Topoformer: Topology-Infused Transformers for Medical Imaging

Sayoni Chakraborty* SAYONI.CHAKRABORTY@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

Deep learning has transformed 2D medical imaging, but scaling to 3D volumes remains difficult due to high compute, scarce annotations, and the loss of global context in patch-based pipelines. We present `Topoformer`, a transformer framework that makes 3D classification both data- and compute-efficient by integrating topological priors. First, we introduce a *sliding-band cubical filtration* that replaces a single global persistent-homology pass with overlapping intensity bands, yielding an *ordered sequence* of Betti tokens (components, tunnels, cavities). These tokens act as transformer inputs, enabling multi-scale topological reasoning without early saturation. Second, we propose *Topological Supervised Contrastive Learning* (TopoSupCon), which treats the image and its label-preserving topological view as complementary modalities, reducing reliance on brittle geometric or generative augmentations. A lightweight *TopoGate* further lets the image softly weight multiple band widths per case. On 3D brain MRI tumor grading and chest CT benchmarks in low-data regimes, `Topoformer` achieves consistent gains over strong 3D CNN and ViT baselines, including improvements up to 12 AUC points and 8 accuracy points. Our results show that sequential, topology-aware representations provide a powerful inductive bias for volumetric medical image analysis.

Keywords: 3D medical imaging, brain tumor grading, MRI, CT, cubical persistent homology, topological data analysis, volumetric image classification

Data and Code Availability All datasets used in this study are publicly available, and source details are given in Section 4. The complete implementation of our method, including preprocessing scripts and model training code, are provided in the following link: github.com/philmorefkoung/Topoformer

Institutional Review Board (IRB) This research does not require IRB approval, as it uses only publicly available, de-identified datasets.

1. Introduction

3D medical imaging (MRI, CT) is central to diagnosis and treatment planning across neurology, oncology, and cardiology (Litjens et al., 2017; Shen et al., 2017), yet automated analysis of volumetric scans remains difficult due to high dimensionality, scarce annotations, and the cost of 3D computation (Greenspan et al., 2016). While CNNs and ViTs excel in 2D, their 3D variants demand heavy memory/compute (Singh et al., 2020); common workarounds, patching (Isensee et al., 2018), aggressive downsampling (Chen et al., 2021), and large-scale pretraining (Tang et al., 2022), often sacrifice global context and struggle in low-data regimes (Wang et al., 2025). Persistent homology (PH) offers complementary, shape-aware descriptors (Skaf and Laubenbacher, 2022; Wang et al., 2021a; Rieck et al., 2020; Qaiser et al., 2019), but prevailing pipelines rely on *global* filtrations and ad-hoc vectorizations that saturate early and discard the sequential evolution of topology (Hofer et al., 2017). Moreover, standard augmentation, crucial for representation learning, can be risky in medicine, potentially distorting anatomy and labels.

We propose `Topoformer`, a transformer-based framework that tackles both issues with two key innovations. First, we introduce a *sliding-band cubical filtration* that replaces a single global sublevel filtration with overlapping intensity bands. This yields an ordered sequence of Betti tokens (components,

* These authors contributed equally

tunnels, cavities) that preserves late-emerging structure and global context, providing a compact, learnable topological "timeline" for transformers. Second, we develop *Topological Supervised Contrastive Learning* (TopoSupCon), which treats the image and its label-preserving *topological* view as two modalities for supervised contrastive learning (Khosla et al., 2020). This sidesteps brittle geometric or generative augmentations by using topology as a faithful complementary view. A lightweight *TopoGate* further lets the image softly weight multiple band widths, enhancing robustness without manual tuning.

Across brain MRI and chest CT benchmarks, `Topoformer` improves AUC/accuracy over strong 3D CNN and ViT baselines, with the largest gains on challenging tasks (e.g., MGMT methylation and vertebral fracture classification), and maintains balanced sensitivity/specificity where pure image models can collapse. These results suggest that topology, as a sequential, label-stable signal, provides an effective inductive bias for volumetric learning.

Contributions.

- A **sliding-band cubical filtration** that produces *sequential* topological embeddings, capturing multi-scale morphological evolution without early saturation.
- **TopoSupCon**: a *topological* supervised contrastive fusion of images and sliding-band topology, enabling augmentation-robust representation learning.
- A **transformer architecture** that ingests ordered Betti sequences directly; with *TopoGate* to softly combine multiple band widths per case.
- **Stability guarantees** for the sequential encodings and state-of-the-art results on public 3D MRI/CT benchmarks under low-data conditions.

2. Background

2.1. Related Work

Deep Learning for 3D Medical Image Classification. CNNs power much of 2D medical imaging (e.g., ResNet, U-Net) (He et al., 2016a; Ronneberger et al., 2015); 3D variants (3D ResNet, C3D) leverage volumetric context but are memory/compute intensive, so practitioners resort to patching or downsampling that can erode global structure on small datasets like BraTS (Hara et al., 2017; Tran et al., 2015; Suk et al., 2014; Menze et al., 2015). Lightweight volumetric CNNs reduce cost (El-Assy et al., 2024; Akindele et al., 2024) but retain local inductive biases and often rely on heavy augmentation or pretraining.

Transformers adapted to 3D treat patches as tokens and improve global modeling (e.g., TransMed, M3T, MedViT, joint designs) (Dai et al., 2021; Jang and Hwang, 2022; Manzari et al., 2023; Alp et al., 2024), yet token counts in 3D inflate memory/compute and performance typically hinges on large-scale pretraining (Li et al., 2023). These challenges motivate approaches that preserve global information while remaining efficient in low-data regimes. Our Topoformer extracts sequential topological descriptors from sliding intensity bands, yielding compact, informative sequences that a transformer can process efficiently, improving accuracy and resource use on 3D MRI tumor grading.

Contrastive Learning and Data Augmentation in Medical Imaging. Contrastive learning paired with heavy augmentations has delivered strong image-classification results, from long-tailed recognition to whole-slide pathology (Wang et al., 2021b; Li et al., 2021; Wang et al., 2022), by pulling positives together and pushing negatives apart. It is attractive for few-shot medical imaging, but augmentations are problematic: best practices are unclear, synthetic/affine transforms can be unrealistic, and naive transforms may distort pathology and induce label drift (Goceri, 2023). To retain contrastive benefits without ambiguous augmentations, we propose *Topological Supervised Contrastive Loss* (TopoSupCon), a variant of the original supervised contrastive loss (SupCon) (Khosla et al., 2020) that contrasts the original image with a *label-preserving* topological complement derived via sliding-band filtration. This topological view leaves the image unchanged yet provides a complementary representation, encouraging a more structured embedding space with minimal dependence on augmentation.

Topological Machine Learning in Medical Imaging. Topological data analysis, particularly persistent homology (PH), has gained traction in medical imaging for its ability to capture multi-scale structural features beyond traditional pixel-level representations. Early studies successfully applied PH to cell development (McGuirl et al., 2020),

tumor morphology (Crawford et al., 2020; Wang et al., 2021a), brain connectivity (Caputi et al., 2021; Rieck et al., 2020), and histopathology (Qaiser et al., 2019). These efforts laid the groundwork for integrating topological features into deep learning workflows, which has become a growing focus in recent years (Skaf and Laubenbacher, 2022).

Recent studies have embedded topological descriptors into CNNs and transformers to improve classification and segmentation. PHG-Net (Peng et al., 2024) introduces lightweight PH modules for end-to-end training, and 3D persistence image methods have been shown effective for volumetric benchmarks like MedMNIST (Zhu et al., 2024). Topology-aware supervision has also improved anatomical fidelity in segmentation tasks (Santhirasekaram et al., 2023; Gupta et al., 2022; Demir et al., 2023), while hybrid CNN–PH models have enhanced tumor detection and disease grading (Stucki et al., 2023; Somasundaram et al., 2021; Yadav et al., 2023).

2.2. Cubical Persistence

Persistent homology (PH) is a key topological data analysis tool for capturing multi-scale patterns in complex data (Dey and Wang, 2022). In image analysis, we use its cubical variant, which operates directly on pixel/voxel grids via cubical complexes. The PH pipeline can be outlined in three steps; for full details, see (Coskunuzer and Akçora, 2024).

Filtration. From a 3D image $\mathcal{V} \in \mathbb{R}^{p\times q\times r}$, select a sequence of intensity thresholds $0 = \tau_1 < \tau_2 < \cdots < \tau_M = 255$. Let γ_{ijk} represent the voxel intensity at Δ_{ijk} for a fixed color channel (e.g., grayscale). At each τ_n, form the cubical complex $\mathcal{V}_n = \{\Delta_{ij} \subset \mathcal{V} \mid \gamma_{ijk} \leq \tau_n\}$, which yields the nested sequence of binary images $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_N$ known as a *sublevel filtration*. For 2D images, the construction is similar (see Figure 1 for a toy example).

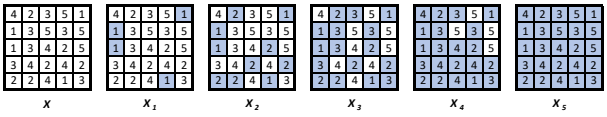

Figure 1: **PH filtration.** For the 5×5 image $\mathcal{X}$ with the given pixel values, **the sublevel filtration** is the sequence of binary images $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \mathcal{X}_3 \subset \mathcal{X}_4 \subset \mathcal{X}_5$.

Persistence diagrams. As we sweep through the binary volumetric images $\{\mathcal{V}_n\}$, topological features, i.e., connected components (0-cycles), tunnels (1-cycles), and voids (2-cycles), are born and later merge or vanish in this sequence. Each feature σ is recorded by its birth threshold $b_\sigma = \tau_m$ and death threshold $d_\sigma = \tau_n$, forming the pair (b_σ, d_σ). The set of all such pairs in homological dimension k is the k^{th} persistence diagram $\mathrm{PD}_k(\mathcal{X}) = \{(b_\sigma, d_\sigma)$. i.e., $\mathrm{PD}_k(\mathcal{X}) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\mathcal{V}_s) \text{ for } b_\sigma \leq s < d_\sigma\}$ where $H_k(\mathcal{V}_s)$ is the k^{th} homology group of $\mathcal{V}_s$.

