Medical Image Analysis 113 (2026) 104174

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

TomoGraphView: 3D medical image classification with omnidirectional slice representations and graph neural networks

Johannes Kiechle [a,b,c,g,*], Stefan M. Fischer [a,b,c,g], Daniel M. Lang [a,c], Cosmin I. Bercea [a,c,g], Matthew J. Nyflot [f], Lina Felsner [a,c], Julia A. Schnabel [a,c,e,g,1], Jan C. Peeken [b,d,1]

[a] Institute for Computational Imaging and AI in Medicine, School of Computation, Information and Technology, Technical University of Munich (TUM), Germany
[b] Department of Radiation Oncology, TUM School of Medicine, TUM University Hospital Rechts der Isar, Technical University of Munich (TUM), Germany
[c] Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Neuherberg, Germany
[d] Institute of Radiation Medicine, Helmholtz Munich, Neuherberg, Germany
[e] School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom
[f] Department of Radiation Oncology, University of Washington, Seattle, United States of America
[g] Munich Center for Machine Learning (MCML), Munich, Germany

ARTICLE INFO

Dataset link: http://github.com/compai-lab/2025-MedIA-kiechle, https://pypi.org/project/OmniSlicer

Keywords:
3D medical image classification
Omnidirectional volume slicing
Graph neural networks

ABSTRACT

The sharp rise in medical tomography examinations has created a demand for automated systems that can reliably extract informative features for downstream tasks such as tumor characterization. Although 3D volumes contain richer information than individual slices, effective 3D classification remains difficult: volumetric data encode complex spatial dependencies, and the scarcity of large-scale 3D datasets has constrained progress toward 3D foundation models. As a result, many recent approaches rely on 2D vision foundation models trained on natural images, repurposing them as feature extractors for medical scans with surprisingly strong performance. Despite their practical success, current methods that apply 2D foundation models to 3D scans via slice-based decomposition remain fundamentally limited. Standard slicing along axial, sagittal, and coronal planes often fails to capture the true spatial extent of a structure when its orientation does not align with these canonical views. More critically, most approaches aggregate slice features independently, ignoring the underlying 3D geometry and losing spatial coherence across slices.

To overcome these limitations, we propose *TomoGraphView*, a novel framework that integrates *omnidirectional* volume slicing with *spherical graph-based feature aggregation*. Instead of restricting the model to axial, sagittal, or coronal planes, our method samples both canonical and non-canonical cross-sections generated from uniformly distributed points on a sphere enclosing the volume. Triangulating these viewpoints yields a spherical graph that captures spatial relationships among views, and we use a graph neural network to aggregate their features accordingly. Experiments across six oncology 3D medical image classification datasets demonstrate that *omnidirectional* volume slicing improves the average performance in Area Under the Receiver Operating Characteristic Curve (AUROC) from 0.7701 to 0.8154 compared with traditional slicing approaches relying on canonical view planes. Moreover, we can further improve AUROC performance from 0.8198 to 0.8372 by leveraging our proposed graph neural network-based feature aggregation. Notably, *TomoGraphView* also surpasses large-scale pretrained 3D medical imaging models across all datasets and tasks, underscoring its effectiveness as a powerful framework for volumetric analysis and therefore represents a key step toward bridging the gap until fully native 3D foundation models become available in medical image analysis. We provide a user-friendly library for *omnidirectional* volume slicing at https://pypi.org/project/OmniSlicer.

1. Introduction

Deep learning methods have become a central driver of progress in medical image analysis, owing to their ability to learn complex hierarchical and semantically rich feature representations from large-scale imaging data. These capabilities enable a broad spectrum of clinical applications, including early disease detection and accurate diagnosis (Rana and Bhushan, 2023), malignancy prediction (Prakash

* Corresponding author.
E-mail address: johannes.kiechle@tum.de (J. Kiechle).
[1] These senior authors contributed equally to this work.

https://doi.org/10.1016/j.media.2026.104174
Received 17 December 2025; Received in revised form 10 June 2026; Accepted 11 June 2026
Available online 22 June 2026

and Visakha, 2020; Naser and Deen, 2020; Navarro et al., 2021), and lesion characterization (Herent et al., 2019; Yin et al., 2023; Dadoun et al., 2022), as well as robust image segmentation (Ronneberger et al., 2015; Isensee et al., 2021; Cao et al., 2022; Erdur et al., 2025; Buchner et al., 2023) and both prognostic and predictive modeling (Zhu et al., 2020; Schulz et al., 2021; Spohn et al., 2023). While these capabilities may not only facilitate personalized treatment strategies, they also hold the potential to improve therapeutic effectiveness and patient outcomes (Rasuli, 2025). As advances in imaging hardware continue to accelerate the acquisition of high-resolution tomographic data, the need for automated systems capable of distilling meaningful imaging biomarkers from complex 3D volumes to support clinicians has grown correspondingly (Thrall et al., 2018). Effective extraction of such biomarkers is not only essential for alleviating clinician workload but also constitutes a foundational step for enhancing the efficacy of subsequent downstream analytical tasks (Rasuli, 2025).

One promising direction to extract semantically rich and versatile imaging features from 2D images lies in the use of vision foundation models, which have demonstrated remarkable generalization capabilities across diverse 2D computer vision tasks (Radford et al., 2021; Kirillov et al., 2023; Zou et al., 2023; Oquab et al., 2023; Fang et al., 2024; Tschannen et al., 2025; Venkataramanan et al., 2025; Siméoni et al., 2025). These models excel at producing both substantially meaningful and robust feature embeddings and can be readily adapted to a wide range of downstream tasks, including segmentation (Ayzenberg et al., 2024), classification (Veasey and Amini, 2024), or registration (Song et al., 2024). However, despite their widespread adoption in 2D natural image analysis, efforts to translate foundation model principles to 3D medical imaging have progressed more slowly. A primary constraint remains the absence of large and publicly available multimodal medical imaging datasets required for training at scale (Azad et al., 2023). According to Paschali et al. (2025), a radiology foundation model should couple large-scale architectures with multimodal imaging data, rely on self-supervised training to limit annotation burdens, and exhibit emergent capabilities beyond its explicit training objectives. Yet, despite recent progress and rapid development of large-scale pretrained 3D models, the performance of 3D radiology foundation models still trails that of state-of-the-art 2D vision model counterparts (van Veldhuizen et al., 2025). Nevertheless, despite challenges such as the limited availability of large-scale 3D datasets and the computational demands of training 3D models, ongoing research in these areas is likely to lead to substantial breakthroughs and enhanced integration of native 3D models into clinical workflows (Azad et al., 2023; van Veldhuizen et al., 2025).

Despite the current performance gap between large-scale 2D and 3D pretrained models, recent studies have demonstrated that 2D vision foundation models trained on natural images, particularly the DINO (Caron et al., 2021) model family, can be transferred to medical applications with strong results (Huix et al., 2024; Liu et al., 2025). Their use, however, necessitates decomposing 3D medical volumes into 2D slices, a process that inevitably disrupts the intrinsic spatial coherence of volumetric data.

A common strategy to combine the resulting slice-wise information extracted by a 2D foundation model is to use a dimension-wise average thereof or a shallow multi-layer perceptron (MLP) (Kiechle et al., 2024; Müller-Franzes et al., 2025; Liu et al., 2025). While this enables end-to-end learning for downstream tasks such as classification, there is no inherent mechanism that explicitly accounts for the relative positions among the slices. More structured approaches attempt to mitigate this limitation: LSTM-based aggregation (Donahue et al., 2015; Nguyen et al., 2020; Lilhore et al., 2025) treats sequential slices analogously to temporal data, capturing inter-slice associations, while transformer-based models (Müller-Franzes et al., 2025; Meneghetti et al., 2025) leverage self-attention to more directly model spatial relationships across slices. Still, these methods ultimately remain constrained by the inherent loss of geometry introduced during slice-based decomposition.

With the emergence of graph neural networks (GNNs) (Scarselli et al., 2008), effective modeling of naturally graph-structured data has become possible. GNNs are particularly well-suited for problems involving pairwise interactions or modeling complex relational data by encoding relationships in the graph topology (Wu et al., 2020). In the context of 3D data, GNNs have demonstrated a strong ability to preserve an underlying spatial structure when the graph topology is derived on the basis of the relative position of 2D representations (Wei et al., 2020, 2023). In medical imaging, this makes GNNs well-suited to capture spatial relationships between encoded slices, thereby preserving aspects of the underlying 3D context that are lost in traditional slice-wise representations (Kiechle et al., 2024; Di Piazza et al., 2025).

In this work, we introduce *TomoGraphView*, a novel framework that preserves the spatial structure of volumetric medical images while leveraging the representational power of a 2D foundation model encoder. The name reflects its core components: *Tomo* refers to tomography data, *Graph* denotes the use of GNN-based feature aggregation, and *View* signifies the integration of multi-view information from several perspectives. More specifically, our framework integrates two key components: *omnidirectional* volume slicing and *spherical graph-based feature aggregation*. First, rather than relying solely on canonical axial, coronal, or sagittal planes, we sample cross-sectional views by means of uniformly distributed points on a sphere surrounding the 3D volume, capturing a diverse set of canonical and non-canonical perspectives. These viewpoints are triangulated into a spherical mesh that defines the graph topology, with each node corresponding to features extracted from a 2D vision foundation model applied to a specific slice. Edges between nodes are weighted according to their relative distance, enabling the model to efficiently integrate both local and global context while explicitly preserving spatial structure. By aggregating node features through a GNN, our framework effectively models 3D context from 2D representations, demonstrating the strength of graph-based feature aggregation for volumetric data analysis. The primary contributions and key findings of this work can be summarized as follows:

- We propose a novel out-of-plane slicing technique, termed *omnidirectional* volume slicing, which extends traditional slicing beyond canonical anatomical axes and yields consistently superior downstream performance compared with standard in-plane axial-coronal-sagittal volume slicing approaches.
- We introduce a spherical mesh-based graph topology derived from the *omnidirectional* sampling strategy, enabling faithful preservation of spatial relationships between multi-view 2D cross-sections within a unified *spherical graph-based feature aggregation* framework.
- Extensive experiments on six oncology 3D medical imaging datasets demonstrate that the proposed *TomoGraphView* framework, integrating *omnidirectional* volume slicing with *spherical graph-based feature aggregation*, consistently outperforms slice-wise aggregation baselines and large-scale pretrained 3D medical models across all datasets and tasks, establishing *TomoGraphView* as a powerful and generalizable framework for 3D medical image classification.

2. Related works

In the context of preserving the spatial structure of volumetric data when decomposed into 2D slices compatible with 2D encoders, we review prior work in two main areas. In Section 2.1, we begin with 2D slice-based methods for 3D medical image analysis, focusing specifically on approaches that extract slices directly from volumetric data. Projection-based methods, such as maximum intensity projections (Antropova et al., 2018; Hu et al., 2021; Takahashi et al., 2022), which collapse 3D volumes into 2D representations, are therefore excluded from this discussion. Subsequently, in Section 2.2, we turn our attention to slice-wise feature aggregation methods.

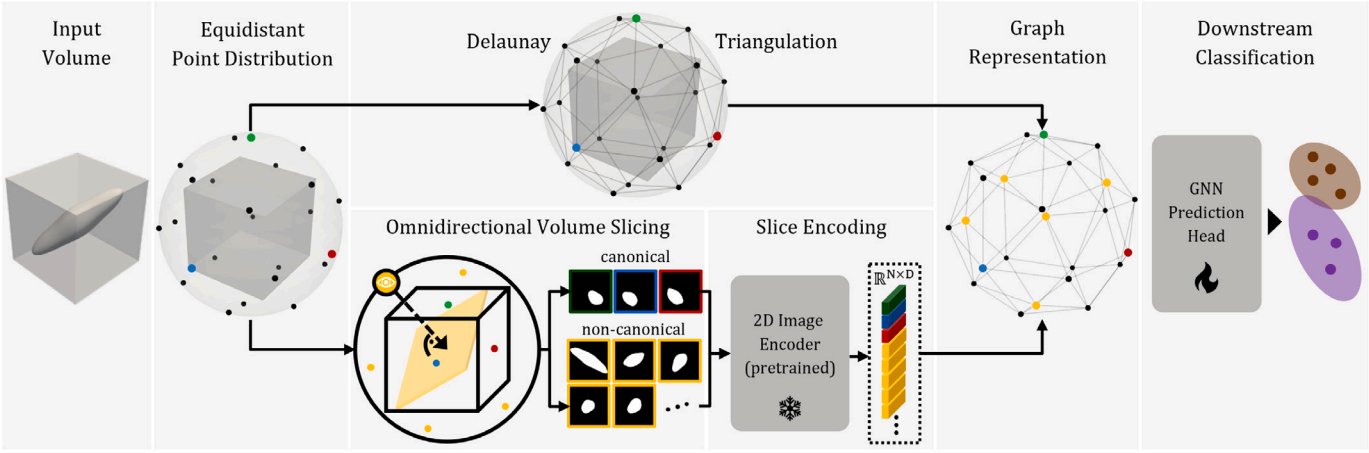

Fig. 1. Visual overview of the *TomoGraphView* framework. The input volume is enclosed within a sphere with N equidistant surface points, three corresponding to canonical planes indicated as green (axial), blue (coronal), red (sagittal), and the remaining $N-3$ distributed via repulsion-based optimization. From this distribution, two processing branches emerge: Delaunay triangulation forms a spherical mesh defining the graph topology (done only once, for any N node graph topology), and *omnidirectional* volume slicing generates canonical and non-canonical slices along the directions of the sampled points. The resulting slices are encoded into 1D vector representations that serve as node features within the spherical graph. A GNN prediction head performs the downstream classification.

