TopoCAM: ROI-Driven Topological Signatures in 3D Medical Imaging

Brighton Nuwagira BRIGHTON.NUWAGIRA@UTDALLAS.EDU
University of Texas at Dallas, USA

Philmore Koung PHILMORE.KOUNG@UTSOUTHWESTERN.EDU
University of Texas Southwestern Medical Center, USA

Baris Coskunuzer COSKUNUZ@UTDALLAS.EDU
University of Texas at Dallas, USA

Abstract

Accurate classification of 3D medical images is challenging due to the high dimensionality of volumetric data and the scarcity of well-annotated clinical datasets. We propose a hybrid framework that couples explainable deep learning with topological data analysis (TDA). First, we compute layer-weighted Grad-CAM across multiple network layers, upsample and normalize the maps to the input grid, and threshold them to produce a binary region-of-interest (ROI) mask. We then apply this mask to the input volume to obtain a segmented image that suppresses irrelevant anatomy while preserving clinically salient structures. Within these attention-derived ROIs and segmented images, we compute cubical persistent homology to derive compact topological descriptors that capture diagnostically meaningful features. Across both 3D volumes and 2D medical imaging benchmarks, this segmentation-guided TDA pipeline surpasses strong 3D CNN and Transformer baselines, yielding higher accuracy and improved robustness in limited-data settings while providing localized, interpretable evidence for clinical decision support.

Keywords: Grad-CAM, Medical Image Analysis, Topological Data Analysis, Cubical Persistence, Computer-aided Diagnosis, MedMNIST,

Data and Code Availability Our datasets are publicly available and links are provided in Section 4. Our implementation is available at

https://github.com/BrightonNuwagira/TopoCAM

Institutional Review Board (IRB) Our study does not require IRB approval.

1. Introduction

Accurate and efficient classification of 3D medical images remains a significant challenge for current deep learning methods. While convolutional neural networks (CNNs) and Vision Transformers (ViTs) have achieved remarkable successes in 2D image analysis tasks, translating these successes into the domain of volumetric medical scans (e.g., MRI, CT) has proven difficult. The computational complexity associated with processing high-dimensional 3D medical images, combined with the frequent scarcity of adequately annotated datasets, limits the performance and practicality of even advanced 3D architectures in clinical scenarios (Litjens et al., 2017; Shen et al., 2017). This gap highlights the need for methods that can deliver strong diagnostic performance in 3D under realistic data and resource constraints.

Topological data analysis (TDA), particularly persistent homology, has recently emerged as a promising alternative or complementary approach for medical imaging due to its intrinsic robustness, interpretability, and ability to succinctly encode geometric structures (Brito-Pacheco et al., 2025; Singh et al., 2023). Persistent homology captures shape and connectivity patterns through concise topological summaries, offering robust and interpretable descriptors even with limited data (Nuwagira et al., 2025). Yet, naive application of TDA methods to entire volumetric datasets suffers from significant noise introduced by irrelevant anatomical structures, thereby diluting diagnostic signals and limiting predictive performance.

Motivated by these complementary strengths and limitations, we propose a novel hybrid approach that integrates the localization strengths of explainable deep learning methods with the robustness of

© 2025 B. Nuwagira, P. Koung & B. Coskunuzer.

topological data analysis (TDA). Specifically, our framework utilizes a multi-scale Grad-CAM attention mechanism (Selvaraju et al., 2017) to efficiently isolate clinically relevant regions of interest (ROIs) within volumetric scans, effectively filtering out irrelevant anatomical structures. We then apply cubical persistent homology exclusively to these ROIs, generating concise topological descriptors that capture diagnostically meaningful geometric and structural features.

Our framework explicitly bridges the gap between visual attention-driven deep learning and rigorous topological summarization, addressing critical weaknesses inherent in standalone DL and naive TDA methods. We validate our approach comprehensively on multiple challenging 3D and 2D medical imaging benchmarks covering diverse anatomical targets and imaging modalities. Experimental results consistently show that our model outperforms leading 3D CNN and transformer architectures, achieving state-of-the-art diagnostic accuracy, robustness, and computational efficiency, even under limited data conditions. **Our contributions are**

- A novel hybrid framework integrating Grad-CAM-driven ROI localization with cubical persistent homology for high-performance medical image classification.
- An efficient approach that significantly improves diagnostic accuracy and robustness by focusing topological analysis on clinically relevant anatomical regions identified via deep attention.
- Extensive empirical evaluation on diverse 3D and 2D medical imaging datasets, demonstrating consistent performance gains over state-of-the-art CNN and transformer baselines in realistic, limited supervision scenarios.

Our findings demonstrate the promise of combining topological and deep learning methods to enable accurate, reliable, interpretable, and clinically relevant solutions in medical imaging diagnostics.

2. Background

2.1. Related Work

Grad-CAM. While CNNs have achieved state-of-the-art performance in image classification, their lack of interpretability limits deployment in high-stakes applications. To address this, Selvaraju et al. (Selvaraju et al., 2017) introduced *Grad-CAM*, which generates class-discriminative localization maps by leveraging the gradient information of target classes flowing into the final convolutional layers. These visual explanations have enabled model introspection, error analysis, and guided region-of-interest (ROI) selection in complex visual domains. Several variants have extended Grad-CAM to improve localization and robustness. Grad-CAM++ (Chattopadhyay et al., 2018) incorporates higher-order derivatives for better handling of multiple object instances and diffuse activations. Score-CAM (Wang et al., 2020) replaces gradient dependence with forward-passed class scores to yield sharper and more faithful saliency maps. Grad-CAM techniques have found extensive use in medical image analysis, where explainability is critical. For instance, they have been used to localize retinal lesions in fundus images (Zhao et al., 2024) and to segment tumor regions in CT scans (Schlemper et al., 2019). During the COVID-19 pandemic, Grad-CAM++ was employed to visualize diagnostic cues for pneumonia and COVID-19 from chest X-rays (Karim et al., 2020). A recent study (Suara et al., 2023) provides an overview of explainable deep learning methods in medical imaging, emphasizing the role of Grad-CAM in enhancing interpretability and diagnostic reliability.

Topological Data Analysis in Medical Imaging. Over the past two decades, topological methods, most notably persistent homology (PH), have demonstrated remarkable efficacy in pattern recognition for image and shape analysis. In medical imaging, PH-based techniques have driven advances in characterizing cellular development processes (McGuirl et al., 2020), detecting tumor structures (Crawford et al., 2020), quantifying features in histopathological assays (Qaiser et al., 2019; Yadav et al., 2023) and interpreting genomic profiles (Rabadán and Blumberg, 2019). Comprehensive reviews of TDA techniques in biomedical contexts can be found in (Skaf and Laubenbacher, 2022; Singh et al., 2023). More recently, integrating topological priors into deep learning frameworks has gained traction (Papamarkou et al., 2024), with evidence that topological descriptors enhance convolutional neural network performance in tasks such as image segmentation (Santhirasekaram et al., 2023; Stucki et al., 2023). Concurrently, persistent topological signatures have emerged as critical biomarkers in diagnostic contexts across

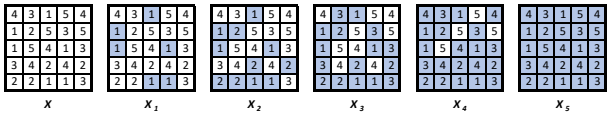

Figure 1: For the 5×5 image $\mathcal{X}$, **the sublevel filtration** is the sequence of binary images $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \mathcal{X}_3 \subset \mathcal{X}_4 \subset \mathcal{X}_5$.

various clinical domains (Somasundaram et al., 2021; Hajij et al., 2021; Khramtsova et al., 2022). These developments underscore the growing importance of topological insights in advancing deep learning–based medical image analysis.

