A Multi-Dataset Benchmark of Multiple Instance Learning for 3D Neuroimage Classification

Ethan Harvey[1*] ETHAN.HARVEY@TUFTS.EDU
Dennis Johan Loevlie[1*] DENNIS.LOEVLIE@TUFTS.EDU
Amir Ali Satani[2] AMIR.SATANI@TUFTS.EDU
Wansu Chen[3] WANSU.CHEN@KP.ORG
David M. Kent[2] DAVID.KENT@TUFTSMEDICINE.ORG
Michael C. Hughes[1] MICHAEL.HUGHES@TUFTS.EDU
[1] *Department of Computer Science, Tufts University, Medford, MA, USA*
[2] *Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston, MA, USA*
[3] *Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, USA*

Abstract

Despite being resource-intensive to train, 3D convolutional neural networks (CNNs) have been the standard approach to classify CT and MRI scans. Recent work suggests that deep *multiple instance learning* (MIL) may be a more efficient alternative for 3D brain scans, especially when the pre-trained image encoder used to embed each 2D slice is frozen and only the pooling operation and classifier are trained. In this paper, we systematically compare simple MIL, attention-based MIL, 3D CNNs, and 3D ViTs across three CT and four MRI datasets, including two large datasets of at least 10,000 scans. Our goal is to help resource-constrained practitioners understand which neural networks work well for 3D neuroimages and why. We further compare design choices for MIL, including different encoders, pooling operations, and architectural orderings. We find that simple mean pooling MIL, without any learnable attention, matches or outperforms recent MIL or 3D NN alternatives on 4 of 6 moderate-sized tasks. This baseline remains competitive on two large datasets while being 25x faster to train. To explain mean pooling's success, we examine per-slice attention quality and a semi-synthetic dataset where we can derive the best possible classifier via a Bayes estimator. This analysis reveals the limits of existing MIL approaches and suggests routes for future improvements.

Data and Code Availability. We use open datasets like ADNI1 (Mueller et al., 2005), OASIS-3 (LaMontagne et al., 2019), and RSNA (Flanders et al., 2020), as well as non-shareable Proprietary data (Sec. 4). Our code is publicly available: github.com/tufts-ml/neuroimage-classifiers.

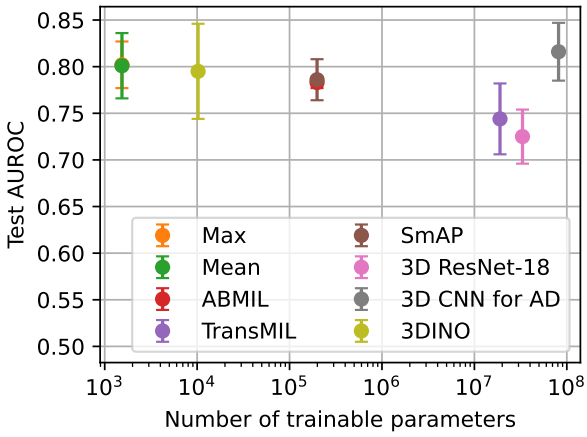

Figure 1: Test AUROC on OASIS-3 MRI (higher is better) vs. number of trainable parameters (lower is better) for MIL and 3D NN methods. **Takeaway: Mean pooling MIL is almost as good and far more efficient than the best MIL or 3D NN method.** All MIL methods use the same pre-trained ViT encoder.

Institutional Review Board (IRB). Our study of Proprietary datasets of deidentified neuroimages with associated diagnostic labels was approved by our local Institutional Review Board (Tufts Health Science IRB #3977 and #4374). Other data are open-access and deidentified; no approval is needed.

[*] Equal contribution.

1. Introduction

We investigate the problem of predicting a single binary label given a 3D image of the brain. We focus on two common imaging modalities: computed tomography (CT) and magnetic resonance imaging (MRI). Such images provide high-quality clinical evidence for critical neurological conditions such as Alzheimer's or brain lesions. Our work is especially motivated by the potential for effective detection of small vessel diseases like covert brain infarction or white matter disease. Reliable automatic detection on routine scans may improve long-term risk assessment for stroke, dementia, and other serious consequences (Pasi and Cordonnier, 2020).

Deep neural networks have become widespread for neuroimaging classification tasks. One approach allows the network direct access to an input 3D volume (Wen et al., 2020). Another increasingly common approach is *multiple instance learning* (MIL). Here, the 3D volume is divided into slices. Each slice or "instance" is processed separately by an encoder, then the MIL architecture pools instance-level results into a whole-scan prediction. We consider here both simplistic pooling strategies as well as recent pooling methods like TransMIL (Shao et al., 2021) and SmAP (Castro-Macías et al., 2024) that account for interaction between instances.

Two aspects of MIL make it appealing for neuroimage classification. First, current MIL processing broadly parallels how human clinical experts make classification judgments from 3D images. Experts today page through axial slices to find a specific slice or set of slices with visual evidence for an abnormality, synthesizing information across slices with clinical context. The ability to find one "smoking gun" instance to justify a positive classification of the entire scan is a fundamental assumption of MIL (Dietterich et al., 1997; Raff and Holt, 2023). Second, MIL neural networks can be more *runtime and storage efficient* for training and evaluation than 3D neural networks, especially if using a pre-trained encoder.

We make several contributions to improve understanding of the strengths and weaknesses of MIL for 3D brain scan classification:

- We provide a systematic study of MIL architectural approaches, encoders, and pooling operations on 9 classification tasks using 7 datasets of MRI or CT scans (Tab. 2). This multi-dataset, apples-to-apples comparison is missing in earlier work trying MIL for brain imaging (Castro-Macías et al., 2024), which compare advanced MIL methods but not 3D neural networks or simple baselines like mean pooling on one relatively small CT dataset.
- We provide evidence across multiple MRI and CT datasets that a simple MIL baseline using mean pooling, rather than learned attention, can often match or outperform recent MIL methods published in top conferences, despite the substantial additional complexity of learned attention. Even on larger datasets with 10,000+ scans, we find recent variants of attention-based MIL like TransMIL (Shao et al., 2021) or SmAP (Castro-Macías et al., 2024) never provide an absolute gain in AUROC greater than 0.025 (see Tab. 5), while requiring 25x longer training times. We thus argue for including mean pooling MIL as a strong baseline in future work.
- We directly examine the quality of learned attention on real data. The RSNA CT dataset (Flanders et al., 2020) uniquely provides instance-level (per-slice) labels for lesion presence/absence, enabling evaluation of how well learned attention predicts instance-level labels. We compare ABMIL, TransMIL, and SmAP to a simple center-focused Gaussian baseline that *ignores the image entirely* and allocates attention solely by slice position. Surprisingly, on this large dataset no learned attention method outperforms this trivial baseline on attention correctness, AUROC, or AUPRC (see Tab. 6).
- To better explain the lack of strong gains from recent MIL methods designed to account for instance interaction, in Sec. 6 we create a semi-synthetic dataset intended to match the statistics of slice-level labels in the RSNA CT dataset. Knowledge of the data-generating process allows us to define the best possible classifier via a Bayes estimator. This analysis reveals that even at large data sizes, recent MIL approaches score substantially worse than the best possible classifier.

Altogether, our work suggests the potential for future methods innovation in the MIL design space and provides a reproducible platform for verifying how such innovations impact overall classifier quality.

2. Background: MIL Methods

MIL (Dietterich et al., 1997; Maron and Lozano-Pérez, 1997; Quellec et al., 2017) is a branch of weakly supervised learning where the goal is to train a predictor that, given a variable-sized set of instances each with its own feature vector, can predict a single binary label for the entire set. The training dataset for a generic MIL problem, denoted $\{(x_i, y_i)\}_{i=1}^{N}$, con-

sists of N labeled bags of data. Each bag is a set of S_i instance feature vectors $x_i = \{x_{i,1}, \ldots, x_{i,S_i}\}$ with a single binary label $y_i \in \{0, 1\}$. For our work on brain scans, each bag is a 3D CT or MRI, and an instance is the 2D image of one axial slice.

In this work, we study three aspects of MIL design for 3D brain scans: architectural approach, encoder choice, and pooling operation. These are summarized in Tab. 1 and outlined in the three subsections below.

Table 1: We provide a systematic comparison of MIL architectural approaches, encoders, and pooling operations.

Architectural ordering	*Embedding-aggregation*, *prediction-aggregation*
Encoders	ViT-B/16, ConvNeXt-Tiny, MedSAM
Pooling	
MIL without instance interaction	Max, Mean, ABMIL
MIL with instance interaction	TransMIL, SmAP

2.1. Architectural ordering approaches

Among deep neural networks for MIL, there are two main paradigms, which we refer to as *embedding-aggregation* and *prediction-aggregation*. Both approaches consist of three parts: an encoder, a pooling operation, and a classifier. They differ in the ordering of these parts, as shown in Fig. 2.

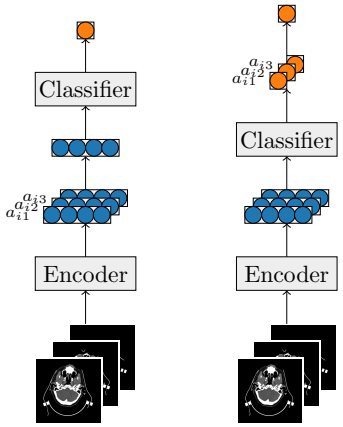

Figure 2: Architectural ordering approaches: *embedding-aggregation* (left) pools before classifying; *prediction-aggregation* (right) classifies before pooling.