Vectorization. Persistence diagrams are multisets of interval pairs and must be transformed into fixed-length representations for downstream ML tasks. At this stage, one may choose among a variety of techniques, e.g., Betti curves, persistence images, persistence landscapes, silhouettes, kernel embeddings, and more (Ali et al., 2023). Each method maps the diagram's birth–death pairs $\{(b_\sigma, d_\sigma)\}$ into a vector or function feature: $\Phi(\mathrm{PD}_k) \in \mathbb{R}^D$, where Φ denotes the chosen embedding and D its dimensionality. This flexibility allows practitioners to select the representation best suited to their model architecture and computational budget. For example, in Figure 1, $\mathrm{PD}_0(\mathcal{X}) = \{(1,\infty), (1,2), (1,3)\}$ and $\mathrm{PD}_1(\mathcal{X}) = \{(2,4), (3,4), (3,5)\}$ while corresponding Betti vectors $\boldsymbol{\beta}_0 = [3,3,1,1,1]$ and $\boldsymbol{\beta}_1 = [0,1,3,1,0]$ where i^{th} entry in $\boldsymbol{\beta}_0$ is the number of components in $\mathcal{X}_i$, while i^{th} entry in $\boldsymbol{\beta}_1$ is the number of holes in $\mathcal{X}_i$, i.e., $\beta_0(\mathcal{X}_i)$ and $\beta_1(\mathcal{X}_i)$.

3. Topoformers for Medical Imaging

Standard cubical persistence on a global sublevel filtration can saturate quickly in volumetric scans: once bright regions activate, large components form early and later features are suppressed. In addition, converting persistence diagrams to fixed-length features discards the natural ordering of topological changes and introduces ad hoc design choices. We address these issues with *sliding-band filtration*: replace the single global filtration by a sliding sequence of overlapping intensity bands, compute per-band Betti numbers, and treat the resulting values as an ordered sequence of *topological tokens*. This preserves global context, exposes late-emerging morphology, and removes the need for diagram assembly or handcrafted vectorization. The construction is simple and scalable, applies to both 2D and 3D, and yields a compact sequence of fixed length $N = 50$ that is well matched

to transformer encoders. In Sec. 3.3 we show these sequences are stable under small intensity perturbations.

3.1. Sliding-Band Filtrations for Sequential Topological Embeddings

We reformulate cubical persistence to yield a sequence of local topological descriptors suitable for transformer encoders by replacing a single global sublevel filtration with overlapping intensity bands. The pipeline has two stages.

Stage 1: Banded cubical complexes. Let $V \in \mathbb{R}^{p\times q\times r}$ be a volumetric image with voxel intensities γ_{ijk} defined on a fixed range $[A, B]$. We specify a band width $\omega > 0$ and a desired number of slices N. The stride in intensity space is then set adaptively as

$$\Delta = \frac{B - A - \omega}{N - 1}.$$

For $s = 1, \ldots, N$, we define the sliding-band cubical complex

$$V_s = \{\, \Delta_{ijk} \subset V \;\big|\; \ell_s \le \gamma_{ijk} < u_s \,\},$$

with $\ell_s = A + (s-1)\Delta$ and $u_s = \ell_s + \omega$, except for the final band which is closed at B. This produces a chain of N partially overlapping complexes $\{V_1, \ldots, V_N\}$, each isolating voxels whose intensities fall within a sliding band of width ω. For 2D images $X \in \mathbb{R}^{p\times q}$ with pixel values γ_{ij}, the definition is analogous (see Figure 2).

$$X_s = \{\, \Delta_{ij} \subset X \;\big|\; \ell_s \le \gamma_{ij} < u_s \,\}.$$

Next, to define our output sequence, we compute k^{th} Betti numbers

$$\beta_k(\mathcal{V}_s) = \operatorname{rank}(H_k(\mathcal{V}_s)), \quad k \in \{0, 1, 2\},$$

tracking the number of *components* (β_0), *tunnels* (β_1), and *cavities* (β_2) in the sequence of binary volumetric images $\{\mathcal{V}_s\}$. We control sequence granularity by fixing the number of slices N. Given a band width ω, the stride Δ in the intensity space ensures that the entire range $[A, B]$ is covered in exactly N overlapping bands.

Stage 2: Ordered topological encoding. For each band $\mathcal{V}_s$, compute Betti numbers $\beta_k(\mathcal{V}_s)$ and optionally simple slice statistics (for example active-pixel or active-voxel count). Define a per-band token

$$\psi_s \;=\; \begin{cases} \big[\beta_0(\mathcal{V}_s),\, \beta_1(\mathcal{V}_s)\big] \in \mathbb{R}^2, & \text{2D} \\ \big[\beta_0(\mathcal{V}_s),\, \beta_1(\mathcal{V}_s),\, \beta_2(\mathcal{V}_s)\big] \in \mathbb{R}^3, & \text{3D} \end{cases}$$

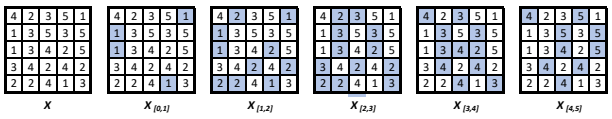

Figure 2: **Sliding Band Filtration.** For the toy example in Figure 1, we give the sliding band filtration with band-width=1 and stride=1.

and stack them into an ordered sequence

$$\Psi(\mathcal{V}) \;=\; [\psi_1, \psi_2, \ldots, \psi_L] \in \mathbb{R}^{L\times D},$$

where $D = 2$ for 2D and $D = 3$ for 3D. We use $\Psi(\mathcal{V})$ directly as the transformer input, bypassing persistence-diagram assembly and ad hoc vectorization, so the model can attend to topological changes across intensity bands without premature saturation.

PH vs. SB Filtration. In Figures 1 and 2, we illustrate the standard sublevel filtration in persistent homology (PH) and our more flexible sliding band (SB) filtration for the same image. In sublevel filtration, images often saturate early, limiting the ability to detect topological features that appear at higher thresholds. By contrast, SB filtration continues tracking changes across higher intensity ranges, allowing it to capture topological variations throughout the full color scale. For example, in the sublevel filtration (Figure 1), the Betti vectors are $\boldsymbol{\beta}_0 = [3, 3, 1, 1, 1]$ and $\boldsymbol{\beta}_1 = [0, 1, 3, 1, 0]$, whereas with SB filtration (Figure 2), they are $\boldsymbol{\beta}_0 = [3, 3, 1, 1, 2]$ and $\boldsymbol{\beta}_1 = [0, 1, 2, 2, 2]$.

3.2. Topoformer Architecture

Our model consists of two stages, (i) TopoGate: we first use the images to guide which topological bandwidth is the most useful, and (ii) TopoSupCon: treat the image and its topological representation as complementary views that agree for the same class.

Pseudocodes are given in Appendix D.

Inputs. Each case provides a grayscale 3D volume $V \in \mathbb{R}^{64\times64\times64}$ and W topology vectors $\{\mathbf{t}^{(w)} \in \mathbb{R}^L\}_{w=1}^{W}$ computed from sliding-band filtrations (e.g., widths 20 and 40). We z-score topology features per dataset.

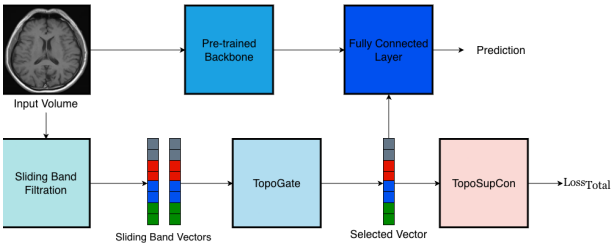

Figure 3: **Topoformer framework.** The 3D image branch extracts semantic features via a pretrained CNN backbone, while the topology branch encodes shape information using sliding-band filtrations of Betti features. `TopoGate` adaptively chooses multi-width topological embeddings, and `TopoSupCon` aligns representations through supervised contrastive learning.

3.2.1. TopoGate

A 3D encoder turns V into patch tokens; a tiny gating module uses those tokens to softly weight the W topology vectors; a lightweight transformer reads the resulting topological sequence and predicts the class. This lets the model *learn* which filtration width is most informative, without manual selection (See Figure 4).

3D Visual Encoder. We use an R3D-18 backbone (pretrained on Kinetics-400) adapted to single-channel inputs by replacing the RGB stem with a 1→64 convolution (initialized by channel-averaging the pretrained filters). The encoder outputs a feature map that we flatten into P patch tokens $\mathbf{X} \in \mathbb{R}^{P\times C}$ (details in §4).

HyperGate: Image-conditioned bandwidth selection From the mean of the image tokens we form a query vector. We attend this query to W learnable prototypes and pass the result through a small MLP+softmax to obtain weights $\mathbf{w} \in \mathbb{R}^W$ with $\sum_w w_w = 1$. The gated topology is a convex combination $\bar{\mathbf{t}} = \sum_{w=1}^{W} w_w \, \mathbf{t}^{(w)} \in \mathbb{R}^L$, so the image softly "chooses" which sliding-band widths to emphasize.

TopoTransformer head We view $\bar{\mathbf{t}}$ as a length-L sequence. Each scalar goes through a small embedding layer; a shallow transformer encoder produces contextual token features that we mean-pool and feed to a linear classifier.

3.2.2. TopoSupCon: supervised contrastive fusion

We also use a contrastive variant that treats the volume and its topology as two views of the same case (See Appendix C for details). An image encoder $f_\theta(\cdot)$ and a topology encoder $g_\phi(\cdot)$ produce embeddings that are (i) fused for classification, and (ii) passed through projection heads for a supervised contrastive loss. Positives are all pairs that share the same class label, both within a modality and across modalities; negatives are different-class pairs. The total loss is a simple sum of cross-entropy and SupCon,

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

which encourages class-consistent alignment between image and topology while keeping the classifier objective unchanged.