2.2. Cubical Persistence

Persistent homology (PH) is commonly applied to image data through cubical persistence, where a grayscale image $\mathcal{X} \in \mathbb{R}^{r \times s}$ is treated as a piecewise-constant function $\gamma : \Delta_{ij} \mapsto \gamma_{ij}$ defined over pixel regions. Using an increasing sequence of intensity thresholds $0 = t_1 < t_2 < \cdots < t_N = 255$, one constructs a sublevel filtration of cubical complexes $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$, where $\mathcal{X}_n = \{\Delta_{ij} \mid \gamma_{ij} \leq t_n\}$. At each step, pixels are activated when their grayscale value falls below the corresponding threshold, producing a nested sequence of binary images that encodes topological changes in structure (see Figure 1). While we describe the construction for 2D grayscale images for clarity, the same filtration process extends naturally to color and volumetric data by applying similar operations across channels or spatial dimensions (Coskunuzer and Akçora, 2024; Brito-Pacheco et al., 2025).

PH then encodes the evolution of connected components, loops, and higher-dimensional voids in a persistence diagram (PD), which records the birth time b_σ and death time d_σ of each k-dimensional homological feature σ. Formally, $\mathrm{PD}_k(\mathcal{X}) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\mathcal{X}_n) \text{ for } b_\sigma \leq t_n < d_\sigma\}$, so that each point (b_σ, d_σ) reflects the thresholds at which the feature σ appears and disappears.

Since persistence diagrams are unordered multisets of point pairs $\{(b_\sigma, d_\sigma)\} \subset \mathbb{R}^2$, a vectorization step is necessary to enable their effective use in ML models. A simple yet effective approach is the Betti function $\beta_k(t_n)$, which records the number of k-dimensional topological features (e.g., connected components or loops) alive at each threshold t_n. Evaluating β_k across all thresholds yields a vector representation $\boldsymbol{\beta}_k = [\beta_k(t_1), \ldots, \beta_k(t_N)]$, offering a compact and interpretable topological summary. Alternative vectorizations, such as persistence images (Adams et al., 2017), persistence landscapes (Bubenik and Dłotko, 2017), silhouettes (Chazal et al., 2014), and kernel-based embeddings (Ali et al., 2023), provide richer feature encodings, often at increased computational cost. In this work, we primarily adopt Betti-sequence embeddings due to their efficiency and natural compatibility with sequence-based architectures such as transformers.

3. TopoCAM: ROI focused Topological Signatures

We propose a modular framework that combines volumetric feature encoding, explainable attention, and topological summarization for efficient and interpretable 3D medical image classification under limited supervision. Let $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{M}$ denote a training set of volumetric scans $\mathbf{X}_i \in \mathbb{R}^{D \times H \times W}$ with corresponding labels $y_i \in \{1, \ldots, C\}$. Our pipeline is structured as a composition of four stages:

$$\hat{y} = g_\psi \circ \phi \circ A_{w,\tau} \circ f_\theta(\mathbf{X}),$$

where f_θ extracts multi-resolution volumetric features, $A_{w,\tau}$ constructs a spatial attention mask via weighted Grad-CAM, ϕ computes topological summaries from the attention-guided segmented volume, and g_ψ is a lightweight classifier (See Fig. 2). Below, we detail each component and its design motivation.

Volumetric Feature Encoding. We employ a pretrained 3D ResNet-18 backbone f_θ (`r3d_18`) with weights initialized from large-scale video dataset pretraining (Kinetics-400) for volumetric data analysis. This mirrors our approach for 2D images, where we utilize a standard 2D ResNet-18 backbone pretrained on ImageNet, ensuring architectural consistency across both dimensional domains while leveraging appropriate pretraining strategies for each modality. This initialization provides strong spatiotemporal priors that effectively mitigate overfitting in our limited clinical data setting. The model retains its stem and four residual stages (layer 1–4) to encode volumetric input $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ (replicated to 3 channels if needed) into a hierarchy of spatial features $\{A^l\}_{l \in \mathcal{L}}$, where $A^l \in \mathbb{R}^{C_l \times D_l \times H_l \times W_l}$[1].

1. We use $\mathcal{L} = \{\texttt{layer2}, \texttt{layer3}, \texttt{layer4}\}$ for multi-scale features.

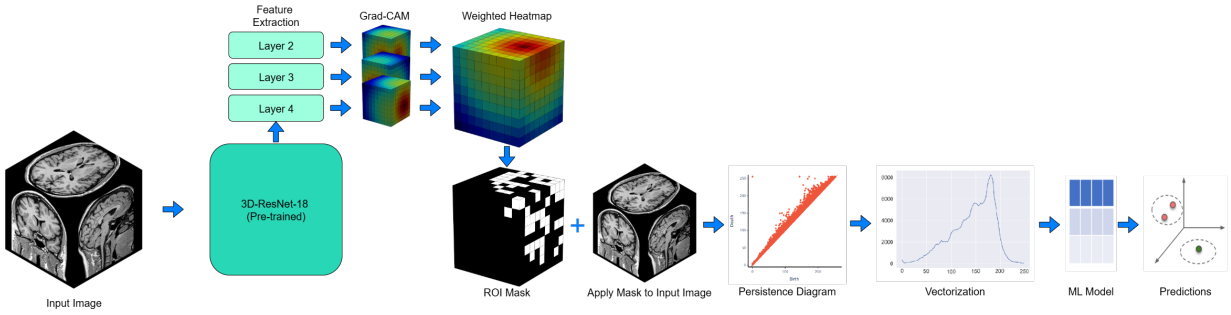

Figure 2: **TopoCAM Flowchart.** The input volume is processed by a 3D CNN (e.g., 3D-ResNet) to extract volumetric features, followed by explainable ROI localization using Grad-CAM. Topological features are then computed from the localized regions using persistent homology. These features are summarized and passed to a classifier to produce the final prediction.

The model is fine-tuned for 100 epochs using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss. The relatively shallow depth of `r3d_18` maintains computational efficiency during both fine-tuning and inference while leveraging transfer learning to achieve robust feature extraction.

