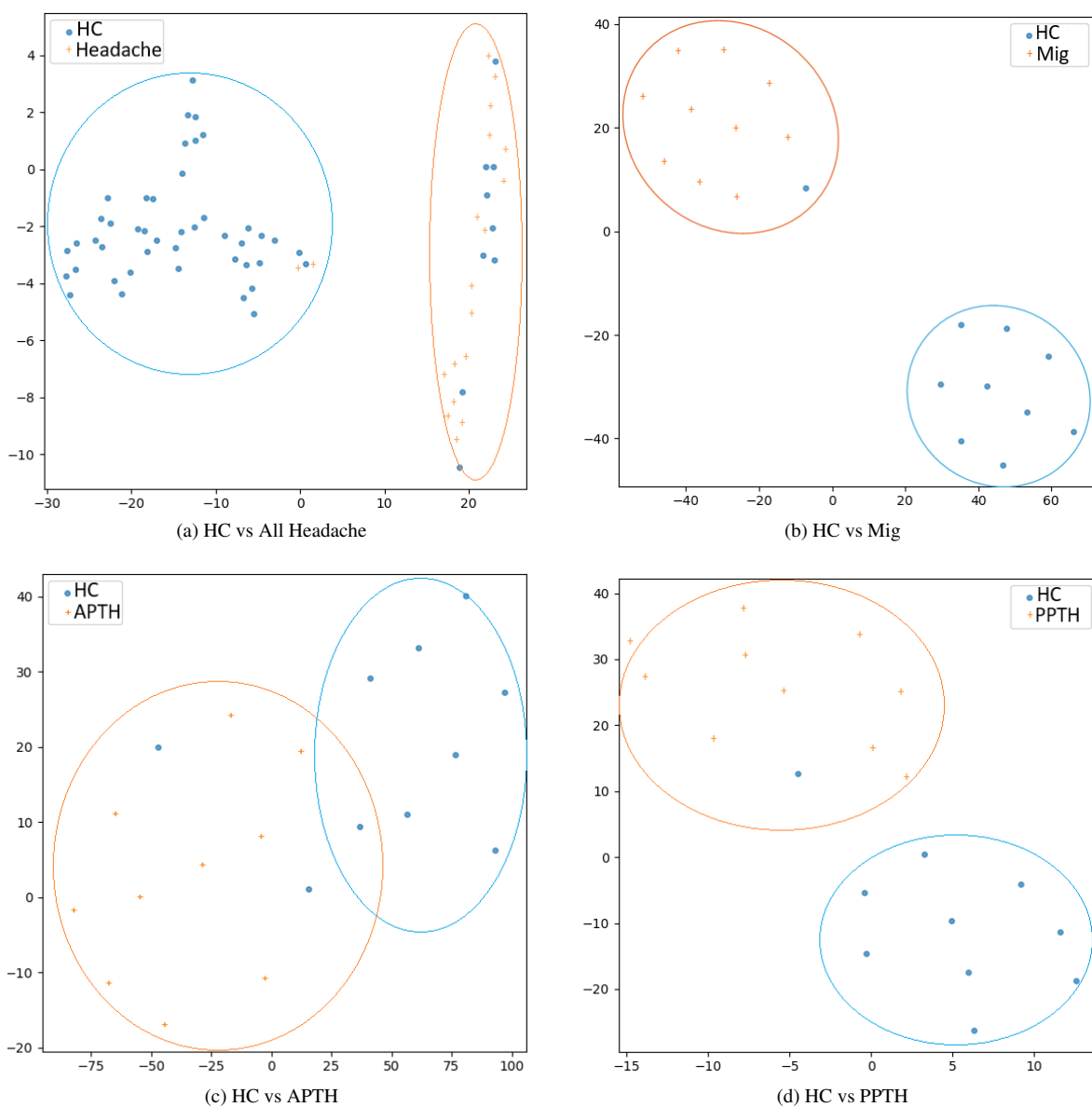


Figure 5. t-SNE plots of aggregated embeddings for Headache dataset test subjects.

adaptive pretraining strategies to further enhance performance in specialized clinical applications.

Acknowledgment: This work was supported by the United States Department of Defense W81XWH-15-1-0286 and W81XWH1910534, National Institutes of Health K23NS070891, National Institutes of Health—National Institute of Neurological Disorders and Stroke, Award Number 1R61NS113315-01, and Amgen Investigator Sponsored Study 20187183.