In the *embedding-aggregation* approach, the order is encode, pool, then classify. First, each instance's C-channel 2D image $x_{i,j} \in \mathbb{R}^{C \times H \times W}$ is encoded to an instance-specific representation vector $h_{i,j} = f(x_{i,j}) \in \mathbb{R}^M$. Second, a pooling operation $\sigma(\cdot)$ (e.g., max, mean, or attention-based pooling) aggregates all S_i instance representations $h_i = \{h_{i,1}, \ldots, h_{i,S_i}\}$ into a single representation vector $z_i = \sigma(h_i) \in \mathbb{R}^M$. Finally, the bag level representation vector z_i is classified into a predicted probability, $g(z_i) \in [0, 1]$. We can denote the ultimate prediction as $\hat{y}_i = g(\sigma(f(x_i))$. In this notation, applying f to a set yields another set containing a mapping of each instance.

In the prediction-aggregation approach, the ordering of $g(\cdot)$ and $\sigma(\cdot)$ is swapped. A separate prediction score vector containing logits or probabilities is produced for each of the S_i instances separately, and then pooling determines the final prediction, $\hat{y}_i = \sigma(g(f(x_i))))$.

In either approach, model parameters for all parts (encoder, pooling, and classifier) can be trained to minimize binary cross entropy averaged across all data: $\frac{1}{N}\sum_{i=1}^{N} \ell^{\text{BCE}}(y_i, \hat{y}_i)$, where $\hat{y}_i$ is a function of input features and parameters.

2.2. Encoders

In both architectural approaches, an encoder $f(\cdot)$ embeds each 2D slice into a representation vector. We compare a vision transformer (ViT, Dosovitskiy et al. 2021) pre-trained on ImageNet-1k (Deng et al., 2009), a convolutional encoder (Liu et al., 2022b) also pre-trained on ImageNet-1k, and a medical image encoder from the medical segment anything model (MedSAM, Ma et al. 2024), a variant of a recent method (Kirillov et al., 2023) fine-tuned for medical image segmentation on a large medical image dataset.

2.3. Pooling

The design of the pooling operation $\sigma(\cdot)$, which aggregates across instances, is generally most important for understanding how spatial context is incorporated. We describe several architectures below. We focus on *embedding-aggregation* for concreteness; translation to *prediction-aggregation* is straightforward. Here, we take as input a set of embeddings $h_i = \{h_{i,1}, \ldots, h_{i,S_i}\}$ for bag i. Each instance j in the bag is encoded as a representation vector $h_{i,j} \in \mathbb{R}^M$.

Max and mean pooling. Two simple pooling operations find the maximum or mean ***element-wise*** of the given M-dimensional vectors:

$$z_i = \max_{j=1,\ldots,S_i} h_{ij}, \quad \text{or} \quad z_i = \operatorname{mean}_{j=1,\ldots,S_i} h_{ij}. \tag{1}$$

Early deep MIL methods (Pinheiro and Collobert, 2015; Zhu et al., 2017; Feng and Zhou, 2017) used such simple, non-trainable operations to aggregate instance representations.

Attention-based pooling. Attention-based pooling (ABMIL) (Ilse et al., 2018) assigns an attention weight a_{ij} to each instance via

$$a_{ij} = \frac{\exp\left(u^\top \tanh\left(Uh_{ij}\right)\right)}{\sum_{k=1}^{S_i} \exp\left(u^\top \tanh\left(Uh_{ik}\right)\right)}, \quad (2)$$

then forms bag-level embedding vector z_i via a weighted average: $z_i = \sum_{j=1}^{S_i} a_{ij} h_{ij}$. The weights a_{ij} are non-negative and sum to one: $a_{ij} \geq 0$ for all j; $\sum_j a_{ij} = 1$. In the equation above, vector $u \in \mathbb{R}^L$ and matrix $U \in \mathbb{R}^{L \times M}$ are trainable parameters.

Smooth attention pooling. Smooth attention pooling (SmAP) (Castro-Macías et al., 2024) uses a smoothing operation to add local interactions between instance embeddings. The smoothed embeddings $g_i \in \mathbb{R}^{S_i \times M}$ for all S_i instances are obtained by solving an optimization problem

$$\texttt{Sm}(h_i) = \underset{g_i}{\operatorname{argmin}}\, \alpha \mathcal{E}_D(g_i) + (1-\alpha)\|h_i - g_i\|_F^2, \quad (3)$$

where $\alpha \in [0, 1)$ controls the amount of smoothness, $\|\cdot\|_F$ denotes the Frobenius norm, and

$$\mathcal{E}_D(g_i) = \tfrac{1}{2} \sum_{j=1}^{S_i} \sum_{k=1}^{S_i} A_{ijk} \|g_{ij} - g_{ik}\|_2^2. \quad (4)$$

Here, $A_i \in \mathbb{R}^{S_i \times S_i}$ is an adjacency matrix defining local relationships between instances and $\|\cdot\|_2^2$ denotes the squared Euclidean norm aka "sum of squares". Following Castro-Macías et al. (2024), we use an adjacency matrix that links each slice to its adjacent slices in scan order. The resulting smoothed embedding g_{ij} then replaces embedding h_{ij} in Eq. (2).

Transformer-based pooling. Shao et al. (2021)'s transformer-based correlated MIL (TransMIL) allows instance interactions to inform pooling. First, TransMIL uses convolutions over instances in a pyramidal position encoding to model dependencies. Second, interactions between *all pairs* of instances are captured via multi-head self-attention. For layer ℓ and head h, there's a $S_i{+}1 \times S_i{+}1$ attention matrix, where rows sum to one and weight j, k is:

$$a_{i,j,k}^{(\ell,h)} \propto \exp\left(\left(q_{i,j}^{(\ell,h)}\right)^\top k_{i,k}^{(\ell,h)} / \sqrt{D}\right). \quad (5)$$

Here, each instance j has embeddings of size D for query $q_{i,j}^{(\ell,h)} = W_Q^{(\ell,h)} h_{i,j}^{\ell-1}$, key $k_{i,j}^{(\ell,h)} = W_K^{(\ell,h)} h_{i,j}^{\ell-1}$, and value $v_{i,j}^{(\ell,h)} = W_V^{(\ell,h)} h_{i,j}^{\ell-1}$. Propagating embeddings via attention-weighted value averages over several layers and heads allows instance features to interact flexibly to inform the ultimate bag-level embedding.

3. Related Work

CNNs for 3D neuroimaging. Deep learning has been widely applied to neuroimaging classification tasks (Huang et al., 2023; Dorfner et al., 2025). Several works examine 2D vs. 3D CNN architectures. Wen et al. (2020) provide a comprehensive overview and reproducible evaluation of CNNs for Alzheimer's disease (AD) classification using MRI neuroimages. They compared 2D slice-based and 3D volumetric approaches on the ADNI, AIBL and OASIS datasets and found that the performance of 2D slice-based approaches pre-trained on ImageNet-1k (Deng et al., 2009) was generally lower compared to 3D approaches. On a small subset, Dufumier et al. (2021) showed that randomly initialized 3D CNNs outperform randomly initialized 2D CNNs using mean pooling with a *prediction-aggregation* architecture. Recent work further demonstrates that the performance of 3D CNNs strongly depends on design choices related to normalization, downsampling, and depth (Liu et al., 2019, 2022a). In contrast to these works that focus only on 2D and 3D CNN approaches, we systematically compare MIL methods.

MIL for neuroimaging. The popularity of MIL for neuroimaging classification tasks has grown in recent years (Tong et al., 2014; López Pérez et al., 2022; Harvey et al., 2023; Perez-Cano et al., 2024), yet quality benchmarking especially to methods outside MIL remains underexplored. For example, Wu et al. (2021) in their Table 2 compare their MIL method's ICH classification results on a subset of the RSNA dataset to 3D CNN ICH classification numbers pulled directly from other papers that evaluate on *different datasets*. This comparison is "apples-to-oranges", making it difficult to draw conclusions about the relative rankings of the two approaches.

Recent MIL for neuroimaging work has focused on improving classification and localization results compared to conventional MIL by introducing smooth attention (Wu et al., 2023; Castro-Macías et al., 2024), described earlier in Sec. 2. However, despite growing interest, there are no systematic evaluations of different MIL encoders, aggregation approaches, and pooling operations. Moreover, MIL's relative strengths and weaknesses compared to more traditional approaches like 3D CNNs remain poorly understood.

Brain foundation models. Work has begun to develop brain foundation models that pre-train large encoders on neuroimaging data and then adapt them to downstream neuroimaging tasks. For example, Deng et al. (2026) propose a brain foundation model

that they adapt for brain disease segmentation and classification tasks, and Wang et al. (2024) introduce Vote-MI, an unsupervised representative slice selection method that selects representative 2D slices from 3D brain MRIs to enable effective use of 2D vision–language models (VLMs). Related multimodal clinical foundation models, such as Dai et al. (2025a), also include 3D neuroimages in their training data (Dai et al., 2025b). Overall, these efforts are relevant but largely tangential to our study. Many brain foundation model pipelines ultimately process 3D neuroimages in a slice-wise manner, which requires aggregating information across slices to produce scan-level predictions. Therefore, our systematic evaluation of different MIL architectural orderings and pooling operators can inform the design of future MIL that uses a brain foundation model encoder.

4. Experiment Design

4.1. Datasets of 3D Brain Scans

We study classification on the seven datasets of 3D brain scans in Tab. 2. For all CT datasets, we follow recommended preprocessing (Muschelli, 2019), leaving the original number of slices which varies across scans. For all MRI datasets, we follow recommended preprocessing (Wen et al., 2020; Routier et al., 2021). Each available modality (T1 and T2, if included) is mapped to a fixed-size template with 179 slices. Across CT and MRI, the same preprocessed 3D image is fed into all 3D and MIL NNs. MIL encoders process axial slices where each 2D image is resized to 224×224 (except for MedSAM, which uses 1024 × 1024). See App. A for more details preprocessing.