Multi-Scale Grad-CAM Attention. To guide downstream topological analysis toward clinically relevant structures, we employ a multi-layer attention mechanism based on Grad-CAM (Selvaraju et al., 2017). For a given input volume and its predicted class, we extract activation maps from the deeper target convolutional layers $\mathcal{L} = \{\text{layer2}, \text{layer3}, \text{layer4}\}$, which capture more complex and class-specific features. For each layer $l \in \mathcal{L}$, the algorithm computes a coarse localization map by first obtaining neuron importance weights α_c^l for each channel c via global average pooling of the gradients:

$$\alpha_c^l = \frac{1}{D_l H_l W_l} \sum_{d,h,w} \frac{\partial S_y}{\partial A_{c,d,h,w}^l},$$

where S_y is the logit score for class y. A weighted combination of activations followed by a ReLU yields the attention map for each layer:

$$M^l = \text{ReLU}\left(\sum_c \alpha_c^l A_c^l\right).$$

This results in multi-resolution maps $\{M^2, M^3, M^4\}$ highlighting discriminative regions across different abstraction levels.

A key innovation is our data-driven fusion of these maps. Rather than using a simple average, we optimize the fusion coefficients $\mathbf{w} = [w_2, w_3, w_4]$ to maximize clinical relevance:

$$M = \sum_{l=2}^{4} w_l \cdot \text{Upsample}(M^l), \quad \text{subject to } \sum_l w_l = 1.$$

The weights $\mathbf{w}$ are optimized by maximizing the Area Under the Curve (AUC) of the fused map M. This ensures the final attention map best identifies structures predictive of the class label. The optimized map M is normalized and thresholded at level τ to produce a binary mask $\mathbf{1}_{M \geq \tau}$. This mask is *applied directly to the original scan* $\mathbf{X}$ to obtain an attention-guided segmented subvolume:

$$\mathbf{X}' = \mathbf{X} \odot \mathbf{1}_{M \geq \tau}.$$

Topological Feature Extraction. Cubical persistence (Section 2.2) captures the intrinsic shape and connectivity of structures in the segmented volume $\mathbf{X}'$. We define a sublevel filtration over grayscale intensities using N thresholds $0 = t_1 < \cdots < t_N = \max(\mathbf{X}')$, producing a sequence of cubical complexes

$$F_n = \{v \mid \mathbf{X}'(v) \leq t_n\}, \quad n = 1, \ldots, N.$$

For each homology dimension $k \in \{0, 1, 2\}$, we compute the persistence diagram $\text{PD}_k(\mathbf{X}') = \{(b_i^k, d_i^k)\}$, where (b, d) denotes the birth and death of a k-dimensional topological feature across the filtration. These diagrams are vectorized via Betti curves:

$$\beta_k(t_n) = \text{rank}\, H_k(F_n), \; \boldsymbol{\beta}_k = [\beta_k(t_1), \ldots, \beta_k(t_N)] \in \mathbb{N}^N$$

We concatenate $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, normalize by the number of activated voxels in $\mathbf{X}'$, and append the voxel count itself, resulting in a compact topological descriptor $\phi(\mathbf{X}') \in \mathbb{R}^{3N+1}$.

Classification and Optimization. The topological vector $\phi(\mathbf{X}')$ is passed to a multilayer perceptron $g_\psi : \mathbb{R}^{3N+1} \to [0,1]^C$ trained with cross-entropy loss. While the backbone f_θ remains fixed, we jointly optimize the Grad-CAM fusion weights $\{w_l\}$, threshold τ, and classifier parameters ψ to minimize the classification objective:

$$\min_{\psi,\tau,\{w_l\}} \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}\left(g_\psi\left(\phi(\mathbf{X}_i')\right), y_i\right).$$

This optimization is performed using black-box strategies (i.e differential evolution) due to the non-differentiable nature of topological features. This formulation ensures that the attention mechanism is explicitly guided by task performance, and that topology is extracted from semantically meaningful segmented regions of the original scan.

In summary, our method is interpretable by design, focusing analysis on class-discriminative subregions of the *original* volume, summarizing their geometric complexity via persistent homology, and enabling robust classification with minimal supervision. Each module contributes inductive bias: volumetric priors from pretraining, attention from Grad-CAM, and topological abstraction from persistent homology.

4. Experiments

4.1. Experimental Setup.

Datasets. We validate our framework on a diverse collection of eight 3D and two 2D medical imaging benchmarks to demonstrate its generality across modalities and anatomies. Five 3D datasets are drawn from MedMNIST3D (Yang et al., 2023), namely NoduleMNIST3D (Armato et al., 2011), AdrenalMNIST3D (Yang et al., 2022), FractureMNIST3D (Jin et al., 2020), VesselMNIST3D (Yang et al., 2020) , and SynapseMNIST3D (Wei et al., 2020), while the Harvard OCT (Luo et al., 2023) dataset provides volumetric optical coherence tomography. We further include two brain MRI corpora, BraTS 2019 (Bakas et al., 2017, 2018) and BraTS 2021 (Baid et al., 2021), for glioma classification and prediction of the O[6]-methylguanine-DNA methyltransferase (MGMT) promoter methylation status. For 2D evaluation, we use two MedMNIST2D sets: BreastMNIST (Al-Dhabyani et al., 2020) and PneumoniaMNIST (Liu et al., 2022). Table 1 summarizes their key characteristics.

Table 1: Summary of 2D and 3D datasets.

Dataset	Dim.	Modality	Class	# Images
NoduleMNIST3D	3D	CT (lung nodules)	2	1,633
AdrenalMNIST3D	3D	CT (adrenal glands)	2	1,584
FractureMNIST3D	3D	CT (bone fractures)	3	1,370
VesselMNIST3D	3D	MRA (vasculature)	2	1,908
SynapseMNIST3D	3D	EM (synapses)	2	1,759
Harvard OCT	3D	OCT (retinal volumes)	2	1,000
BraTS 2019	3D	MRI (brain tumors)	2	335
BraTS 2021	3D	MRI (brain tumors)	2	585
BreastMNIST	2D	Ultrasound	2	780
PneumoniaMNIST	2D	Chest X-Ray	2	5,856

Implementation. We developed a unified deep learning pipeline, `TopoCAM`, for both 2D and 3D medical image classification. The pipeline integrates a convolutional backbone, Grad-CAM–based attention to identify salient regions, persistent homology for topological feature computation, and a lightweight MLP classifier for the final prediction. For 2D tasks, we used a ResNet18 backbone pretrained on ImageNet (with grayscale inputs replicated to three channels), and for 3D tasks, a 3D ResNet18 (`R3D_18`) pretrained on Kinetics-400. The backbones were fine-tuned on the target dataset, and their feature maps were used to generate multi-scale Grad-CAM attention maps. Inputs were resized to 224×224 (2D) or standardized to $64 \times 64 \times 64$ (3D). The attention maps were then used to segment relevant structures, from which topological features were extracted to train the downstream classifier.

Grad-CAM maps from layers 2–4 were fused with learned weights optimized via differential evolution (population size 15, 50 iterations) to maximize AUC on validation data. The fused maps were thresholded at 0.6 to obtain binary masks highlighting discriminative regions. Cubical persistent homology was then computed on these masks. We extracted Betti numbers for dimensions 0–2 (3D) or 0–1 (2D) and converted persistence diagrams into Betti curve vectors with 50 bins using Giotto-TDA. Curves were normalized by nonzero pixels/voxels, and this count was appended as an additional feature.

