This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Integrating Multi-scale and Multi-filtration Topological Features for Medical Image Classification*

Pengfei Gu*
The University of Texas Rio Grande Valley
pengfei.gu01@utrgv.edu

Huimin Li*
The University of Texas Rio Grande Valley
huimin.li01@utrgv.edu

Haoteng Tang
The University of Texas Rio Grande Valley
haoteng.tang@utrgv.edu

Dongkuan Xu
North Carolina State University
dxu27@ncsu.edu

Erik Enriquez
The University of Texas Rio Grande Valley
erik.enriquez01@utrgv.edu

DongChul Kim
The University of Texas Rio Grande Valley
dongchul.kim@utrgv.edu

Bin Fu
The University of Texas Rio Grande Valley
bin.fu@utrgv.edu

Danny Z. Chen
University of Notre Dame
dchen@nd.edu

Abstract

Modern deep neural networks have shown remarkable performance in medical image classification. However, such networks either emphasize pixel-intensity features instead of fundamental anatomical structures (e.g., those encoded by topological invariants), or they capture only simple topological features via single-parameter persistence. In this paper, we propose a new topology-guided classification framework that extracts multi-scale and multi-filtration persistent topological features and integrates them into vision classification backbones. For an input image, we first compute cubical persistence diagrams (PDs) across multiple image resolutions/scales. We then develop a "vineyard" algorithm that consolidates these PDs into a single, stable diagram capturing signatures at varying granularities, from global anatomy to subtle local irregularities that may indicate early-stage disease. To further exploit richer topological representations produced by multiple filtrations, we design a cross-attention-based neural network that directly processes the consolidated final PDs. The resulting topological embeddings are fused with feature maps from CNNs or Transformers. By integrating multi-scale and multi-filtration topologies into an end-to-end architecture, our approach enhances the model's capacity to recognize complex anatomical structures. Evaluations on three public datasets show consistent, considerable improvements over strong baselines and state-of-the-art methods, demonstrating the value of our comprehensive topological perspective for robust and interpretable medical image classification.

1. Introduction

Deep neural networks (DNNs) have become commonly-used methods for medical image analysis (e.g., classification), powered by expressive visual representations learned by convolutional networks (CNNs) [11, 13, 15, 18, 19, 23, 33, 35] and, more recently, by Vision Transformers (ViTs) [8, 25, 26] that model long-range dependencies via self-attention. Despite these advances, mainstream architectures rely predominantly on encoding intensity-driven appearance and local texture features, while *global, robust anatomical structures* — such as connected components and loops that characterize tissues, organs, and lesions — remain weakly represented. This gap is consequential to medical practice, where object topology often conveys clinically meaningful patterns (e.g., lesion cavities, branches, and enclosure) [1, 12, 14, 34, 37].

Topological data analysis (TDA) [9], and in particular *persistent homology* (PH) [10], offers a powerful approach to quantify such structures. PH tracks the birth and death of topological features along a filtration, summarizing them as

*⋆ indicates equal contribution. This research was supported in part by NSF Grants CCF-2523787, IIS-2334389, 2018900, BCS-2416846, and OAC-2417850, and DoD Grant W911NF2110169.

a multi-set of points in a *persistence diagram* (PD). However, efforts of leveraging PDs in deep learning (DL) methods are non-trivial. (1) Established vectorization methods: Previous methods such as persistent images [2], persistent landscapes [4], PersLay [5], PLLay [22], and projection layers [16, 17] convert a PD (a variable-size multi-set) into a fixed-length vector via a mathematically prescribed encoding. These processes are often arduous and task-specific (e.g., choices of kernels, grids, and smoothing), and can suffer information bottlenecks that discard fine-grained geometry or higher-order interactions among PD points. (2) Data-driven PD encoding: More recently, PHG-Net [30] learns a permutation-invariant encoder that directly consumes PD points and co-trains with a vision backbone, thereby reducing hand-crafted design and retaining task-relevant signals. Yet, PHG-Net, like most prior work, operates on *single-parameter* persistence from a single filtration, and does not explicitly target *stable multi-scale* or *multi-filtration* topological signatures.

We hypothesize that exploiting stable, complementary topological cues can enhance medical image classification by aligning learned representations with clinically meaningful structures across scales and filtrations. Hence, in this paper, we propose a new approach that (1) extracts *stable* multi-scale, multi-filtration topological signatures and (2) injects them throughout the backbone to improve classification performance. Specifically, given an image, we compute cubical PDs at multiple resolutions (scales) under multiple filtration functions (e.g., intensity and gradient magnitude). We then introduce a "vineyard" tracking algorithm that associates homologous points across scales and selects reliable trajectories, producing a *stable* PD that captures topological signatures ranging from organ-level shapes to subtle boundary irregularities. To further leverage multi-filtration topological features, we generate stable PDs under different filtrations and introduce a *PDs encoder* to fuse complementary cues across filtrations via bidirectional cross-attention. The resulting topological embedding is then injected into the backbone at *multiple layers* via a channel-wise gating module with residual connections, enabling topology-aware refinement of intermediate features in the backbone models.

We evaluate the proposed method with standard DL classification networks on three public medical image datasets. Extensive experiments show that our method yields significant improvements over state-of-the-art (SOTA) methods, demonstrating the effectiveness of our new approach.

Our contributions are three-fold: (1) We propose a new framework for extracting stable multi-scale, multi-filtration topological features for medical image classification. (2) Our method is model-agnostic and can incorporate topological features into any vision backbone, including CNNs and Transformers. (3) We achieve considerable performance improvements on three public datasets compared to SOTA medical image classification methods.