ADNI1. The ADNI1 Complete 1Y 1.5T dataset (Mueller et al., 2005) includes 2,294 T1 MRI scans from 639 patients. We use the diagnostic cohorts assigned upon enrollment for binary classification of Alzheimer's disease (AD).

OASIS-3 CT. The OASIS-3 CT dataset (LaMontagne et al., 2019) has 662 CT scans from 495 patients. We use the clinical dementia rating (CDR) for binary classification of AD. Positive labels indicate the patient has a diagnosis at most 80 days before or 365 days after the MRI scan date.

OASIS-3 MRI. The OASIS-3 MRI dataset (LaMontagne et al., 2019) includes 1,620 T1 and T2 MRI scans from 903 patients. Label definitions for AD are the same as in OASIS-3 CT.

RSNA-1,149. Prior work in MIL for neuroimage classification (Wu et al., 2021; López Pérez et al., 2022; Perez-Cano et al., 2024; Castro-Macías et al., 2024) uses a subset of 1,150 CT scans from the RSNA 2019 Brain CT Hemorrhage Challenge (Flanders et al., 2020). In the released subset's train and test sets (Castro-Macías et al., 2025), we found one scan appears in both the training and test set. We removed the duplicate scan, resulting in 1,149 unique CT scans which we subdivide as described below.

RSNA-21,744. The RSNA 2019 Brain CT Hemorrhage Challenge (Flanders et al., 2020) includes 752,803 slices from 21,744 CT scans with released labels for presence/absence of any intracranial hemorrhage (ICH). We subdivide the released training set (no other release has labels) into our own training, validation, and test sets as described below.

Proprietary-800. The Proprietary-800 dataset contains de-identified axial T1 and T2 MRI scans from 800 patients 50+ years of age who received an MRI in 2009-2019 during the course of routine care within the Kaiser Permanente health system in southern California, USA. Our study's IRB approval is documented on page 1.

Two separate binary labels are of interest: white matter disease (WMD) and covert brain infarction (CBI). These are often incidental findings with no outward symptoms. Yet the presence of either may be predictive of future stroke or dementia (Pasi and Cordonnier, 2020).

Of the 800 scans here, 640 scans were randomly sampled from all eligible scans. For the remainder, we deliberately sample 160 scans directly from CBI-positive patients in the eligible cohort (regardless of WMD status) to overcome CBI's rarity. Our 800 scan dataset contains 251 CBI cases and 541 WMD cases.

Proprietary-10k. The Proprietary-10k dataset is a larger dataset of T1 and T2 MRIs from 10,000 patients 50+ years of age who received a routine-care MRI in 2009-2019 from the same southern California health system. Our study's IRB approval is documented on page 1.

This larger dataset has the same binary labels for WMD and CBI. All scans with CBI (regardless of WMD status) were included and the remaining scans were randomly sampled from the eligible cohort. This yielded a total dataset of 474 CBI cases and 4,291 WMD cases.

WMD/CBI label extraction. WMD and CBI labels for Proprietary scans are extracted from routine text reports provided by an clinical expert during routine image interpretation. We use an NLP tool developed by Fu et al. (2019) for the extraction. The tool achieved 1.00 positive predictive value

Table 2: Dataset statistics: Patient-scan counts by class.

Dataset	Modality	Label	Num Neg.	Num Pos.	Total Scans	Instances/Scan
ADNI1	MRI	AD	1,818	476	2,294	179-179
OASIS-3	CT	AD	556	106	662	74-111
	MRI	AD	1,239	381	1,620	179-179
RSNA-1,149	CT	ICH	666	483	1,149	24-57
RSNA-21,744	CT	ICH	12,862	8,882	21,744	20-60
Proprietary-800	MRI	CBI	549	251	800	179-179
		WMD	259	541	800	179-179
Proprietary-10k	MRI	CBI	9,526	474	10,000	179-179
		WMD	5,709	4,291	10,000	179-179

(PPV) and 0.99 negative predictive value (NPV) for CBI and 0.99 PPV and 0.99 NPV for WMD on 1,000 manually annotated reports across two sites with high interannotator agreement (Cohen's $\kappa = 0.87$ and 0.91 for CBI at the two sites).

Spatial extent of labels. The relevant spatial extent of a 3D scan needed to correctly classify the different labels here differs in important ways that impact our later analysis. The binary label for AD arises from broader clinical knowledge not just observed imaging. In imaging, the signs of neurodegenerative diseases like AD are likely diffuse throughout many axial slices. In contrast, ICH lesions are visible in a focal region of the brain but can be volumetrically extensive: in RSNA the mean number of contiguous slices for a lesion is 12. CBI lesions are also focal, lacunar infarcts in particular are often less than 1 cm in size. WMD is typically bilateral and regionally patterned (periventricular and deep white matter), often visible across multiple slices. It is regionally concentrated, but not as focal as a lacunar infarct.

4.2. 3D Methods

We compare two families of 3D baselines.

3D CNNs. 3D CNNs are the standard approach to classify CT and MRI scans. We include a 3D ResNet-18 that can process a variable number of inputs (Tran et al., 2018) and a 3D CNN designed for AD (Liu et al., 2019, 2022a) as baselines.

3D ViTs. We additionally evaluate 3DINO (Xu et al., 2025), a ViT-L/16 pre-trained on a large corpus of 3D medical data via self-supervised learning. As recommended by the creators, in our implementation the embedding concatenates the class tokens from the last 4 transformer blocks with the mean-pooled patch tokens of the final block. For MRI scans with both T1 and T2 available, the per-modality embeddings are also concatenated. A single linear classification head is then trained on this scan-level embedding. 3DINO's pre-training data includes images from ADNI1, OASIS-3, and the large RSNA-21,744 dataset, but no labels y_i from these datasets informed the pre-training of 3DINO. Only our Proprietary datasets provide a truly leakage-free evaluation of this encoder.

4.3. Training and Evaluation Procedures

For all non-Proprietary datasets, we randomly assign images at a 4:1:1 ratio into training, validation, and testing sets. We ensure each patient's data belongs to exactly one set to avoid leakage. We stratify by class to ensure comparable class frequencies. We repeat this process with three data-split random seeds; each seed selects a different partition into training, validation, and test sets.

Our Proprietary data comes from different hospital sites within an integrated healthcare system. To better assess cross-site generalization, we assign scans to training, validation, and test sets using site IDs, ensuring each site's data belongs to exactly one split. We repeat this data splitting 5 times in a leave-one-site-group-out design.

Performance metrics. Ultimately, all tables and figures report the test set mean of a specific metric across 3 data splits (5 for Proprietary), with uncertainty quantified via "+/-" one standard deviation.

All our classification tasks are binary, so we primarily use area under the receiver operating characteristic curve (AUROC) to measure discriminative quality. We further report area under the precision-recall curve (AUPRC) in the supplement. We find that relative method rankings are typically preserved across these metrics.

Training and hyperparameter tuning. As much as possible, we ensure an apples-to-apples treatment of training and hyperparameter search for fair benchmarking across diverse methods (Huang et al.,

2024). All methods are trained with minibatch stochastic gradient descent (SGD) to minimize binary cross entropy loss with L1 or L2 regularization. We use SGD with a momentum parameter of 0.9. We set batch size to 64 for frozen MIL and 4 for 3D CNNs (more than 4 risks memory errors on our commodity GPUs). We train frozen MIL for 1,000 epochs; we train the far more expensive 3D CNNs for 100 epochs, following Liu et al. (2019, 2022a). After training, we select the checkpoint that maximizes validation AUROC, subject to the constraint that validation AUROC is lower than training AUROC, to mitigate unreliable model selection on small datasets.

For both MIL and 3D CNNs, we select learning rate from {0.1, 0.01, 0.001, 0.0001} and L1 or L2 regularization strength from {1.0, 0.1, 0.01, 0.001, 0.0001, 1e-5, 1e-6, 0.0}.

5. Results and Analysis

5.1. Performance on Moderate-Sized Data

Here, we report results and analysis from experiments on moderately-sized CT and MRI datasets (600-3,000 total scans). For such data, thorough experiments were affordable for understanding the impact of different MIL design choices (architectural ordering, encoders, pooling, etc.) as well as tradeoffs between 3D CNNs, and MIL methods. In a later subsection, we analyze the two bigger datasets to see if the best-performing methods still work well.

What architectural ordering works best? In Fig. 3, we compare *embedding-aggregation* and *prediction-aggregation* versions of MIL pooling methods, all using a common ViT encoder. We find overall that usually the two orderings yield roughly the same AUROC. Though some individual results can differ by up to 0.05 AUROC, the scale of uncertainty is also high. Tables of AUROC and AUPRC numbers for *embedding-aggregation* (Tab. 4 and D.2) and *prediction-aggregation* (Tab. C.1 and D.3) show that relative ranking of pooling approaches is not greatly altered by the ordering choice. For the rest of this main paper, we focus on embedding aggregation as it allows inclusion of TransMIL, which does not have a *prediction-aggregation* approach due to the nature of its multi-head self-attention.

What encoder works best? Tab. 3 and D.1 report AUROC and AUPRC on OASIS-3 MRI for three different encoders: ViT-B/16, ConvNeXt-Tiny, and MedSAM. Encoder performance varied for each pooling strategy. ViT-B/16 performed better for simple pooling operations (Max and Mean) while ConvNeXt-Tiny performed better for learnable pooling operations (ABMIL, SmAP, and TransMIL). MedSAM consistently underperformed both ViT-B/16 and ConvNeXt-Tiny for all pooling operations. We chose a frozen ViT-B/16 for all subsequent experiments because of its superior performance with simple pooling operations, which was not significantly exceeded by any other encoder with any other pooling strategy.