The resulting topological feature vectors were classified using a lightweight multilayer perceptron (MLP). A consistent MLP architecture was employed for both 2D and 3D datasets, comprising two hidden layers with 128 and 64 units, respectively. This design balances expressiveness and computational efficiency, enabling robust feature abstraction across modalities. All hidden layers employed ReLU activation functions. The models were trained for 100 epochs using the Adam optimizer with a cross-entropy loss func-

tion. No gradients were propagated to the ResNet backbones. We followed official train/val/test splits for MedMNIST2D/3D and Harvard OCT, and used 70:10:20 splits for BraTS 2019 and BraTS 2021.

Performance Metrics. We evaluated model performance using accuracy and ROC-AUC as primary metrics, and additionally report F1 score, specificity, and sensitivity in Appendix.

Computational Complexity and Runtime. Our framework is designed for computational efficiency by leveraging modular components and focusing processing on diagnostically relevant regions. The volumetric feature extraction stage relies on a lightweight 3D ResNet-18 backbone, which maintains a manageable memory footprint and runtime compared to deeper architectures. The multi-scale Grad-CAM attention mechanism requires only a small number of backward passes and channel-weighted feature aggregations at selected layers, resulting in minimal overhead beyond standard forward and backward propagation. Crucially, by applying the attention-guided mask, we restrict topological feature extraction to a small fraction of the input volume, significantly reducing the cost of cubical persistent homology, which otherwise scales with the number of voxels and filtration thresholds. The complexity of computing Betti curves for a masked subvolume is $\mathcal{O}(NV')$, where N is the number of filtration steps and V' is the number of voxels within the ROI, typically much smaller than the full scan. The final classification stage uses a shallow MLP with negligible cost. Overall, our pipeline scales linearly with the number of training examples and is well suited for GPU acceleration. By integrating attention-driven focus and topological summarization, our method delivers robust performance with substantially reduced computational and memory requirements compared to conventional 3D deep learning approaches that process entire volumes. As an example, our complete pipeline for the VesselMNIST3D dataset which includes Grad-CAM generation, Betti vector extraction, and MLP-based classification was executed on a single node equipped with an NVIDIA H100 GPU. The job was submitted using SLURM with one node and one task, and completed in approximately 23 hours and 33 minutes. This highlights the practicality of our approach even on large volumetric datasets.

4.2. Results

3D Baselines. We compare against seven strong volumetric classification architectures from both the convolutional and transformer literatures. To ensure fair comparison, we use open source implementations which can be found in our code page. First, we include three standard 3D-CNNs, R3D-18, MC3-18, and R(2+1)D-18, which employ full 3D convolutions, mixed 3D/2D convolutions, and factorized (2D+1D) spatio-temporal filters, respectively (Tran et al., 2018). We also evaluate EfficientNet3D, a direct volumetric adaptation of the compound-scaled EfficientNet family (Tan and Le, 2019). To assess the benefits of global attention, we benchmark three transformer-style models: M3T, which fuses multi-plane, multi-slice transformer encoding with a 3D-CNN backbone (Jang and Hwang, 2022); 3D-CCT, which injects convolutional tokenization into a compact transformer architecture for volumetric inputs (Hassani et al., 2021); and 3D-ViT, a 3D adaptation of the original vision transformer (Dosovitskiy et al., 2020). Models were trained for 100 epochs with the Adam optimizer (learning rate = $1e^{-4}$); the checkpoint that achieved the highest validation AUC was subsequently evaluated on the test set.

2D Baselines. For 2D tasks, we benchmark seven pretrained backbones: ResNet-18 (He et al., 2016), DenseNet-121 (Huang et al., 2017), VGG-16 (Simonyan and Zisserman, 2015), EfficientNet-B0 (Tan and Le, 2019), and three transformer based models, DaViT (Ding et al., 2022), Swin Transformer (Liu et al., 2021), and MobileViT (Mehta and Rastegari, 2022). These span classic CNNs, a compound scaled network, and attention based architectures. All 2D baselines use the same training protocol and hyperparameters as the 3D setups.

Results. Table 2 shows that `TopoCAM` delivers strong gains across a broad set of 3D benchmarks. Out of eight 3D benchmark datasets, it attains the best AUC on four: Vessel3D, Nodule3D, FractureMNIST3D, and BraTS 2019, exceeding the strongest baselines by 1.7, 3.8, 6.3, and 7.6 points respectively. On Synapse3D and Adrenal3D, performance is close to the best CNN and transformer models, within 0.1 to 2.1 AUC. On Harvard OCT, `TopoCAM` achieves state-leading accuracy 81.4 with AUC 78.4. On BraTS 2021, it reaches AUC 64.7 and accuracy 58.9, remaining *significantly above* the *Topo-Original* baseline that computes topology over the full volume.

Table 2: **Performance on 3D Datasets.** AUC and accuracy of CNN, ViT-based, and topological models across five MedMNIST3D and three other 3D benchmarks. Models marked with * were trained from scratch, while all others used Kinetics-400 pretrained weights. Further performance metrics are provided in Table 9.

	Vessel3D		Synapse3D		Nodule3D		Fracture3D		Harvard OCT		Adrenal3D		BraTS 2019		BraTS 2021	
Model	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
Topo-Original	69.6	88.7	81.2	79.8	74.4	81.3	59.1	52.5	53.6	49.9	78.8	78.2	42.1	77.6	58.0	52.5
R3D-18	91.9	92.7	**98.5**	92.9	90.9	86.8	66.1	50.0	81.5	92.7	90.7	85.2	85.8	83.6	62.6	**64.4**
MC3-18	93.5	93.2	96.8	**94.3**	90.3	86.1	67.8	51.2	**93.5**	93.2	85.2	81.9	88.1	82.1	65.3	47.5
R(2+1)D-18	86.3	87.2	97.5	91.8	89.3	86.5	69.2	50.4	82.1	87.2	**97.5**	**91.8**	89.3	86.5	64.5	62.7
EfficientNet3D*	52.4	88.7	49.9	73.0	75.8	84.5	54.8	44.2	48.8	88.7	60.2	76.8	47.6	76.5	57.3	55.9
3D-CCT*	85.2	90.6	58.8	73.0	81.8	83.9	62.5	47.1	79.6	90.6	84.8	77.8	85.2	82.3	65.4	52.5
3D-ViT*	78.6	89.0	57.8	70.2	89.4	85.5	66.1	50.8	76.6	89.0	78.7	75.5	81.0	79.4	**74.7**	57.6
M3T	97.2	95.0	95.8	93.2	88.6	87.7	70.9	52.9	84.2	**95.0**	83.7	80.5	86.9	85.3	62.7	62.7
TopoCAM	**98.9**	**97.1**	98.4	81.0	**99.6**	**98.1**	**77.2**	**66.3**	78.4	81.4	95.4	91.6	**94.2**	**98.5**	64.7	58.9

Table 3: **Performance on 2D Datasets.** Performance of pre-trained CNN, ViT baselines, and topological models across two 2D medical imaging benchmarks. Further performance metrics are given in Table 10.