2. Method

Fig. 1(a) shows the overall pipeline of our approach. Generally, given an input image I, we first compute PDs from multi-scale versions of I using different filtration functions. Stable PDs are then generated from these PDs using Algorithm 1. A PDs encoder is applied to transform the stable PDs into a topological feature vector, which is integrated with the feature maps produced by the backbone DL classification model at each stage to enable feature refinement.

This section begins with a brief review of how PDs are constructed from images. We then describe the "vineyard" algorithm for generating stable PDs from multi-scale images. Next, we present the design of our PDs encoder for embedding stable PDs into topological feature vectors. Finally, we discuss how these vectors are integrated to guide the DL classification model in feature refinement for medical image classification.

2.1. Constructing Persistence Diagrams of Images

Given an input image I, we model its pixel grid as a *cubical complex* C that takes each pixel as a 0-cell (vertex) and encodes neighborhood connectivity through 1-cells (edges), 2-cells (squares), and higher-dimensional cubes when appropriate [21, 36]. Let $f : I \to \mathbb{R}$ be a *filtration function* (e.g., the raw intensity or a derived scalar field). Varying a threshold τ induces a nested family of super-level subcomplexes $C^{(\tau)} := \{ x \in I \mid f(x) \geq \tau \}$, and, as τ decreases through finitely many critical values $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_m$, we obtain the *super-level set filtration* $\emptyset \subseteq C^{(\tau_1)} \subseteq C^{(\tau_2)} \subseteq \cdots \subseteq C^{(\tau_m)} = I$.

PH [10] tracks how homology classes evolve along this filtration. For 2D images, we mainly consider $k \in \{0, 1\}$, corresponding to connected components (0-D) and cycles/loops (1-D) in homology classes. A topological feature that first *appears* (is born) at threshold value b and then *disappears/merges* (dies) at threshold value $d < b$ is encoded by a pair (b, d). Its *persistence (lifespan)* is $pers = b - d > 0$, quantifying the range of scales over which the feature remains topologically salient (longer lifespans usually indicate more meaningful structures). PH provides a principled multi-scale description of image topology.

Collecting all birth–death pairs in a fixed dimension k yields the *k-th PD*, $D_I^k = \{ (b_i^{(k)}, d_i^{(k)}) \in \mathbb{R}^2 \mid d_i^{(k)} < b_i^{(k)} \}, k \in \{0, 1\}$, a multi-set of points that compactly summarizes when features occur and how long they persist across the filtration. PDs serve as stable, interpretable shape descriptors that encode the births and deaths of connected components and loops across thresholds. For visualization, one often plots points with $b < d$ by switching to an equivalent sub-level convention (or by replacing f with $-f$). This

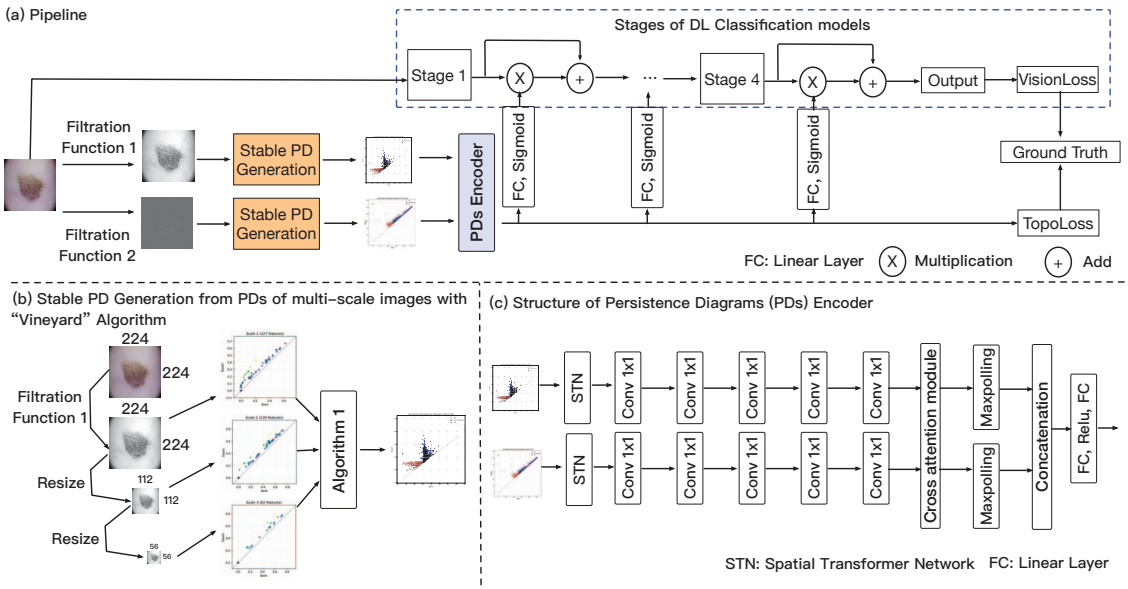


Figure 1. (a) Illustrating the pipeline of our proposed approach. (b) Illustrating the generation of a stable persistence diagram from the multi-scale persistence diagrams using the "vineyard" tracking Algorithm 1 (using a single filtration function as an example). (c) The architecture of the proposed persistence diagrams (PDs) encoder. VisionLoss and TopoLoss represent the vision loss and topological loss, respectively. We use an image from the ISIC 2018 dataset and a CNN backbone for illustration. For simplicity, the batch normalization and rectified linear unit that follow each 1D convolution layer are omitted.

monotone reparameterization keeps the underlying topology and persistences unchanged. Fig. 2 shows some examples of PDs in images from the ISIC 2018 dataset [6] using the intensity filtration function (i.e., magnitude values of the three RGB channels of pixels) and gradient filtration function (i.e., average gradient magnitude values computed by the Laplacian operator across the three RGB channels).