3.3. Stability of Sliding-Band Sequences

We show that sliding-band Betti sequences are stable under small perturbations of the intensity function. Let $\mathcal{V}$ be a 3D (or 2D) image and $\gamma^1, \gamma^2 : \mathcal{V} \to \mathbb{R}$ two voxel intensity functions. Fix an intensity range $[A, B]$, a bandwidth $\omega > 0$, and a number of slices N. Set the stride in intensity space as

$$\Delta = \frac{B - A - \omega}{N - 1}.$$

For $m = 1, 2$ and $s = 1, \ldots, N$, define the overlapping banded cubical complexes

$$\mathcal{V}_s^m \;=\; \big\{ \Delta_{ijk} \subset \mathcal{V} \;\big|\; \ell_s \le \gamma_{ijk}^m < u_s \big\},$$

with $\ell_s = A + (s-1)\Delta$ and $u_s = \ell_s + \omega$, except for the final band which is closed at B. For $k \ge 0$, let

$$\Psi_k(\mathcal{V}, \gamma^m) \;=\; \big[\beta_k(\mathcal{V}_1^m), \ldots, \beta_k(\mathcal{V}_N^m)\big].$$

Theorem 1 *For any $k \ge 0$,*

$$\big\|\Psi_k(\mathcal{V}, \gamma^1) - \Psi_k(\mathcal{V}, \gamma^2)\big\|_{\ell^1} \;\le\; C_k \, \|\gamma^1 - \gamma^2\|_{L^1},$$

where $\|\cdot\|_{\ell^1}$ denote ℓ^1-norm on $\mathbb{R}^L$ and $\|\cdot\|_{L^1}$ denote the L^1-norm on $\mathcal{F}(\mathcal{V}, \mathbb{R})$, and C_k depends only on L and the cubical neighborhood structure.

The proof is given in Appendix A.

4. Experiments

4.1. Setup.

Datasets. We evaluate on six publicly available 3D datasets spanning brain and breast MRI and chest CT (Table 1). For brain MRI, we use three benchmarks: BraTS 2019 (HGG vs. LGG) (Menze et al., 2015; Bakas et al., 2018), BraTS 2021 MGMT (methylated vs. unmethylated) (Baid et al., 2021), and the MRI subset of the RSNA 2025 Aneurysm challenge (aneurysm presence) (Rudie et al., 2025). For breast MRI, we include ODELIA 2025 with three-class lesion labels (no lesion/benign/malignant) (Consortium, 2025; Müller-Franzes et al., 2025). For chest CT, we adopt two standardized MedMNIST v2 tasks (Yang et al., 2023, 2024): NoduleMNIST3D, derived from LIDC–IDRI for nodule detection (Armato et al., 2011), and FractureMNIST3D, derived from FracNet for vertebral fracture classification (Jin et al., 2020); both provide 28^3 volumes and fixed train/val/test splits. Summary counts and classes are given in Table 1. For completeness, we also report results on three 2D COVID CXR/CT datasets (See Appendix B for details).

Table 1: **3D Datasets.** Summary statistics for brain MRI and chest CT datasets.

Dataset	Modality	# Images	# Classes
BraTS 2019	Brain MRI	335	2
BraTS 2021 MGMT	Brain MRI	585	2
RSNA 2025 Aneurysm	Brain MRI	1123	2
ODELIA 2025	Breast MRI	1022	3
FractureMNIST3D	Chest CT	1370	3
NoduleMNIST3D	Chest CT	1633	2

Preprocessing. All MR volumes are resampled to 1.0 mm isotropic spacing (B-spline), resized to $64 \times 64 \times 64$, clipped to the 1st–99th percentiles, and z-scored per volume. Multi-modal scans (FLAIR, T1w, T1wCE, T2w) are processed independently. MedMNIST (NoduleMNIST3D, FractureMNIST3D) volumes are already normalized; we only resize them to $64 \times 64 \times 64$ for consistency.

Topological features. On z-scored volumes clipped to $[-5, 5]$, we construct sliding-band Betti sequences with 50 overlapping bands of widths $w \in \{0.2, 0.4\}$ (Corresponding to SB20 and SB40). For each band, we form a binary image and record $(\beta_0, \beta_1, \beta_2)$, yielding a length-50 sequence and $50 \times 3 = 150$ features per width. As a PH baseline, we use `CubicalPersistence` (dims 0–2) followed by `BettiCurve` with 50 bins (Giotto-TDA), producing 150 features per modality; across M modalities this gives a $150M$-dimensional descriptor.

Hyperparameters. We used the Adam optimizer with a learning rate of 1e-4, batch size of 32, and cross-entropy loss for all experiments. For the MedMNIST datasets, we use the predefined training, validation, and test splits. For the BraTS 2019, BraTS 2021, and RSNA 2025 Aneurysm datasets, we use identical random stratified 70:10:20 train/val/test split across all models. Each model was trained for 100 epochs and the epoch with the highest validation AUC was subsequently used for testing. In our baseline 3D ResNet 18+SupCon model, we use the best augmentation techniques for lung CT and brain MRI images identified in (Goceri, 2023) for the two views.

Specifically, a combination of translation and shearing for lung CT images and rotation with shearing, and translation for brain MRI images. Our MLP consists of 3 hidden layers with 128, 64, and 64 units with ReLU activations. Our Transformer encoder has a model dimension of 128 with 8 attention heads, 3 layers, a 512-dimensional feed-forward, GELU activations, and 0.1 dropout. Sliding Band Vectors are tokenized via patching with patch size 5 and stride 2; we add learnable positional embeddings (length 1024). The encoded sequence is mean-pooled and fed to a classifier head with one hidden layer of 64 units with GELU and 0.1 dropout.

Computational Complexity and Runtime. We report runtimes for representative 3D datasets below; other datasets follow similar trends in runtime scaling with volume count.

Table 3: Average runtime (in seconds) per dataset and module.

Dataset	#Volumes	Betti(64 core CPU)	TopoGate	TopoSupCon
NoduleMNIST3D	1633	166	388	1153
BRATS 2019	335	101	148	253
BRATS 2021	585	174	226	435

TopoGate and TopoSupCon evaulated on an NVIDIA L40S (8 vCPU, 62 GB RAM, 48 GB VRAM)

Let $V \in \mathbb{R}^{p \times q \times r}$ have $|V|{=}pqr$ voxels and fix N sliding bands (50 in our experiments). For band s, let $A_s{=}|V_s|$ with $\sum_{s=1}^{N} A_s = O(|V|)$. Per band, Betti computation (union–find for β_0, Euler-characteristic cell counts for β_1, and one background pass for β_2 in 3D) costs $O(A_s\,\alpha(A_s))$ time and $O(A_s)$ memory, yielding total $O(N|V|\,\alpha(|V|))$ time and $O(\max_s A_s)$ peak memory across bands processed sequentially. The resulting sequence $\Psi(V) \in \mathbb{R}^{N \times D}$ is encoded

Table 2: Baseline comparison of 3D CNN and Transformer models with our proposed topological hybrid approach.

	BRATS 2019 (Binary)					BRATS 2021 (Binary)					RSNA 2025 (Binary)				
Model	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**
R3D-18	85.8	83.6	83.6	98.1	90.3	57.3	54.2	53.9	88.7	67.1	59.1	74.7	37.5	97.1	9.5
R3D-18+SC	82.3	80.9	66.2	66.2	68.1	57.3	55.9	54.5	54.5	51.4	56.2	70.2	50.8	50.8	49.6
MC3-18	88.1	82.1	82.3	98.1	89.5	52.4	45.8	48.9	72.6	58.4	67.5	73.3	45.1	83.5	43.4
R(2+1)D-18	86.8	86.6	**85.2**	**100.0**	**92.0**	45.3	48.3	50.5	75.8	60.7	65.9	57.3	32.2	54.1	43.5
EfficientNet3D	47.6	76.5	38.2	50.0	43.3	48.5	52.5	52.5	**100.0**	**68.9**	50.7	**75.6**	0.0	**100.0**	0.0
M3T	86.9	85.3	75.8	84.5	78.6	51.1	47.5	50.0	54.8	52.3	48.7	**75.6**	0.0	**100.0**	0.0
ViT-3D	81.0	79.4	67.9	65.9	66.7	49.8	51.7	53.7	58.1	55.8	60.8	**75.6**	50.0	96.5	17.9
CCT-3D	85.2	82.3	75.3	62.4	64.8	49.7	47.5	50.0	47.7	46.6	55.4	72.0	33.3	90.6	20.2
Topoformer	**91.3**	**88.2**	82.9	82.9	82.9	**69.1**	**62.7**	**68.3**	61.2	57.9	**69.9**	73.8	**63.6**	62.4	**62.9**

	ODELIA (3-class)					NoduleMNIST (Binary)					FractureMNIST (3-class)				
Model	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**
R3D-18	65.3	62.6	51.8	44.8	**46.2**	90.9	86.8	67.7	91.5	68.2	66.1	50.0	51.7	72.7	46.5
R3D-18+SC	59.9	62.1	43.2	71.6	44.2	90.9	87.1	84.4	84.4	81.7	65.2	49.2	47.7	72.6	48.3
MC3-18	66.5	63.1	48.8	40.5	40.8	90.3	86.1	65.2	90.2	67.7	67.8	51.2	49.9	74.5	50.2
R(2+1)D-18	60.8	62.6	**56.1**	44.4	45.5	89.3	86.5	68.3	**92.3**	66.1	69.2	50.4	56.8	73.0	47.2
EfficientNet-3D	58.4	56.3	37.7	36.5	36.2	75.8	84.5	78.3	69.4	72.3	54.8	44.2	29.6	68.9	32.6
M3T	46.4	63.1	21.0	33.3	25.8	88.6	87.7	67.6	90.2	72.5	70.9	52.9	53.3	75.5	52.9
ViT-3D	60.3	63.1	21.0	33.3	25.8	89.4	85.5	77.7	79.3	78.5	66.1	50.8	47.7	72.9	44.3
CCT-3D	60.1	58.3	40.0	39.2	38.0	81.8	83.9	79.9	65.0	68.1	62.6	47.1	**64.6**	70.3	36.0
Topoformer	**69.0**	**64.6**	50.7	**45.2**	45.9	**93.2**	**88.1**	**81.7**	82.1	**81.9**	**75.5**	**60.8**	63.0	**78.2**	**58.4**

by a Transformer with T layers and width E in $O(TN^2E)$ time and $O(TNE)$ memory. Overall:

- time $= O\big(N|V|\,\alpha(|V|)\big) + O(TN^2E)$, and
- memory $= O(|V|) + O(NE)$,

with N small and fixed in practice.