	Breast2D		Pneumo2D	
Model	**AUC**	**Acc.**	**AUC**	**Acc.**
Topo-Original	78.8	78.2	84.5	78.8
ResNet18	92.8	89.7	88.3	63.8
DenseNet121	87.8	87.8	97.5	71.3
VGG16	87.1	87.2	98.1	93.6
EfficientNetB0	87.5	85.9	83.5	71.5
DaViT	94.7	90.4	98.6	89.9
MobileViT	86.2	84.6	95.7	89.7
Swin v2	91.0	88.5	99.1	93.4
TopoCAM	**99.9**	**98.7**	**100**	**99.4**

These results support our central claim: full-volume topological analysis is confounded by irrelevant anatomy, whereas Grad-CAM-guided ROIs concentrate persistence signals on clinically meaningful structures. The consistently high performance across CT, MRI, and OCT indicates that attention-guided topology generalizes well across modalities. On 2D MedMNIST benchmarks (Table 3), `TopoCAM` *matches or exceeds* strong pretrained CNN and ViT baselines, underscoring applicability to both volumetric and planar settings. For context, our MedMNIST3D runs use 64×64×64 inputs, while results on the official webpage are reported for 28×28×28.

In summary, `TopoCAM` combines explainable attention with topological descriptors to improve predictive performance and interpretability in 3D medical image classification, with large gains on several challenging datasets and competitive results on the remainder.

4.3. Ablation Studies

We evaluate two design choices within our TopoCAM pipeline that computes topological descriptors on *segmented subvolumes* defined by a Grad-CAM mask. First, we study how the region-of-interest (ROI) binarization threshold t affects performance *when using TopoCAM with fixed weighted multi-layer Grad-CAM fusion*. Second, we compare using only the last Grad-CAM layer with the weighted multi-layer fusion when constructing the attention map that defines the ROI. Unless specified, topological features are extracted exclusively from voxels retained by the ROI mask.

Let $H^{(l)} \in [0,1]^{D\times H\times W}$ denote the Grad-CAM heatmap from layer l (upsampled and normalized). The single-layer condition uses $H_{\text{single}} = H^{(L)}$. The multi-layer condition forms a fused map $H_{\text{fused}} = \sum_{l\in\mathcal{L}} w_l \tilde{H}^{(l)}$ with nonnegative weights $\sum w_l = 1$ and normalized $\tilde{H}^{(l)}$. Given a chosen map $H \in \{H_{\text{single}}, H_{\text{fused}}\}$, we build a binary ROI via $M_t(x) = \mathbf{1}_{[H(x)\geq t]}$ and define the *segmented subvolume* $\Omega_t = \{x : M_t(x) = 1\}$; TopoCAM computes topology only on Ω_t.

For the threshold sensitivity in Tables 4 and 11, vary t. The results show that $t = 0.6$ yields the best trade-off between coverage and precision across representative datasets and is adopted in the final configuration.

To isolate the benefit of multi-scale attention, we also compare single-layer versus weighted multi-layer Grad-CAM in Tables 5 and 12. Weighted fusion consistently improves AUC and accuracy. These results suggest that earlier layers contribute fine-

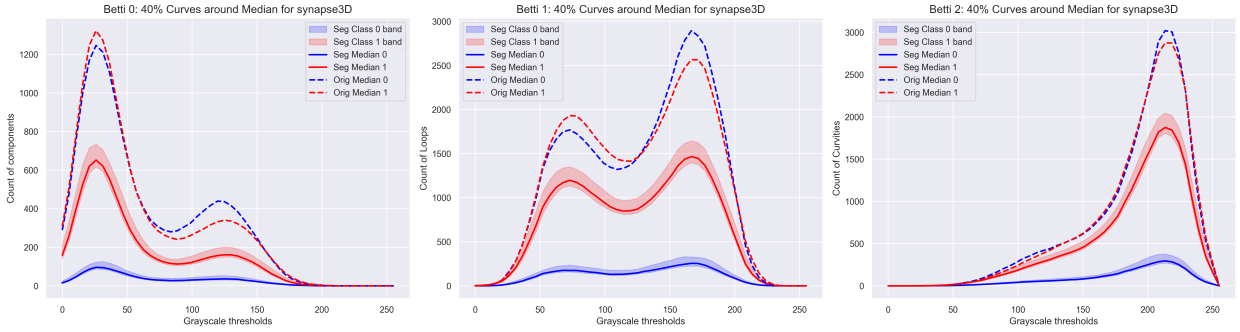

Figure 3: **Betti curves on ROIs.** For Synapse3D, classwise Betti-0 (left), Betti-1 (center), and Betti-2 (right) curves are shown. Solid lines denote medians within Grad-CAM ROIs with 40% confidence bands, while dotted lines are the corresponding medians computed on full volumes. Unlike the overlapping full-volume curves, ROI-based curves and bands show clear class separation.

Table 4: **ROI threshold** Impact of threshold t on TopoCAM measured on segmented subvolumes with $H = H_{\text{fused}}$.

Threshold t	Fracture3D		Breast2D		Vessel3D	
	AUC	Acc.	AUC	Acc.	AUC	Acc.
0.5	72.6	55.0	98.1	94.9	94.6	92.4
0.6	**77.2**	**63.3**	**99.9**	**98.7**	**98.9**	**97.1**
0.7	72.1	55.4	**99.9**	97.4	97.6	95.5

Table 5: **Grad-CAM Layers.** Performance comparison of TopoCAM using ROIs from the final Grad-CAM layer versus a weighted combination of all layers.

Configuration	Nodule3D		Adrenal3D		Pneum2D	
	AUC	Acc.	AUC	Acc.	AUC	Acc.
Single layer	97.2	97.1	88.7	85.2	99.8	95.7
Weighted layers	**99.6**	**98.1**	**95.4**	**91.6**	**100.0**	**99.4**

grained structural cues that help define a more reliable ROI for topology than the last layer alone.

To ensure a fair assessment against pretrained baselines, we evaluate TopoCAM under two conditions: with the 3D backbone trained from scratch and with Kinetics-400 pretrained weights. Notably, in Table 2, all baseline models were initialized with pretrained weights except 3D-ViT, 3D-CCT, M3T, and EfficientNet3D. As summarized in Table 6, pretraining the TopoCAM backbone consistently improves AUC and accuracy by facilitating faster convergence and more stable optimization. However, the major performance gain originates from TopoCAM's ROI-driven topological encoding, which refines feature learning by focusing topology computation on clinically salient regions rather than from weight initialization alone.

We performed an ROI generator ablation on three representative 3D datasets (VesselMNIST3D, SynapseMNIST3D, and NoduleMNIST3D) under identical training conditions, varying only the attention mechanism. As shown in Table 7, Grad-CAM consistently achieves the best overall performance, attaining the highest AUC on Vessel3D and synapse3D while remaining competitive on Nodule3D.

To ensure a fair comparison, we implemented a segment-then-analyze baseline on the BraTS 2019 dataset, which includes both segmentation masks and class labels. In this setup, a MONAI 3D U-Net (Cardoso et al., 2022) was used to generate tumor masks, from which Betti-vector features were extracted. Classification was then performed using the same MLP architecture as in TopoCAM to isolate the effect of segmentation. As shown in Table 8, this baseline yields notably lower performance compared to TopoCAM.