2.2. Computing Stable Multi-scale Topological Features with the "Vineyard" Algorithm

A single-scale topological analysis risks being either too coarse — missing subtle, localized cues such as early-stage lesions or boundary irregularities — or too fine, thereby overlooking global organ-level structures and anatomical context. By computing PDs across multiple resolutions, $\{PD_1, \ldots, PD_n\}$, we capture topological signatures at varying granularities: from large-scale anatomical shapes to fine-scale locally irregular patterns that may be clinically meaningful. Hence, this multi-scale representation is better aligned with the heterogeneous nature of medical imagery, where pathology can manifest at multiple spatial scales simultaneously.

However, not all multi-scale features are equally reliable for downstream tasks (e.g., classification). Some features are transient (noise or resolution artifacts), while others persist consistently across scales and thus reflect stable anatomical or pathological structures. To identify and retain the most reliable topological features, we propose to explicitly track putative correspondences of PD points across adjacent scales (forming "vines"), and then select those vines whose trajectories are both smooth (small inter-scale movement) and topologically significant (high persistence).

Given ordered PDs across the scales, we conduct a "vineyard" tracking procedure to assemble trajectories and select stable features. Its main steps are as follows. **(1) Multi-Scale Diagrams:** Load PDs at n image resolutions $\{PD_i\}_{i=1}^{n}$, where each PD_i contains birth–death pairs (b, d) for a fixed homology degree. **(2) Persistence-Aware Costs:** For each consecutive pair (PD_i, PD_{i+1}), we define a distance between points $p = (b_p, d_p) \in PD_i$ and $q = (b_q, d_q) \in PD_{i+1}$. **(3) Bipartite Matching:** Construct the pairwise distance matrix D on (PD_i, PD_{i+1}), and compute an optimal matching assignment with the Hungarian algorithm [3]. Keep only matches with $D[a, b] \leq \tau_m$ (the matching threshold) to avoid implausible jumps. **(4) Vine Assembly:** Each accepted match creates a segment from $p \in PD_i$ to $q \in PD_{i+1}$, annotated with its distance and endpoint persistences. If the source point is already tracked, we *extend* its vine; otherwise, we *start* a new vine at scale i. Tracked points not matched at the next scale are *terminated* (the vine "dies"). **(5) Stability Scoring:** For a vine v with segments s, we compute a persistence-weighted stability $\sigma(v) = \frac{\sum_{s \in v} w_s \cdot \frac{1}{1+\mathrm{dist}(s)}}{\sum_{s \in v} w_s}$, $w_s = \max\left(0.1, \frac{\mathrm{pers}_{\mathrm{from}}(s) + \mathrm{pers}_{\mathrm{to}}(s)}{2}\right)$, which rewards smooth trajectories while emphasizing high-persistence structure. We

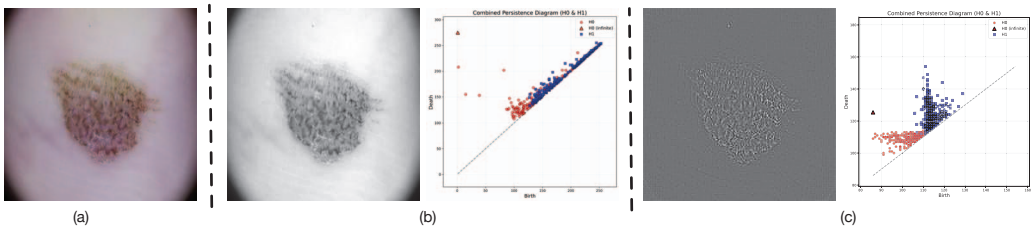


Figure 2. Examples of persistence diagrams in an image from the ISIC 2018 dataset under different filtrations. (a) An original image; (b) image with intensity values and its corresponding persistence diagram; (c) image with gradient magnitude values and its corresponding persistence diagram. In the PDs, red circles denote 0-D persistent homology (H0), and blue squares denote 1-D persistent homology (H1).

keep vines with $\sigma(v) \geq \tau_s$. **(6) Stable Diagram Construction:** Each retained vine is summarized by a representative birth–death pair taken at its *medial* scale (middle segment in practice), yielding the final stable PD PD_{stable} to be used as a compact, reliable descriptor for downstream learning.

Algorithm 1 gives the procedure for computing the final stable PD for one homology degree (e.g., H_0, H_1, or H_2), and the final stable diagram is formed by a degree-wise union: $PD_{\text{stable}} = \bigcup_{r\in\{0,1,2\}} PD^{(r)}_{\text{stable}}$. In our experiments, we compute the PDs at three image resolutions (i.e., 224×224, 112×112, and 56×56) to generate the final stable PD (see examples in Fig. 3). Unlike [7], which studied vineyards along a continuous family of filter functions, our approach constructs "vineyards" by matching PDs computed at different image resolutions.

2.3. Exploiting Multi-filtration Topological Features with the Persistence Diagrams Encoder

A single filtration (e.g., raw pixel intensity) captures only one projection of image structures. Applying *multiple* filtrations to the same image, such as intensity and gradient magnitude, accentuates complementary aspects of anatomy [14]. One filtration may sharpen tissue interfaces, while another may emphasize gradual transitions or pathology-related gradients. Formally, for given filtration functions $\{f^{(t)} : I \to \mathbb{R}\}_{t=1}^{T}$ on image I, each $f^{(t)}$ induces a stable PD using Algorithm 1, providing complementary multi-scale and multi-filtration topological features.