Baselines. We compare against widely used 3D CNN and Transformer families. For convolutional 3D and video backbones, we include `R3D-18`, `MC3-18`, and `R(2+1)D-18` following the spatiotemporal designs of Tran et al.(Tran et al., 2018), and an `EfficientNet-3D` variant built from the EfficientNet scaling rules (Tan and Le, 2019). For transformer-style baselines, we evaluate a patchified `ViT-3D` (Dosovitskiy et al., 2021), `CCT-3D`, a lightweight compact convolutional transformer adapted for volumetric inputs (Sun et al., 2022), and `M3T`, a multi-plane multi-slice transformer for 3D medical image classification (Jang and Hwang, 2022). As a contrastive-learning baseline, we include `ResNet3D-18+SupCon`, which augments a 3D ResNet-18 with supervised contrastive learning.

4.2. Results

Across six public datasets spanning brain MRI, chest CT, and breast MRI, our topological hybrids outperform strong 3D CNN and Transformer baselines on most tasks (Table 2). `Topoformer` achieves the best AUC on all six datasets and the best or near–best accuracy on five of them. The gains are most pronounced on the MGMT methylation task in BraTS 2021 and on FractureMNIST, where improvements reach roughly ten–plus AUC points and about eight accuracy points over the strongest non–topological baselines. On RSNA 2025, `Topoformer` delivers the top AUC with competitive accuracy, and on BraTS 2019 and NoduleMNIST it edges out the best baselines in both AUC and accuracy. Notably, several pure 3D baselines exhibit degenerate behavior on RSNA 2025 with extreme specificity and near–zero sensitivity, while our methods maintain balanced performance.

Table 4: **2D COVID benchmarks.** We compare our topology-aware hybrids with recent CNN/ViT baselines. The details and other performance metrics can be found in Appendix B.

	COVID CT		COVID QU		COVID Rad-3		COVID Rad-4	
Method	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
ResNet18	87.6	82.0	98.7	94.0	98.9	89.4	99.1	94.5
DenseNet121	96.4	90.7	99.3	95.3	96.3	89.4	95.3	95.3
Xception	96.0	89.3	99.4	95.9	98.8	86.4	99.4	95.2
SwinV2	96.6	90.7	99.3	95.2	**100.0**	**98.5**	99.3	94.7
DeiT	93.0	81.3	99.4	96.8	99.2	84.9	99.1	93.5
DaViT	96.4	89.3	98.8	92.9	**100.0**	95.5	99.4	95.2
DaViT+SC	97.9	92.7	**99.5**	94.9	99.9	95.5	99.4	94.9
Topoformer	**98.2**	**93.3**	**99.5**	**97.2**	**100.0**	**98.5**	**99.5**	**95.4**

Table 5: **Filtration types.** Performance comparison using Sliding-Band filtrations (SB20 & SB40) and standard sublevel filtrations (PH), evaluated with MLP and Transformer (TR) classifiers. Additional metrics are reported in Table 11.

	BRATS 2019		BRATS 2021		RSNA 2025		ODELIA 2025		NoduleMNIST		FractureMNIST	
Model	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
PH+MLP	81.6	**82.3**	53.6	54.2	54.2	72.8	54.6	48.1	72.3	76.8	59.4	45.4
PH+TR	**83.0**	77.9	53.1	51.6	59.9	72.8	55.4	55.3	69.4	**78.7**	57.5	42.1
SB20+MLP	82.9	**80.9**	57.2	53.4	**64.4**	**77.3**	**61.3**	**55.8**	**75.6**	78.4	62.9	**51.7**
SB40+MLP	**85.0**	**79.4**	**62.0**	**61.9**	**67.2**	**74.7**	**58.6**	52.9	**72.9**	75.2	**64.0**	**51.3**
SB20+TR	82.1	73.5	**57.8**	**56.8**	62.3	73.3	56.7	**56.3**	72.8	**80.0**	**66.1**	**51.7**
SB40+TR	**85.3**	**80.9**	**61.8**	**59.3**	**66.1**	**75.1**	**55.9**	**59.7**	**73.7**	**79.4**	**65.4**	**52.1**

Table 6: **Fusion types.** Performance comparison using different fusion types with an R3D-18 backbone. SB = sliding-band Betti; PH = sublevel Betti; Concat = feature concatenation; +SupCon = supervised contrastive fusion. Additional metrics are given in Table 12.

	BRATS 2019		BRATS 2021		RSNA 2025		ODELIA 2025		NoduleMNIST		FractureMNIST	
Model	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
R3D-18	85.8	83.6	57.3	54.2	59.1	74.7	65.3	62.6	90.9	86.8	66.1	50.0
R3D-18+SupCon	82.3	80.9	57.3	55.9	56.2	70.2	59.9	62.1	90.9	87.1	65.2	49.2
R3D-18+SB (Concat)	90.9	73.5	58.0	52.5	61.6	72.4	64.3	64.6	91.0	87.4	67.1	52.5
R3D-18+PH+SupCon	90.7	77.9	60.2	56.8	61.1	75.6	66.4	62.6	87.2	85.5	70.2	52.1
R3D-18+SB+SupCon	**91.3**	**88.2**	**69.1**	**62.7**	**69.9**	**73.8**	**68.9**	**64.6**	**93.2**	**88.1**	**75.5**	**60.8**

Ablations indicate that topology helps even without contrastive fusion: `Topo-R3D18` already improves on its 3D counterpart, and `Topo-BV` (simple concatenation) is stronger than most non–topological models. The full `Topoformer` variant, which models the sliding–band Betti sequences and applies supervised contrastive fusion, is consistently the best among our approaches across datasets. These trends support our design choices of sliding–band filtrations, sequence modeling over topological tokens, and supervised contrastive fusion as complementary ingredients for robust 3D medical image classification under limited data.

4.3. Ablation Studies

We conduct three ablation studies to assess the contribution of key components in our framework. First, we compare *sliding-band filtrations* with the standard global sublevel filtration to evaluate the benefit of sequential topological encoding. Second, we examine how topological outputs are consumed: either aggregated as a single vector with an MLP classifier, or modeled as an ordered sequence with a Transformer. Finally, we test our *fusion strategy* by contrasting simple feature concatenation with our proposed supervised contrastive fusion (SupCon).

Filtration type. Across brain/breast MRI and chest CT (Table 5), sliding-band filtrations (SB20/SB40) consistently outperform standard sublevel PH under both MLP and Transformer heads. Gains are most pronounced on the harder benchmarks (e.g., BraTS-2021 MGMT and FractureMNIST), reaching ∼8 points in AUC and accuracy, indicating that *sliding bands recover late-emerging structure that global filtrations miss*. The only deviation is a single setting where PH attains the top accuracy while sliding-band still yields the best AUC, suggesting more reliable ranking overall.

Fusion types. We ablate four ways of combining topology with a fixed R3D-18 backbone (Table 6). Adding a supervised contrastive head alone (R3D-18+SupCon, no topology) yields limited or inconsistent gains, indicating that SupCon benefits most from a complementary view. Introducing topology via simple feature concatenation (R3D-18+SB (Concat)) already improves over image-only baselines on most datasets, showing that our sliding-band (SB) descriptors carry discriminative signal even without contrastive coupling.

To test both components of our method, we compare contrastive fusion with standard PH features (R3D-18+PH+SupCon) against our full model (R3D-18+SB+SupCon, i.e., `Topoformer`). Two design choices drive the consistent gap in favor of `Topoformer`: (i) the *SB representation*, modeled as an ordered sequence, captures late-emerging structure that global PH misses; and (ii) *TopoGate* uses

the image to softly weight multiple widths, leveraging both SB20 and SB40 per case rather than committing to a single filtration. In combination with supervised contrastive multi-view learning, this yields the strongest performance across datasets.

Analysis of Slidingband Width While Section 4.1 describes the construction of sliding-band Betti sequences (N=50, $w \in \{0.2, 0.4\}$), here we empirically examine how these choices influence classification performance. Each 3D volume is z-scored and clipped to $[-5, 5]$, and we vary the sliding-band width (w) between 0.1 and 0.5 while keeping N=50 fixed. This configuration controls the density of sampled intensity levels: within a standardized 10-unit range, 50 bands yield ≈0.2-unit intensity steps, balancing coverage and efficiency. Narrower bands (e.g., w=0.1) produce sparse and fragmented masks, whereas broader ones (e.g., w=0.6) oversmooth structural transitions. As shown in Table 7, a moderate width of w=0.4 consistently provides the best trade-off between sensitivity and stability across BRATS 2019 and BRATS 2021. These findings validate our choice of dual-width configurations (SB20/SB40) and motivate the adaptive multi-width design in `TopoGate`, which learns soft weights across different bandwidths.

Table 7: AUC (%) of Betti-vector classifiers under varying sliding-band widths (w) for BRATS 2019 and BRATS 2021.