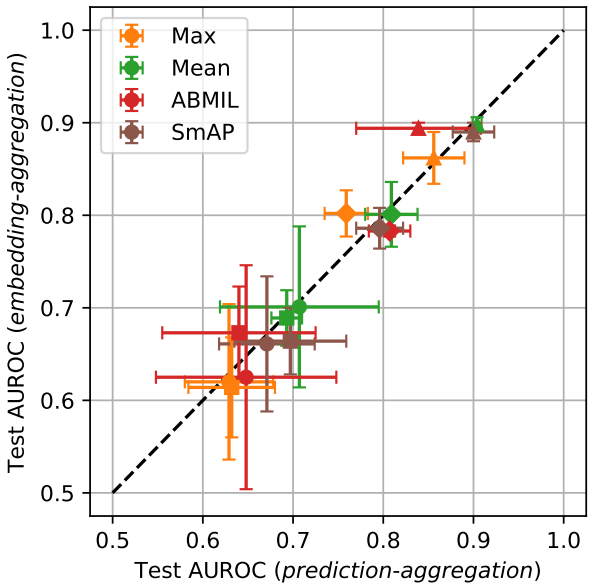

Figure 3: Test AUROC for *embedding-aggregation* and *prediction-aggregation* on OASIS-3 CT (●), RSNA-1,149 CT (▲), ADNI1 MRI (■), and OASIS-3 MRI (◆). **Takeaway: Results are concentrated around the $y = x$ line, indicating the performance of *embedding-aggregation* and *prediction-aggregation* is comparable across pooling operations and datasets.**

What pooling operation works best? Tab. 4 and D.2 report AUROC and AUPRC for *embedding-aggregation* approaches for five different pooling operations: Max, Mean, ABMIL, TransMIL, and SmAP. On all non-Proprietary datasets, mean pooling performed comparable with and in many cases outperforms, more complicated pooling approaches. Mean pooling also was competitive in separate evaluations of *prediction-aggregation* approaches (see Tab. C.1 and D.3). Notably, many prior works on MIL for brain scans (Wu et al., 2021; López Pérez et al., 2022; Perez-Cano et al., 2024; Castro-Macías et al., 2024) do not include mean pooling as a baseline, despite its strong and consistent performance observed here.

Table 3: Encoder comparison: Test AUROC on OASIS-3 MRI. All MIL methods use *embedding-aggregation*.

	MIL without instance interaction			*MIL with instance interaction*	
	Max	**Mean**	**ABMIL**	**TransMIL**	**SmAP**
ViT-B/16	$0.802_{\pm 0.025}$	$0.801_{\pm 0.035}$	$0.783_{\pm 0.006}$	$0.744_{\pm 0.038}$	$0.786_{\pm 0.022}$
ConvNeXt-Tiny	$0.789_{\pm 0.039}$	$0.788_{\pm 0.024}$	$0.801_{\pm 0.019}$	$0.792_{\pm 0.032}$	$0.800_{\pm 0.025}$
MedSAM	$0.769_{\pm 0.038}$	$0.767_{\pm 0.003}$	$0.791_{\pm 0.031}$	$0.775_{\pm 0.010}$	$0.780_{\pm 0.026}$

Table 4: Pooling comparison: Test AUROC on medium datasets. All MIL methods use a frozen ViT and *embedding-aggregation*. †3DINO was pre-trained on ADNI1, OASIS-3, and half of the RSNA ICH full dataset.

		CT		**MRI**			
		OASIS-3	**RSNA-1,149**	**ADNI1**	**OASIS-3**	**Proprietary-800**	
		AD	ICH	AD	AD	CBI	WMD
MIL without instance interaction	Max	$0.620_{\pm 0.084}$	$0.862_{\pm 0.028}$	$0.614_{\pm 0.054}$	$0.802_{\pm 0.025}$	$0.648_{\pm 0.020}$	$0.644_{\pm 0.052}$
	Mean	$0.701_{\pm 0.087}$	$0.898_{\pm 0.008}$	$0.689_{\pm 0.030}$	$0.801_{\pm 0.035}$	$0.610_{\pm 0.050}$	$0.640_{\pm 0.053}$
	ABMIL	$0.625_{\pm 0.121}$	$0.894_{\pm 0.006}$	$0.673_{\pm 0.050}$	$0.783_{\pm 0.006}$	$0.609_{\pm 0.038}$	$0.648_{\pm 0.020}$
MIL with instance interaction	TransMIL	$0.648_{\pm 0.097}$	$0.893_{\pm 0.011}$	$0.593_{\pm 0.067}$	$0.744_{\pm 0.038}$	$0.565_{\pm 0.060}$	$0.681_{\pm 0.054}$
	SmAP	$0.661_{\pm 0.073}$	$0.890_{\pm 0.010}$	$0.664_{\pm 0.036}$	$0.786_{\pm 0.022}$	$0.610_{\pm 0.063}$	$0.692_{\pm 0.077}$
	3D ResNet-18	$0.537_{\pm 0.079}$	$0.850_{\pm 0.014}$	$0.567_{\pm 0.081}$	$0.725_{\pm 0.029}$	$0.591_{\pm 0.050}$	$0.563_{\pm 0.063}$
	3D CNN for AD	*Variable-sized input not supported*		$0.680_{\pm 0.043}$	$0.816_{\pm 0.031}$	$0.689_{\pm 0.054}$	$0.708_{\pm 0.077}$
	3DINO	$^{\dagger}0.622_{\pm 0.038}$	$^{\dagger}0.895_{\pm 0.014}$	$^{\dagger}0.581_{\pm 0.015}$	$^{\dagger}0.795_{\pm 0.051}$	$0.645_{\pm 0.032}$	$0.691_{\pm 0.063}$

In contrast, for WMD and CBI classification on Proprietary-800, mean pooling did notably worse than the best 3D CNN and best attention-based MIL.

Why does mean pooling perform well on AD or ICH but not CBI/WMD? A possible explanation for the strong performance of mean pooling is that, for several neuroimaging tasks, signal relevant to classification may be distributed across many slices rather than concentrated in a narrow highly informative region. Therefore, averaging across slices can yield a stable scan-level representation and reduce sensitivity to noise or spurious slice-level artifacts.

We tentatively hypothesize that mean pooling performs relatively poorly for CBI and WMD because, unlike the AD label, the imaging signal for these covert cerebrovascular abnormalities can often be focal or regional rather than diffuse. CBI in particular might appear as a single subcortical lesion that is less than 1cm, particular when due to a lacunar infarct, the most common form of CBI. Mild WMD (the most common subtype in our data) is often apparent in periventricular regions rather than diffusely throughout the brain. In these settings, diagnostically informative signal is anatomically concentrated, and thus averaging across all slices may dilute signal. An alternative explanation could be the imperfect nature of the NLP tools used to obtain WMD and CBI labels.

How do our reported numbers compare to other published efforts? All numbers here were done by our team using our released code. It is useful to verify the discriminative performance we report is at least comparable with (if not better than) other efforts published elsewhere. On RSNA-1,149 dataset, our ABMIL, TransMIL, and SmAP results are comparable to recently published work (Castro-Macías et al., 2024). Notably, unlike that work we include the competitive mean pooling baseline.

As a final sanity check, specifically for the white matter disease (WMD) task on the Proprietary-800 dataset, we compare to SAMSEG (Puonti et al., 2016; Cerri et al., 2021, 2023), a FreeSurfer tool for white matter lesion segmentation. Using this tool, we can estimate the total volume of white matter as a thresholdable score for WMD classification. SAMSEG scores $0.699_{\pm 0.072}$ AUROC, within 0.01 of the best results from deep MIL and 3D CNNs in Tab. 4.

5.2. Performance on Large Datasets

We now examine the two largest datasets, RSNA-21,744 CT and Proprietary-10k MRI. Here, the size of the datasets made exhaustive comparisons of all methods with large hyperparameter search grids infeasible. We thus manually selected an appropriate subset of methods, always using a frozen ViT for MIL.

What works on big CT? Here, we focused on deep MIL comparisons. The *3D CNN for AD* cannot

Table 5: Pooling comparison: Test AUROC on **Large** datasets. All MIL methods use a frozen ViT and *embedding-aggregation*. Time is for training one NN at one hyperparameter on one train/test split. Results average over 3 data splits for RSNA-21,744 and 5 data splits for Proprietary-10k and show best of a grid search across many hyperparameters.

		CT		**MRI**	
		RSNA-21,744		**Proprietary-10k**	
		ICH	Time	CBI	WMD
MIL without instance interaction	Max	$0.888_{\pm 0.009}$	20 min.	$0.644_{\pm 0.053}$	$0.661_{\pm 0.020}$
	Mean	$0.920_{\pm 0.012}$	25 min.	$0.666_{\pm 0.017}$	$0.689_{\pm 0.027}$
	ABMIL	$0.919_{\pm 0.009}$	36 min.	$0.647_{\pm 0.055}$	$0.713_{\pm 0.033}$
MIL with instance interaction	TransMIL	$0.925_{\pm 0.014}$	13 hr. 22 min.	$0.637_{\pm 0.056}$	$0.715_{\pm 0.030}$
	SmAP	$0.925_{\pm 0.012}$	12 hr. 37 min.	$0.669_{\pm 0.050}$	$0.714_{\pm 0.031}$
	3D CNN for AD	*Variable-sized input not supported*		$0.651_{\pm 0.042}$	$0.728_{\pm 0.033}$

be used out-of-the-box on RSNA-21,744 because each CT scan has a variable number of slices. The 3D ResNet-18 performed poorly on moderately-sized CT data, so we elected to skip it here.

Tab. 5 shows that the core message of our Fig. 1 is true even with 10,000+ scans: more flexible MIL methods like TransMIL can outperform simplistic MIL, but only by a modest margin (less than 0.01 AUROC gain on CT) while requiring 25x the compute. For many practitioners, it may be question whether such modest gains are worth the effort over just using the strong 0.920 AUROC from *Mean pooling MIL*.