PDs are the most prevalent descriptors in PH, and numerous handcrafted mappings (e.g., persistence images, landscapes, Betti curves) transform PDs into vectors for downstream learning. However, these mappings impose fixed functional forms and may discard task-relevant information. To avoid such information drawbacks, we propose a *learning*, end-to-end PDs encoder that directly encodes PDs.

First, for each stable PD, we encode the sets of 2D points in the PD into expressive vectors. A naive approach for this would aggregate $\{(b, d)\}$ by max-/avg-pooling; but such pooling is order invariant yet *interaction-poor*, risking loss of higher-order relations. Therefore, we first *lift* each 2D point to a higher-dimensional feature space that enables point-point interactions before applying a symmetric aggregation. Specifically, we adapt a PointNet style [32] Spatial Transformer Network (STN) [20] to learn spatial transformations that make our network invariant to rigid transformations of 3D point clouds. Given an input point cloud represented as a tensor of shape $(\text{channels}, \text{points})$, where the channels correspond to coordinate dimensions (typically 3 for xyz coordinates), the STN predicts a single 3×3 transformation matrix. The STN architecture consists of a feature extraction backbone followed by a regression head. The backbone uses three 1D convolutional layers with channel dimensions $\{64, 128, 1024\}$, each followed by batch normalization and ReLU activation: $x^{(1)} = \text{ReLU}(\text{BN}(\text{Conv1d}_{c\to64}(x^{(0)})))$, $x^{(2)} = \text{ReLU}(\text{BN}(\text{Conv1d}_{64\to128}(x^{(1)})))$, $x^{(3)} = \text{ReLU}(\text{BN}(\text{Conv1d}_{128\to1024}(x^{(2)})))$, where $x^{(0)}$ is the input tensor and c is the number of input channels. To achieve permutation invariance across the points, we apply symmetric max-pooling, $f = \max_i x_i^{(3)}$, yielding a global feature vector $f \in \mathbb{R}^{1024}$. This global feature is then passed through a three-layer MLP regression head with dimensions $\{512, 256, 9\}$: $h^{(1)} = \text{ReLU}(\text{BN}(\text{Linear}_{1024\to512}(f)))$, $h^{(2)} = \text{ReLU}(\text{BN}(\text{Linear}_{512\to256}(h^{(1)})))$, $\hat{T} = \text{Linear}_{256\to9}(h^{(2)})$. To ensure that the transformation matrix is initialized as an identity and remains close to it during early training, we add an identity matrix residual connection, $T = \hat{T} + I_{3\times3}$, where $I_{3\times3}$ is the 3×3 identity matrix flattened to a 9-dimensional vector and then reshaped back to matrix form. The final transformation matrix $T \in \mathbb{R}^{3\times3}$ is applied to the input points via matrix multiplication, enabling the network to learn the invariance to rotations, translations, and uniform scaling of the input point cloud.

Then, we process the vectors through multiple 1D convolutions with batch normalization and ReLU activations, and fuse complementary information across the filtrations. Stable PDs produced by different filtrations contain non-redundant cues (e.g., a loop persistent in the gradient-filtration but not in intensity). An effective encoder must discover correspondences and complementarities across the PD sets and fuse them without collapsing distinct sig-

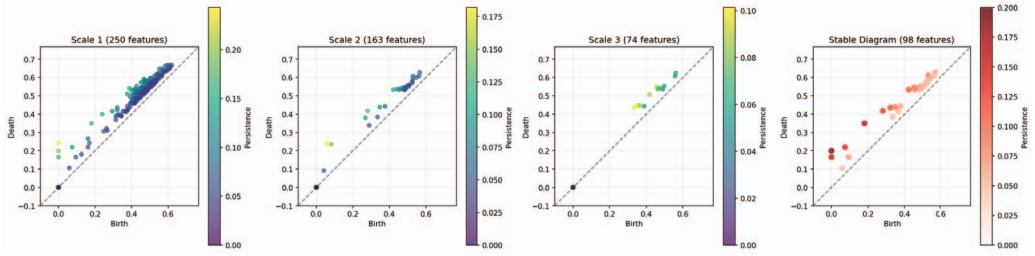


Figure 3. Examples of stable persistence diagrams generated from multi-scale persistence diagrams in images of the Kvasir dataset. Here, scale 1 corresponds to 224×224, scale 2 to 112×112, and scale 3 to 56×56.

nals. To capture correspondences among different feature sets, we employ a multi-head cross-attention mechanism. Given query features $Q \in \mathbb{R}^{N_q \times D}$, key features $K \in \mathbb{R}^{N_k \times D}$, and value features $V \in \mathbb{R}^{N_v \times D}$, we compute attended features through scaled dot-product attention. The input features are first projected through learned linear transformations, $Q' = QW_Q, K' = KW_K$, and $V' = VW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ are learnable parameter matrices. For multi-head attention with H heads, we reshape the projected features into head-specific representations with dimension $d = D/H$: $Q_h = Q'[:, h \cdot d : (h+1) \cdot d], K_h = K'[:, h \cdot d : (h+1) \cdot d]$, and $V_h = V'[:, h \cdot d : (h+1) \cdot d]$. The attention mechanism computes similarity scores between queries and keys, applies softmax normalization, and aggregates values: $\text{Attention}_h(Q, K, V) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d}}\right) V_h$. The outputs from all attention heads are concatenated and projected through a final linear layer, $\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Attention}_1, \ldots, \text{Attention}_H) W_O$, where $W_O \in \mathbb{R}^{D \times D}$ is the output projection matrix. This cross-attention mechanism enables selective information exchange among different feature sets, allowing the model to identify and leverage correspondences across heterogeneous inputs. The multi-head structure captures diverse types of relationships simultaneously, while the scaled attention prevents gradient vanishing in high-dimensional feature spaces. Finally, we concatenate the feature vectors and pass it to two MLP layers to produce the final topological feature vector, capturing richer topological representations. Fig. 1(c) illustrates the details of the PDs encoder that we design.