4.4. Discussion

Refining Topological Signatures with Grad-CAM ROIs. In Tables 2 and 3, the baseline *Topo-Original* computes Betti curves over the full volume, while *TopoCAM* improves specificity by first generating an attention map M via multi-layer Grad-CAM fusion. The scan is segmented into ROIs as $I_{\text{roi}} = I \odot \mathbf{1}_{\{M \geq \tau\}}$, and Betti curves for $k \in \{0, 1, 2\}$ are computed only on I_{roi}. These curves are normalized by ROI voxel count, which is also appended as a feature. Across datasets, *TopoCAM* consistently improves AUC and accuracy, showing that full-volume

Table 6: **Effect of Pretraining.** Comparison of TopoCAM trained from scratch (*) and with Kinetics-400 pretrained initialization on two representative 3D datasets. Models marked with * indicate that the backbone was trained from scratch, while the other used Kinetics-400 pretrained weights.

	Vessel3D					Adrenal3D				
Configuration	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**
TopoCAM*	94.6	95.4	88.4	99.1	90.4	95.0	94.3	90.5	94.1	94.0
TopoCAM	**98.9**	**97.1**	76.7	**99.7**	85.7	**95.4**	91.6	79.7	**95.2**	81.5

Table 7: **ROI Generator Ablation.** Comparison of ROI generation methods on three 3D datasets using identical training and evaluation settings. Grad-CAM shows superior AUC and computational efficiency.

ROI Generator	Vessel3D	Synapse3D	Nodule3D
Grad-CAM	**0.989**	**0.984**	0.996
Grad-CAM++	0.935	0.892	**0.997**
Score-CAM	0.901	0.881	0.854

Table 8: **Segment-then-Analyze Baseline.** Comparison between the segmentation-based Betti-vector pipeline (Segment+Betti) and TopoCAM on the BraTS 2019 dataset. TopoCAM achieves higher discrimination without voxel-level supervision.

Model	AUC	Acc.	Sens.	Spec.	F1
TopoCAM	**94.2**	**98.5**	**100.0**	**93.3**	**99.0**
Segment+Betti	83.0	88.6	86.8	46.2	78.5

topology is confounded by irrelevant anatomy, while Grad-CAM-guided ROIs concentrate persistence signals on diagnostically meaningful structures.

Interpretability. Our pipeline is interpretable at two levels: spatially, Grad-CAM highlights *where decisions are made*, producing human-readable ROIs; structurally, Betti curves summarize the *topological patterns* driving predictions. As shown in Figures 3 and 5, Betti curves of masked regions (solid) exhibit clearer class separation than unmasked ones (dashed). For Betti-0, masking suppresses background micro-components, leaving peaks tied to lesion and vessel fragmentation, while Betti-1 and Betti-2 peaks shift to clinically relevant intensity ranges with reduced overlap between classes. The 40% envelopes reveal lower variance and sharper discriminative peaks, and the t-SNE plots in Figures 4 and 6 further confirm tighter, more distinct clustering of ROI-based features. Together, these results show that `TopoCAM` decisions are both localized and structurally grounded, offering transparent rationale for classification.

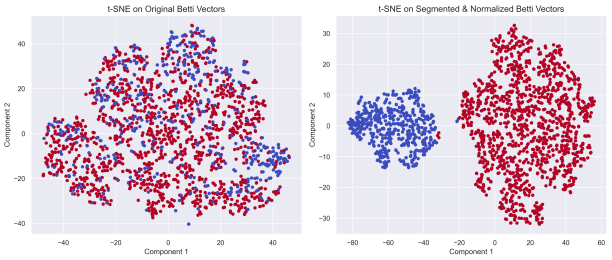

Figure 4: **t-SNE on Synapse3D.** Left: t-SNE of Betti vectors computed on the full volumes. Right: t-SNE of our normalized Betti vectors (TopoCAM) on GradCAM ROIs. TopoCAM features yield tighter, better-separated class clusters, consistent with the gains in Tables 2 and 3.

5. Conclusion

We presented `TopoCAM`, a hybrid framework that combines Grad-CAM–guided localization with cubical persistence to deliver accurate, interpretable medical image classification in both 2D and 3D. By restricting topological descriptors to task-relevant regions, `TopoCAM` consistently outperforms state-of-the-art CNN and transformer baselines across diverse benchmarks, achieving near-perfect AUC on several tasks and substantial gains on challenging datasets such as Harvard OCT, FractureMNIST3D, and BraTS 2021. Its design not only improves predictive performance but also enhances interpretability by linking visual attention to structural descriptors. Future directions include extending TopoCAM to multimodal imaging, adapting it to cross-institutional and domain-shifted datasets, exploring differentiable topological layers for end-to-end learning, and applying the framework to broader clinical tasks where reliability and transparency are essential.

Acknowledgments

This work was partially supported by National Science Foundation under grants DMS-2220613, and DMS-2229417. The authors acknowledge the Texas Advanced Computing Center (TACC) at UT Austin for providing computational resources that have contributed to the research results reported within this paper.

References

Henry Adams et al. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1):218–252, 2017.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. A survey of vectorization methods in topological data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Samuel G. Armato, Greg McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Bo Zhao, Denise R. Aberle, Ciira I. Henschke, Eric A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(3):915–931, 2011. doi: 10.1118/1.3528204.

Ujjwal Baid et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamidreza Akbari, Aristeidis Sotiras, Michael Bilello, Mary Rozycki, Justin S Kirby, et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Matthias Rempfler, Arianna Crimi, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Daniel Brito-Pacheco, Panos Giannopoulos, and Constantino Carlos Reyes-Aldasoro. Persistent homology in medical image processing: A literature review. *medRxiv*, pages 2025–02, 2025.

Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, pages 474–483, 2014.

Baris Coskunuzer and Cüneyt Gürcan Akçora. Topological methods in machine learning: A tutorial for practitioners. *arXiv preprint arXiv:2409.02901*, 2024.

Lorin Crawford, Anthea Monod, Andrew X Chen, Sayan Mukherjee, and Raúl Rabadán. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *Journal of the American Statistical Association*, 115(531): 1139–1150, 2020.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mustafa Hajij, Ghada Zamzmi, and Fawwaz Batayneh. Tda-net: fusion of persistent homology and deep learning features for covid-19 detection from chest x-ray images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4115–4119. IEEE, 2021.

Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *CVPR*, pages 20718–20729, 2022.

Liang Jin et al. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. *EBioMedicine*, 61:103026, 2020.

Md. Rezaul Karim, Till Döhmen, Michael Cochez, Oya Beyan, Dietrich Rebholz-Schuhmann, and Stefan Decker. Deepcovidexplainer: Explainable covid-19 diagnosis from chest x-ray images. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1034–1037, 2020. doi: 10.1109/BIBM49941.2020.9313304.

Ekaterina Khramtsova, Guido Zuccon, Xi Wang, and Mahsa Baktashmotlagh. Rethinking persistent homology for visual recognition. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pages 206–215. PMLR, 2022.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

Yan Luo, Min Shi, Yu Tian, Tobias Elze, and Mengyu Wang. Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20471–20482, 2023.