What works on big MRI? The *3D CNN for AD* was competitive on moderately-sized MRI datasets, so we compared it to MIL on the Proprietary-10k MRI dataset. Tab. 5 shows that 3D CNNs can outperform attention-based MIL, but they are more expensive to train. Notably, mean pooling is competitive and much more efficient to train.

5.3. Evaluating Per-Slice Attention Quality

Learnable attention is what differentiates the advanced MIL methods of recent years from simplistic mean pooling MIL. Here, we take advantage of a unique aspect of the large RSNA dataset: every scan has instance-level labels indicating the positive/negative status of each slice of the 3D scan for the ICH binary task. While the attention value at a slice is not necessarily intended to exactly mean the predicted positive-class probability in any MIL network, the basic logic of MIL prediction suggests that when a 3D scan is positive, at least some attention should be paid to positive slices. Past works have thus assessed per-instance attention as a predictor of instance-level class label (Castro-Macías et al., 2024).

Metrics. To study how well attention values a_{ij} produced by MIL may predict the binary instance-level ICH labels y_{ij}, we report three higher-is-better metrics: (1) attention correctness (Liu et al., 2017), (2) AUROC, and (3) AUPRC. Our evaluations follows prior work that used instance-level labels to assess whether attention aligns with known positive instances (e.g., Fig. 5 in Ilse et al. (2018); Fig. 6 in Shao et al. (2021); Fig. 4, 5, 9 in Castro-Macías et al. (2024)).

Center-focused baseline. Beyond the MIL methods, we compare to a simple baseline that always allocates attention via a Gaussian bell-shaped curve centered at the middle slice of the 3D scan $a_{ij} \propto \mathcal{N}(j \mid \frac{S_i}{2}, 1)$, regardless of the input image. This is meant to provide a "center-focused" inductive bias, as a brief exploratory analysis of RSNA per-slice labels suggests positive instances often occur close to the middle of the axial scan, reflecting the typical neuroanatomic distribution of ICH lesions.

Result: MIL attention quality does not exceed simple baseline. Results are found in Tab. 6. Among attention-based MIL methods, we find that TransMIL and SmAP substantially outperform ABMIL on instance-level AUROC and AUPRC. However, the center-focused baseline, which does not depend on the input neuroimage at all, surprisingly achieves the best results overall, exceeding the next-best method by 0.154 attention correctness and 0.019 AUROC.

We use bootstrapping (Foody, 2009) to access the statistical significance of some AUROC difference.

6. Semi-Synthetic Experiments

To better understand the effectiveness of mean pooling, we design a semi-synthetic dataset and com-

Table 6: Per-slice attention quality metrics on test set of RSNA-21,744. All MIL methods use a frozen ViT and *embedding-aggregation*.

	Centered Gaussian	Max	Mean	ABMIL	TransMIL	SmAP
Attention correctness	$0.701_{\pm 0.011}$	N/A	$0.359_{\pm 0.009}$	$0.547_{\pm 0.087}$	$0.435_{\pm 0.028}$	$0.509_{\pm 0.010}$
AUROC	$0.850_{\pm 0.001}$	N/A	$0.500_{\pm 0.000}$	$0.736_{\pm 0.052}$	$0.792_{\pm 0.031}$	$0.831_{\pm 0.003}$
AUPRC	$0.710_{\pm 0.007}$	N/A	$0.359_{\pm 0.009}$	$0.634_{\pm 0.049}$	$0.659_{\pm 0.032}$	$0.713_{\pm 0.007}$

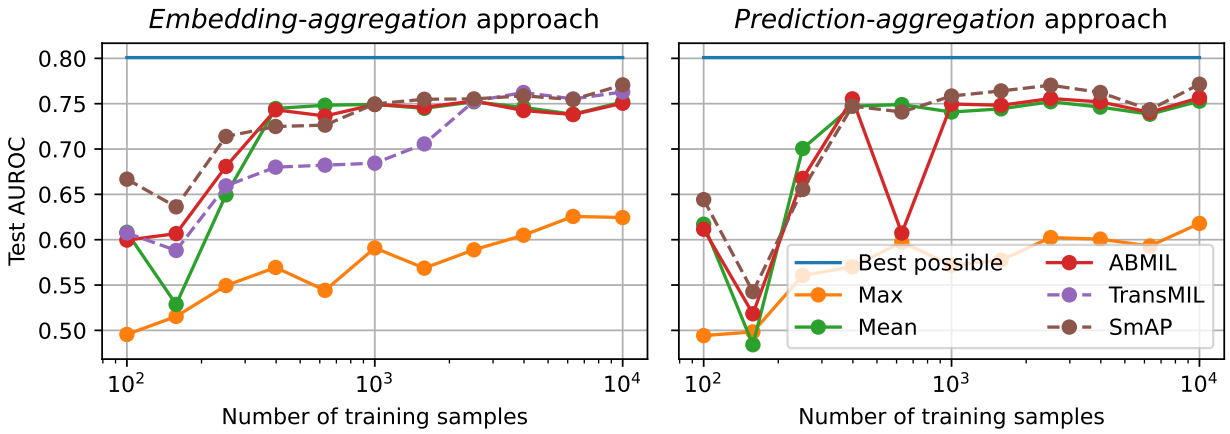

Figure 4: Test AUROC vs. train set size for semi-synthetic Shifted Mean task (Sec. 6), where aggregating subtle signals across 12 instances is needed. **Takeaway: Even with 10,000 training scans, SmAP and TransMIL barely outperform mean pooling and fall at least 0.04 AUROC below the best possible Bayes estimator.**

pare MIL variants on it. We intend this data to represent key challenges in MIL for 3D brain scans: (1) only some features in an embedding are discriminative, (2) only a few instances in a bag signal whether it should have a positive label, and (3) context from nearby instances matters, as the information from an individual instance may be statistically ambiguous. We set key data statistics to match the RSNA dataset: we match the range of the number of instances in a bag: $S_i \in \{20, \dots 60\}$, set the mean number of contiguous positive instances to $R = 12$ as in RSNA, and use 768 features like a ViT embedding.

This effort is similar in spirit to work on *algorithmic unit tests* for MIL (Raff and Holt, 2023). That paper uses synthetic classification tasks designed to reveal whether learned models violate key MIL assumptions, such as a bag is positive if and only if one or more instances have a positive label. Our new data focuses instead on assessing context from nearby instances and quantifying best possible performance.

Data generation. We define a data-generating process we call *Shifted Mean MIL* that jointly samples $h_{i,1:S_i}, y_i$ embedding-label pairs. We draw $y_i \sim \text{Bern}(0.5)$, so 50% of scans are positive, then draw the number of instances $S_i \sim \text{Unif}(\{20, \dots, 60\})$. If the scan is negative, all embeddings for all instances are drawn from a mean=0, variance=1 Gaussian. If positive, we select a contiguous block of $R = 12$ instances to indicate positivity, and draw their first feature (of many) from a Gaussian with mean $\Delta = 0.5$; other features are drawn from a zero-mean Gaussian. We intentionally set Δ low so that no one instance will be a "smoking gun," but reasoning over R instances should reliably indicate which bags are positives. See App. F for dataset details.

Given the true distribution, we characterize the best possible classifier for this data by deriving a Bayes estimator (DeGroot, 1970; Murphy, 2022), the best probabilistic predictor of y_i given h_i for this data. See App. G for details.

Experiments. We intend the provided h_i already represent encoder-provided embeddings. Thus, we assess how well MIL pooling and classification variants can classify as a function of the total number of training samples, which we step from 100 to 10,000. At each size, we randomly assign bags using an 4:1 ratio into training and validation sets. We report results on a fixed separate test set of 1,000 bags.

Results: Can current MIL methods capture instance interaction? Across training sample sizes in Fig. 4, even modern interaction-aware MIL approaches may fail to match Bayes-optimal performance. Neither TransMIL nor SmAP offers any noticeable gains over Mean-pooling, and all fall well short (at least 0.05 AUROC below) the best possible blue line of the Bayes estimator. Capturing subtle contextual dependence across instances appears challenging for current off-the-shelf MIL even at large data sizes. This helps explain why mean pooling works well for 3D brain classification tasks that have subtle signals at individual slices yet many instances (a dozen or more) that could indicate positivity.

7. Conclusion

We presented a benchmark of simple MIL, attention-based MIL, 3D CNNs, and 3D ViTs on 9 classification tasks across 7 datasets. Our results highlight that MIL with mean pooling is often a strong baseline that takes less than an hour to train even on the largest dataset we tested (21,744 scans), while the latest attention-based models require 8x or more training time per run to deliver modest gains of roughly +0.01-0.03 to the AUROC value. Instead of immediately reaching for bigger models, practitioners may explore what else they could do with those hours, such as pursue embedding-based augmentation strategies (Verma et al., 2019), consider deep ensembles (Lakshminarayanan et al., 2017) that average over encoders, or seek additional data sources.

Another key message of our work, supported by the per-slice attention quality experiments in Sec. 5.3, is that at least for 3D brain scans, the learned attention of modern well-published methods like TransMIL or SmAP does not seem to find reliable per-image signals. Future work might seek to inform attention-based MIL via the inductive bias of our Centered Gaussian baseline. Our semi-synthetic results in Sec. 6 further suggest that, especially for brain tasks where careful examination of many relevant slices is necessary, current MIL scores substantially below what is possible even with transformer-based attention and 10,000 training scans. Improved inductive biases and regularization strategies are needed.