Following [14], we use intensity and gradient filtration functions in our experiments.

2.4. Integrating Topological Features in DL Classification Models

A straightforward integration strategy is to concatenate the topological embedding $t \in \mathbb{R}^M$ (from the PDs encoder) with the penultimate deep features and train a joint classifier. However, this *late fusion* has two drawbacks: (i) gradients for the topology and image branches are largely computed in isolation during back-propagation, limiting cross-branch interaction; (ii) concatenation at the final stage prevents topology from shaping intermediate, multi-scale representations where semantic and anatomical cues emerge.

To address this issue, we employ an attention-based residual gating method [30]. Let $\mathcal{L}$ index a set of intermediate layers of a CNN or Transformer backbone. For a layer $\ell \in \mathcal{L}$, denote the features as $F^{(\ell)} \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$ for CNNs (or $F^{(\ell)} \in \mathbb{R}^{N_\ell \times C_\ell}$ for token features). We inject a topology-conditioned *channel gate* $\mathbf{g}^{(\ell)} \in \mathbb{R}^{C_\ell}$ computed from t: $\mathbf{g}^{(\ell)} = \sigma\Big(W_1^{(\ell)} \,\text{ReLU}\big(W_2^{(\ell)} t\big)\Big)$, where $W_2^{(\ell)} \in \mathbb{R}^{\frac{C_\ell}{r} \times M}$ reduces t to a bottleneck of width C_ℓ/r, $W_1^{(\ell)} \in \mathbb{R}^{C_\ell \times \frac{C_\ell}{r}}$ restores channel dimension, σ denotes sigmoid, and $r \geq 1$ balances accuracy and cost. The gate modulates features multiplicatively (broadcast over spatial positions or tokens): $\hat{F}^{(\ell)} = F^{(\ell)} \odot \mathbf{g}^{(\ell)}$. To preserve identity mapping and stabilize training, we adopt a residual update with a learnable scalar α_ℓ: $F_{\text{ref}}^{(\ell)} = F^{(\ell)} + \alpha_\ell \hat{F}^{(\ell)}$, $\alpha_\ell \geq 0$. We optimize the network end-to-end with the following combination losses: $\mathcal{J} = \mathcal{L}_{\text{cls}}^{\text{vis}}(\hat{y}^{\text{vis}}, y) + \lambda \mathcal{L}_{\text{cls}}^{\text{topo}}(\hat{y}^{\text{topo}}, y)$, where $\mathcal{L}_{\text{cls}}^{\text{vis}}$ and $\mathcal{L}_{\text{cls}}^{\text{topo}}$ are cross-entropy terms for the vision backbone and topological branch respectively, $\hat{y}^{\text{vis}}$ and $\hat{y}^{\text{topo}}$ are the class-score outputs of the two branches, y is ground truth, and λ is sets as 0.1 to control the contribution of the topology-guided supervision (see Fig. 1(a) for more details).

3. Experiments

3.1. Datasets

ISIC 2018 [6]: This skin lesion classification dataset contains 10,015 training images and 193 validation images across 7 classes of skin disease lesions. For fair comparison, we follow the dataset split strategy in [39] using five-fold cross validation.

Kvasir [31]: This dataset contains 4,000 images, annotated and verified by experienced endoscopists, including 8 classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract, with hundreds of images for each class. Following [38], we split the dataset into 2,408/392/1,200 images for train-

Algorithm 1 "Vineyard" Tracking Algorithm for a Stable Persistence Diagram

Input: Ordered persistence diagrams (PDs) $\{PD_1, \ldots, PD_n\}$, matching threshold τ_m, stability threshold τ_s, distance metric dist
Output: Stable PD PD_{stable}

$V_{\text{active}} \leftarrow \emptyset$ {map: current-point index $\mapsto$ vine}
$V_{\text{complete}} \leftarrow \emptyset$
for $i = 1$ **to** $n - 1$ **do**
  $A \leftarrow PD_i[:, (b, d)]$, $B \leftarrow PD_{i+1}[:, (b, d)]$
  Compute pairwise distance matrix D between points of A and B using dist
  Compute optimal matching assignment M by the Hungarian algorithm on D
  $V_{\text{next}} \leftarrow \emptyset$; $S_{\text{curr}} \leftarrow \emptyset$; $S_{\text{next}} \leftarrow \emptyset$
  for each matched pair $(a \in A,\ b \in B)$ in M **do**
    if $D[a, b] \leq \tau_m$ **then**
      Form segment $s = (i,\ a \to b,\ \text{distance} = D[a, b],\ \text{pers}_a, \text{pers}_b)$
      if $a \in V_{\text{active}}$ **then**
        Append s to vine $V_{\text{active}}[a]$; $V_{\text{next}}[b] \leftarrow V_{\text{active}}[a]$
      else
        Create new vine v with segment s; $V_{\text{next}}[b] \leftarrow v$
      end if
      $S_{\text{curr}} \leftarrow S_{\text{curr}} \cup \{a\}$; $S_{\text{next}} \leftarrow S_{\text{next}} \cup \{b\}$
    end if
  end for
  for each a such that $a \in V_{\text{active}}$ and $a \notin S_{\text{curr}}$ **do**
    Set death_scale of $V_{\text{active}}[a]$ to $i{+}1$ and move it to V_{complete}
  end for
  $V_{\text{active}} \leftarrow V_{\text{next}}$
end for
Move all vines remaining in V_{active} to V_{complete} with death_scale $= n$
$PD_{\text{stable}} \leftarrow \emptyset$
for each vine $v \in V_{\text{complete}}$ **do**
  Compute stability score $\sigma(v)$
  if $\sigma(v) \geq \tau_s$ **then**
    Let $m \leftarrow \lfloor |v.\text{segments}|/2 \rfloor$ (middle segment index)
    Add the middle segment's source point (b, d) to PD_{stable}
  end if
end for
return PD_{stable}

ing/validation/test, and apply five-fold cross validation.