Melissa R McGuirl, Alexandria Volkening, and Björn Sandstede. Topological data analysis of zebrafish patterns. *Proceedings of the National Academy of Sciences*, 117(10):5113–5124, 2020.

Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022.

Brighton Nuwagira, Caner Korkmaz, Philmore Koung, and Baris Coskunuzer. Topological machine learning for low data medical imaging. In *Machine Learning for Health (ML4H)*, pages 824–838. PMLR, 2025.

Theodore Papamarkou, Tolga Birdal, Michael Bronstein, Gunnar Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Liò, Paolo Di Lorenzo, et al. Position paper: Challenges and opportunities in topological deep learning. *arXiv preprint arXiv:2402.08871*, 2024.

Talha Qaiser, Yee-Wah Tsang, Daiki Taniyama, Naoya Sakamoto, Kazuaki Nakane, David Epstein, and Nasir Rajpoot. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Medical image analysis*, 55:1–14, 2019.

Raúl Rabadán and Andrew J Blumberg. *Topological data analysis for genomics and evolution: topology in biology*. Cambridge University Press, 2019.

Ainkaran Santhirasekaram, Mathias Winkler, Andrea Rockall, and Ben Glocker. Topology preserving compositionality for robust medical image segmentation. In *CVPR*, pages 543–552, 2023.

Jo Schlemper, Ozan Oktay, Marcel Schaap, Matt Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient image regions. *Medical Image Analysis*, 53:197–207, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Yashbir Singh et al. TDA in Medical Imaging: Current state of the art. *Insights into Imaging*, 14(1): 1–10, 2023.

Yara Skaf and Reinhard Laubenbacher. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, page 104082, 2022.

Eashwar Somasundaram, Adam Litzler, Raoul Wadhwa, Steph Owen, and Jacob Scott. Persistent homology of tumor ct scans is associated with survival in lung cancer. *Medical physics*, 48(11):7043–7051, 2021.

N. Stucki et al. Topologically faithful image segmentation via induced matching of persistence barcodes. In *ICML*, 2023.

Subhashis Suara, Aayush Jha, Pratik Sinha, and Arif Ahmed Sekh. Is grad-cam explainable in medical images? In *International Conference on Computer Vision and Image Processing*, pages 124–135. Springer, 2023.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPR workshops*, pages 24–25, 2020.

Donglai Wei et al. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *MICCAI*, volume 12264 of *Lecture Notes in Computer Science*, pages 66–76. Springer, 2020.

Ankur Yadav, Faisal Ahmed, Ovidiu Daescu, Reyhan Gedik, and Baris Coskunuzer. Histopathological cancer detection with topological signatures. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1610–1619. IEEE, 2023.

Jiancheng Yang, Rui Shi, Udaranga Wickramasinghe, Qikui Zhu, Bingbing Ni, and Pascal Fua. Neural annotation refinement: Development of a new 3d dataset for adrenal gland analysis. *arXiv preprint arXiv:2206.15328*, 2022.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. https://medmnist.com.

Xi Yang, Ding Xia, Taichi Kin, and Takeo Igarashi. Intra: 3d intracranial aneurysm dataset for deep learning. In *CVPR*, pages 10154–10163, 2020. URL https://github.com/intra3d2019/IntrA.

Zhongchen Zhao, Huai Chen, Yu-ping Wang, Deyu Meng, Qi Xie, Qi Yu, and Lisheng Wang. Retinal disease diagnosis with unsupervised grad-cam guided contrastive learning. *Neurocomputing*, 593: 127816, 2024.

Appendix

Further Performance Metrics. While the main text (Tables 2 and 3) focuses on accuracy and AUC for clarity, here (Tables 9 and 10) we provide additional evaluation metrics, specificity, sensitivity, and F1 score, for both our models and the baselines. Furthermore, in Tables 11 and 12 below, we give the additional performance metrics for our ablation studies given in the main text.

Table 9: **Performance Metrics on 3D Datasets.** Diagnostic performance of CNN, ViT-based, and topological models evaluated across four 3D medical imaging datasets.

	VesselMNIST3D					SynapseMNIST3D					NoduleMNIST3D					FractureMNIST3D				
Model	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**
Topo-Original	69.6	88.7	2.3	**99.7**	4.4	81.2	79.8	93.0	44.2	87.1	74.4	81.3	40.6	91.9	47.3	59.1	43.3	42.4	69.7	40.5
R3D-18	91.9	92.7	51.2	97.9	61.1	**98.5**	92.9	**97.3**	81.0	95.2	90.9	86.8	68.8	91.5	68.2	66.1	50.0	46.6	72.7	46.5
MC3-18	93.5	93.2	55.8	97.9	64.9	96.8	**94.3**	96.9	87.4	**96.1**	90.3	86.1	70.3	90.2	67.7	67.8	51.2	50.8	74.5	50.2
R(2+1)D-18	86.3	87.2	51.2	91.7	47.3	97.5	91.8	94.5	84.2	94.4	89.3	86.5	64.1	92.3	66.1	69.2	50.4	48.2	73.0	47.2
EfficientNet3D	52.4	88.7	50.0	50.0	47.0	49.9	73.0	50.0	50.0	42.2	75.8	84.5	69.4	69.4	72.3	54.8	44.2	36.5	68.9	32.6
3D-CCT	85.2	90.6	64.2	64.2	68.4	58.8	73.0	50.0	50.0	42.2	81.8	83.9	65.0	65.0	68.1	62.5	47.1	39.5	70.3	35.9
3D-ViT	78.6	89.0	51.2	51.2	49.4	57.8	70.2	49.7	49.7	45.4	89.4	85.5	79.3	79.3	78.5	66.1	50.8	44.4	72.9	44.3
M3T	97.2	95.0	69.8	98.2	75.9	95.8	93.2	96.9	83.2	95.4	88.6	87.7	78.1	90.2	72.5	70.9	52.9	55.1	75.5	52.9
TopoCAM	**98.9**	**97.1**	**76.7**	**99.7**	**85.7**	98.4	81.0	74.3	**98.9**	85.1	**99.6**	**98.1**	**96.9**	**98.4**	**95.4**	**77.2**	**63.3**	**58.9**	**80.3**	**59.0**