2.1. Slice-based methods in 3D medical imaging

Two-dimensional slice-based approaches remain an appealing alternative to computationally intensive 3D architectures, as they offer substantially lower memory and compute requirements, benefit from the decomposition of 3D volumes into numerous 2D slices, and can leverage powerful pretrained backbones from natural image analysis. These models have repeatedly demonstrated strong performance in medical image classification tasks, underscoring their practical relevance despite the lack of explicit volumetric reasoning (Huix et al., 2024; Liu et al., 2025).

A substantial body of work highlights the effectiveness of 2D paradigms across diverse clinical settings. Navarro et al. (2021) fine-tuned a DenseNet161 pretrained on ImageNet using axial T1- and T2-weighted MRI slices of soft-tissue sarcomas for tumor grading. Similarly, Ahamed et al. (2023) trained a ResNet18 on axial PET/CT slices of lymphoma to classify whether a slice intersected tumor tissue. Wang et al. (2024) proposed a two-stage slice-to-volume feature representation framework for Alzheimer's disease, where informative axial MRI slices were selected and aggregated into volume-level features. In parallel, self-supervised foundation models have also been evaluated for slice-level tasks. Huang et al. (2024) and Baharoon et al. (2023) applied DINOv2 to axial brain MRIs and chest CTs for glioma grading and COVID-19 diagnosis, respectively, while Liu et al. (2025) assessed DINOv3 on non-contrast axial CT slices for detecting 18 clinically significant abnormalities.

Building on the limitations of single-plane analysis, several studies have explored multi-plane, or so-called 2.5D, strategies that integrate axial, coronal, and sagittal views. Saint-Esteven et al. (2022) employed an Xception model trained on 2.5D representations to predict HPV status in oropharyngeal cancer. Similarly, Meneghetti et al. (2025) explored foundation model-based multiple instance learning for head and neck cancer outcomes, comparing single-plane, multi-plane, and 3D sub-volume representations. A related hybrid approach by Jang and Hwang (2022) first extracted volumetric features using a 3D CNN, followed by multi-plane slicing of the feature tensor and 2D CNN and transformer processing for Alzheimer's disease classification.

Although empirically successful, these methods rely exclusively on canonical anatomical planes aligned with the axial, coronal, and sagittal axes, potentially missing structures that are obliquely or irregularly oriented. Beyond the slicing strategy, the aggregation of slice-wise features is critical for preserving volumetric context, as it determines how well spatial coherence is maintained. This motivates a closer examination of aggregation strategies in the following section.

2.2. Slice-wise feature aggregation in 3D medical imaging

Slice-wise feature aggregation constitutes a promising paradigm for 3D medical image analysis, enabling the use of powerful 2D foundation models while avoiding the computational cost of end-to-end 3D training. However, because slicing discards volumetric structure, performance depends on how effectively inter-slice spatial dependencies are modeled. A common strategy collapses slice embeddings into a single volume-level descriptor using non-learnable operations such as mean pooling. For example, Müller-Franzes et al. (2025) encoded axial slices with DINOv2 and averaged embeddings for breast MRI, lung CT, and knee MRI classification, while Liu et al. (2025) similarly averaged DINOv3 features from axial chest CT slices for anatomical classification. Although efficient, these approaches ignore inter-slice relationships and provide limited volumetric context.

Learnable fusion mechanisms offer a natural extension. Multi-layer perceptrons (MLPs) have been widely adopted to integrate slice-level features, as in Truong et al. (2021), who fine-tuned self-supervised ImageNet models with an MLP head for several medical classification tasks. Baharoon et al. (2023) scaled this evaluation to 200 radiology experiments across diverse modalities (X-ray, CT, and MRI), and Kiechle et al. (2024) benchmarked MLP aggregation across MedMNIST3D structures. Despite their flexibility, MLPs typically lack explicit awareness of slice order or spatial layout unless augmented with positional conditioning.

Recurrent architectures address this limitation by treating slices as ordered sequences. LSTM-based aggregation has been shown to enhance volumetric reasoning, as demonstrated by Zhang et al. (2019) for multiple sclerosis lesion segmentation and by Nguyen et al. (2020) for intracranial hemorrhage detection. Bidirectional variants have also been explored, such as the EfficientNet-B0 + BiLSTM hybrid for breast cancer classification (Lilhore et al., 2025). Slice-by-slice CNN–LSTM hybrids further illustrate the utility of sequential modeling for capturing long-range spatial dependencies (Miron et al., 2021).

Transformer-based models provide an alternative that leverages attention to learn global inter-slice relationships. The M3T framework (Jang and Hwang, 2022) fused multi-plane representations for Alzheimer's diagnosis, while Meneghetti et al. (2025) introduced a foundation model–based multiple instance learning (MIL) transformer for predicting clinical outcomes in head and neck cancer, encoding 2D CT slices with a foundation model and aggregating them using transformer-based attention. More recently, the Medical Slice Transformer (MST) (Müller-Franzes et al., 2025) combined transformer aggregation with self-supervised 2D encoders, achieving strong results

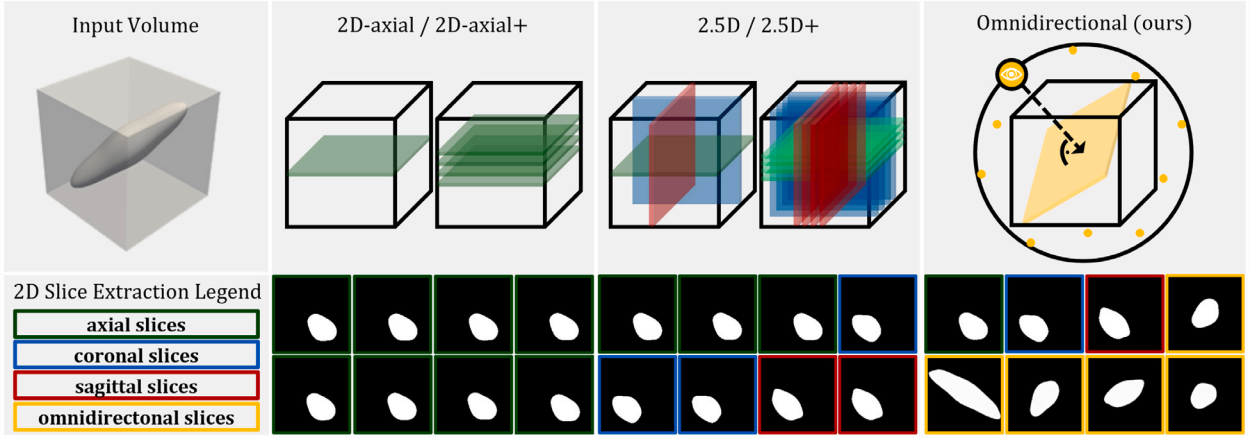

Fig. 2. Comparison of volume slicing strategies for a synthetic 3D volume containing an elongated structure oriented obliquely to the canonical axes (i.e. axial, coronal, sagittal). From left to right: 2D-axial and 2D-axial+, 2.5D and 2.5D+, and the proposed *omnidirectional* slicing approach. The bottom panel illustrates the corresponding extracted slices for the synthetic 3D volume encapsulating an elongated out-of-plane structure. Axial, coronal, and sagittal slices are highlighted in green, blue, and red, respectively, while yellow denotes oblique, out-of-plane slices enabled by our proposed *omnidirectional* volume slicing strategy.

across breast MRI, lung CT, and knee MRI classification downstream tasks.

A complementary direction represents volumetric data as structured graphs. Kiechle et al. (2024) and Di Piazza et al. (2025), Piazza et al. (2025) modeled tomographic volumes where nodes correspond to slice features and edges encode spatial adjacency. Kiechle et al. (2024) showed that graph representations improve accuracy and runtime over MLP-based aggregation but depend heavily on selecting suitable graph convolution operators and topologies, limiting general applicability. Di Piazza et al. (2025) and Piazza et al. (2025) constructed graphs from axial slice stacks to capture both local and global inter-slice dependencies along the axial axis. However, this design remains restricted to a single orientation, omitting valuable coronal and sagittal context. Moreover, evaluation limited to chest CTs leaves the generalizability of this approach across modalities and anatomical sites uncertain.

3. Methodology

In this work, we introduce *TomoGraphView*, a novel framework designed to preserve the spatial structure of volumetric medical imaging data while exploiting the representational power of 2D foundation models. The proposed framework is composed of two key components: *omnidirectional* volume slicing and *spherical mesh–based graph topology* construction, coupled with GNN-based feature aggregation and downstream prediction. An overview of our framework is illustrated in Fig. 1, which we detail in the following sections. Section 3.1 introduces *omnidirectional* volume slicing and contrast it to conventional slicing approaches. Section 3.2 describes the encoding of the resulting *omnidirectional* 2D cross-sectional slices into compact 1D feature representations. Section 3.3 presents the *spherical mesh–based graph topology* construction, while Section 3.4 details the graph representation learning procedure employed for downstream classification tasks.

3.1. Omnidirectional volume slicing

Traditional volume slicing strategies typically rely on canonical planes such as axial, coronal, and sagittal views, or a combination thereof, referred to as 2.5D. However, before decomposing a volume into 2D slices, it is common practice in 3D medical image classification to first extract a subvolume enclosing the target structure of interest, commonly defined using a segmentation mask. Accordingly, the most representative slice along each direction is defined as the one containing the largest visible lesion area. As illustrated in Fig. 2, in the 2D-axial setting, a single slice is extracted along the axial direction. The 2D-axial+ variant further includes adjacent slices symmetrically around the axial largest-lesion slice, resulting in a total of N views. The 2.5D approach combines information from all the canonical planes by selecting the most representative slice along the axial, coronal, and sagittal directions, whereas $2.5D+$ extends this strategy by additionally including adjacent slices symmetrically around the largest-lesion slice in each direction.

In this work, additionally to utilizing information from canonical planes, we propose to extract slices from non-canonical orientations to obtain *omnidirectional* slice representations, thereby enabling a richer characterization of volumetric structures that are not naturally aligned with the canonical axes. To obtain N *omnidirectional* slice representations, we visually embed the volume within a bounding sphere S, centered at the origin, which is formally described as

$$S = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 \leq r\}, \tag{1}$$

where r is the radius such that the sphere fully contains the 3D volumetric scan as illustrated in Fig. 3. Next, as we intend to make use of both, canonical and non-canonical perspectives, we first of all fix three points on the surface of the sphere representing the canonical view planes (i.e., axial, coronal, and sagittal) with their known coordinates $(r,0,0),(0,r,0)$ and $(0,0,r)$, as illustrated in Fig. 3 (step I). We then optimize the location of the remaining $(N-3)$ points, representing non-canonical view planes, such that all points on the surface are approximately equally distributed, as illustrated in Fig. 3 (step II). To this end, we make use of the repulsion-based optimization outlined in Algorithm 1, which is inspired by the Thomson problem (Thomson, 1904). Therein, points x_i are interpreted as identical charged particles constrained to lie on the spherical surface S, and their positions are optimized by minimizing the system's total electrostatic energy. Formally, given N points $\{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^3$ and $\|x_i\| = 1$, the pairwise simplified Coulomb energy $E(N)$, assuming unit charges and ignoring the Coulomb constant, is used without loss of generality, and mathematically defined as

$$E(N) = \sum_{i \neq j} \frac{1}{\|x_i - x_j\|^2}. \tag{2}$$

The gradient of the Coulomb energy function $\nabla E(N)$ yields repulsive forces between all point pairs which is used to iteratively update the position of the $N-3$ non-fixed points using gradient descent. After each update step, points are re-projected onto the sphere to enforce the unit-norm constraint. By visually enclosing the volume in a sphere, both

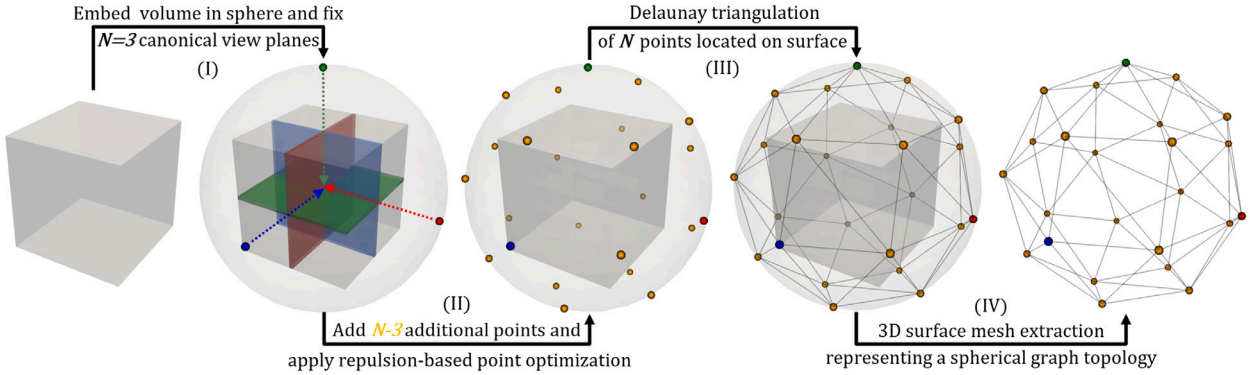

Fig. 3. Sequential construction of the spherical mesh-based graph topology. (I) Three canonical view directions (axial, coronal, sagittal) are fixed on the unit sphere. (II) The remaining $(N-3)$ points are initialized on the sphere and optimized via repulsion-based sampling to achieve an approximately uniform surface distribution among all points. (III) Next, Delaunay triangulation is performed on the resulting equidistant point set to define triangular neighborhood connectivity. (IV) The corresponding 3D surface mesh is extracted, yielding a spherical mesh-based graph structure for GNN downstream prediction.