CBIS-DDSM [24]: The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) consists of 1,566 participants and 6,775 studies, categorized as benign or malignant. We adopt the train/test split provided by the dataset, with 20% of the cases reserved for testing and the rest for training.

3.2. Implementation Details

Our experiments are conducted with the PyTorch framework. We employ the Cubical Ripser software [21] to compute PDs. Model training is performed on an NVIDIA Tesla V100 Graphics Card with 32GB GPU memory using the AdamW optimizer [27] with a weight decay = 0.005. The learning rate is 0.0001, and the number of training epochs is 400 for the experiments. The batch size for each case is set as the maximum size allowed by the GPU. We apply standard data augmentation, such as rotation, random flip, random crop, etc. to avoid overfitting.

3.3. Results and Analysis

We evaluate three strong vision backbones: ResNet152 [15], SENet154 [18], and SwinV2-B [26]. For each backbone, we conduct a controlled ablation by progressively enabling: (i) a topology branch following PHG [30], (ii) + *multi-scale* vineyard stabilization across image resolutions, and (iii) + *multi-filtrations* (full model) which fuses intensity- and gradient-derived stable PDs via our PDs encoder and applies multi-scale topology-guided refinement (Sec. 2). We report results in Accuracy, AUC, Sensitivity, and Specificity.

We compare our method against PHG-Net [30], a SOTA topology-based DL classification method. Note that we do not additionally compare to two common PD vectorization approaches, PersLay [5] and PLLay [22], because PHG-Net has already shown that it outperforms these vectorizers on medical image classification tasks. We also do not compare to TopoImages [14], as it is a hand-crafted data augmentation method that encodes local topology rather than learning from PDs for medical image classification. Besides, we do not include CuMPerLay [29], a differentiable layer for cubical multiparameter persistence, since our method builds on *stable single-parameter PDs* aggregated across scales and filtrations rather than full multiparameter persistence.

Tables 1, 2, and 3 present the experimental results. We observe the following. (1) Our method consistently outperforms all three backbones and PHG-Net [30] across the three datasets and the performance metrics. In particular, we improve the accuracy by 4.55% over SENet154 on the ISIC 2018 dataset and by 3.08% over PHG-Net [30] on the CBIS-DDSM dataset. These results show the effectiveness of the proposed method. (2) Adding the "vineyard" algorithm across resolutions (while using a single filtration) yields average gains of $+0.64$ Acc (range $[+0.33, +0.85]$), $+0.64$ AUC ($[+0.10, +1.99]$), $+1.48$ Sens ($[+0.33, +2.58]$), and $+0.47$ Spec ($[+0.11, +1.39]$)

Table 1. Experimental results of different methods on the ISIC 2018 dataset. The best results are highlighted in **bold**.

Method	Accuracy	AUC	Sensitivity	Specificity
ResNet152 [15]	87.68	96.70	70.82	96.60
ResNet152 + PHG [30]	89.63	97.47	76.62	97.07
ResNet152 + PHG + multi-scale	90.16	98.34	78.97	97.31
ResNet152 + PHG + multi-scale + multi-filtrations (Ours)	91.56	98.81	81.11	97.48
SENet154 [18]	88.03	97.03	75.94	96.83
SENet154 + PHG [30]	90.47	97.75	78.55	97.31
SENet154 + PHG + multi-scale	91.13	98.20	81.13	97.43
SENet154 + PHG + multi-scale + multi-filtrations (Ours)	92.58	98.95	82.73	97.82
SwinV2-B [26]	89.03	97.77	79.63	97.03
SwinV2-B + PHG [30]	90.87	98.13	82.36	97.39
SwinV2-B + PHG + multi-scale	91.60	98.74	83.97	97.68
SwinV2-B + PHG + multi-scale + multi-filtrations (Ours)	**92.94**	**99.10**	**85.56**	**98.12**

Table 2. Experimental results of different methods on the Kvasir dataset. The best results are highlighted in **bold**.

Method	Accuracy	AUC	Sensitivity	Specificity
ResNet152 [15]	75.92	96.33	75.92	96.99
ResNet152 + PHG [30]	77.52	97.03	77.52	97.59
ResNet152 + PHG + multi-scale	78.27	97.18	78.27	97.70
ResNet152 + PHG + multi-scale + multi-filtrations (Ours)	79.52	97.63	79.52	**97.88**
SENet154 [18]	78.00	97.45	78.00	96.50
SENet154 + PHG [30]	79.00	97.85	79.00	97.00
SENet154 + PHG + multi-scale	79.33	98.01	79.33	97.19
SENet154 + PHG + multi-scale + multi-filtrations (Ours)	80.42	**98.11**	80.42	97.43
SwinV2-B [26]	78.58	97.35	78.58	97.08
SwinV2-B + PHG [30]	79.78	97.05	79.78	97.09
SwinV2-B + PHG + multi-scale	80.36	97.15	80.36	97.27
SwinV2-B + PHG + multi-scale + multi-filtrations (Ours)	**81.52**	97.25	**81.52**	97.40

Table 3. Experimental results of different methods on the CBIS-DDSM dataset. The best results are highlighted in **bold**.