Limitations. Our work has several limitations. All analyses use recommended data preprocessing steps (App. A) and did not consider alternatives. Our results mostly focus on encoders pre-trained on non-brain images without further fine-tuning. In preliminary experiments, fine-tuning the ViT encoder on the OASIS-3 MRI dataset improved performance a few percentage points, but the average time per epoch for mean pooling MIL increased from 1 minute to 46 minutes, making it prohibitively expensive to compare all methods using the same hyperparameter search space. We leave this to future work. We do compare to a few well-published 3D CNNs and one recent 3D ViT, but our list of 3D NN methods is not exhaustive and others may perform better. We did not use data augmentation for any method because we wanted our experiments to focus on the difference in model architectures while avoiding the additional computational cost of repeated forward passes in MIL experiments with frozen encoders.

Future work. Future work could include parameter-efficient fine-tuning (Hu et al., 2022), advanced regularization techniques (Li et al., 2018), top-k pooling (Yu et al., 2025), and cross-dataset generalization (Shao et al., 2025).

Outlook. In future years, practitioners are likely to rely more and more on pre-trained foundation models to provide embeddings for 3D brain scan classification. We hope our work provides a path forward for resource-constrained analysts using such models to quickly reach competitive prediction quality. We also hope it inspires method developers to improve per-slice attention mechanisms and overall pooling strategies.

Acknowledgments

This work is supported by the U.S. National Institutes of Health (grant # R01NS134859) and the Alzheimer's Drug Discovery Foundation. Author MCH is also supported in part by the U.S. National Science Foundation (NSF) via IIS CAREER grant # 2338962. Author AAS is supported by an NIH postdoctoral T32 award (T32TR004418). We are grateful for resources and support from the Tufts High-Performance Computing Cluster. This paper's content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

We are thankful for computing infrastructure support provided by Research Technology Services at Tufts University, with hardware funded in part by NSF award OAC CC* # 2018149.

References

Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric Diffeomorphic Image Registration with CrossCorrelation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. *Medical Image Analysis*, 12(1): 26–41, 2008.

Brian B. Avants, Nicholas J. Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C. Gee. The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, 8:44, 2014.

Francisco M. Castro-Macías, Pablo Morales-Álvarez, Yunan Wu, Rafael Molina, and Aggelos K. Katsaggelos. `Sm`: enhanced localization in Multiple Instance Learning for medical imaging classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Francisco M. Castro-Macías, Francisco J. Sáez-Maldonado, Pablo Morales-Álvarez, and Rafael Molina. RSNA_ICH_MIL. https://huggingface.co/datasets/torchmil/RSNA_ICH_MIL, 2025. Accessed: 2025-12-30.

Stefano Cerri, Oula Puonti, Dominik S. Meier, Jens Wuerfel, Mark Mühlau, Hartwig R. Siebner, and Koen Van Leemput. A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *Neuroimage*, 225:117471, 2021.

Stefano Cerri, Douglas N. Greve, Andrew Hoopes, Henrik Lundell, Hartwig R. Siebner, Mark Mühlau, and Koen Van Leemput. An open-source tool for longitudinal whole-brain and white matter lesion segmentation. *NeuroImage: Clinical*, 38:103354, 2023.

Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. QoQ-Med: Building Multimodal Clinical Foundation Models with Domain-Aware GRPO Training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Wei Dai, Peilin Chen, Malinda Lu, Daniel Li, Haowen Wei, Hejie Cui, and Paul Pu Liang. CLIMB: Data Foundations for Large Scale Multimodal Clinical Foundation Models. In *International Conference on Machine Learning (ICML)*, 2025b.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Zhongying Deng, Haoyu Wang, Ziyan Huang, Lipei Zhang, Angelica I. Aviles-Rivero, Chaoyu Liu, Junjun He, Zoe Kourtzi, and Carola-Bibiane Schönlieb. Brain foundation models with hypergraph dynamic adapter for brain disease analysis. *Pattern Recognition*, 172:112595, 2026.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

Felix J. Dorfner, Jay B. Patel, Jayashree Kalpathy-Cramer, Elizabeth R. Gerstner, and Christopher P. Bridge. A review of deep learning for brain tumor analysis in MRI. *npj Precision Oncology*, 9(1):2, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Benoit Dufumier, Pietro Gori, Ilaria Battaglia, Julie Victor, Antoine Grigis, and Edouard Duchesnay. Benchmarking CNN on 3D Anatomical Brain MRI: Architectures, Data Augmentation and Deep Ensemble Learning. *arXiv preprint arXiv:2106.01132*, 2021.

Ji Feng and Zhi-Hua Zhou. Deep MIML Network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Geetika Choudhary, Luciano Cala, Luís Coelho, Mads Mogensen, Fátima Morón, Eric Miller, Ichiro Ikuta, Vahe Zohrabian, Oran McDonnell, Christoph Lincoln, Luciano Shah, Devon Joyner, Ashish Agarwal, Richard K. Lee, and Jayashree Nath. Construction of a Machine Learning Dataset

through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence*, 2(3), 2020.

Vladimir S. Fonov, Alan C. Evans, Robert C. McKinstry, C. Robert Almli, and D. Louis Collins, Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009.

Vladimir S. Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, and D. Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1): 313–327, 2011.

Giles M. Foody. Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113(8):1658–1663, 2009. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425709000923.

Sunyang Fu, Lester Y. Leung, Yanshan Wang, Anne-Olivia Raulli, David F. Kallmes, Kristin A. Kinsman, Kristoff B. Nelson, Michael S. Clark, Patrick H Luetmer, Paul R. Kingsbury, David M. Kent, and Hongfang Liu. Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. *JMIR Medical Informatics*, 7(2), 2019.

Ethan Harvey, Wansu Chen, David M. Kent, and Michael C. Hughes. A Probabilistic Method to Predict Classifier Accuracy on Larger Datasets given Small Pilot Data. In *Machine Learning for Health (ML4H)*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6 (1):74, 2023.

Zhe Huang, Ruijie Jiang, Shuchin Aeron, and Michael C. Hughes. Systematic Comparison of Semi-supervised and Self-supervised Learning for Medical Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning (ICML)*, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pamela J. LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *MedRxiv*, 2019. URL https://doi.org/10.1101/2019.12.13.19014902.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. In *International Conference on Machine Learning (ICML)*, 2018.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention Correctness in Neural Image Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Sheng Liu, Chhavi Yadav, Carlos Fernandez-Granda, and Narges Razavian. On the design of convolutional neural networks for automatic detection of Alzheimer's disease. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, 2019.

Sheng Liu, Arjun V. Masurkar, Henry Rusinek, Jingyun Chen, Ben Zhang, Weicheng Zhu, Carlos Fernandez-Granda, and Narges Razavian. Generalizable deep learning model for early Alzheimer's

disease detection from structural MRIs. *Scientific Reports*, 12(1):17106, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Miguel López Pérez, Arne Schmidt, Yunan Wu, Rafael Molina, and Aggelos K. Katsaggelos. Deep Gaussian processes for multiple instance learning: Application to CT intracranial hemorrhage detection. *Computer Methods and Programs in Biomedicine*, 219:106783, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment Anything in Medical Images. *Nature Communications*, 15:654, 2024.

Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1997.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*, chapter 5.1 Bayesian decision theory. MIT Press, 2022.

John Muschelli. Recommendations for processing head CT data. *Frontiers in Neuroinformatics*, 13: 61, 2019.

Marco Pasi and Charlotte Cordonnier. Clinical Relevance of Cerebral Small Vessel Diseases. *Stroke*, 2020.

Jose Perez-Cano, Yunan Wu, Arne Schmidt, Miguel Lopez-Perez, Pablo Morales-Alvarez, Rafael Molina, and Aggelos K. Katsaggelos. An end-to-end approach to combine attention feature extraction and Gaussian Process models for deep multiple instance learning in CT hemorrhage detection. *Expert Systems with Applications*, 240: 122296, 2024.

Pedro O. Pinheiro and Ronan Collobert. From Image-Level to Pixel-Level Labeling With Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Oula Puonti, Juan Eugenio Iglesias, and Koen Van Leemput. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. *NeuroImage*, 143:235–249, 2016.

Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-Instance Learning for Medical Image and Video Analysis. *IEEE Reviews in Biomedical Engineering*, 10:213–234, 2017.

Edward Raff and James Holt. Reproducibility in Multiple Instance Learning: A Case For Algorithmic Unit Tests. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Sabrina Fontanella, Pietro Gori, Jérémy Guillon, Alexis Guyot, Ravi Hassanaly, Thomas Jacquemont, Pascal Lu, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. *Frontiers in Neuroinformatics*, Volume 15 - 2021, 2021.

Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do Multiple Instance Learning Models Transfer? In *International Conference on Machine Learning (ICML)*, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V. Hajnal, and Daniel Rueckert. Multiple instance learning for classification of dementia in brain MRI. *Medical Image Analysis*, 18 (5):808–818, 2014.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States. In *International Conference on Machine Learning (ICML)*, 2019.

Yuli Wang, Jian Peng, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Jing Wu, Wen-Chi Hsu, Maliha Imami, Zhicheng Jiao, Paul Zhang, and Harrison Bai. Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020.

Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, and Aggelos K. Katsaggelos. Combining Attention-Based Multiple Instance Learning and Gaussian Processes for CT Hemorrhage Detection. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

Yunan Wu, Francisco M. Castro-Macías, Pablo Morales-Álvarez, Rafael Molina, and Aggelos K. Katsaggelos. Smooth Attention for Deep Multiple Instance Learning: Application to CT Intracranial Hemorrhage Detection. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023.

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G. Krishnan, Anne L. Martel, and Maged Goubran. A generalizable 3D framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 2025.

Sicheng Yu, Xingshu Chen, Fangzhou Cao, and Ting Tian. Tka-mil: Top-k attention multiple instance learning for whole slide image classification and instance probability derivation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017.

Appendix A. Preprocessing

A.1. CT

For CT images, we convert images into Hounsfield Units (HU) using each image's rescale slope and intercept; use intensity windowing to exclude the skull, other bones, and calcifications (only including -100 to 300 HU) (Muschelli, 2019); resize each 2D slice to 224×224 pixels for VIT-B/16 and ConvNeXt-Tiny, and 1024×1024 pixels for MedSAM; and normalize images with the training set mean and standard deviation of each channel.