References

- [1] Mohammed Baharoon, Waseem Qureshi, Jiahong Ouyang, Yanwu Xu, Abdulrhman Aljouie, and Wei Peng. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. *arXiv preprint arXiv:2312.02366*, 2023. 1, 3
- [2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1
- [3] Brain Development. Ixi dataset. <https://brain-development.org/ixi-dataset/>. 5
- [4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2
- [5] Marc-Andre Carbonneau, Veronika Cheplygina, Eric

- Granger, and Guillaume Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 2
- [6] Yiming Che, Fazle Rafsani, Jay Shah, Md Mahfuzur Rahman Siddiquee, and Teresa Wu. Anofpdm: Anomaly detection with forward process of diffusion models for brain mri. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1113–1122, 2025. 1, 2
- [7] Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):8894, 2021. 2
- [8] Leonardo Crespi, Daniele Loiacono, and Pierandrea Sartori. Are 3d better than 2d convolutional neural networks for medical imaging semantic segmentation? In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 2
- [9] Giannis Daras, Yuyang Tang, Shiyu Tang, et al. How much is a noisy image worth? data scaling laws for ambient diffusion. *arXiv preprint arXiv:2401.03196*, 2024. 2
- [10] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in biology and medicine*, 111:103345, 2019. 2
- [11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2
- [12] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017. 2
- [13] Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7634–7643, 2024. 3
- [14] Max Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2
- [15] Sajid Iqbal, M Usman Ghani, Tanzila Saba, and Amjad Rehman. Brain tumor segmentation in multi-spectral mri using convolutional neural networks (cnn). *Microscopy research and technique*, 81(4): 419–427, 2018. 2
- [16] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*, 2017. 2
- [17] Haonan Li, Tete Yu, Xinyang Zhang, et al. Pruning then reweighting: Towards data-efficient training of diffusion models. *arXiv preprint arXiv:2401.04224*, 2024. 2
- [18] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014. 2
- [19] Xiang Li, Ming Xu, Bing Wang, Licheng Wang, Kaicheng Ma, and Shaoting Zhang. Sc-mil: Supervised contrastive multiple instance learning for imbalanced classification in pathology. In *International Conference on Computer Vision (ICCV)*, pages 14971–14981, 2021. 2
- [20] Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou, and Delong Chen. Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, 57(10):268, 2024. 1
- [21] Ming Y Lu, Drew F K Williamson, Tiffany Y Chen, Ronald J Chen, Milena Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 2
- [22] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3
- [23] Aliasghar Mortazi and Ulas Bagci. Automatically designing cnn architectures for medical image segmentation. In *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*, pages 98–106. Springer, 2018. 2
- [24] Hanna Mykula, Lisa Gasser, Silvia Lobmaier, Julia A Schnabel, Veronika Zimmer, and Cosmin I Bercea. Diffusion models for unsupervised anomaly detection in fetal brain ultrasound. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 220–230. Springer, 2024. 2
- [25] Jes Olesen. The international classification of headache disorders. *Headache: The Journal of Head and Face Pain*, 48(5):691–693, 2008. 5
- [26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3
- [27] Plabon Paul, Md Nazmul Islam, Fazle Rafsani, Pegah Khorasani, and Shovito Barua Soumma. Efficient feature extraction and classification architecture for mri-based brain tumor detection. *arXiv preprint arXiv:2410.22619*, 2024. 2
- [28] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

- [30] Md Mahfuzur Rahman Siddiquee, Jay Shah, Teresa Wu, Catherine Chong, Todd Schwedt, and Baoxin Li. Healthygan: Learning from unannotated medical images to detect anomalies associated with human disease. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 43–54. Springer, 2022. 1, 2
- [31] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahim Alabdullah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023. 2
- [32] Thomas Schlegl, Philipp Seebock, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2:2, 2017. 1, 2
- [33] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. 2
- [34] Md Mahfuzur Rahman Siddiquee, Jay Shah, Teresa Wu, Catherine Chong, Todd J Schwedt, Gina Dumkrieger, Simona Nikolova, and Baoxin Li. Brainomaly: Unsupervised neurologic disease detection utilizing unannotated t1-weighted brain mr images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7573–7582, 2024. 1, 2
- [35] Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging—a mini review. In *International Data Science Conference*, pages 33–38. Springer, 2021. 2
- [36] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022. 1
- [37] Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025. 1
- [38] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36:72137–72154, 2023. 2
- [39] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 650–656, 2022. 1, 2
- [40] Weilin Yan, Haoran Song, Ming Xu, Bing Wang, Licheng Wang, Kaicheng Ma, and Shaoting Zhang. Mil3d: Multi-instance learning for 3d medical image classification. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. 2, 5
- [41] Haibo Zhang, Wenping Guo, Shiqing Zhang, Hongsheng Lu, and Xiaoming Zhao. Unsupervised deep anomaly detection for medical images using an improved adversarial autoencoder. *Journal of Digital Imaging*, 35(2):153–161, 2022. 1
- [42] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024. 3
- [43] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 3