Method	Accuracy	AUC	Sensitivity	Specificity
ResNet152 [15]	72.16	77.81	73.72	70.93
ResNet152 + PHG [30]	74.29	79.16	72.24	71.96
ResNet152 + PHG + multi-scale	75.14	79.86	74.05	73.35
ResNet152 + PHG + multi-scale + multi-filtrations (Ours)	76.43	80.29	75.19	74.47
SENet154 [18]	70.88	77.44	72.74	70.70
SENet154 + PHG [30]	74.15	80.23	73.44	72.94
SENet154 + PHG + multi-scale	74.95	80.95	75.07	73.31
SENet154 + PHG + multi-scale + multi-filtrations (Ours)	77.23	82.43	76.80	74.91
SwinV2-B [26]	73.15	80.12	72.27	73.22
SwinV2-B + PHG [30]	75.74	83.98	75.13	75.03
SwinV2-B + PHG + multi-scale	76.28	85.97	76.84	76.35
SwinV2-B + PHG + multi-scale + multi-filtrations (Ours)	**78.40**	**87.22**	**78.35**	**77.89**

when averaged over all the datasets/backbones. This supports the premise that *scale-stable* topological signatures suppress resolution-specific artifacts while retaining clinically salient structures. (3) Fusing stable PDs from different filtrations (intensity and gradient) provides the largest improving boost. For example, on the CBIS-DDSM dataset, the accuracy is improved by 2.28% and 2.12% using SENet154 and SwinV2-B, respectively. In addition, improvements are particularly pronounced in Sensitivity (crucial for screening) while Specificity remains high, indicat-

Table 4. Accuracy results for different matching and stability thresholds in Algorithm 1, evaluated on the ISIC 2018 dataset with ResNet152.

	Stability Threshold τ_s		
Matching Threshold τ_m	0.6	0.7	0.8
0.2	89.00	90.67	88.83
0.3	90.50	**91.56**	90.17
0.4	89.83	90.33	89.58

ing that topology-guided refinement improves true-positive detection without sacrificing false-positive control. Together, these parts yield consistent Accuracy/AUC gains and marked Sensitivity improvements across the backbones/datasets, validating the effectiveness of multi-scale, multi-filtration topologies for medical image classification.

3.4. Ablation Study

The "vineyard" algorithm (Algorithm 1) has two key hyperparameters: the *matching threshold* τ_m, which controls how aggressively PD points are linked across adjacent scales, and the *stability threshold* τ_s, which filters completed vines based on their persistence-weighted smoothness score $\sigma(v)$. Intuitively, a small τ_m risks breaking valid correspondences (vine fragmentation), while a large τ_m may link unrelated points (spurious vines). Similarly, a small τ_s retains noisy or unstable vines, whereas a large τ_s may prune away informative but moderately stable structures.

Our experiments evaluate $\tau_m \in \{0.2, 0.3, 0.4\}$ and $\tau_s \in \{0.6, 0.7, 0.8\}$ on the ISIC 2018 dataset using ResNet152, with all the other settings fixed. Table 4 reports the accuracy results. The highest accuracy, 91.56%, occurs at $(\tau_m, \tau_s) = (0.3, 0.7)$. Therefore, we set $\tau_m = 0.3$ and $\tau_s = 0.7$ as the default values and adopt this setting in our experiments, as it balances recall (sufficient cross-scale linking) with precision (filtering to stable vines).

Furthermore, we conduct experiments to investigate the effects of different distance metrics in the "vineyard" algorithm. We compare three metrics between points $p = (b_p, d_p)$ and $q = (b_q, d_q)$ with persistence $\text{pers}_p = \max(0, d_p - b_p)$: (1) Wasserstein Distance [28]: $d_{\text{W}}(p,q) = \sqrt{(b_p - b_q)^2 + (d_p - d_q)^2}$, (2) Persistence-scaled Distance: $d_{\text{PS}}(p,q) = d_{\text{W}}(p,q) \Big/ \left(1 + \frac{\text{pers}_p + \text{pers}_q}{2}\right)$, and (3) Relative Persistence Distance: $d_{\text{RP}}(p,q) = d_{\text{W}}(p,q) \left(1 + \frac{|\text{pers}_p - \text{pers}_q|}{\max(\text{pers}_p, \text{pers}_q)}\right)$. We evaluate on the CBIS-DDSM dataset with SwinV2-B. Table 5 presents the accuracy results. We observe that the Relative Persistence Distance attains the best performance (78.40%), outperforming Wasserstein (73.30%; $\uparrow$ 5.10%) and Persistence-Scaled (73.72%; $\uparrow$ 4.68%). Therefore, we use the Relative Persistence Distance (d_{RP}) as the default distance metric in Algorithm 1 for our experiments.

Table 5. Results of different distance metrics in Algorithm 1, evaluated on the CBIS-DDSM dataset with SwinV2-B.

Method	Accuracy
Wasserstein Distance	73.30
Persistence-Scaled Distance	73.72
Relative Persistence Distance	**78.40**

Table 6. Parameter complexity comparison of different methods.

Method	Number of Parameters
SwinV2-B	86.908 M
SwinV2-B + PHG	94.455 M
Ours	144.238 M

3.5. Time and Parameter Complexity

Our pipeline for computing stable PDs consists of the following main steps: (i) constructing cubical PDs at three image scales (224×224, 112×112, and 56×56) with two filtrations (intensity and gradient), and (ii) applying the "vineyard" algorithm across the PDs of adjacent scales to compute stable PDs. Here, we report the execution time on the Kvasir dataset (4,000 images). For the intensity filtration, computing PDs takes 32 seconds, 28 seconds, and 20 seconds at the 224×224, 112×112, and 56×56 resolutions, respectively. The "vineyard" algorithm requires 22 minutes and 56 seconds to generate stable PDs from these PDs. For the gradient filtration, computing PDs takes 38 seconds, 30 seconds, and 18 seconds at the three scales, respectively, and the "vineyard" algorithm takes 1 hour, 13 minutes, and 8 seconds for computing the stable PD. It is worth noting that PD computation is performed offline, and this process needs to be conducted only once.