A.2. MRI

For MRI images, we follow the `t1-linear` pipeline from Clinica (Wen et al., 2020; Routier et al., 2021). We correct bias field inhomogeneities using the N4ITK method (Tustison et al., 2010); register each image to the MNI space with the ICBM 2009c nonlinear symmetric template (Fonov et al., 2009, 2011) using the SyN algorithm (Avants et al., 2008) from ANTs (Avants et al., 2014); crop each image to remove background; resize each 2D slice to 224×224 pixels for VIT-B/16 and ConvNeXt-Tiny, and 1024×1024 pixels for MedSAM; and normalize images with the training set mean and standard deviation of each channel.

Appendix B. Architectures

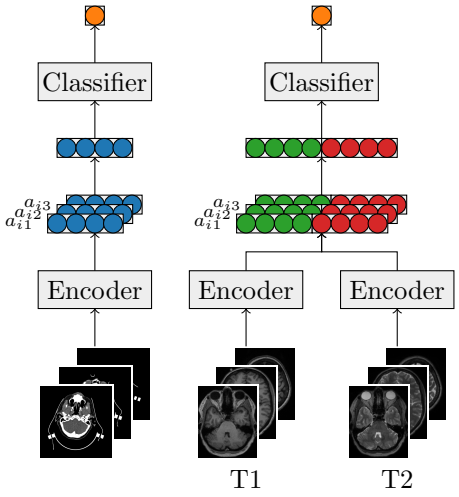

Figure B.1: Visualization of *embedding-aggregation* approach for CT (left) and T1 and T2 MRI (right) scans.

Appendix C. *Prediction-Aggregation* AUROC Results

Table C.1: Test AUROC on moderately-sized datasets. All MIL methods use the *prediction-aggregation* approach.

		CT		**MRI**			
		OASIS-3	**RSNA Subset**	**ADNI1**	**OASIS-3**	**Proprietary-800**	
		AD	ICH	AD	AD	CBI	WMD
MIL without instance interaction	Max	$0.629_{\pm 0.049}$	$0.856_{\pm 0.034}$	$0.632_{\pm 0.048}$	$0.759_{\pm 0.024}$	$0.553_{\pm 0.063}$	$0.642_{\pm 0.059}$
	Mean	$0.707_{\pm 0.088}$	$0.904_{\pm 0.005}$	$0.693_{\pm 0.017}$	$0.809_{\pm 0.029}$	$0.620_{\pm 0.034}$	$0.627_{\pm 0.070}$
	ABMIL	$0.648_{\pm 0.100}$	$0.839_{\pm 0.069}$	$0.640_{\pm 0.085}$	$0.807_{\pm 0.023}$	$0.599_{\pm 0.038}$	$0.664_{\pm 0.066}$
MIL with instance interaction	SmAP	$0.671_{\pm 0.053}$	$0.900_{\pm 0.023}$	$0.697_{\pm 0.062}$	$0.796_{\pm 0.026}$	$0.602_{\pm 0.016}$	$0.664_{\pm 0.063}$

Appendix D. AUPRC Results

Table D.1: Test AUPRC on OASIS-3 MRI. All MIL methods use the *embedding-aggregation* approach.

	MIL without instance interaction			*MIL with instance interaction*	
	Max	**Mean**	**ABMIL**	**TransMIL**	**SmAP**
ViT-B/16	$0.562_{\pm 0.081}$	$0.613_{\pm 0.079}$	$0.555_{\pm 0.028}$	$0.498_{\pm 0.085}$	$0.570_{\pm 0.065}$
ConvNeXt-Tiny	$0.562_{\pm 0.086}$	$0.531_{\pm 0.037}$	$0.531_{\pm 0.037}$	$0.571_{\pm 0.115}$	$0.579_{\pm 0.064}$
MedSAM	$0.509_{\pm 0.102}$	$0.510_{\pm 0.013}$	$0.547_{\pm 0.007}$	$0.522_{\pm 0.027}$	$0.515_{\pm 0.044}$

Table D.2: Test AUPRC moderately-sized datasets. All MIL methods use the *embedding-aggregation* approach. †3DINO was pre-trained on ADNI1, OASIS-3, and half of the RSNA ICH full dataset.

		CT		**MRI**			
		OASIS-3	**RSNA Subset**	**ADNI1**	**OASIS-3**	**Proprietary-800**	
		AD	ICH	AD	AD	CBI	WMD
MIL without instance interaction	Max	$0.303_{\pm 0.095}$	$0.821_{\pm 0.032}$	$0.315_{\pm 0.066}$	$0.562_{\pm 0.081}$	$0.452_{\pm 0.111}$	$0.769_{\pm 0.114}$
	Mean	$0.347_{\pm 0.078}$	$0.872_{\pm 0.008}$	$0.326_{\pm 0.038}$	$0.613_{\pm 0.079}$	$0.413_{\pm 0.108}$	$0.776_{\pm 0.097}$
	ABMIL	$0.305_{\pm 0.057}$	$0.874_{\pm 0.015}$	$0.325_{\pm 0.030}$	$0.555_{\pm 0.028}$	$0.412_{\pm 0.113}$	$0.787_{\pm 0.070}$
MIL with instance interaction	TransMIL	$0.292_{\pm 0.061}$	$0.866_{\pm 0.001}$	$0.260_{\pm 0.025}$	$0.498_{\pm 0.085}$	$0.392_{\pm 0.122}$	$0.807_{\pm 0.090}$
	SmAP	$0.356_{\pm 0.029}$	$0.867_{\pm 0.029}$	$0.323_{\pm 0.062}$	$0.570_{\pm 0.065}$	$0.422_{\pm 0.114}$	$0.809_{\pm 0.092}$
	3D ResNet-18	$0.233_{\pm 0.051}$	$0.789_{\pm 0.037}$	$0.245_{\pm 0.053}$	$0.439_{\pm 0.027}$	$0.393_{\pm 0.087}$	$0.729_{\pm 0.104}$
	3D CNN for AD	*Variable-sized input not supported*		$0.303_{\pm 0.016}$	$0.636_{\pm 0.016}$	$0.520_{\pm 0.133}$	$0.817_{\pm 0.114}$
	3DINO	$^{\dagger}0.283_{\pm 0.039}$	$^{\dagger}0.887_{\pm 0.015}$	$^{\dagger}0.284_{\pm 0.032}$	$^{\dagger}0.530_{\pm 0.103}$	$0.597_{\pm 0.027}$	$0.693_{\pm 0.061}$

Table D.3: Test AUPRC on moderately-sized datasets. All MIL methods use the *prediction-aggregation* approach.

		CT		**MRI**			
		OASIS-3	**RSNA Subset**	**ADNI1**	**OASIS-3**	**Proprietary-800**	
		AD	ICH	AD	AD	CBI	WMD
MIL without instance interaction	Max	$0.321_{\pm 0.134}$	$0.823_{\pm 0.029}$	$0.290_{\pm 0.039}$	$0.511_{\pm 0.082}$	$0.392_{\pm 0.103}$	$0.784_{\pm 0.102}$
	Mean	$0.328_{\pm 0.077}$	$0.876_{\pm 0.008}$	$0.334_{\pm 0.043}$	$0.618_{\pm 0.078}$	$0.418_{\pm 0.108}$	$0.761_{\pm 0.112}$
	ABMIL	$0.348_{\pm 0.038}$	$0.819_{\pm 0.076}$	$0.322_{\pm 0.079}$	$0.596_{\pm 0.046}$	$0.406_{\pm 0.104}$	$0.791_{\pm 0.095}$
MIL with instance interaction	SmAP	$0.330_{\pm 0.029}$	$0.888_{\pm 0.024}$	$0.366_{\pm 0.023}$	$0.579_{\pm 0.033}$	$0.404_{\pm 0.098}$	$0.788_{\pm 0.099}$

Table D.4: Test AUPRC on large datasets. All MIL methods use the *embedding-aggregation* approach.

		CT	**MRI**	
		RSNA-21,744	**Proprietary-10k**	
		ICH	CBI	WMD
MIL without instance interaction	Max	$0.859_{\pm 0.013}$	$0.073_{\pm 0.020}$	$0.711_{\pm 0.068}$
	Mean	$0.903_{\pm 0.011}$	$0.093_{\pm 0.022}$	$0.741_{\pm 0.059}$
	ABMIL	$0.903_{\pm 0.007}$	$0.078_{\pm 0.023}$	$0.772_{\pm 0.051}$
MIL with instance interaction	TransMIL	$0.912_{\pm 0.012}$	$0.088_{\pm 0.028}$	$0.773_{\pm 0.052}$
	SmAP	$0.910_{\pm 0.011}$	$0.102_{\pm 0.039}$	$0.773_{\pm 0.054}$
	3D CNN for AD	*Variable-sized input not supported*	$0.104_{\pm 0.031}$	$0.784_{\pm 0.053}$

Appendix E. Instance-Level Results

E.1. Attention Quality Metric Definitions

For each positive bag i ($y_i = 1$) with S_i instances, let $y_{ij} \in \{0, 1\}$ denote the instance-level label and a_{ij} the attention weight for instance j. We compute three metrics using attention weights as predictors of instance labels: (1) *attention correctness* (Liu et al., 2017) $\sum_{j=1}^{S_i} a_{ij} \cdot y_{ij}$, (2) AUROC treating a_{ij} as scores for predicting y_{ij}, and (3) AUPRC similarly. Each metric is computed per positive bag, then averaged across all positive bags in the test set.