Algorithm 1: Repulsion-based point optimization

Input: N total points, F fixed points on sphere

Place fixed points at given coordinates;
Randomly sample $(N-F)$ additional points on sphere;

while *not converged* **do**
  foreach *point* i **do**
    if i *is fixed* **then**
      continue;
    Compute Coulomb force between pairs $j \neq i$;
    $E_i \leftarrow \sum_{j \neq i} \frac{1}{\|x_i - x_j\|^2}$;
    Update position: $x_i \leftarrow x_i - \eta \cdot \nabla E_i$;
    Reproject: $x_i \leftarrow x_i / \|x_i\|$;

Output: Optimized set of N points on the sphere

centered at the origin, we guarantee that all possible slice orientations, which are defined by points on the sphere, will intersect the volume. For each volume, canonical and non-canonical slices are then extracted by first aligning the slicing plane such that its normal vector points to the respective point sampled on the surface of the sphere, and second by identifying the largest lesion slice from the respective canonical and non-canonical plane. Repeating this process for all $i = 1, \ldots, N$ yields a set of N uniformly distributed *omnidirectional* slices, thereby effectively resembling the spatial structure of the input volume.

3.2. Cross-sectional image encoding

After *omnidirectional* volume slicing, we transfer the obtained 2D volume slices into 1D feature representations by means of DINOv2 (Oquab et al., 2023). Generally, the DINO (Caron et al., 2021) framework is a self-supervised learning method based on a student–teacher architecture designed to extract semantically rich and meaningful visual representations. Building on this, Oquab et al. (2023) introduced DINOv2, an extension of DINO (Caron et al., 2021) trained with self-supervised contrastive learning on a large-scale dataset of 142 million curated images. In this work, we employ a pretrained DINOv2 vision transformer (ViT) model in its small variant, featuring 21M parameters. The encoder is kept frozen during all experiments and processes input slices of size 224×224 pixels, obtained by linearly upscaling the input images if they are not initially matched. The model produces a 384-dimensional feature representation for a single input image corresponding to a single view.

3.3. Spherical mesh-based graph topology construction

Given the N view points, which are equidistantly distributed on the spherical surface (Section 3.1) and their corresponding cross-sectional volume slice encodings (Section 3.2), we proceed by describing the construction of the *spherical mesh-based graph topology*. Therein, nodes represent canonical and non-canonical view planes and slice encodings their corresponding feature vector. We denote the set of points uniformly distributed on the spherical surface as $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$. To derive a graph structure from the sampled points, we apply Delaunay triangulation to $\mathbf{P}$, which guarantees that no point lies inside the circumcircle of any spherical triangle (Delaunay, 1934). The triangulation finally defines the graph edge set $\mathbf{E}$ using points $\mathbf{p}_i, \mathbf{p}_j$ as

$$\mathbf{E} = \{\, (i,j) \mid \mathbf{p}_i, \mathbf{p}_j \text{ are adjacent in the triangulation} \,\}, \tag{3}$$

yielding an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A})$, where the vertices $\mathbf{V} = \mathbf{P}$ and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where all connections $\mathbf{e}_{ij}$ resulting from the triangulation have an edge weighting $\mathbf{A}_{ij} = \mathbf{w}_{ij} = 1$. This procedure results in a mesh-based graph topology with near-uniform node distribution and well-defined local neighborhoods, both of which have been shown to be advantageous for subsequent mesh-based graph learning tasks (Bronstein et al., 2017). A graphical visualization of the Delaunay triangulation (step III) and 3D surface mesh extraction (step IV) can be found in Fig. 3.

Besides the local connectivity pattern, where points $\mathbf{p}_i$ and $\mathbf{p}_j$ are connected if they are adjacent in the triangulation, we also introduce weighted cross-connections between all remaining node pairs, effectively forming a complete graph. The corresponding edge weights $\mathbf{w}_{ij}$ of cross-connections are defined as the inverse node distance. The distance between two nodes is measured as hop distance, which can be defined as the shortest path length between two nodes $(\mathbf{v}_i, \mathbf{v}_j \in \mathbf{V})$ along the graph topology using Dijkstra algorithm (Dijkstra, 2022). Let as $\mathbf{P}(\mathbf{v}_i, \mathbf{v}_j)$ denote the set of all paths from node $\mathbf{v}_i$ to node $\mathbf{v}_j$, the goal is to find the path $\mathbf{p}$ which shows the minimum distance $\mathbf{d_{ij}}$ from node $\mathbf{v}_i$ to node $\mathbf{v}_j$, outlined as

$$\mathbf{d_{ij}}(\mathbf{v}_i, \mathbf{v}_j) = \min_{\mathbf{p} \in \mathbf{P}(\mathbf{v}_i, \mathbf{v}_j)} |\mathbf{p}|, \tag{4}$$

where $|\mathbf{p}|$ refers to the number of hops in path $\mathbf{p}$ along the spherical graph topology. The spherical graph topology is depicted in Fig. 3 (rightmost). While we retain all edge connections e_{ij} from the triangulation and keep their edge weights $\mathbf{w_{ij}} = 1$, we also add additional connections between nodes that are not part of the triangulation edge set $\mathbf{E}$ (see Eq. (3)). The weights of these additional edges are computed as their inverse node hop distance

$$\mathbf{w}_{ij} = \frac{1}{\mathbf{d_{ij}}(\mathbf{v}_i, \mathbf{v}_j)}. \tag{5}$$

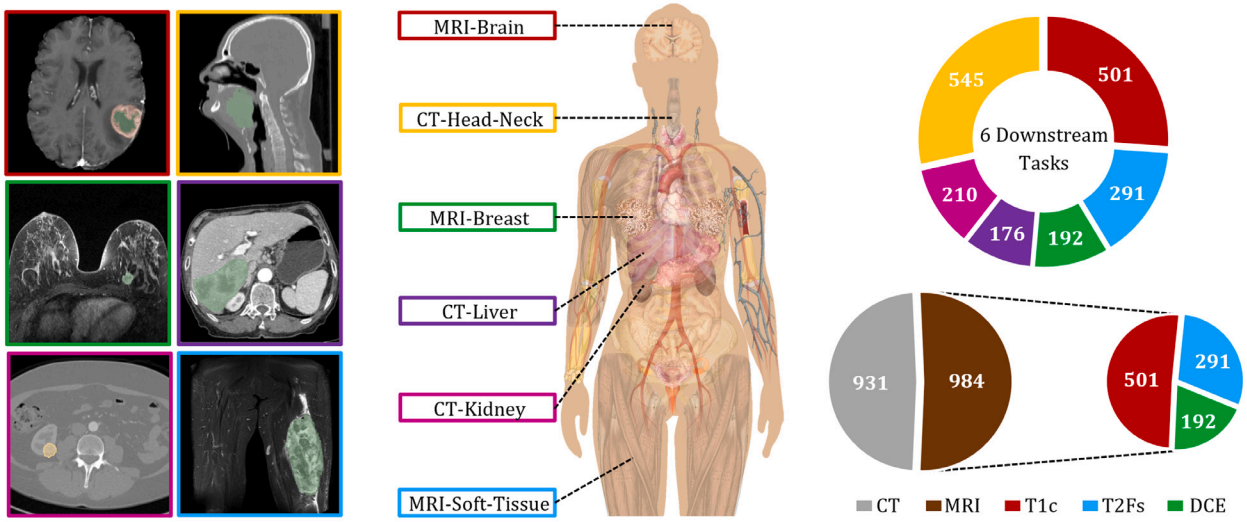

Fig. 4. Visual summary of the collected datasets from various target anatomical structures across the human body. The overview includes a visual example from each dataset (left), as well as the number of images under different acquisition criteria (right). Here, CT refers to computed tomography images, MRI refers to magnetic resonance imaging, T1c refers to T1-weighted post-contrast, T2Fs refers to T2-weighted fat-saturated, and DCE refers to dynamic contrast-enhanced.

The resulting graph topology and its weighting scheme effectively integrates local and global information, while assigning different levels of importance. Nearby view points receive higher weights because they provide complementary information that expands the receptive field. In contrast, antipodal view points, meaning geometrically opposite points, still contribute due to their strong semantic relationship, but with lower importance since they mainly offer a flipped representation and therefore limited additional information.

3.4. Graph representation learning

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A})$ denote a spherical mesh-based graph, as introduced in Section 3.3, with node attributes $\mathbf{X}_\mathbf{v}$ for $\mathbf{v} \in \mathbf{V}$ and edge weights $\mathbf{w}_{i,j} = \mathbf{A}_{ij}$ for $(i,j) \in \mathbf{E}$. Given a set of graphs $\{\mathbf{G}_N\}$ and their labels $\{\mathbf{y_N}\}$, the task of graph supervised learning is to learn a representation vector $\mathbf{h_G} \in \mathbb{R}^n$ that is used to predict the class of the entire graph. Generally, graph representation learning can be derived from three different flavors of GNN layers, namely convolutional, attentional, and message-passing (Bronstein et al., 2021). In this work, we focus on GNNs that follow message-passing, as it provides a flexible and conceptually intuitive framework where node features are aggregated among neighborhoods, following the underlying graph structure (Gilmer et al., 2017). For a single iteration, this is generically represented as

$$\mathbf{h}_i = \phi\left(\mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} \psi(\mathbf{x}_i, \mathbf{x}_j)\right), \tag{6}$$

where $\mathbf{h}_i$ represents the updated embedding of node $\mathbf{v}_i$ given its initial feature vector $\mathbf{x}_i$, and $\mathcal{N}_i$ represents its set of neighboring nodes $\mathbf{v}_j$. The function $\psi(\mathbf{x}_i, \mathbf{x}_j)$ computes a message from neighbor j to node i, and the operator $\bigoplus$ aggregates all messages using a permutation-invariant function (e.g. mean). The update function ϕ combines the ith node feature $\mathbf{x}_i$ with the aggregated neighborhood representation to finally produce the updated embedding $\mathbf{h}_i$. More specifically, for GraphSAGE (Hamilton et al., 2017), the message passing update to obtain $\mathbf{h}_i$ is formally given as

$$\mathbf{h}_i = \phi\left(\mathbf{W}\left[\mathbf{x}_i \parallel \bigoplus_{j \in \mathcal{N}_i} \left(\{\mathbf{e}_{ji} \cdot \mathbf{x}_j,\ \forall j \in \mathcal{N}(i)\}\right)\right]\right), \tag{7}$$

where $\mathbf{W}$ is a learnable weight matrix, $\parallel$ represents the concatenation operation, and $\mathbf{e}_{ji}$ is the scalar weight on the edge from node j to node i. As using multiple node neighborhood aggregation functions has been shown to provide notable improvements (Corso et al., 2020), we leverage both mean and max aggregation. They are particularly suitable to capture the distribution of elements within the graph (i.e., mean) and prove to be advantageous to identify representative elements (i.e, max) (Xu et al., 2018). As our underlying graph topology does not change within a setting of N nodes, we do not use sum aggregation, which has been shown to be especially powerful for learning structural graph properties. After the node updates, the resulting set of N one-dimensional node feature vectors is aggregated using dimension-wise mean pooling δ (i.e., a mean readout function) to obtain a global graph representation $\mathbf{h}_G \in \mathbb{R}^N$. This representation is then passed through a linear layer σ to predict the binary class label $\mathbf{y}_G$ for the entire graph, formally expressed as

$$\mathbf{y}_G = \sigma(\mathbf{h}_G), \quad \text{where} \quad \mathbf{h}_G = \delta\big(\{\,\mathbf{h}_v \mid v \in G\,\}\big). \tag{8}$$

4. Experiments & results

In Section 4.1, we describe the datasets used in this study, followed by details on implementation and training procedures (Section 4.2). Next, we systematically assess the proposed volume-slicing strategy (Section 4.3), before proceeding to a comprehensive evaluation of our proposed *TomoGraphView* framework (Section 4.4). Building on this, we further extend the analysis by investigating the influence of volume slice spacing (Section 4.5), performing a graph-topology ablation study (Section 4.6), and comparing alternative slice-wise feature aggregation methods (Section 4.7). Finally, we benchmark *TomoGraphView* against 3D pretrained models (Section 4.8), taking into account both frozen and fine-tuned 3D backbone configurations to provide a comprehensive comparison.

4.1. Datasets

We evaluate our approach on six clinically relevant oncological tasks. An overview of the datasets is given below, with representative examples and statistics shown in Fig. 4.

Brain Tumors (Calabrese et al., 2022) This open-source dataset comprises $n = 501$ patients with brain gliomas, with a median voxel spacing of 1.00×1.00×1.00 mm. It contains T1-weighted post-contrast (T1c) MRI scans from patients diagnosed with glioblastoma, oligodendroglioma,

or astrocytoma. The cohort was collected at a single institution, the University of California, San Francisco, USA. For our analysis, tumors were categorized into high-grade (G4) and low-grade (G2/G3) according to the WHO CNS classification.