To examine the parameter complexity, Table 6 reports the numbers of parameters for SwinV2-B, SwinV2-B + PHG, and our proposed approach for comparison.

4. Conclusions

In this paper, we introduced a new topology-guided framework for medical image classification that augments intensity-based representations with stable topological signatures which are aggregated across scales and filtrations. We computed cubical PDs at multiple resolutions/scales and consolidate them via a "vineyard" algorithm to capture stable topological signatures ranging from global anatomy to fine local irregularities. To further exploit richer topological representations derived from multiple filtrations, we designed a cross-attention-based neural network specifically for processing the final PDs, producing task-relevant embeddings that are fused with CNN/Transformer features in an end-to-end pipeline. Experiments on three public datasets showed that our method yields consistent, considerable gains over the baselines and SOTA methods, demonstrating the effectiveness of our new approach.

References

[1] Diego Adame, Jose A Nunez, Fabian Vazquez, Nayeli Gurrola, Huimin Li, Haoteng Tang, Bin Fu, and Pengfei Gu. Topo-VM-UNetV2: Encoding topology into Vision Mamba UNet for polyp segmentation. In *Proceedings of the IEEE 38th International Symposium on Computer-Based Medical Systems*, pages 258–263, 2025. 1

[2] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence Images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18 (8):1–35, 2017. 2

[3] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993. 3

[4] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015. 2

[5] Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. PersLay: A neural network layer for persistence diagrams and new graph topological signatures. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 2786–2796, 2020. 2, 6

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. 3, 5

[7] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*, pages 119–126, 2006. 4

[8] 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. 1

[9] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Soc, 2010. 1

[10] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002. 1, 2

[11] Pengfei Gu, Hao Zheng, Yizhe Zhang, Chaoli Wang, and Danny Z Chen. kCBAC-Net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347, 2021. 1

[12] Pengfei Gu, Yejia Zhang, Huimin Li, Chaoli Wang, and Danny Z Chen. Self pre-training with topology-and spatiality-aware masked autoencoders for 3D medical image segmentation. *arXiv preprint arXiv:2406.10519*, 2024. 1

[13] Pengfei Gu, Zihan Zhao, Hongxiao Wang, Yaopeng Peng, Yizhe Zhang, Nishchal Sapkota, Chaoli Wang, and Danny Z Chen. Boosting medical image classification with segmentation foundation model. *arXiv preprint arXiv:2406.11026*, 2024. 1

[14] Pengfei Gu, Hongxiao Wang, Yejia Zhang, Huimin Li, Chaoli Wang, and Danny Chen. TopoImages: Incorporating local topology encoding into deep learning models for medical image classification. *arXiv preprint arXiv:2508.01574*, 2025. 1, 4, 5, 6

[15] 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, 2016. 1, 6, 7

[16] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[17] Christoph Hofer, Florian Graf, Bastian Rieck, Marc Niethammer, and Roland Kwitt. Graph filtration learning. In *Proceedings of the International Conference on Machine Learning*, pages 4314–4323, 2020. 2

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 1, 6, 7

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 1

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial Transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015. 4

[21] Shizuo Kaji, Takeki Sudo, and Kazushi Ahara. Cubical Ripser: Software for computing persistent homology of image and volume data. *arXiv preprint arXiv:2005.12692*, 2020. 2, 6

[22] Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Sik Kim, Frédéric Chazal, and Larry Wasserman. PLLay: Efficient topological layer based on persistence landscapes. *Advances in Neural Information Processing Systems*, 2020. 2, 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[24] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4(1):1–9, 2017. 6

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al.

Swin Transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 1, 6, 7

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[28] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011. 8

[29] Brighton Nuwagira, Caner Korkmaz, Baris Coskunuzer, and Tolga Birdal. CuMPerLay: Learning cubical multiparameter persistence vectorizations. In *Proceedings of the 2025 Fall Southeastern Sectional Meeting*. 6

[30] 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*, pages 7583–7592, 2024. 2, 5, 6, 7

[31] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017. 5

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 4

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[34] Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *Proceedings of the International Conference on Machine Learning*, pages 32698–32727, 2023. 1

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 1

[36] Hubert Wagner, Chao Chen, and Erald Vuçini. Efficient computation of persistent homology for cubical data. In *Proceedings of the Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*, pages 91–106, 2011. 2

[37] Fan Wang, Zhilin Zou, Nicole Sakla, Luke Partyka, Nil Rawal, Gagandeep Singh, Wei Zhao, Haibin Ling, Chuan Huang, Prateek Prasanna, et al. TopoTxR: A topology-guided deep convolutional network for breast parenchyma learning on DCE-MRIs. *Medical Image Analysis*, 99: 103373, 2025. 1

[38] Yubiao Yue and Zhenzhang Li. MedMamba: Vision Mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024. 5

[39] Jiaxin Zhuang, Weipeng Li, Siyamalan Manivannan, Roy Wang, JianGuo Zhang, Jihan Liu, Jiahui Pan, Gongfa Jiang, and Ziyu Yin. Skin lesion analysis towards melanoma detection using deep neural network ensemble. *ISIC Challenge*, 2018(2):1–6, 2018. 5