	Harvard OCT3D					AdrenalMNIST3D					BRATS 2019					BRATS 2021				
Model	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**
Topo-Original	53.6	49.9	21.9	**99.7**	32.2	78.8	78.2	86.8	78.2	85.3	42.1	77.6	**100.0**	0.0	87.4	58.0	52.5	**100.0**	0.0	68.9
R3D-18	81.5	92.7	51.2	97.9	61.1	90.7	85.2	44.9	97.4	58.5	85.8	83.6	98.1	33.3	90.3	62.6	**64.4**	75.8	51.8	**69.1**
MC3-18	**93.5**	93.2	55.8	97.9	64.9	85.2	81.9	72.5	84.7	64.9	88.1	82.1	98.1	26.7	89.5	65.3	47.5	0.0	**100.0**	0.0
R(2+1)D-18	82.1	87.2	51.2	91.7	47.3	**97.5**	**91.8**	**94.5**	84.2	**94.4**	89.3	86.5	64.1	92.3	66.1	64.5	62.7	40.3	87.5	53.2
EfficientNet3D	48.8	88.7	50.0	50.0	47.0	60.2	76.8	50.0	50.0	43.5	47.6	76.5	50.0	50.0	43.3	57.3	55.9	50.0	50.0	35.9
3D-CCT	79.6	90.6	64.2	64.2	71.7	84.8	77.8	60.8	60.8	62.1	85.2	82.3	62.4	62.4	64.8	65.4	52.5	54.3	54.3	52.2
3D-ViT	76.6	89.0	51.2	51.2	49.4	78.7	75.5	64.8	64.8	65.0	81.0	79.4	65.9	65.9	66.7	**74.7**	57.6	50.0	50.0	36.6
M3T	84.2	**95.0**	**69.8**	98.2	75.9	83.7	80.5	20.3	**98.7**	32.6	86.9	85.3	84.5	84.5	78.6	62.7	62.7	62.6	62.6	62.6
TopoCAM	78.4	81.4	68.2	83.2	**77.4**	95.4	91.6	79.7	95.2	81.5	**94.2**	**98.5**	**100.0**	**93.3**	**99.0**	64.7	58.9	62.0	56.0	59.0

Table 10: **Performance Metrics on 2D Datasets.** Diagnostic performance of CNN, ViT-based, and topological models evaluated across three 2D medical imaging datasets.

	BreastMNIST2D					PneumoniaMNIST2D				
Model	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**
Topo-Original	78.8	78.2	86.8	78.2	85.3	84.5	78.8	**94.6**	78.8	84.8
ResNet18	92.8	89.7	98.3	66.7	93.3	88.3	63.8	51.7	51.7	42.1
DenseNet121	87.8	87.8	92.1	76.2	91.7	97.5	71.3	61.8	61.8	59.7
VGG16	87.1	87.2	94.7	66.7	91.5	98.1	93.6	92.4	92.4	93.1
EfficientNetB0	87.5	85.9	92.1	69.1	90.5	83.5	71.5	62.4	62.4	61.0
DaViT	94.7	90.4	85.9	85.9	87.3	98.6	89.9	86.8	86.8	88.6
Swin	91.0	88.5	85.3	85.3	85.3	**99.1**	93.4	91.3	91.3	92.7
MobileViT	86.2	84.6	79.7	79.7	80.2	95.7	89.7	89.2	89.2	89.1
TopoCAM	99.9	98.7	98.2	99.1	98.4	100.0	99.4	100.0	98.7	99.5

Table 11: **ROI threshold.** Effect of ROI threshold on the performance of TopoCAM.

	Fracture3D					Breast2D					Vessel3D				
Threshold t	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**
0.5	72.6	55.0	49.2	76.1	48.5	98.1	94.9	91.8	85.7	93.3	94.6	92.4	44.2	98.5	56.7
0.6	77.2	63.3	58.9	80.3	59.0	99.9	98.7	98.2	99.1	98.4	98.9	97.1	76.7	99.7	85.7
0.7	72.1	55.4	57.5	76.0	57.3	99.9	97.4	92.9	96.0	96.7	97.6	95.5	76.7	97.9	79.5

Table 12: **GradCAM Layers.** Further performance metrics of TopoCAM using ROIs from the final GradCAM layer versus a weighted combination of all layers.

	Nodule3D					Adrenal3D					PneumoniaMNIST2D				
Configuration	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F_1**
Single layer	97.2	97.1	95.3	97.7	**95.6**	88.7	85.2	59.4	93.0	77.6	99.8	95.7	**99.5**	89.3	95.3
Weighted layers	**99.6**	**98.1**	**97.0**	**98.4**	95.4	**95.4**	**91.6**	**79.7**	**95.2**	**81.5**	**100.0**	**99.4**	99.1	**98.7**	**99.3**

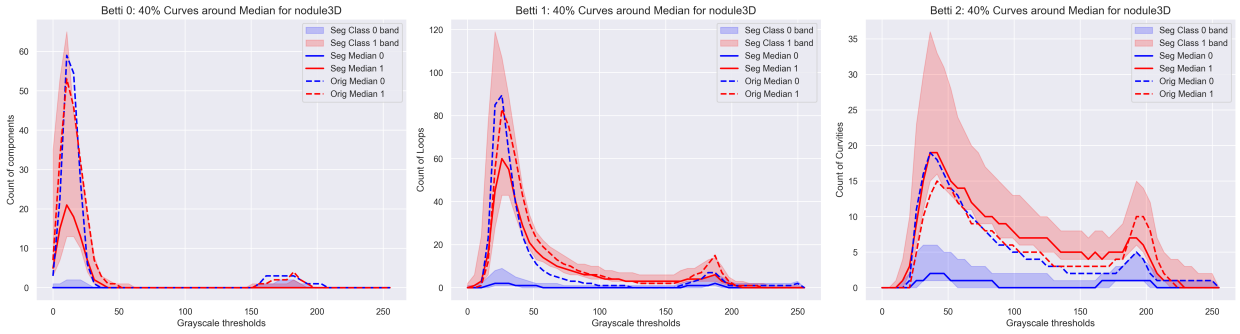

Figure 5: **Betti curves on ROIs.** For Nodule3D, classwise Betti-0 (left), Betti-1 (center), and Betti-2 (right) curves are shown. Solid lines denote medians within Grad-CAM ROIs with 40% confidence bands, while dotted lines are the corresponding medians computed on full volumes. Unlike the overlapping full-volume curves, ROI-based curves and bands show clear class separation.

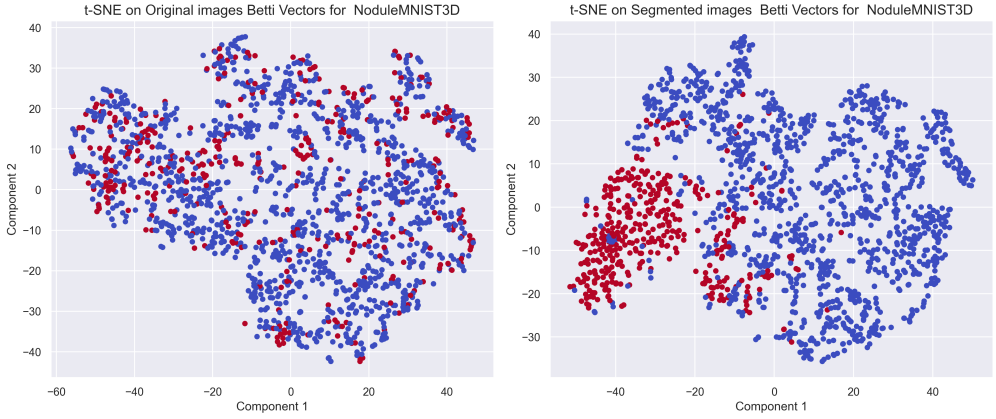

Figure 6: **t-SNE of TopoCAM descriptors on Nodule3D.** Left: t-SNE of Betti vectors computed on the full volumes. Right: t-SNE of our normalized Betti vectors (TopoCAM) on GradCAM ROIs. TopoCAM features yield tighter, better-separated class clusters, consistent with the gains in Tables 2 and 3.