Dataset	Model	PH	SW10	SW20	SW30	SW40	SW50
BRATS 2019	MLP	81.6	77.9	82.9	77.9	**85.0**	80.6
	Transformer	83.0	83.3	82.1	82.7	**85.3**	84.6
BRATS 2021	MLP	53.6	60.8	57.2	52.7	**62.0**	54.5
	Transformer	53.1	56.3	57.8	53.5	**64.0**	53.9

Effect of Sequence Modeling Head. Given the sequential nature of Betti tokens, recurrent models such as LSTM or GRU present natural alternatives to the Transformer. To assess this, we compared BiLSTM, GRU, and 1D-CNN heads against the Transformer and a simple MLP baseline using the same sliding-band representation (w=0.4). As shown in Table 8, the Transformer consistently achieved the highest AUC and F1 on both BRATS 2019 and BRATS 2021, outperforming all RNN and CNN variants. While recurrent models can capture short-range dependencies, the Transformer's global self-attention better models long-range relationships across intensity bands essential for capturing extended morphological transitions in 3D volumes. These results empirically justify our choice of a Transformer backbone for Betti-sequence encoding.

Table 8: Comparison of sequence modeling heads on Betti sequences (w=0.4).

Dataset	Model	AUC	Acc.	Sens.	Spec.	F1
BRATS 2019	SW40+MLP	85.0	79.4	90.6	40.0	87.3
	SW40+BiLSTM	77.4	74.6	92.3	13.3	84.9
	SW40+GRU	52.2	76.1	98.1	0.0	86.4
	SW40+1D-CNN	74.9	76.1	90.4	26.7	85.5
	SW40+Transformer	**85.3**	**80.9**	**90.6**	**46.7**	**88.1**
BRATS 2021	SW40+MLP	62.0	61.9	59.7	64.3	62.2
	SW40+BiLSTM	60.9	55.6	72.1	37.5	62.9
	SW40+GRU	63.6	59.8	75.4	42.9	66.2
	SW40+1D-CNN	62.8	52.1	95.1	5.4	67.4
	SW40+Transformer	**64.0**	**62.7**	**67.7**	**57.1**	**65.6**

5. Conclusion

We introduced `Topoformer`, a transformer-based framework for 3D medical image classification that operates directly on *sequential* topological signatures from sliding-band cubical filtrations. By replacing global PH vectorization with compact, ordered Betti sequences, and by using *TopoGate* to softly combine multiple band widths, our approach preserves late-emerging topology while remaining lightweight. We further paired the image with a label-preserving topological view via Topological Supervised Contrastive Learning, providing multi-view supervision without risky spatial augmentations. Across brain MRI and chest CT benchmarks, `Topoformer` consistently improves AUC and accuracy over strong 3D CNN and ViT baselines, with marked gains on the harder tasks, and maintains balanced sensitivity/specificity where pure image models can collapse. Future work includes extending to multi-modal imaging (e.g., MRI+PET), developing end-to-end differentiable topology layers, and exploring richer TDA descriptors (e.g., multiparameter persistence) to further enhance performance and interpretability in clinical settings.

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

Romoke Grace Akindele, Samuel Adebayo, Paul Shekonya Kanda, and Ming Yu. Alzhinet: Traversing from 2dcnn to 3dcnn, towards early detection and diagnosis of alzheimer's disease. *arXiv preprint arXiv:2410.02714*, 2024.

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.

Sait Alp, Taymaz Akan, Md. Shenuarin Bhuiyan, Elizabeth A. Disbrow, Steven A. Conrad, John A. Vanchiere, Christopher G. Kevil, and Mohammad A. N. Bhuiyan. Joint transformer architecture in brain 3d mri classification: its application in alzheimer's disease classification. *Scientific Reports*, 14:8996, 2024. doi: 10.1038/s41598-024-59578-3. URL https://www.nature.com/articles/s41598-024-59578-3.

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, 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(2):915–931, 2011. doi: 10.1118/1.3528204.

Ujjwal Baid, Sonam Ghodasara, Sharut Mohan, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. URL https://arxiv.org/abs/2107.02314.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *Frontiers in Neuroscience*, 12:125, 2018. doi: 10.3389/fnins.2018.00125.

Luigi Caputi, Anna Pidnebesna, and Jaroslav Hlinka. Promises and pitfalls of topological data analysis for brain connectivity analysis. *NeuroImage*, 238:118245, 2021.

Jieneng Chen, Yutong Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1251–1258, 2017.

Grand Challenge ODELIA Consortium. Odelia: Breast mri lesion classification challenge. *arXiv preprint arXiv:2506.00474*, 2025. URL https://odelia2025.grand-challenge.org.

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.

Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021. doi: 10.3390/diagnostics11081384. URL https://doi.org/10.3390/diagnostics11081384.

Andac Demir, Elie Massaad, and Bulent Kiziltan. Topology-aware focal loss for 3d image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 580–589, 2023.

Tamal Krishna Dey and Yusu Wang. *Computational topology for data analysis*. Cambridge University Press, 2022.

Mingyu Ding, Xin Xia, Xiaoyi Chu, Xiaohua Zhang, Weidi Xie, Zhiding Yu, Ping Luo, and Jian Wang. Davit: Dual attention vision transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 74–92, 2022.

Paweł Dłotko and Davide Gurnari. Euler characteristic curves and profiles: a stable shape invariant for big data problems. *GigaScience*, 12:giad094, 2023.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

A. M. El-Assy, Hanan M. Amer, H. M. Ibrahim, and M. A. Mohamed. A novel cnn architecture for accurate early detection and classification of alzheimer's disease using mri data. *Scientific Reports*, 14:3463, 2024.

Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial intelligence review*, 56(11):12561–12605, 2023.

Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5):1153–1159, 2016.

Saumya Gupta, Xiaoling Hu, James Kaan, Michael Jin, Mutshipay Mpoy, Katherine Chung, Gagandeep Singh, Mary Saltz, Tahsin Kurc, Joel Saltz, et al. Learning topological interactions for multi-class medical image segmentation. In *European Conference on Computer Vision*, pages 701–718. Springer, 2022.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3154–3162, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016a.

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

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *arXiv preprint arXiv:1707.04041*, 2017.

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

Fabian Isensee, Jens Petersen, André Klein, David Zimmerer, Paul F. Jäger, Simon A. A. Kohl, Jakob Wasserthal, Gregor Köhler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20718–20729, 2022.

Liang Jin, Jiancheng Yang, Kang Kuang, et al. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. *EBioMedicine*, 62:103106, 2020. doi: 10.1016/j.ebiom.2020.103106.

Megan Johnson and Jae-Hun Jung. Instability of the betti sequence for persistent homology and a stabilized version of the betti sequence. *Journal of the Korean Society for Industrial and Applied Mathematics*, 25(4): 296–311, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A. Landman, and S. Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, 85:102762, 2023. doi: 10.1016/j.media.2023.102762. URL https://doi.org/10.1016/j.media.2023.102762.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. 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. doi: 10.1016/j.media.2017.07.005.

Ze Liu, Han Hu, Yutong Lin, Zhiqiang Yao, Zhuliang Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023. doi: 10.1016/j.compbiomed.2023.106791. URL https://doi.org/10.1016/j.compbiomed.2023.106791.

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.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, and et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

Gustav Müller-Franzes, Lorena Escudero Sánchez, Nicholas Payne, Alexandra Athanasiou, Michael Kalogeropoulos, Aitor Lopez, Alfredo Miguel Soro

Busto, Julia Camps Herrero, Nika Rasoolzadeh, Tianyu Zhang, et al. A european multi-center breast cancer mri dataset. *arXiv preprint arXiv:2506.00474*, 2025.

Yaopeng Peng, Hongxiao Wang, Milan Sonka, and Danny Z. Chen. Phg-net: Persistent homology guided medical image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7568–7577, January 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.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, et al. Covid-19 radiography database. Kaggle, 2021. URL https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database. Accessed: 2025-08-16.

Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fMRI data using cubical persistence. *NeurIPS*, 33:6900–6912, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Jeffrey D. Rudie et al. Rsna intracranial aneurysm detection challenge 2025. https://www.kaggle.com/competitions/rsna-intracranial-aneurysm-detection, 2025.

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

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

Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: a review. *Sensors*, 20 (18):5097, 2020.

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

Primoz Skraba and Katharine Turner. Wasserstein stability for persistence diagrams. *arXiv:2006.16824*, 2020.

Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *medRxiv*, May 2020. doi: 10.1101/2020.04.24.20078584. URL https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset.

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.

Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Hierarchical feature representation and fusion for alzheimer's disease diagnosis. *NeuroImage*, 134:313–322, 2014.

Weiwei Sun, Yu Pang, and Guo Zhang. Cct: Lightweight compact convolutional transformer for lung disease ct image classification. *Frontiers in Physiology*, 13: 1066999, 2022.

Anas M. Tahir et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. doi: 10.1016/j.compbiomed.2021.105002. URL https://www.kaggle.com/datasets/anasmohammedtahir/covidqu.

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

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett A. Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Tang_Self-Supervised_Pre-Training_of_Swin_Transformers_for_3D_Medical_Image_Analysis_CVPR_2022_paper.html.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou.

Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018.

Fan Wang, Saarthak Kapse, Steven Liu, Prateek Prasanna, and Chao Chen. Topotxr: a topological biomarker for predicting treatment response in breast cancer. In *International Conference on Information Processing in Medical Imaging*, pages 386–397. Springer, 2021a.

Jiaji Wang, Shuihua Wang, and Yudong Zhang. Deep learning on medical image analysis. *CAAI Transactions on Intelligence Technology*, 10(1):1–35, 2025.

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021b.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

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, Xiaosong Shi, Bingbing Ni, et al. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10: 41, 2023. doi: 10.1038/s41597-022-01821-5.

Jiancheng Yang et al. Medmnist v2 updates. https://medmnist.com, 2024.

Yanfan Zhu, Yash Singh, Khaled Younis, Shunxing Bao, and Yuankai Huo. Persistence image from 3d medical image: Superpixel and optimized gaussian coefficient. *arXiv preprint arXiv:2408.07905*, 2024.