E.2. Attention Visualizations

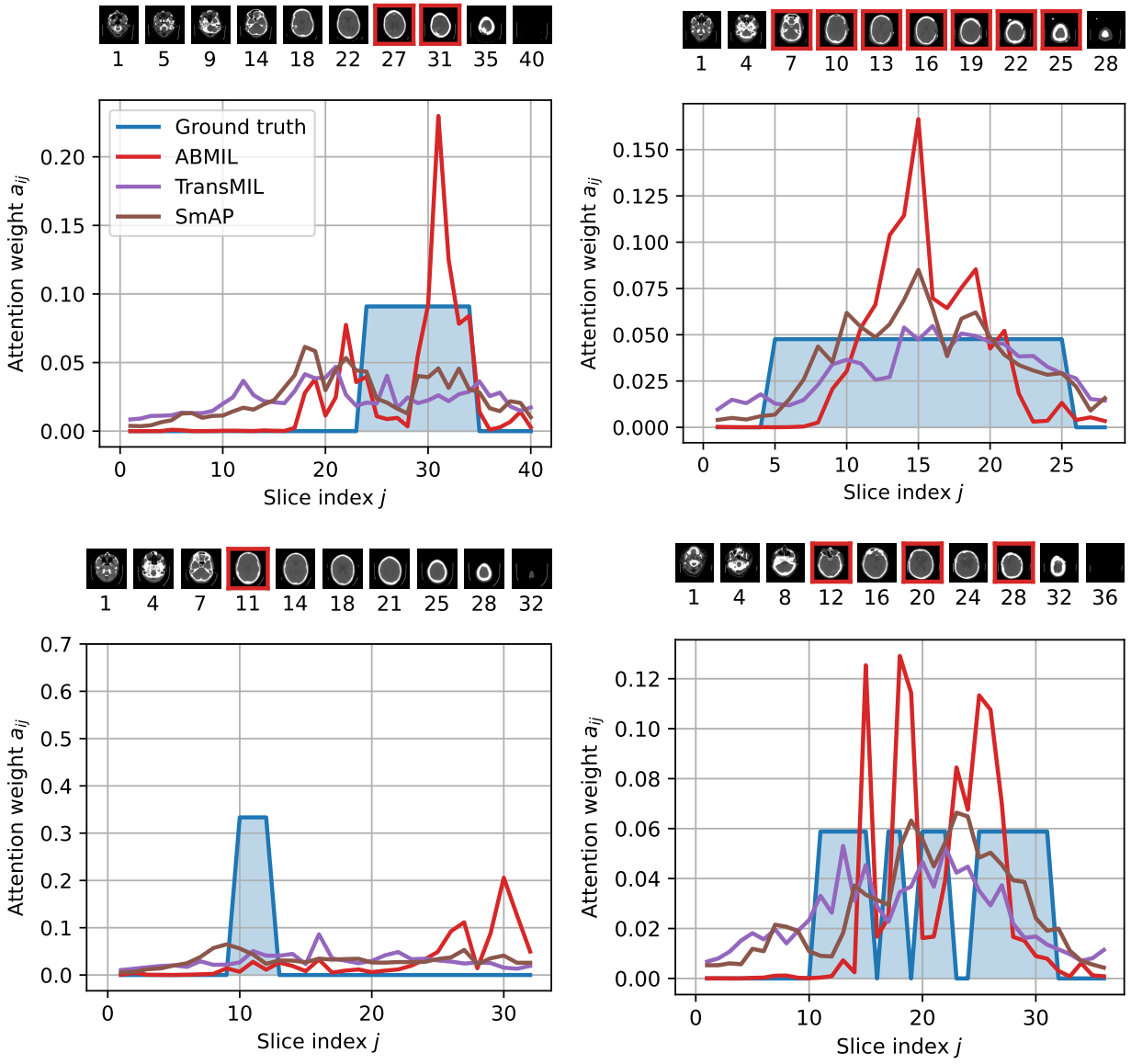

Figure E.1: Slice-level attention analysis on test scans. The ground truth has attention uniformly distributed on positive slices and positive slice images are outlined in red.

Appendix F. Shifted Mean MIL Dataset

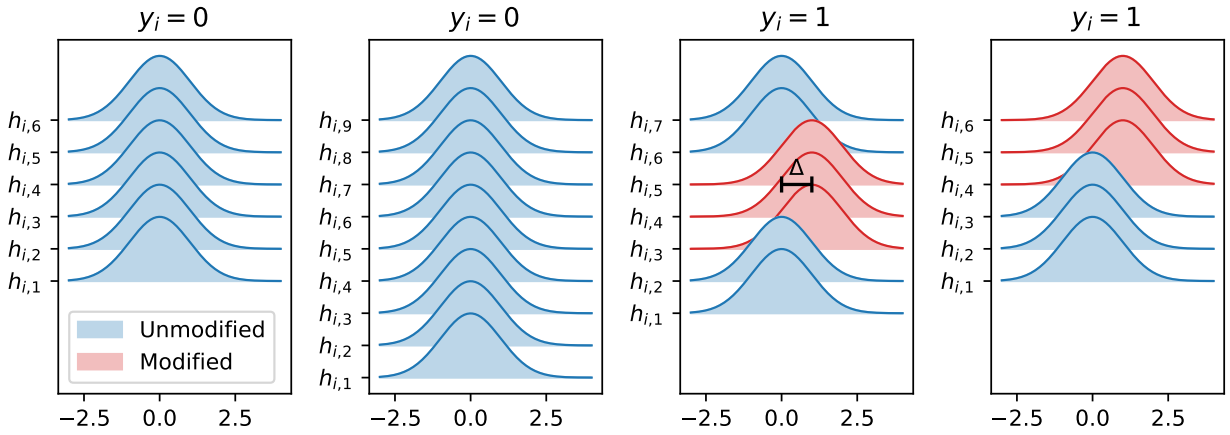

Figure F.1: Example data-generating distributions for a discriminative feature for negative ($y_i = 0$) and positive ($y_i = 1$) bags of S_i instances drawn from our Shifted Mean MIL dataset.

We propose a new data-generating process designed to mimic several key challenges in real-world multiple instance medical imaging tasks:

- Across the whole dataset, only some features are discriminative (K of M).
- For each positively-labeled bag, only a few instances are relevant (R of S_i) and they are adjacent in a known 1D listing of all S_i instances.
- Context matters. Adjacent instances together provide stronger statistical signal than any one relevant instance's discriminative feature value alone.

The generative process for bag i first draws the bag's binary label and the number of instances in the bag

$$y_i \sim \text{Bern}(q_+), \quad S_i \sim \text{Unif}(\{S_{\text{low}}, \ldots, S_{\text{high}}\}). \tag{6}$$

Next, for negative bags we sample all features k for all instances j independently from a common Gaussian:

$$h_{ijk} \mid y_i{=}0 \sim \mathcal{N}(\mu, \sigma^2). \tag{7}$$

For positive bags, most instances and features are sampled from this same Gaussian. However, for the K discriminative features, we select R adjacent instances (using u_i to denote the starting index) and sample these from a Gaussian with *shifted mean*:

$$u_i \mid S_i, y_i{=}1 \sim \text{Unif}(\{1, \ldots, S_i{-}R{+}1\}), \tag{8}$$

$$h_{ijk} \mid u_i, y_i{=}1 \sim \begin{cases} \mathcal{N}(\mu + \Delta, \sigma^2), & \text{if } j \in [u_i, u_i{+}R{-}1] \text{ and } k \text{ is discriminative} \\ \mathcal{N}(\mu, \sigma^2), & \text{otherwise.} \end{cases} \tag{9}$$

Here $\Delta > 0$ indicates the magnitude of shift for discriminative features. Setting $R > 1$ indicates that context helps. Given a fixed μ, bags drawn from this process are more challenging to classify (even with knowledge of the true process) when Δ is smaller, R is smaller, $\frac{K}{M}$ is smaller, and σ is larger.

This data-generating process is illustrated in Fig. F.1, depicting only one feature that is discriminative. In each positive bag, a different contiguous block of $R = 3$ instances draw from the shifted mean Gaussian. If future work wanted to model correlations between features within an instance, the sampling of vector h_{ij} in Eq. (9) could be modified to draw from a multivariate Gaussian with a non-diagonal covariance matrix.

Appendix G. Bayes Estimator

Given a data-generating process, a *Bayes estimator* is a decision rule that minimizes the posterior expected loss with respect to the data-generating distribution (DeGroot, 1970; Murphy, 2022). It is an oracle upper bound on performance. By comparing conventional or recent MIL methods to the Bayes estimator for our synthetic dataset, we can quantify how close they come to the best possible performance.

Given a bag h_i of S_i instances and assuming our data-generating process defined above, the Bayes estimator of the posterior probability is:

$$p(y_i{=}1 \mid h_i, S_i) = \frac{p(h_i \mid S_i, y_i{=}1)p(S_i)p(y_i{=}1)}{p(h_i \mid S_i, y_i{=}0)p(S_i)p(y_i{=}0) + p(h_i \mid S_i, y_i{=}1)p(S_i)p(y_i{=}1)}. \tag{10}$$

Each term on the right-hand side can be computed in closed-form. The class-conditional likelihood for the negative class factors over instances:

$$p(h_i \mid S_i, y_i{=}0) = \prod_{j=1}^{S_i} \prod_{k=1}^{M} \mathcal{N}(h_{ijk} \mid \mu, \sigma^2). \tag{11}$$

Each positive bag has a latent segment of R consecutive relevant instances. The class-conditional likelihood for the positive class marginalizes out the unknown index u:

$$p(h_i \mid S_i, y_i{=}1) = \sum_{u=1}^{S_i-R+1} \left[p(u \mid S_i, y_i{=}1) \prod_{j=1}^{S_i} \prod_{k=1}^{M} p(h_{ijk} \mid u, y_i{=}1) \right]. \tag{12}$$

Eq. (8) and (9) provide the necessary PDF values to evaluate the right hand side.