Head-Neck Tumors (Wee and Dekker, 2019; Grossberg et al., 2018; Kwan et al., 2019) This open-source dataset comprises CT scans of $n = 545$ patients with head and neck squamous cell carcinoma (HNSCC) of the oropharynx, acquired at a median voxel spacing of $0.98 \times 0.98 \times 2.00$ mm. The multi-institutional cohort was assembled from Maastricht University Medical Center, Netherlands, University of Texas MD Anderson Cancer Center, USA, and Princess Margaret Cancer Center, Ontario, Canada. The task involves binary classification of human papillomavirus (HPV) status into HPV-positive and HPV-negative entities.

Breast Tumors (Saha et al., 2021) This open-source dataset includes dynamic contrast-enhanced (DCE) MRI scans of $n = 192$ patients with biopsy-confirmed invasive breast cancer, acquired at a median voxel spacing of $0.74 \times 0.74 \times 1.00$ mm. The single-institutional cohort was retrospectively collected at Duke Hospital, North Carolina, USA. Tumors were stratified into high-grade ("high") and low-grade ("intermediate"/"low") categories according to their Nottingham grade. In our experiments, we use the first post-contrast DCE-MRI volumes.

Liver Tumors (Luo et al., 2025) This open-source dataset consists of contrast-enhanced multi-phase CT scans from $n = 176$ patients with primary liver tumors, acquired at a median voxel spacing of $1.00 \times 1.00 \times 1.00$ mm. The cohort was collected at the Radiology Department of Chongqing Yubei District People's Hospital, China. Binary classification was performed on the venous phase (C2) to distinguish hepatocellular carcinoma (HCC) from combined hepatocellular–cholangiocarcinoma (cHCC-CCA), the latter being clinically more aggressive due to poorer overall survival at comparable stage.

Kidney Tumors (Heller et al., 2019) This open-source dataset contains contrast-enhanced multi-phase CT scans of $n = 210$ patients with kidney tumors, acquired at a median voxel spacing of $0.78 \times 0.78 \times 3.0$ mm. The single-institutional cohort was retrospectively collected at the University of Minnesota, USA. Tumors were graded into high-grade (G3/G4) and low-grade (G1/G2) entities according to the WHO ISUP classification, as determined from post-operative surgical pathology reports.

Soft-Tissue Tumors This in-house dataset comprises MRI scans of $n = 291$ patients with soft-tissue sarcomas (STS) of the extremities and trunk, acquired with T2-weighted fat-saturation (T2FS) sequences at a median voxel spacing of $0.78 \times 0.94 \times 4.80$ mm. Two independent patient cohorts were retrospectively collected at the University of Washington/Seattle Cancer Care Alliance, USA, and the Technical University of Munich, Germany. Tumors were categorized into high-grade (G2/G3) and low-grade (G1) entities based on histopathological grading.

4.2. Implementation and training details

All input volumes were first reoriented to a common positive Right-Anterior-Superior (RAS+) coordinate system, followed by resampling to an isotropic voxel spacing of $(1 \times 1 \times 1)$ mm. After slice extraction (either canonical or non-canonical), the resulting 2D images were cropped to fully encompass the lesion as defined by the segmentation mask, while aiming for symmetric spatial dimensions to accommodate potential irregular aspect ratios. In cases where the target structure was located near the image boundary, padding was applied to prevent truncation. This procedure avoids geometric distortions prior to linear resizing of the 2D image slices to 224×224 pixels, as required by DINOv2. Subsequently, the images were normalized using ImageNet statistics and provided as input to the frozen DINOv2 encoder. Within the proposed *omnidirectional slicing* framework, image volumes were linearly interpolated to generate N views from corresponding N distinct orientations. In contrast, segmentation masks were resampled using nearest-neighbor interpolation to preserve discrete label integrity. For all experiments, we attach a lightweight classification head to the frozen DINOv2 feature extractor. Specifically, the head of each baseline method consists of two layers with a non-linear ReLU activation in between, featuring a total of 100k trainable parameters to ensure a fair comparison across methods. We restrict the prediction head to a shallow design to avoid overfitting, given the limited dataset sizes in medical imaging, which also helps to mitigate graph oversmoothing in the case of GNNs (Keriven, 2022). For the MLP, slice encodings are concatenated. We train and evaluate our models using a 5-fold cross-validation scheme, and stratify splits according to the binary class label. More specifically, we use three splits for training, one for validation, and one for testing. Training is conducted using the SGD optimizer with a weight decay of $1e^{-3}$, a batch size of 16, and 300 epochs. The model that achieves the best AUROC score on the validation set is selected for final evaluation on the test set. A learning rate scheduler first linearly increases the learning rate to 0.001 for 100 warm-up epochs, and decays the learning rate afterwards by a factor of 0.95 when there is no improvement in AUROC validation score for at least 5 epochs, to ensure smooth convergence. All models were trained on an Nvidia RTX A6000 GPU. We publicly share our code repository at http://github.com/compai-lab/2025-MedIA-kiechle. In addition, we provide a user-friendly library for *omnidirectional* volume slicing at https://pypi.org/project/OmniSlicer.

4.3. Impact of volume slicing strategy - MLP comparison

We start our experimental evaluation by comparing baseline volumetric slicing strategies, i.e., *2D-axial*, *2D-axial+*, *2.5D*, and *2.5D+*, against our proposed *omnidirectional* volume slicing method introduced in Section 3.1. To ensure a fair comparison across all datasets and view configurations, our volume slicing strategy benchmark is solely based on identical MLP prediction heads comprising 100k trainable parameters. The number of views is empirically set to 8, 16, and 24, providing a reasonable tradeoff between computational efficiency and captured slice diversity. The corresponding results for all combinations of MLP and the respective volume slicing strategy are summarized in Table 1, with the analysis restricted exclusively to the non-gray-colored rows.

As visible from the results, the *2D-axial* approach performs worse than all alternative slicing methods, likely because the majority of volumetric information is discarded when relying on a single slice. We nevertheless included this experiment as it establishes a meaningful lower bound for the performance evaluation of different volume slicing methods. As indicated by the results, incorporating multiple axial slices represented by the *2D-axial+* approach improves performance by leveraging a greater amount of volumetric information.

Incorporating additional complementary perspectives, as in the *2.5D* slicing strategy, leads to a clear improvement in performance. For example, the average AUROC computed across all datasets increases from 0.7335 for the *2D-axial+* configuration (three slices) to 0.7681 for the *2.5D* setup (three slices). This trend extends to the *2.5D+* variant, which consistently outperforms its *2D-axial+* counterpart for the same number of views. While the performance differences are more pronounced between 3, 8, and 16 views, the gap narrows at 24 views, suggesting that incorporating diverse perspectives is particularly beneficial when only a limited number of slices are available.

Beyond canonical plane slicing, our proposed *omnidirectional* slicing strategy, which integrates both canonical and non-canonical orientations, further improves downstream performance. On average, across all evaluated configurations, the *omnidirectional* approach consistently outperforms both the *2D-axial+* and *2.5D+* methods. Furthermore, for our volumetric slicing strategy, we observe a general trend across

Table 1
Quantitative comparison of conventional volume slicing strategies (2D-axial, 2D-axial+, 2.5D, 2.5D+) with our proposed *omnidirectional* slicing approach using an MLP prediction head, as well as our *TomoGraphView* framework. All methods are evaluated across multiple datasets and numbers of volume slices (views) using 5-fold cross-validation and Area Under the Receiver Operating Characteristic Curve (AUROC) as the performance metric. The rightmost column reports the mean performance over all datasets. Bold values indicate the best result per dataset, and underlined values denote the second-best.

Method (Head – slicing strategy)	Views	Breast	Brain	Head-neck	Kidney	Liver	Soft-tissue	Mean
MLP – 2D-axial slicing	1	0.5844	0.9013	0.6192	0.6541	0.8821	0.7421	0.7305
MLP – 2D-axial+ slicing	3	0.6166	0.8936	0.6550	0.6234	0.8803	0.7322	0.7335
	8	0.6402	0.9164	0.6171	0.7689	0.8738	0.7530	0.7615
	16	0.6467	0.9104	0.5982	0.7348	0.9053	0.7555	0.7584
	24	0.6299	0.9295	0.6277	0.7806	0.8895	0.7634	0.7701
MLP – 2.5D slicing	3	0.6783	0.9062	0.6818	0.6815	0.9662	0.6951	0.7681
MLP – 2.5D+ slicing	8	0.7002	0.9296	0.6555	0.7300	0.9484	0.7488	0.7854
	16	**0.7437**	0.9226	0.6588	0.7606	**0.9667**	0.7473	0.7999
	24	0.6991	0.9265	0.6799	0.7041	0.9468	0.7751	0.7885
MLP – Omnidirectional slicing	8	0.6627	0.9457	0.6790	0.7306	0.9372	0.7702	0.7876
	16	0.6689	0.9491	0.6999	0.7450	0.9352	0.8090	0.8012
	24	0.6794	0.9492	0.7384	0.7229	0.9421	**0.8602**	0.8154
GNN – Omnidirectional slicing **TomoGraphView (ours)**	8	0.7336	**0.9678**	0.6780	0.7878	0.9600	0.8399	0.8279
	16	0.6907	0.9666	0.7244	0.7479	0.9585	0.8389	0.8212
	24	0.7084	0.9655	**0.7527**	**0.7893**	0.9634	0.8441	**0.8372**

datasets. Increasing the number of views typically leads to higher performance. This relationship remains consistent for the *omnidirectional* approach, with the sole exception of the kidney dataset, highlighting a clear positive correlation between the number of views and overall model performance. Furthermore, although the proposed *omnidirectional* slicing strategy does not achieve the highest performance on every individual dataset, it consistently surpasses its *2D-axial+* and *2.5D+* counterparts for the same number of views when considering the mean performance across all datasets. This demonstrates the method's robust and, in most cases, superior capability in capturing versatile volumetric information relevant for downstream classification.

Finally, to demonstrate the robustness of our *omnidirectional* volume slicing approach, we conduct two additional experiments on the brain tumor dataset, which provides a sufficiently large sample size (n = 501) for reliable performance estimation. First, we simulate segmentation variability following Weltens et al. (2001) and Gillett et al. (2025). Specifically, we alter the original tumor segmentation masks using inter-observer variability patterns such as tumor volume and mean center of mass differences, based on expert annotations from brain tumors (Weltens et al., 2001). Visual examples are provided in Fig. F.11. Second, we assess the sensitivity of downstream performance in a segmentation free setting by using bounding boxes and selecting corresponding central slices. To ensure a fair comparison, both experiments use an MLP prediction head, and we compare 2.5D+ slicing with our proposed *omnidirectional* slicing. Quantitative results are shown in Fig. F.10. Across both experiments, including segmentation variability patterns and the use of bounding boxes, our method consistently outperforms 2.5D+ slicing for all evaluated numbers of views (8, 16, and 24). These results highlight both the robustness and versatility of *omnidirectional* volume slicing, even under more challenging conditions.

4.4. Impact of graph neural network prediction head

Building on the results in Section 4.3, where the *omnidirectional* slicing strategy achieved the best performance, we next focus on aggregating slice-level representations using the spherical mesh-based graph topology and the corresponding graph-based feature aggregation method introduced in Section 3.3. To evaluate the impact of the GNN-based prediction head, we compare two models: *MLP – Omnidirectional Slicing* and *GNN – Omnidirectional Slicing*, the latter constituting our *TomoGraphView* framework. Both methods operate on identical *omnidirectional* slice inputs, ensuring that any performance differences arise solely from the aggregation method rather than the underlying representations. The corresponding results are summarized in Table 1 (lower two sections).

Our findings support the advantage of propagating information across slices via a meaningful underlying graph structure. Specifically, when comparing the average performance of *MLP - Omnidirectional Slicing* against our *TomoGraphView* framework (*GNN - Omnidirectional Slicing*) across all six datasets, our proposed framework consistently outperforms the MLP for every view configuration, ultimately increasing from 0.8154 to 0.8372 in AUROC performance for the best performing model in each case. Notably, our *TomoGraphView* framework achieves superior performance across nearly all dataset–view combinations, with the exception of Head-Neck at 8 views and Soft-Tissue at 24 views. These findings strongly highlight the effectiveness of the graph neural network-based feature aggregation approach over the MLP baseline method.

In summary, although the proposed *TomoGraphView* framework consistently improves the overall average performance across all experiments reported in Table 1, it does not necessarily achieve the highest performance on every individual dataset, while generally remaining highly competitive. We attribute these performance variations to dataset-specific characteristics and underlying differences in data composition, which are discussed in more detail in Section 5.1.

4.5. TomoGraphView - Impact of volume slice spacing

A more detailed analysis of the results obtained with our proposed *TomoGraphView* framework (Table 1) reveals distinct, dataset-specific trends. For the head–neck, kidney, liver, and soft-tissue tumor datasets, optimal performance is consistently achieved when using the highest number of views. In contrast, for the breast and brain tumor datasets, configurations with 8 views outperform those employing a larger number of views. To better understand these differing behaviors, we hypothesize that the impact of the number of views is influenced by dataset-specific imaging characteristics, in particular the voxel spacing along the axial dimension. As illustrated in Fig. 5 (left), datasets with larger axial voxel spacing, such as soft-tissue, head–neck, and kidney, benefit more from an increased number of views. This trend is supported by AUROC improvements when increasing the number of views from 8 to 24, for example in soft-tissue (0.8399 vs. 0.8441), head–neck (0.6780 vs. 0.7527), and kidney tumors (0.7878 vs. 0.7893). These findings suggest that additional views are particularly advantageous in low axial resolution settings, where they help compensate for reduced spatial information.