Appendix

Appendix A. Proofs of Stability Theorems

Here, we first recall the main definitions and notation used in our stability analysis. Given a 3D cubical complex $\mathcal{V}$ (e.g. an intensity-band slice) and a real-valued function γ on its voxels, the kth persistence diagram $\mathrm{PD}_k(\mathcal{V}, \gamma)$ is the multiset of birth–death pairs $(b_\sigma, d_\sigma) \in \mathbb{R}^2$ tracking homological features of dimension k for sublevel filtration induced by $\gamma : \mathcal{V} \to \mathbb{R}$. The p-Wasserstein distance between two diagrams PD and PD$'$ is defined by

$$\mathcal{W}_p\big(\mathrm{PD}, \mathrm{PD}'\big) \;=\; \Big(\inf_{\alpha} \sum_{x \in \mathrm{PD}} \|x - \alpha(x)\|_\infty^p\Big)^{1/p},$$

where the infimum is over all bijections α matching points in the two diagrams (with unmatched points paired to the diagonal). For two intensity functions γ^1, γ^2 on the same voxel set $\mathcal{V}$, the discrete L^p-norm is

$$\|\gamma^1 - \gamma^2\|_p \;=\; \Big(\sum_{i,j,k} |f(\Delta_{ijk}) - g(\Delta_{ijk})|^p\Big)^{1/p}.$$

For a given volume $\mathcal{V}$ and two intensity functions $\gamma^1, \gamma^2 : \mathcal{V} \to \mathbb{R}$, fix an intensity range $[A, B]$, a bandwidth $\omega > 0$, and a number of slices N. Set the stride

$$\Delta = \frac{B - A - \omega}{N - 1}.$$

For $m \in \{1, 2\}$ and $s = 1, \ldots, N$, define the overlapping banded cubical complexes

$$\mathcal{V}_s^m \;=\; \big\{\Delta_{ijk} \subset \mathcal{V} \;\big|\; \ell_s \le \gamma_{ijk}^m < u_s\big\},$$

$$\ell_s = A + (s-1)\Delta, \qquad u_s = \ell_s + \omega.$$

For $k \ge 0$, let

$$\Psi_k(\mathcal{V}, \gamma^m) \;=\; \big[\beta_k\big(\mathcal{V}_1^m\big), \ldots, \beta_k\big(\mathcal{V}_N^m\big)\big].$$

These Ψ_k are the sliding-band Betti sequences for which we establish stability.

Let $\mathrm{PD}_k(\mathcal{Y}, f)$ represent the k^{th} persistence diagram for sublevel filtration induced by $f : \mathcal{Y} \to \mathbb{R}$. With these conventions, we now state the two key stability lemmas.

Lemma 2 (Skraba and Turner (2020)) *Let $\mathcal{Y}$ be a compact metric space and $f, g : \mathcal{Y} \to \mathbb{R}$. For any $p \ge 1$,*

$$\mathcal{W}_p\big(\mathrm{PD}_k(\mathcal{Y}, f),\, \mathrm{PD}_k(\mathcal{Y}, g)\big) \;\le\; \|f - g\|_p,$$

where $\mathcal{W}_p$ is the p-Wasserstein distance between persistence diagrams.

In the following, let

$$\boldsymbol{\beta}_k(\mathcal{Y}, f) = [\beta_k(\mathcal{Y}_1), \ldots, \beta_k(\mathcal{Y}_M)]$$

where $\mathcal{Y}_s = f^{-1}(-\infty, \tau_s]$, i.e., $\mathcal{Y}_1 \subset \mathcal{Y}_2 \cdots \subset \mathcal{Y}_M$ is the traditional sublevel filtration induced by $f : \mathcal{Y} \to \mathbb{R}$ for thresholds $\tau_1 < \cdots < \tau_M$.

Lemma 3 (Dłotko and Gurnari (2023)) *Let $\boldsymbol{\beta}_k(\cdot)$ be the Betti curve vectorization of a persistence diagram. Then*

$$\big\|\boldsymbol{\beta}_k(\mathcal{Y}, f) - \boldsymbol{\beta}_k(\mathcal{Y}, g)\big\|_1 \;\le\; 2\,\mathcal{W}_1\big(\mathrm{PD}_k(\mathcal{Y}, f),\, \mathrm{PD}_k(\mathcal{Y}, g)\big).$$

Note that while these lemmas are proven for simplicial complexes, they generalize to cubical complexes in straightforward way Skraba and Turner (2020). Now, we are ready to prove Theorem 1.

Theorem 1 *With notation as above, for any $k \ge 0$,*

$$\big\|\Psi_k(\mathcal{V}, \gamma^1) - \Psi_k(\mathcal{V}, \gamma^2)\big\|_{\ell^1} \;\le\; C_k\, \|\gamma^1 - \gamma^2\|_{L^1},$$

where $\|\cdot\|_{\ell^1}$ denote ℓ^1-norm on $\mathbb{R}^L$ and $\|\cdot\|_{L^1}$ denote the L^1-norm on $\mathcal{F}(\mathcal{V}, \mathbb{R})$, and C_k depends only on L and the cubical neighborhood structure.

Proof Each band complex $\mathcal{V}_s^m$ is a cubical complex on $\mathcal{V}$. The voxel intensity functions γ^m induce filtrations on cubes by taking the maximum vertex intensity. Applying Lemma 2 to each band gives

$$\mathcal{W}_1\big(\mathrm{PD}_k(\mathcal{V}, \gamma^1),\, \mathrm{PD}_k(\mathcal{V}, \gamma^2)\big) \;\le\; \|\gamma^1 - \gamma^2\|_1.$$

Lemma 3 then yields, for each s,

$$\big\|\boldsymbol{\beta}_k(\mathcal{V}, \gamma^1) - \boldsymbol{\beta}_k(\mathcal{V}, \gamma^2)\big\|_1 \;\le\; 2\mathcal{W}_1(\mathrm{PD}_k(\mathcal{V}, \gamma^1), \mathrm{PD}_k(\mathcal{V}, \gamma^2))$$

As for each s, we have $|\beta_k(\mathcal{V}_s^1) - \beta_k(\mathcal{V}_s^2)| \le \big\|\boldsymbol{\beta}_k(\mathcal{V}, \gamma^1) - \boldsymbol{\beta}_k(\mathcal{V}, \gamma^2)\big\|_1$, this gives

$$\begin{aligned}\|\Psi_k(\mathcal{V}, f) - \Psi_k(\mathcal{V}, g)\|_1 &= (M - \omega)\big|\boldsymbol{\beta}_k(\mathcal{V}, \gamma^1) - \boldsymbol{\beta}_k(\mathcal{V}, \gamma^2)\big| \\ &\le 2\,(M - \omega)\,\|\gamma^1 - \gamma^2\|_1.\end{aligned}$$

Hence, one may take $C_k = 2(M - \omega)$. This completes the proof. ∎

Table 9: **PH vs. Sliding-Band (SB) filtrations on 2D COVID datasets.** Performance comparison between Betti sequences from traditional PH sublevel filtration and Sliding-Band filtrations with bandwidths 20 and 40. The first three rows report results with an MLP classifier, and the last three with a Transformer. Best results are in **bold**, second best are underlined.

	COVID CT 2					COVID QU 3					COVID Radio. 3					COVID Radio. 4				
Method	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1
PH+MLP	82.2	76.7	77.1	76.2	75.5	82.5	67.2	64.1	82.8	64.2	93.8	77.3	75.0	88.5	74.8	87.5	70.3	63.9	88.4	65.1
SB20+MLP	85.9	80.0	**85.7**	75.0	80.0	**86.9**	**71.7**	**69.0**	**85.7**	**68.6**	94.6	81.8	81.2	91.1	81.0	90.1	74.3	**73.1**	90.6	71.1
SB40+MLP	82.2	74.0	72.9	75.0	72.3	85.6	**71.7**	67.8	84.9	68.1	94.3	80.3	79.5	90.4	79.2	89.8	72.4	69.5	89.5	71.7
PH+TR	85.4	77.3	80.0	75.0	76.7	84.3	70.3	65.8	83.9	66.5	91.0	75.8	73.3	87.6	73.0	89.6	73.0	68.7	89.6	68.9
SB20+TR	85.4	77.3	75.7	**78.8**	75.7	86.4	71.4	68.4	85.3	68.5	91.2	75.8	74.9	88.2	74.7	**91.6**	**75.9**	73.0	**90.8**	**73.0**
SB40+TR	**87.5**	**81.3**	84.3	**78.8**	**80.8**	85.9	71.1	68.5	85.3	68.3	**96.4**	**84.9**	**83.7**	**92.2**	**84.2**	90.6	74.4	69.2	89.8	71.7

Table 10: **Results across four 2D COVID benchmarks.** We compare our topology-aware hybrids with recent CNN/ViT baselines. *TopoBV* concatenates Betti tokens with DaViT features followed by a classifier. Topoformer integrates Betti tokens with DaViT features using a supervised contrastive (SupCon) objective and a shared head. *DaViT+SC* applies the same SupCon training to DaViT alone (ablation without Betti tokens). Best results are in **bold**; second best are underlined.

	COVID CT (2)					COVID QU (3)					COVID Radio. (3)					COVID Radio. (4)				
Method	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**
ResNet18	87.6	82.0	82.3	82.3	82.0	98.7	94.0	92.5	97.2	92.4	98.9	89.4	88.3	94.9	88.0	99.1	94.5	95.5	97.7	95.6
DenseNet121	96.4	90.7	90.9	90.9	90.7	99.3	95.3	94.2	97.9	94.0	96.3	89.4	88.3	94.8	88.3	95.3	95.3	**96.2**	**98.1**	**96.2**
Xception	96.0	89.3	89.6	89.6	89.3	99.4	95.9	94.9	98.1	94.9	98.8	86.4	85.0	93.5	84.2	99.4	95.2	96.1	**98.1**	96.1
SwinV2	96.6	90.7	91.0	91.0	90.7	99.3	95.2	94.0	97.7	94.0	**100.0**	**98.5**	**98.7**	**99.3**	**98.5**	99.3	94.7	94.7	97.7	95.4
DeiT	93.0	81.3	82.0	82.0	81.3	99.4	96.8	96.1	98.6	96.0	99.2	84.9	83.3	92.8	82.2	99.1	93.5	93.6	97.2	94.2
DaViT	96.4	89.3	89.8	89.8	89.3	98.8	92.9	91.3	96.5	91.4	**100.0**	95.5	95.0	97.8	95.0	99.4	95.2	95.6	98.0	95.8
DaViT+SC	97.9	92.7	93.0	93.0	92.7	**99.5**	94.9	93.4	97.7	93.4	99.9	95.5	95.0	97.8	95.0	99.4	94.9	95.9	98.0	95.6
TopoBV	98.1	92.0	92.3	92.3	92.0	98.9	93.1	91.3	97.0	91.1	99.9	97.0	96.7	98.6	96.7	99.4	94.5	95.1	97.8	95.3
Topoformer	**98.2**	**93.3**	**93.6**	**93.6**	**93.3**	**99.5**	**97.2**	**96.5**	**98.7**	**96.4**	**100.0**	**98.5**	98.3	**99.3**	98.3	**99.5**	**95.4**	95.9	**98.1**	96.0

We note that while Theorem 1 is stated in terms of perturbations to a single intensity function on a fixed volume, the same L^1-stability bound applies when comparing two scans $\mathcal{V}^1$ and $\mathcal{V}^2$ that share the same band thresholds: one simply views the voxel-wise intensity difference $\gamma^1 - \gamma^2$ as arising from scan variation. Crucially, this result relies on the L^1-norm; replacing it with the L^∞-norm can fail to control topological changes, since arbitrarily small, localized intensity shifts may induce large variations in Betti counts Johnson and Jung (2021).

Appendix B. 2D COVID Benchmarks

To evaluate our approach on 2D medical imaging, we conducted experiments on three publicly available, de-identified COVID-19 classification benchmarks covering both CT and chest X-ray (CXR) modalities.

COVID CT-2 (CT). A slice-level chest CT benchmark framed as binary classification (COVID vs. non-COVID), based on the SARS-CoV-2 CT-Scan dataset (Soares et al., 2020). Images are converted to 8-bit, clipped to a robust window, resized to 224×224, and z-score normalized per image. During training we apply mild geometric augmentation (random flips/rotations) and intensity jitter.

COVID QU-3 (CXR). A three-class CXR benchmark (COVID-19, Pneumonia (non-COVID), Normal) derived from the COVID-QU-Ex dataset curated by Qatar University (Tahir et al., 2021). We standardize to 224 × 224 RGB (channel-replicated from grayscale), apply per-image z-score normalization, and use the same augmentation policy.

COVID Radiography (CXR): 3-class and 4-class. We use the COVID-19 Radiography Database (Rahman et al., 2021), reporting both the common 3-class split {COVID-19, Pneumonia, Normal} and the 4-class split {COVID-19, Normal, Bacterial, Viral}. Preprocessing matches COVID QU-3; only label granularity differs.

Evaluation. For each dataset we report ROC-AUC, accuracy, sensitivity, specificity, and macro F1 on the test split, averaged over three random seeds. Class imbalance is handled with stratified sampling and class-weighted loss. Full dataset statistics, ex-

Table 11: **Filtration types.** Performance comparison using Sliding-Band filtrations (SB20 & SB40) and standard sublevel filtrations (PH), evaluated with MLP and Transformer (TR) classifiers.

	BRATS 2019 (Binary)					BRATS 2021 (Binary)					RSNA 2025 Aneurysm (Binary)				
Model	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1
PH+MLP	81.6	**82.3**	86.7	**66.6**	**88.4**	53.6	54.2	56.4	51.7	56.4	54.2	72.8	20.0	90.0	26.5
PH+TR	83.0	77.9	84.9	53.3	85.7	53.1	51.6	53.2	50.0	53.6	59.9	72.8	21.8	89.4	28.2
SB20+MLP	82.9	80.9	88.7	53.3	87.9	57.2	53.4	46.8	60.7	51.3	64.4	**77.3**	18.2	**96.5**	28.2
SB40+MLP	85.0	79.4	**90.6**	40.0	87.3	**62.0**	**61.9**	59.7	**64.3**	**62.2**	**67.2**	74.7	**49.1**	82.9	**48.7**
SB20+TR	82.1	73.5	77.4	60.0	82.0	57.8	56.8	59.7	53.6	59.2	62.3	73.3	27.3	88.2	33.3
SB40+TR	**85.3**	80.9	**90.6**	46.7	88.1	61.8	59.3	**62.9**	55.4	61.9	66.1	75.1	40.0	86.5	44.0

	ODELIA (3-class)					NoduleMNIST (Binary)					FractureMNIST (3-class)				
Model	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1
PH+MLP	54.6	48.1	36.7	68.3	36.6	72.3	76.8	43.8	85.4	43.8	59.4	45.4	40.6	70.3	40.8
PH+TR	55.4	55.3	37.8	68.4	38.0	69.4	78.7	34.4	**90.2**	40.0	57.5	42.1	40.4	69.1	40.8
SB20+MLP	**61.3**	55.8	**43.0**	**71.2**	**43.4**	**75.6**	78.4	45.3	87.0	46.4	62.9	51.7	41.9	72.4	37.3
SB40+MLP	58.6	52.9	42.5	69.6	42.2	72.9	75.2	**59.4**	79.3	**49.7**	64.0	51.3	43.1	73.1	37.5
SB20+TR	56.6	56.3	36.1	68.6	35.5	72.8	**80.0**	46.9	88.6	49.2	**66.1**	51.7	47.3	74.0	47.5
SB40+TR	55.9	**59.7**	39.1	70.8	39.1	73.7	79.4	45.3	88.2	47.5	65.4	**52.1**	**50.3**	**74.8**	**50.3**

Table 12: **Fusion types.** Performance comparison using different fusion types with an R3D-18 backbone.

	BRATS 2019 (Binary)					BRATS 2021 (Binary)					RSNA 2025 (Binary)				
Model	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1
R3D-18	85.8	83.6	**98.1**	33.3	**90.3**	57.3	54.2	**88.7**	16.1	**67.1**	59.1	74.7	5.5	**97.1**	9.5
R3D-18+SupCon	82.3	80.9	66.2	66.2	68.1	57.3	55.9	54.5	54.5	51.4	56.2	70.2	50.8	50.8	49.6
R3D-18+SB (Concat)	90.9	73.5	83.0	83.0	71.0	58.0	52.5	50.9	50.9	45.7	61.6	72.4	56.6	56.6	56.9
R3D-18+PH+SupCon	90.7	77.9	83.5	**83.5**	74.5	60.2	56.8	55.7	55.7	54.0	61.1	**75.6**	50.0	50.0	43.0
R3D-18+SB+SupCon	**91.3**	**88.2**	82.9	82.9	82.9	**69.1**	**62.7**	61.2	**61.2**	57.9	**69.9**	73.8	**62.4**	62.4	**62.9**

	ODELIA (3-class)					NoduleMNIST (Binary)					FractureMNIST (3-class)				
Model	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1	AUC	Acc.	Sens.	Spec.	F1
R3D-18	65.3	62.6	44.8	73.1	46.2	90.9	86.8	68.8	**91.5**	68.2	66.1	50.0	46.6	72.7	46.5
R3D-18+SupCon	59.9	62.1	43.2	71.6	44.2	90.9	87.1	**84.4**	84.4	81.7	65.2	49.2	47.7	72.6	48.3
R3D-18+SB (Concat)	64.3	**64.6**	38.1	69.7	36.2	91.0	87.4	83.4	83.4	81.7	67.1	52.5	46.1	73.6	46.1
R3D-18+PH+SupCon	66.4	62.6	**46.5**	**74.3**	**47.0**	87.2	85.5	76.4	76.4	77.2	70.2	52.1	52.2	74.7	51.5
R3D-18+SB+SupCon	**69.0**	**64.6**	45.2	73.6	45.9	**93.2**	**88.1**	82.1	82.1	**81.9**	**75.5**	**60.8**	**56.9**	**78.2**	**58.4**

act split counts, and download links appear in the repository and Supplement.

PH vs. SB filtration performances. In Table 9, we present the performances of Betti sequences obtained by traditional sublevel (PH) and our Sliding Band (SB) filtrations. SB filtrations consistently outperform traditional PH across all datasets, and the gains are most pronounced when paired with a Transformer head, indicating that attention benefits from the ordered bandwise topology. Even with a simple MLP, SB features surpass PH, showing that the improvement comes primarily from the filtration itself rather than the classifier. Across tasks, the narrower band (SB20) tends to favor sensitivity/recall, whereas the wider band (SB40) more often optimizes overall accuracy/F1 and AUC, highlighting a practical bandwidth trade-off between detecting positives and consolidating global context. Improvements are especially visible on the multi-class CXR benchmarks, where preserving large-scale structure matters, suggesting that global sublevel filtrations saturate too early while sliding bands retain late-emerging topology. Overall, these results support sliding-band persistence as a robust, data-efficient replacement for standard PH, with Transformers further amplifying its advantages.

Topoformer vs Baselines. In Table 10, we present the performance comparison of Topoformer

with baseline DL models. The CNN group includes ResNet18 (He et al., 2016b), DenseNet121 (Huang et al., 2017), and Xception (Chollet, 2017), which have demonstrated strong performance in medical image classification tasks. The Transformer group includes SwinV2 (Liu et al., 2022), DeiT (Touvron et al., 2021), and DaViT (Ding et al., 2022), representing state-of-the-art vision architectures. We also evaluate DaViT+SC, a variant of DaViT trained with a supervised contrastive (SupCon) loss (Khosla et al., 2020), to isolate the effect of contrastive integration without topological tokens.