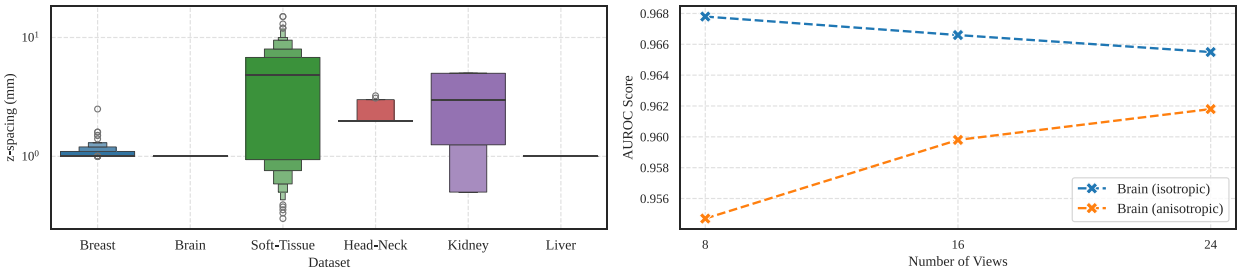

Fig. 5. Box plots illustrating the distribution of axial z-spacings across the datasets used in this study (left), and the effect of slice spacing on the optimal number of views for *TomoGraphView*, shown for the brain tumors dataset (right) using the Area Under the Receiver Operating Characteristic Curve (AUROC) metric.

To further validate this empirical observation, we conducted an additional experiment by artificially increasing the z-spacing in a high-resolution dataset. For this purpose, we utilize the brain tumor dataset, which features isotropic $1 \times 1 \times 1$ mm spacing and provides a sufficiently large sample size ($n = 501$) for robust performance estimation. We resampled the isotropic samples to be anisotropic. Specifically, we increase the axial (z-axis) spacing to 6 mm, while keeping all other dimensions unchanged, and reevaluated our proposed *TomoGraphView* framework for 8, 16, and 24 views.

As visualized in Fig. 5 (right), we observe two effects: First, the overall performance decreased from the isotropic to the anisotropic setting, which is an expected consequence of information loss and resampling artifacts. Second, and more importantly, for the anisotropic setting, the performance across different view counts now follows the same trend as in high z-spacing datasets, such as soft-tissue, head-neck, or kidney, where increasing the number of views consistently improves results, as illustrated in Table 1. This finding suggests that additional views are more suitable for compensating for the loss of information introduced by anisotropic resampling.

4.6. TomoGraphView - Impact of graph topology

We continue by studying the impact of the underlying graph topology, comparing different graph structures to assess the effect of graph connectivity on our proposed *TomoGraphView* framework. In total, we compare five distinct graph topologies, which we refer to as *spherical*, *uniform*, *linear decay*, *inverse*, and *inverse-square*. More specifically, *uniform*, *linear decay*, *inverse*, and *inverse-square* topologies share the same graph connectivity (i.e., complete graph), but show differences in the edge-weights for node cross-connections (i.e., connections to non-immediate neighbors after triangulation as indicated in Section 3.3). The corresponding edge weights are determined according to the node distance in hops as indicated by Eq. (5) and depicted in Fig. 6 (top). In contrast, the *spherical* scheme does not use any node cross-connections, thus only connects nodes that are adjacent in the Delaunay triangulation (see Section 3.3), effectively representing a localized graph structure. Hence, the *spherical* topology results in a substantially sparser graph representation compared to the fully connected variants.

We conducted the graph topology experiments across all datasets and report the average AUROC performance using configurations with 24 nodes, as previous experiments in Section 4.4 demonstrated this to be the overall best-performing setting for *TomoGraphView*. Additionally, a larger number of nodes enables a more informative comparison, since the addition of edges has a greater relative effect in denser graphs. The corresponding results are presented in Fig. 6 (bottom).

Several key observations emerge from these experiments. First, the *spherical* topology achieves comparable performance to the fully connected *uniform* topology, despite relying on substantially fewer edge connections. This highlights the efficiency and relevance of the local spherical graph connectivity pattern. Second, introducing cross-connections between all nodes improves performance only when employing smooth distance-based weighting schemes such as *inverse* and *inverse-square*, which effectively expand the receptive field by aggregating information from diverse perspectives. Among these, the *inverse* weighting outperforms the *inverse-square* variant. In contrast, the *linear-decay* weighting scheme, although also expanding the receptive field, leads to a decline in AUROC performance. This finding highlights the importance of using decaying weighting functions that place a stronger relative emphasis on local connections (i.e., one-hop neighbors) compared to global ones (i.e., node cross-connections). Detailed results, presented on a per-dataset basis for each of the different graph topologies, are provided in the Appendix D.

4.7. Slice-wise feature aggregation method benchmark

In this section, we compare *TomoGraphView* with two established slice-wise aggregation methods: a long short-term memory (LSTM) approach, which interprets a 3D volume as a sequence of consecutive slice embeddings, and a transformer-based approach, represented by the medical slice transformer (MST) (Müller-Franzes et al., 2025), which employs an attention mechanism to fuse intra-slice information. To ensure a fair comparison with previous experiments, the number of trainable parameters is again fixed to 100k across all models. Additionally, we evaluate both LSTM- and transformer-based (MST) aggregation when combined with our proposed *omnidirectional* slicing strategy, thereby disentangling the effects of the aggregation mechanism from those of slice selection. Extending the MST-based aggregation to *omnidirectional* slices does not require any explicit modeling of slice relationships, as the cross-attention mechanism inherently aims to capture inter-slice relationships. In contrast, this adaptation is less intuitive for LSTM models, which are inherently designed to process naturally sequential inputs. Since *omnidirectional* slices, derived from points sampled on a spherical surface, lack a natural ordering, we impose and fix a consistent sequence based on node indices. The mean AUROC and MCC values comparing *TomoGraphView*, LSTM, and MST across all datasets are summarized in Table 2.

We start by evaluating the LSTM and MST approaches in their original formulation using consecutive axial slices (i.e., 2D-axial+). The results show that the best-performing view setting (16) in the MST approach outperforms the best-performing view setting (24) in the LSTM approach, for both evaluation metrics, with AUROC values of 0.7768 vs. 0.7446 and MCC values of 0.3995 vs. 0.3365, highlighting the strong aggregation capability of the MST. Furthermore, this performance advantage is even observable across all view configurations, with the MST again surpassing the LSTM in both evaluation metrics. Nevertheless, both methods fall short of our proposed *TomoGraphView* framework, which overall achieves an AUROC of 0.8372 and an MCC of 0.5191.

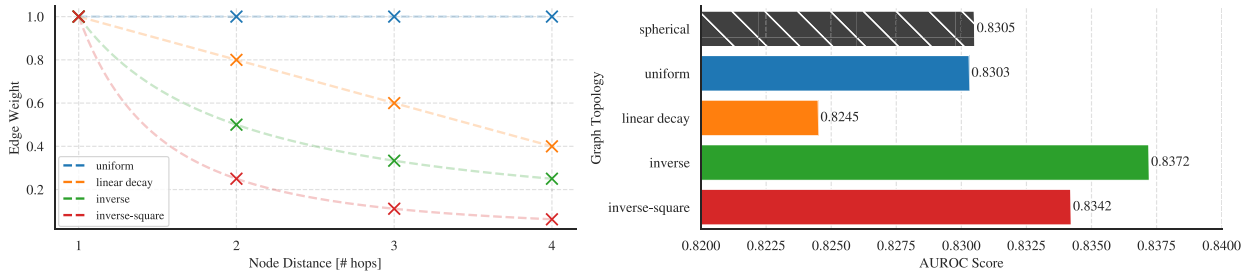

Fig. 6. (Left) Edge-weighting schemes used to define the strength of connections between nodes as a function of their graph distance (number of hops). Four strategies are illustrated: uniform weighting (all connections have weight of 1), linear decay, inverse decay, and inverse-square decay, where edge weights decrease with increasing node distance to different extents. (Right) Corresponding model performance measured by the area under the receiver operating characteristic curve (AUROC) for each graph topology induced by the respective weighting scheme.

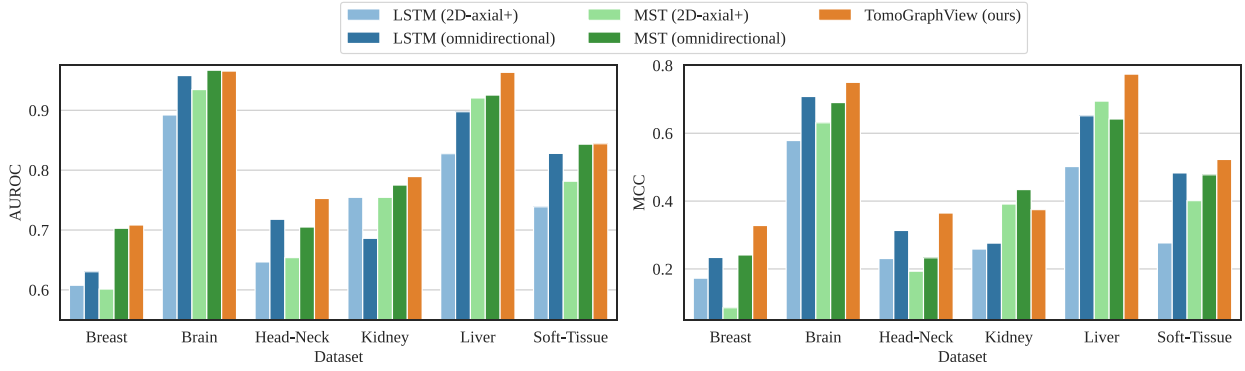

Fig. 7. Detailed results comparing slice-wise aggregation methods: long short-term memory (LSTM), medical slice transformer (MST), and *TomoGraphView*. Besides their original slicing strategy (2D-axial+), LSTM, and MST are also evaluated using *omnidirectional* volume slicing. All methods are evaluated across all six datasets using the Area Under the Receiver Operating Characteristic Curve (AUROC) on the left and Matthews correlation coefficient (MCC) on the right as performance metrics. Bars indicate the average performance across five folds.

Admittedly, comparing LSTM and MST in their original 2D-axial+ slicing configuration to *TomoGraphView* is only partially fair, as differences may not only arise from the aggregation mechanism but also from the contextual information available in the input slices. To address this, we further extend the evaluation by applying *omnidirectional* slicing to both LSTM and MST. Two key observations emerge: (I) MST continues to outperform LSTM across all view configurations, and (II) both models experience substantial performance improvements. Specifically, for LSTM, *omnidirectional* slicing increases AUROC from 0.7446 to 0.7863 and MCC from 0.3365 to 0.4444. Similarly, for MST, AUROC improves from 0.7768 to 0.8198 and MCC from 0.3967 to 0.4732. These results highlight the clear benefit and broad applicability of our proposed *omnidirectional* volume slicing strategy. Remarkably, while both baselines show notable gains, *TomoGraphView* still achieves the highest overall performance.

A more detailed, dataset-wise comparison is presented in Fig. 7, which reports AUROC (left) and MCC (right) scores for 24 views. Comprehensive numeric results across datasets, models, and view configurations are provided in Table E.4. Several trends can be observed: in terms of AUROC, *TomoGraphView* matches the performance of the best baseline method on the breast, brain, and soft-tissue tumor datasets, but clearly surpasses them on the head-neck, kidney, and liver tumor datasets. A similar pattern holds for MCC, where *TomoGraphView* consistently outperforms all competing methods, except for the kidney tumor dataset. Here, both MST variants show superior performance. Nevertheless, the overall results underscore the strong capability of *TomoGraphView* for slice-wise feature aggregation, establishing it as both a competitive and compelling alternative to existing slice-wise approaches.

Table 2
Quantitative results comparing a long short-term memory (LSTM) approach, medical slice transformer (MST), and our *TomoGraphView* framework for slice-wise feature aggregation performance. Besides their original slicing strategy (2D-axial+), LSTM and MST are also evaluated using *omnidirectional* volume slicing. Methods are compared across different numbers of views and performance metrics, including the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Matthews correlation coefficient (MCC). Results are presented as the average across all six datasets and five folds.

Model	Volume slicing	Views	AUROC	MCC
LSTM	2D-axial+	8	0.7305	0.3101
		16	0.7330	0.3129
		24	0.7446	0.3365
LSTM	Omnidirectional	8	0.7764	0.4225
		16	0.7744	0.4178
		24	0.7863	0.4444
MST	2D-axial+	8	0.7583	0.3442
		16	0.7768	0.3967
		24	0.7746	0.3995
MST	Omnidirectional	8	0.8083	0.4691
		16	0.8171	0.4732
		24	0.8198	0.4530
TomoGraphView	Omnidirectional	8	<u>0.8279</u>	0.4891
		16	0.8212	<u>0.5096</u>
		24	**0.8372**	**0.5191**

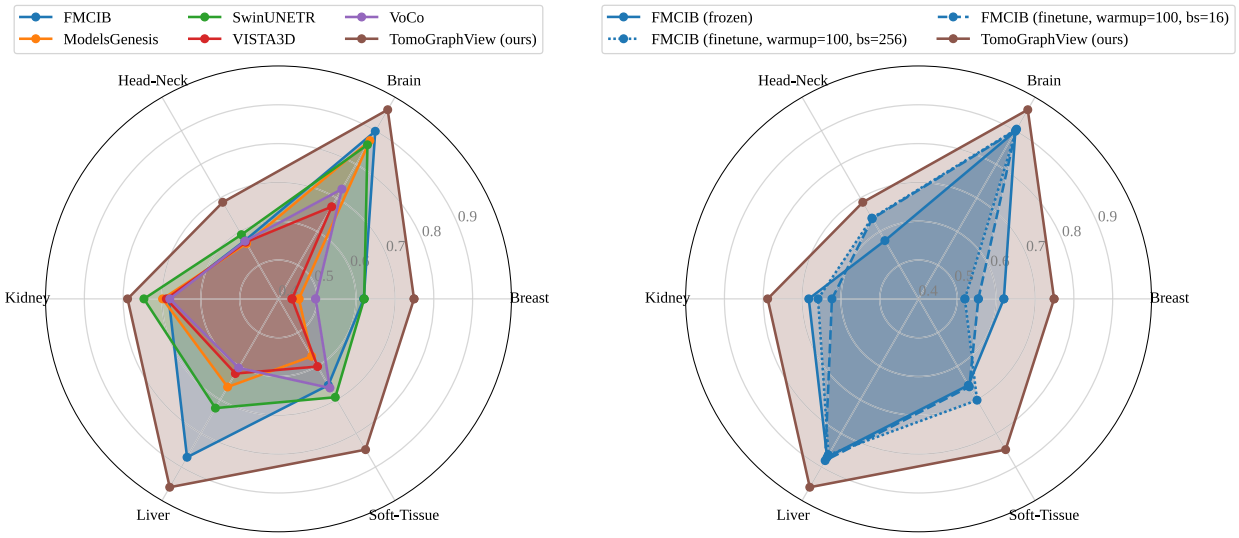

Fig. 8. Radar plots showing quantitative results comparing frozen 3D backbones against our *TomoGraphView* framework (left) and fine-tuning results of the best performing 3D backbone (FMCIB) against our *TomoGraphView* framework (right). Plots show the average Area Under the Receiver Operating Characteristic Curve (AUROC) performance across 5-folds.

4.8. Comparative analysis with 3D pretrained models

In the following, we compare our proposed *TomoGraphView* framework at 24 views with large-scale pretrained 3D models. All baseline models are convolution-based, as 3D vision transformers are substantially more memory-demanding due to the much larger token sequences (i.e., number of patches) compared to 2D inputs. In addition, vision transformers typically require significantly larger datasets to pretrain effectively, whereas medical imaging datasets are often limited in size.

The 3D models used in our comparison with *TomoGraphView* represent the top-performing downstream classification approaches identified by Aerts et al. (2025) and Wald et al. (2024). Specifically, we benchmark our framework against FMCIB (Pai et al., 2024), Vista3D (He et al., 2025), Models Genesis (Zhou et al., 2019), as well as two pretrained models from Wald et al. (2024) based on VoCo (Wu et al., 2024) and SwinUNETR (Tang et al., 2022). A detailed description of these models is provided in Appendix C.

The remainder of this section is structured as follows. In Section 4.8.1, we first evaluate the fixed embeddings obtained from the selected large-scale pretrained 3D models across all datasets, without applying any task-specific finetuning to assess their representational capacity learned during pretraining. Subsequently, in Section 4.8.2, we perform dataset-specific finetuning, which typically improves downstream performance by allowing the pretrained representations to better adapt to the characteristics of each target task. Since all large-scale pretrained 3D models were trained in a self-supervised manner, we attach a classification head to the encoder to enable downstream classification. Consistent with our previous experiments, we fix the number of trainable parameters to 100k to ensure a fair comparison. Notably, prediction heads previously discussed, such as GNN, transformer, or LSTM-based architectures, cannot be integrated with the 3D pretrained models, as they require slice-based inputs rather than full 3D volume embeddings.

4.8.1. TomoGraphView vs. Frozen 3D backbones

We first evaluate the five large-scale pretrained 3D models across all six oncological downstream tasks and compare their performance with our proposed *TomoGraphView* framework. The results are summarized in the left radar plot of Fig. 8. Comprehensive quantitative results for all methods, datasets, and performance metrics are provided in Table F.5.

Across all tasks, our proposed *TomoGraphView* framework achieves an average AUROC of 0.8372, outperforming the on average second-best approach, FMCIB (Pai et al., 2024), which achieves an AUROC of 0.7170. This demonstrates a consistent advantage of *TomoGraphView* over all competing models. Looking more closely, on a per-task level, *TomoGraphView* also yields the highest AUROC for each dataset: breast tumors (0.7084 vs. 0.6207, SwinUNETR), brain tumors (0.9655 vs. 0.8985, FMCIB), head–neck tumors (0.7527 vs. 0.5910, SwinUNETR), kidney tumors (0.7893 vs. 0.7466, SwinUNETR), liver tumors (0.9634 vs. 0.8711, FMCIB), and soft-tissue tumors (0.8441 vs. 0.6926, SwinUNETR). Although the performance gains are less pronounced for kidney tumors (0.7893 vs. 0.7466) and brain tumors (0.9655 vs. 0.8985), *TomoGraphView* consistently outperforms all competing methods by a clear margin across the remaining breast, head-neck, liver, and soft-tissue tumor datasets. This observation, as exemplified by the AUROC results, can be further confirmed by additional evaluation metrics, such as balanced accuracy, F1-Score, and Matthews correlation coefficient, underscoring the great potential of *TomoGraphView*. Detailed results can be found in the Table F.5.

4.8.2. TomoGraphView vs. Finetuned 3D backbones

Building on the frozen backbone results presented in the previous subsection, where, on average, the FMCIB (Pai et al., 2024) approach emerged as the best-performing large-scale pretrained 3D model across all datasets and different performance metrics, we now evaluate its finetuning performance, as it allows the pretrained representations to better adapt to the specific characteristics of each downstream task. Strictly following the scheme introduced by Wald et al. (2024), we finetune FMCIB for 200 epochs in total, using a learning rate of 1×10^{-4}, a cosine annealing scheduler, a gradual warm-up of 100 epochs applied to both the encoder and the classification head, and a batch size of 256. Additionally, we explore a batch size of 16, which is consistent with previous experiments. The corresponding results are illustrated in the right radar plot of Fig. 8, with detailed quantitative results provided in Table F.6, which additionally shows decreased warm-up durations of

20 and 50 epochs to allow for more update steps under higher learning rates.

Overall, fine-tuning with a 100-epoch warm-up and a batch size of 256 yields performance gains across several downstream tasks. In particular, AUROC improves for brain tumors from 0.8985 to 0.9049, soft-tissue tumors from 0.6567 to 0.7013, and head–neck tumors from 0.5733 to 0.6403. However, no measurable improvements are observed for the breast, liver, or kidney tumor datasets compared to the frozen FMCIB backbone. Given that the datasets showing no improvements are also those with relatively small sample sizes, we conduct an additional experiment using a reduced batch size of 16 to allow for more update steps during finetuning. This adjustment improves AUROC performance for the liver tumor dataset from 0.8711 to 0.8811, whereas performance for the breast and kidney datasets remains inferior. These results suggest that the effectiveness of fine-tuning is highly dataset-dependent, and a single universal configuration is unlikely to generalize across datasets without extensive hyperparameter exploration. Notably, our proposed *TomoGraphView* framework continues to outperform the FMCIB model even under finetuning conditions.

5. Discussion

In this work, we have introduced *TomoGraphView*, a novel framework for 3D medical image classification that preserves the spatial structure of volumetric data while enabling compatibility with powerful 2D vision foundation models. We conducted a comprehensive evaluation across six diverse datasets, encompassing multiple anatomical regions (e.g., brain, breast, head and neck, kidney, and liver), different imaging modalities (CT and various MRI sequences), and a range of clinically relevant classification tasks primarily concerning tumor characterization. Building on our two key contributions, *omnidirectional* volume slicing and *spherical mesh-based feature aggregation*, our experimental evaluations show that *TomoGraphView* achieves SOTA results across downstream classification tasks, and outperforms large-scale pretrained 3D models.

While our study focuses on binary classification in lesion-centered settings, this choice aligns with the primary motivation of the framework, namely the robust characterization of localized and heterogeneous structures. Importantly, this design does not limit the general applicability of the method. The framework is inherently task-agnostic and can be extended to any global prediction (i.e. image-level) task, including multi-class classification, regression, or survival prediction, with moderate modifications (e.g., adapting the prediction head and learning objective), as none of its components impose constraints on the downstream task. As our framework aggregates information across several slices into a global volume-level representation, it is not well aligned with dense prediction tasks such as segmentation which rely on representations at pixel level. Furthermore, both omnidirectional slicing and spherical graph aggregation are independent of the semantic nature of the prediction target and are therefore, in principle, applicable beyond tumor-related scenarios to non-pathological tasks such as anatomical or functional categorization. However, many such problems are commonly formulated differently, for example, as segmentation (e.g. categorization of brain regions into hippocampus vs. amygdala or lung lobes into upper, middle or lower lobes) or temporal modeling tasks (e.g. cardiac phase classification into end-diastole vs. end-systole vs. intermediate phases) which are beyond the scope of the present work. Nevertheless, evaluating the framework across a broader range of applications remains an well-motivated research direction. In the following, we provide a detailed discussion of the results and key findings of this study.

5.1. Advancing volume slicing using omnidirectional views

Generally, our experimental results shown in Table 1 suggest that more expressive tumor characterization, achieved by using richer and more representative volume slices from complementary directions, improves the description of the underlying 3D target structures. Specifically, expanding from purely axial 2D slices to a 2.5D multi-plane strategy consistently yields better downstream performance. As a result, the average AUROC performance across all evaluation datasets improves by 4.71% (from 0.7335 to 0.7681) when comparing 2.5D slicing with its corresponding 2D-axial+ configuration (i.e., with three axial slices). This trend also persists for a larger number of views: with 24 views, AUROC increases by 2.39% (from 0.7701 to 0.7885). However, as the first key contribution of this work, we presented a novel *omnidirectional* volumetric slicing strategy that extends beyond the three conventional orientations axial, coronal, and sagittal and enables a richer characterization of volumetric structures that are not inherently aligned with these canonical axes. As a consequence, further gains are achieved with *omnidirectional* volume slicing, which improves AUROC by 3.41% (from 0.7885 to 0.8154) compared to $2.5D+$ slicing, and by 5.88% (from 0.7701 to 0.8154) relative to pure 2D-axial slicing. It is worth highlighting, that the benefit of *omnidirectional* volume slicing generalizes beyond MLP prediction heads across various backbones, including LSTM-, transformer-, and GNN-based architectures, as outlined in Section 4.7. Notably, although our *omnidirectional* slicing strategy cannot be directly converted into a naturally meaningful slice ordering, as compared to consecutive axial slicing, LSTMs trained on *omnidirectional* slices outperform those trained on consecutive axial slices, with a notable AUROC increase of 5.60% (from 0.7446 to 0.7863). Similarly, transformer-based models show an AUROC improvement of 5.83% (from 0.7746 to 0.8198).

While our *omnidirectional* volume slicing strategy consistently improves the average performance across all experiments in Table 1, it does not always yield the single best score for every individual dataset, although it generally remains very close. We attribute these differences as reflections of intrinsic dataset-specific characteristics. In particular, as all datasets offer their highest in-plane resolution when extracting axial slices (i.e., finer spacing in the coronal and sagittal directions), we examine the alignment of target structures along the z-axis. This axis exhibits the greatest anisotropy across all datasets and therefore governs the amount of low-resolution lesion information captured when slices are taken from predominantly coronal or sagittal orientations. To quantify anatomical z-axis alignment, we apply principal component analysis (PCA) to the 3D segmentation masks to identify each target lesion's major axis. The resulting alignment score is computed as the absolute cosine similarity between this major axis and the anatomical z-axis, yielding a value between 0 and 1, corresponding to no alignment and perfect alignment, respectively. Z-axis alignment scores across all datasets are visible in the Table F.7. Overall, we observe the following trend: For datasets with high z-axis alignment such as head-neck (0.92) and soft-tissue tumors (0.93), *omnidirectional* volume slicing strongly improves downstream performance, owing to its inherent slicing plane variability. However, datasets which show a lower z-axis alignment such as breast (0.40) and liver (0.10) tumors naturally benefit from several high-resolution axial slices as offered by e.g., 2.5D+ volume slicing.

Moreover, Table 1 shows that, under *omnidirectional* volume slicing, increasing the number of slice views does not consistently improve performance for either prediction head (MLP or GNN). To better understand this behavior, we perform an additional analysis examining the relationship between radiomics shape features (Van Griethuysen et al., 2017) of the target lesions and model performance. Specifically, for each dataset, we compute the correlation between the median value of each shape feature and the performance difference between 24 and 8 views, considering both backbones (MLP and GNN) under *omnidirectional* slicing. The full set of results is provided in Figs. F.12

and F.13. Overall, no consistent relationship is observed between most radiomics shape features and downstream performance, indicating no general advantage of using more or fewer views. The only exception is elongation (Fig. F.12, top right), which exhibits a clear trend consistent with the properties of *omnidirectional* volume slicing illustrated in Fig. 2. Radiomics shape elongation ranges from 0 to 1, where values close to 1 correspond to compact, approximately spherical shapes, and lower values indicate increasingly elongated structures. For both MLP and GNN models, elongation shows a negative correlation with the performance difference (24 minus 8 views), suggesting that fewer views are preferable for compact lesions, whereas a larger number of views benefits more elongated shapes. This relationship is statistically significant for the MLP ($p = 0.03$) and shows a similar trend for the GNN ($p = 0.07$). We argue that a higher number of views is particularly beneficial for elongated structures, as additional *omnidirectional* slicing planes increase the likelihood of capturing the lesion along its principal axis. This results in larger cross-sectional areas, which are more likely to capture discriminative features.