Topoformer consistently matches or surpasses the strongest vision-only baselines across all four COVID benchmarks and metrics, indicating that topology-aware tokens provide complementary signal beyond what modern CNN/ViT features capture. The naïve fusion (TopoBV) already lifts performance over the same backbone, showing that Betti-based descriptors are informative; however, Topoformer's supervised-contrastive integration reliably pushes further, outperforming both DaViT and its SupCon variant (DaViT+SC). This gap is most evident on the multi-class CXR settings, where preserving global structure and late-emerging patterns matters, but the advantage also holds on CT, suggesting the approach is modality-agnostic. Relative gains are especially consistent in sensitivity and F1, a clinically relevant profile that prioritizes correct positive detection without sacrificing specificity. Taken together, the results support the view that sliding-band topological sequences capture discriminative morphology that standard patch embeddings underutilize, and that principled fusion, not mere concatenation, is key to unlocking that signal.

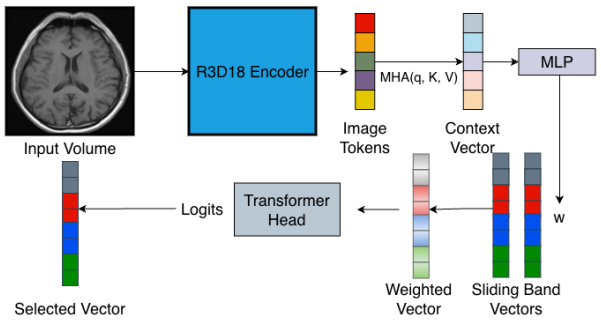

Figure 4: **TopoGate.** A 3D input volume is encoded with an R3D-18 backbone to image tokens, which are pooled by multi-head attention into a context vector. An MLP maps this context to weights (w) over a bank of sliding-band topological vectors; their weighted sum yields a task-adaptive topo vector. A transformer head produces logits, and at inference the gate can select the highest-weight band vector for interpretation.

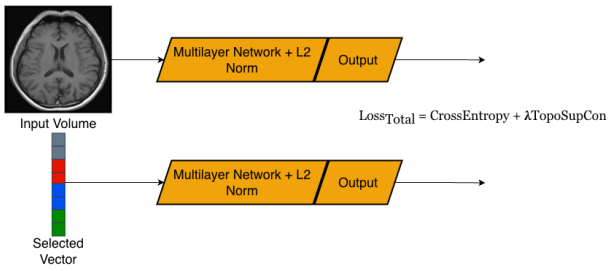

Figure 5: **TopoSupCon.** The image pathway and the selected sliding-band topological vector each pass through an identical MLP head with L2 normalization to produce embeddings and logits. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{TopoSupCon}}$, which pulls same-class image and topology embeddings together and pushes different classes apart. At inference, only the image pathway is used.

Appendix C. Topological Supervised Contrastive Learning

Conventional supervised contrastive learning relies on stochastic data augmentations (cropping, elastic warps, strong intensity jitter, etc.) to create multiple "views" of an image. In medical imaging, however, such transformations can corrupt the very signal that defines the ground truth, e.g., a crop that truncates a tumor, a deformation that alters lesion morphology, or intensity shifts that erase subtle findings. This label drift is especially problematic in low–data regimes and for case-level labels. To avoid augmentation-induced label mismatch, we generate additional views via *topological signatures* of the same image. Our sliding-band filtration converts the volume into an ordered sequence of Betti summaries that preserve global pathology cues without spatially distorting anatomy. These topology tokens act as a label-preserving second view of the original case, providing complementary information while maintaining clinical semantics (see Fig. 5).

Practically, given a volume V and its sliding-band Betti sequence $\Psi(V)$, we compute embeddings $f_\theta(V)$ and $g_\phi(\Psi(V))$, project them through a small MLP head with ℓ_2 normalization, and apply a supervised

contrastive loss with positives defined by shared class labels *within* each view and *across* views. We also fuse the two embeddings for standard cross-entropy classification, yielding a joint objective that aligns image and topology while preserving discriminative power. This design supplies multi-view supervision without risky spatial perturbations, encourages invariance to acquisition/contrast variations, and leverages the stability of our topological sequences to small intensity changes, resulting in more robust representations under limited data.

Appendix D. Pseudocodes

Algorithm 1 TopoGate: training (single step)

Require: batch of volumes $\{V_i\}_{i=1}^{B}$, labels $\{y_i\}$; per-sample topology sets $\{\mathbf{t}_i^{(w)} \in \mathbb{R}^L\}_{w=1}^{W}$; dataset stats (μ_t, σ_t); params $\Theta = \{\phi, W_q, \mathrm{MHA}, \mathrm{MLP}_g, \mathrm{TopoTransformer}, \mathrm{Classifier}\}$

for all i **in batch do**
  // Standardize topology features (per dataset)
  $\tilde{\mathbf{t}}_i^{(w)} \leftarrow (\mathbf{t}_i^{(w)} - \mu_t)/\sigma_t \ \ \forall w$
  // 3D encoder ϕ: tokens from volume
  $\mathbf{X}_i \leftarrow \mathrm{FlattenTokens}(\phi(V_i)) \in \mathbb{R}^{P\times C}$
  // HyperGate: image-conditioned weights over widths
  $\mathbf{q}_i \leftarrow W_q\Big(\frac{1}{P}\sum_{p=1}^{P}\mathbf{X}_{i,p}\Big) \in \mathbb{R}^C$
  $\mathbf{c}_i \leftarrow \mathrm{MHA}\big(\mathbf{q}_i, \mathbf{K}\in\mathbb{R}^{W\times C}, \mathbf{V}\in\mathbb{R}^{W\times C}\big)$
  $\boldsymbol{\alpha}_i \leftarrow \mathrm{MLP}_g(\mathbf{c}_i) \in \mathbb{R}^W; \quad \mathbf{w}_i \leftarrow \mathrm{softmax}(\boldsymbol{\alpha}_i)$
  // Gated topology sequence
  $\bar{\mathbf{t}}_i \leftarrow \sum_{w=1}^{W}(\mathbf{w}_i)_w \tilde{\mathbf{t}}_i^{(w)} \in \mathbb{R}^L$
  // TopoTransformer head
  $\mathbf{Z}_i \leftarrow \mathrm{TopoTransformer}(\bar{\mathbf{t}}_i)$ ▷ embed scalars, add positions, transformer
  $\hat{\mathbf{y}}_i \leftarrow \mathrm{Classifier}\big(\mathrm{MeanPool}(\mathbf{Z}_i)\big)$
end for
$\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^{B}\mathrm{CE}(\hat{\mathbf{y}}_i, y_i)$
update Θ by backprop on $\mathcal{L}$

Algorithm 2 TopoSupCon: training (single step)

Require: batch $\{(V_i, \mathbf{t}_i, y_i)\}_{i=1}^{B}$; encoders f_θ (image), g_ϕ (topology); projectors q_ψ (shared or separate); fusion/classifier h_ω; temperature τ; weight λ

for all i **do**
  $\mathbf{h}_i^{\mathrm{img}} \leftarrow f_\theta(V_i) \in \mathbb{R}^d$ ▷ R3D-18 w/ adapted stem
  $\mathbf{h}_i^{\mathrm{topo}} \leftarrow g_\phi(\mathbf{t}_i) \in \mathbb{R}^d$ ▷ e.g., 2-layer MLP
  // Projection heads (contrastive branch)
  $\mathbf{u}_i^{\mathrm{img}} \leftarrow \mathrm{norm}\big(q_\psi(\mathbf{h}_i^{\mathrm{img}})\big)$; $\mathbf{u}_i^{\mathrm{topo}} \leftarrow \mathrm{norm}\big(q_\psi(\mathbf{h}_i^{\mathrm{topo}})\big)$
  // Fusion for classification (no projections)
  $\hat{\mathbf{y}}_i \leftarrow h_\omega\big([\mathbf{h}_i^{\mathrm{img}} \,\|\, \mathbf{h}_i^{\mathrm{topo}}]\big)$
end for
// Supervised contrastive loss over $2B$ views
$\mathcal{L}_{\mathrm{SupCon}} \leftarrow \mathrm{SupConLoss}\big(\{\mathbf{u}_i^{\mathrm{img}}, \mathbf{u}_i^{\mathrm{topo}}\}_{i=1}^{B}, \{y_i\}_{i=1}^{B}, \tau\big)$
// Cross-entropy on fused logits
$\mathcal{L}_{\mathrm{CE}} \leftarrow \frac{1}{B}\sum_{i=1}^{B}\mathrm{CE}(\hat{\mathbf{y}}_i, y_i)$
update $\{\theta, \phi, \psi, \omega\}$ on $\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\,\mathcal{L}_{\mathrm{SupCon}}$

Algorithm 3 Supervised contrastive loss (helper)

function SupConLoss($\{\mathbf{u}_k\}_{k=1}^{2B}, \{y_k\}_{k=1}^{2B}, \tau$)
  // Inputs are ℓ_2-normalized projections from both modalities; labels duplicated **for** $i = 1$ *to* $2B$ **do**
    end
    $P(i) \leftarrow \{\, p \neq i \mid y_p = y_i \,\}$ ▷ positives
  $S(i) \leftarrow \sum_{a\neq i}\exp(\mathbf{u}_i^\top \mathbf{u}_a/\tau)$
  $\ell_i \leftarrow -\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{u}_i^\top \mathbf{u}_p/\tau)}{S(i)}$
  return $\frac{1}{2B}\sum_{i=1}^{2B}\ell_i$
end function

Algorithm 4 Inference

TopoGate: given $(V, \{\mathbf{t}^{(w)}\})$, compute HyperGate weights $\mathbf{w}$, form $\bar{\mathbf{t}} = \sum_w w_w \mathbf{t}^{(w)}$, run TopoTransformer → logits $\hat{\mathbf{y}}$.
TopoSupCon: given $(V, \mathbf{t})$, encode $\mathbf{h}^{\mathrm{img}}, \mathbf{h}^{\mathrm{topo}}$, fuse with h_ω (no projection heads) → logits $\hat{\mathbf{y}}$.