Taken together, the strong results throughout this work show that *omnidirectional* slicing captures volumetric structures more effectively than conventional slicing strategies, leading to consistent, architecture-independent performance gains. These outcomes underscore the flexibility and added benefit of our approach as a compelling alternative to traditional methods such as 2D-axial or 2.5D slicing. We argue that *omnidirectional* slicing is highly relevant for the medical image analysis community working with slice-based pipelines, as it provides a simple yet elegant way to represent 3D targets in a spatially coherent manner.

5.2. Spatially coherent graph-based feature aggregation

Beyond the choice of representing a volume as a set of 2D slices, the way information from cross-sectional images is aggregated is of central importance. To this end, we presented our second core contribution of this work: GNN-based feature aggregation, which relies on a *spherical mesh-based graph topology* to explicitly model the spatial relationships of encoded 2D slices extracted from 3D volumes. Our spherical graph design is derived from the *omnidirectional* slicing strategy, enabling the integration of information across a set of in-plane (i.e. canonical) and out-of-plane (i.e., non-canonical) perspectives while explicitly encoding their spatial relationships within the graph topology. Our experimental evaluations in Sections 4.4 and 4.7 underscore the advantage of our graph-based feature aggregation method across the investigated datasets.

While MLPs do not offer any built-in capability to model relationships among slices, LSTMs treat consecutive slices as temporal sequences, thereby preserving their spatial ordering within a volume. However, such an ordering is only naturally defined along a single plane of consecutive slices, restricting the representation to one direction and potentially overlooking complementary information from other orientations. Nevertheless, although *omnidirectional* volume slicing improves LSTM downstream performance, there is no natural ordering of these slices, potentially hindering further gains.

We attribute the superior performance of our GNN-based approach to its stronger inductive bias. Specifically, the node feature aggregation and node connectivity are predefined, whereas for transformers, such as MST-based aggregation (Müller-Franzes et al., 2025), spatial relationships and global dependencies must be learned through their self-attention mechanism, introducing additional complexity. Moreover, transformers typically perform best in large-scale data settings, which are uncommon in medical imaging, limiting their practical application (Xia and Wang, 2023). We claim that simplifying the learning task by the topological constraints introduced in our GNN-based approach leads to improved overall average performance while handling limited data more efficiently.

Furthermore, we argue that the performance improvements also lie in the GNN's ability to explicitly integrate both local and global information by propagating slice-level features across nodes through edge connections defined by the graph structure (see Section 3.3). This mechanism is not available e.g., to the MLP, which instead processes all slice information through fully connected layers without explicit relational modeling. The adopted graph structure, which uses hop-distance-based edge weighting to emphasize local over distant connections, is empirically supported by the results in Section 4.6 (Fig. 6, right panel). Specifically, the spherical topology, which restricts connections to immediate neighbors, achieves performance comparable to a fully connected (uniform) graph where all edges are equally weighted, despite requiring substantially fewer connections. This finding supports the effectiveness of the proposed hop-distance-based weighting scheme to emphasize local over distant connections.

Furthermore, given that the employed graph topology approximates a three-dimensional geometric structure (i.e., a sphere), we argue that a graph-based representation is better suited to preserve the inherent spatial relationships among slices compared to dense connections in an MLP. Consequently, the graph structure provides a more faithful representation of the 3D volume when represented through 2D slices and their corresponding relative position.

5.3. Balancing data dimensionality and data scarcity

Current approaches to 3D medical image classification face a fundamental trade-off between data dimensionality (2D slices vs. 3D volumes) and data quantity (i.e., sample size). Although 2D slice-based methods achieve strong performance, they inherently lose anatomical context and spatial coherence when 3D volumes are decomposed into individual slices. Conversely, models that directly process 3D data preserve spatial structure but are constrained by the limited availability of pretraining data. Both, data dimensionality and data scarcity substantially impact a model's representational capacity for volumetric analysis, two crucial aspects for reliable downstream classification.

While foundation models have advanced rapidly in domains such as histopathology (Filiot et al., 2024; Chen et al., 2024; Koch et al., 2024), where large-scale 2D data and sophisticated pretraining strategies drive strong transferability, radiology data-based 3D foundation models remain comparatively underdeveloped (van Veldhuizen et al., 2025). Our findings in Section 4.8 support this observation.

We began by comparing *TomoGraphView* with five large-scale pretrained 3D models across six oncology datasets, using frozen encoder backbones. Evaluating raw embeddings enables a resource-efficient and faithful assessment of the representational quality learned during pretraining. Ideally, a robust pretrained model should produce embeddings that capture rich image semantics and correlate well with downstream performance. As *TomoGraphView* also relies solely on frozen DINOv2 features, this comparison remains fair. Across all datasets, the FMCIB model (Pai et al., 2024) performs best among 3D approaches but consistently trails *TomoGraphView*, with relative improvements of 15.5% in AUROC and 16.8% in balanced accuracy (see Table F.5). Interestingly, despite FMCIB (Pai et al., 2024) being pretrained exclusively on body CT data (DeepLesion Yan et al., 2018), it outperforms MRI-pretrained models on brain tumor MRI classification and matches the performance of the MRI-pretrained SwinUNETR (Wald et al., 2024). We attribute this cross-modality generalization to the model's ability to learn universal imaging features such as textures, edges, and spatial hierarchies. We further evaluated fine-tuning performance to assess how well FMCIB adapts pretrained representations to specific downstream tasks. Closely following the fine-tuning setup of Wald et al. (2024), we observe improvements for brain, soft-tissue, and head–neck tumors, but no gain for breast, kidney, or liver datasets. This highlights that the effectiveness of fine-tuning is highly dataset-dependent and that no single configuration generalizes universally without extensive hyperparameter optimization. Nevertheless, overall fine-tuning performance still falls short of *TomoGraphView*. Notably, in this study, we focused exclusively on finetuning large-scale pretrained 3D models, as

transfer learning has consistently demonstrated superior performance over training from scratch, particularly in medical image classification tasks where labeled data is often scarce. Existing studies show that leveraging prior knowledge from large-scale datasets enables faster convergence and improved generalization to domain-specific medical data (Tajbakhsh et al., 2016).

5.4. Limitations and outlook

While our proposed *omnidirectional* volume slicing consistently improves performance across tasks, datasets, and prediction heads, its current design benefits from the availability of tumor segmentation masks, but is not restricted to them, as effective slice selection relies on identifying the largest lesion area. Notably, this property is shared for all other volumetric slicing methods. Fortunately, many established algorithms can generate sufficiently accurate tumor delineations, making our approach broadly applicable. Nevertheless, it is worth noting that 3D medical image classification is usually performed on subvolumes, typically defined by bounding box coordinates around the structure of interest, to reduce the complexity of the problem. Given such a bounding box, we argue and show that selecting the central slice of the resulting subvolume provides adequate information, thereby bypassing the need for tumor segmentation, which we also show in Section 4.3. Nevertheless, a limitation of our approach is that, as demonstrated in Fig. F.10, optimal downstream performance depends on the availability of high-precision segmentation masks, whereas the use of bounding boxes leads to inferior results. Consequently, the quality of the segmentation masks may constitute a potential performance bottleneck.

Our *omnidirectional* volume slicing strategy introduces a modest preprocessing overhead, as generating N views requires linearly interpolating the volume N times from different orientations. In a quantitative benchmark comparison (Table F.8) of 2D axial+, 2.5D+, and *omnidirectional* slicing, the average preprocessing time for 24 views is 1.99 s for 2D axial+, 2.04 s for 2.5D+, and 4.90 s for *omnidirectional* slicing. Since these methods are not intended for real-time applications and preprocessing is performed only once per sample prior to training, the additional cost remains modest compared to 3D models operating directly on volumetric data. While the overhead is dataset-dependent and largely determined by the number of voxels, we note that none of the methods are optimized for computational efficiency and could be further accelerated. In addition, the *TomoGraphView* framework requires the construction of a spherical graph topology, which takes 1.1 s for 8 views, 4.3 s for 16 views, and 9.7 s for 24 views. Nevertheless, this step is performed only once per view configuration in a preprocessing step and can be reused across all datasets. Therefore, it does not constitute a substantial computational burden as the graph does not to be re-constructed during training. Finally, we report end-to-end inference times and memory consumption for both MLP and GNN prediction heads in Table F.9, demonstrating that both exhibit comparable computational efficiency.

In future work, we plan to extend *TomoGraphView* by integrating a detection module, enabling a fully end-to-end 3D medical image classification pipeline. A promising direction in this regard is leveraging multimodal large language models, which can utilize in-context samples and textual prompts to predict bounding boxes for structures of interest, without requiring costly parameter updates or finetuning (Ye and Tang, 2025). Furthermore, our current analysis of graph topology focuses on edge weighting schemes based on hop distance. A promising direction for future work is the systematic exploration of alternative graph constructions, such as topologies with edge weights based on e.g. angular distance of view directions or multi-level graph designs that combine multiple slices per view direction, as prior work suggests that the predictive performance of GNNs can be strongly affected by the underlying graph structure (Müller et al., 2024). Beyond graph topology, the choice of graph convolution operator has also been shown to impact downstream performance (Kiechle et al., 2024), motivating future investigation into the sensitivity of our framework to different GNN operator design choices. Nevertheless, as *TomoGraphView* is designed as a flexible framework compatible with any 2D foundation model, we anticipate that its performance will continue to improve in tandem with future advancements in 2D backbone architectures and training paradigms.

6. Conclusion

In this work, we present *TomoGraphView*, a unified framework for 3D medical image classification that combines *omnidirectional* volume slicing with *spherical graph-based feature aggregation*. Using six diverse datasets spanning multiple anatomical regions, imaging modalities, and clinical tasks, we demonstrated that (I) our *omnidirectional* slicing approach consistently outperforms standard slicing strategies based on axial, sagittal, coronal, or combined views, and (II) *spherical graph-based feature aggregation* improves performance beyond existing slice-wise feature aggregation baselines.

While 3D models are naturally suited for tomographic data as they capture spatial context across all axes natively, their use in practice remains challenging: they require substantial computational resources, substantially larger datasets to achieve robust generalization due to their high parameter counts, and offer limited possibilities for transfer learning, as publicly available pretrained 3D networks remain comparatively scarce and less established than 2D alternatives (van Veldhuizen et al., 2025). As a result, slice-based 2D methods constitute an appealing alternative: they are more efficient to train, benefit from the large number of slices extracted from each volume, and can directly leverage powerful 2D foundation models developed for natural images.

TomoGraphView builds on these advantages. Our *omnidirectional* slicing strategy provides a flexible and effective alternative to standard slicing along canonical planes, improving downstream performance across a wide range of prediction heads, including MLPs, LSTMs, and transformer-based approaches. Moreover, by integrating *omnidirectional* volume slicing with spherical graph-based feature aggregation, our *TomoGraphView* framework achieves superior performance levels when compared against both existing slice-wise feature aggregation methods and large-scale pretrained 3D models. The results across our experimental evaluations underscore the value of combining powerful 2D encoders with expressive methods capable of sampling and integrating versatile information from volumetric scans. Therefore, *TomoGraphView* offers a practical and powerful approach while the field continues to progress toward fully native 3D foundation models.

In summary, this work highlights both the relevance and flexibility of 2D slice-based strategies for 3D medical image classification. By integrating powerful 2D vision foundation models with improved volumetric slicing and graph-based aggregation, *TomoGraphView* contributes a key step toward bridging the gap until robust and widely available 3D foundation models become standard in medical image analysis.

CRediT authorship contribution statement

Johannes Kiechle: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stefan M. Fischer:** Writing – review & editing, Validation. **Daniel M. Lang:** Writing – review & editing, Validation. **Cosmin I. Bercea:** Writing – review & editing, Validation. **Matthew J. Nyflot:** Writing – review & editing, Validation, Data curation. **Lina Felsner:** Writing – review & editing, Validation, Conceptualization. **Julia A. Schnabel:** Writing – review & editing, Validation, Supervision. **Jan C. Peeken:** Writing – review & editing, Validation, Supervision, Funding acquisition.

Ethical standards

Ethical clearance for this study utilizing retrospectively collected in-house soft-tissue sarcoma data was obtained from the respective ethics committees of the data-providing institutions (i.e., Technical University of Munich, University of Washington/Seattle Cancer Care Alliance).

Declaration of AI Assistance

The authors acknowledge the use of ChatGPT to assist in improving the readability and language of this manuscript. The authors thoroughly reviewed and edited the text following its use and accept full responsibility for the content and integrity of the final publication.

Declaration of competing interest

JCP holds shares in Mevidence and has received honoraria from AstraZeneca and Brainlab, all of which are outside the scope of the submitted work. The remaining authors declare that they have no known conflict of interest nor competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors JK, SMF, CIB, and JAS gratefully acknowledge financial funding received by Munich Center for Machine Learning, sponsored by the German Federal Ministry of Education and Research. Authors JCP and JK gratefully acknowledge funding from the Wilhelm Sander Foundation in Cancer Research (2022.032.1). Author SMF has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 515279324/SPP 2177. SMF acknowledges funding from the EVUK program (“Next-generation AI for Integrated Diagnostics”) of the Free State of Bavaria. JCP acknowledges funding from the Else Kröner-Fresenius Stiftung (2023_EKES.01). SMF and LF acknowledge funding from the Free State of Bavaria (StMGP) as part of the project KI-gestützte multi-modale Diagnostik und stratifizierte Therapie für Endometriose (EndoKI).

Appendix A. Volume slicing comparison

Whereas Fig. 2 in Section 3.1 shows the resulting eight cross-sectional slices for a synthetically generated input volume, displaying a target structure that is not naturally aligned with canonical axes, we extend the visualization to 24 views in Fig. A.9. While the slice variability for both, 2D-axial+ (top) and 2.5D+ (middle), is limited, our *omnidirectional* volume slicing strategy (bottom) excels by representing the target structure with all its facets, thereby depicting a more faithful representation of the target 3D structure.

Appendix B. Dataset task descriptions

For all datasets except the head-neck and liver cohorts, this study focuses on tumor grading, where cases are categorized into entities of differing malignancy. Generally, tumor grade provides critical insight into how aggressive a tumor is and its potential for progression or metastasis, thereby serving as an essential factor for cancer staging and treatment decisions. For the liver dataset, we perform tumor subtype prediction, whereas for the Head-Neck dataset, we instead predict human papillomavirus (HPV) status, as HPV infection constitutes a major risk factor for oropharyngeal cancer (OPC). Importantly, HPV-positive OPCs have been shown to exhibit greater radio-sensitivity and lower cancer-related mortality compared to HPV-negative cases. Thus, determining HPV status constitutes a key diagnostic marker with direct implications for prognosis and treatment decisions.

Appendix C. 3D pretrained model description

Models Genesis (Zhou et al., 2019) is a collection of models built from unlabeled 3D imaging data using a restorative reconstruction-based self-supervised method. It aims to generate powerful application-specific target models through transfer learning. The models demonstrate strong performance across various medical imaging tasks, emphasizing their potential for broad applicability in clinical settings.

VOCO (Wu et al., 2024) is a large-scale 3D medical image pre-training framework that leverages geometric context priors to learn consistent semantic representations. It is built on a substantial dataset of CT volumes and employs a novel pretext task for contextual position predictions. VOCO has demonstrated superior performance across various downstream tasks, establishing itself as a leading model in the field of medical imaging.

VISTA3D (He et al., 2025) is a versatile imaging segmentation and annotation model that supports both automatic and interactive segmentation of 3D medical images. It is the first unified foundation model to achieve state-of-the-art performance across 127 classes and is designed to facilitate efficient human correction through its interactive features. VISTA3D integrates a novel supervoxel method to enhance zero-shot performance, marking a significant advancement in 3D medical imaging.

FMCIB (Pai et al., 2024) is a foundation model designed to distinguish between lesions and non-lesions at the patch level in medical imaging. It aims to enhance the detection and characterization of cancerous lesions by leveraging self-supervised learning techniques. It is based on SimCLR (Chen et al., 2020) (Simple Framework for Contrastive Learning of Visual Representations), a self-supervised learning approach that utilizes contrastive learning to pretrain deep neural networks without the need for labeled data. The core idea behind SimCLR is to maximize the similarity between differently augmented views of the same image while minimizing the similarity between views of different images.

SwinUNETR (Cao et al., 2022) was proposed as a pretraining method for the identically named SwinUNETR architecture and is composed of three components: image inpainting, Rotation prediction, and a contrastive training objective. The inpainting itself is a simple L1 loss applied to a masked-out image region, the rotation is a rotation of 0°, 90°, 180°, or 270° degrees along the z-axis, with an MLP used to classify the applied rotation. Lastly, the contrastive coding ensures that the linearly projected representations of the encoder are highly similar or dissimilar, depending on whether the two sub-volumes belong to the same or a different image, respectively. These three losses are combined with an equal weighting to form the SwinUNETR pretraining.

Appendix D. Graph topology ablation

Detailed quantitative results of our graph topology ablation experiment performed in Section 4.6 can be found in Table D.3. In this context, we compare our proposed *TomoGraphView* framework across various graph topologies and edge connection weighting schemes. For graph topology, we differentiate between local and complete, where local represents a pure spherical structure, while complete represents the spherical structure with additional cross-connections between all nodes in the graph. To this end, we evaluate different edge weighting schemes, including uniform, linear decay, inverse, and inverse-square, which indicate decaying cross-node weights with increasing hop distances.

Appendix E. Slice-wise feature aggregation methods

Detailed quantitative results of our slice-wise feature aggregation benchmark experiment performed in Section 4.7 can be found in Table E.4. Therein, we compare different slice-wise feature aggregation methods, including long short-term memory (LSTM) and medical slice transformer (MST), with our *TomoGraphView* framework across various volume slicing strategies and numbers of views. Numbers are indicated as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds.

Appendix F. 3D pretrained backbones results

Detailed quantitative results of our comparative large-scale 3D pretrained model analysis performed in Section 4.8 can be found in Table F.5 (frozen backbone performance) and Table F.6 (finetuning performance).

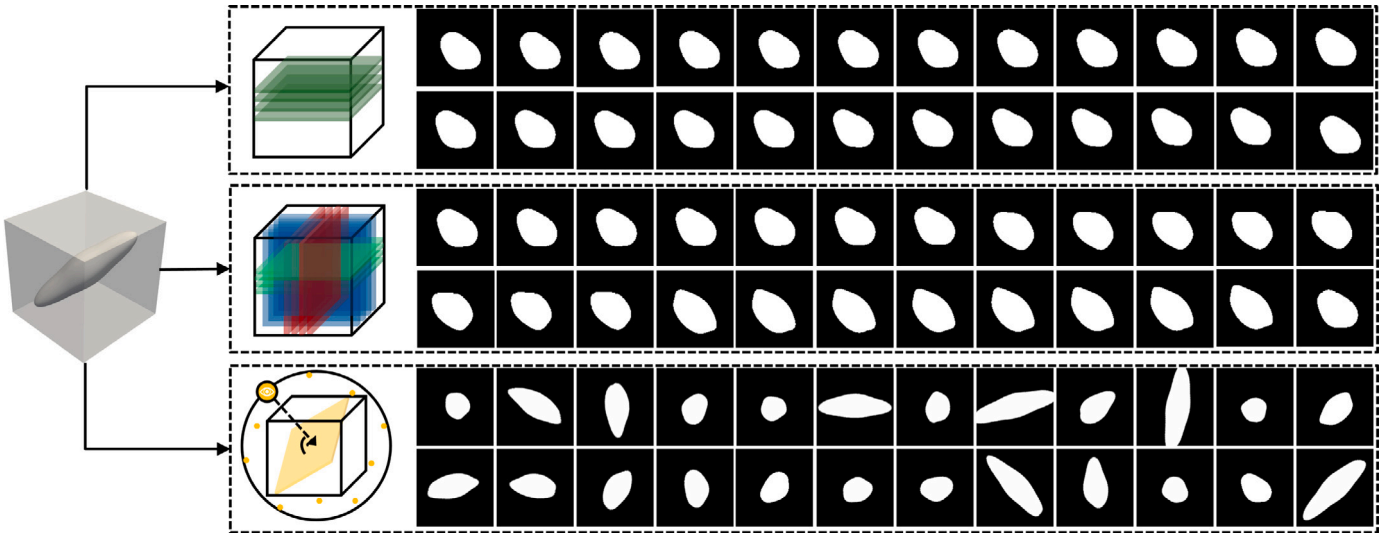

Fig. A.9. A visual comparison of the three different volume slicing techniques and their resulting cross-sectional slices for a synthetically generated input volume, showing a target structure of interest which is not aligned with any canonical axis (i.e, axial, coronal, sagittal). The plot shows 2D-axial+ slicing (top), 2.5D+ slicing (middle), and our proposed *omnidirectional* slicing (bottom), for 24 views each.

Table D.3
Quantitative results for the graph topology analysis comparing different graph topologies and edge weighting schemes. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Method	Topology	Weighting	Breast	Brain	Head–neck	Kidney	Liver	Soft-tissue	Mean
TomoGraphView	Local	Uniform	0.6997	<u>0.9667</u>	**0.7770**	0.7526	**0.9666**	0.8204	0.8305
	Complete	Uniform	0.6939	**0.9668**	0.7480	0.7672	0.9582	**0.8475**	0.8303
	Complete	Linear decay	0.6857	0.9653	0.7503	0.7474	0.9614	0.8371	0.8245
	Complete	Inverse	<u>0.7084</u>	0.9655	0.7527	**0.7893**	0.9634	<u>0.8441</u>	**0.8372**
	Complete	Inverse-square	**0.7111**	0.9621	<u>0.7549</u>	<u>0.7878</u>	<u>0.9646</u>	0.8249	<u>0.8342</u>

Table E.4
Quantitative results for the slice-wise feature aggregation benchmark comparing different methods, such as long short-term memory (LSTM) and medical slice transformer (MST) with *TomoGraphView* across volume slicing strategies (i.e., 2D-axial+ and *omnidirectional*) and number of views (i.e., 8, 16, and 24). Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Method	Volume slicing	Views	Breast	Brain	Head–neck	Kidney	Liver	Soft-tissue	Mean
LSTM	2D-axial+	8	0.6437	0.8826	0.6048	0.6669	0.8923	0.7254	0.7305
		16	0.6505	0.8957	0.6207	0.7072	0.7866	0.7374	0.7330
		24	0.6077	0.8922	0.6467	0.7546	0.8275	0.7388	0.7446
LSTM	Omnidirectional	8	0.6980	0.9419	0.6510	0.6705	0.9133	0.7837	0.7764
		16	0.6731	0.9497	0.6670	0.6733	0.9137	0.7696	0.7744
		24	0.6303	0.9580	0.7179	0.6861	0.8976	0.8280	0.7863
MST	2D-axial+	8	0.6497	0.9232	0.5949	0.7124	0.9002	0.7696	0.7583
		16	0.6539	0.9308	0.6294	0.7563	0.9081	0.7824	0.7768
		24	0.6014	0.9346	0.6539	0.7558	0.9206	0.7815	0.7746
MST	Omnidirectional	8	0.7037	0.9655	0.6443	0.7573	0.9497	0.8295	0.8083
		16	0.7164	0.9663	0.7225	0.7286	0.9238	**0.8452**	0.8171
		24	0.7029	0.9669	0.7050	0.7751	0.9254	0.8434	0.8198
TomoGraphView	Omnidirectional	8	**0.7336**	**0.9678**	0.6780	0.7878	0.9600	0.8399	0.8279
		16	0.6907	0.9666	0.7244	0.7479	0.9585	0.8389	0.8212
		24	0.7084	0.9655	**0.7527**	**0.7893**	**0.9634**	0.8441	**0.8372**

Table F.5
Quantitative results for the frozen backbone benchmark comparing different methods relying on different pretraining schemes, such as Foundation Model for Cancer Imaging Biomarkers (FMCIB), Models Genesis, SwinUNETR, Versatile Imaging SegmenTation and Annotation model (VISTA3D), Volume Contrast (VoCo) with *TomoGraphView*. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC), balanced accuracy (ACC), F1-score, and Matthews correlation coefficient (MCC), across 5 folds. The best-performing method per dataset is indicated in bold, while the second-best is underlined. The last section, indicated as 'Mean', averages the performance across all datasets and performance metrics.

Dataset	Modality	Method	AUROC	ACC	F1-score	MCC
Breast tumors	MRI	FMCIB	0.6197	0.5427	0.5057	0.1016
		ModelsGenesis	0.4527	0.5017	0.3537	0.0024
		SwinUNETR	0.6207	0.5445	0.5303	0.0876
		VISTA3D	0.4349	0.4592	0.3950	−0.0844
		VoCo	0.4957	0.5149	0.5049	0.0321
		TomoGraphView (ours)	**0.7084**	**0.6573**	**0.6553**	**0.3177**
Brain tumors	MRI	FMCIB	0.8985	0.7921	0.8587	0.5716
		ModelsGenesis	0.8708	0.6257	0.7895	0.3037
		SwinUNETR	0.8583	0.7694	0.7979	0.4646
		VISTA3D	0.6741	0.5505	0.5843	0.0795
		VoCo	0.7267	0.6719	0.7962	0.3569
		TomoGraphView (ours)	**0.9655**	**0.9131**	**0.9066**	**0.7443**
Head–neck tumors	CT	FMCIB	0.5733	0.5407	0.5448	0.1027
		ModelsGenesis	0.5645	0.5040	0.2805	0.0240
		SwinUNETR	0.5910	0.5609	0.5532	0.1424
		VISTA3D	0.5686	0.5258	0.4875	0.0452
		VoCo	0.5719	0.5759	0.6035	0.1499
		TomoGraphView (ours)	**0.7527**	**0.6464**	**0.6711**	**0.2924**
Kidney tumors	CT	FMCIB	0.6828	0.6392	0.6785	0.2706
		ModelsGenesis	0.6985	0.5667	0.5181	0.1279
		SwinUNETR	0.7466	0.7112	**0.7520**	0.4282
		VISTA3D	0.6888	0.6102	0.5508	0.2614
		VoCo	0.6801	0.6535	0.6973	0.3111
		TomoGraphView (ours)	**0.7893**	**0.7245**	0.7291	**0.4412**

(*continued on next page*)

Table F.5 (*continued*).

Dataset	Modality	Method	AUROC	ACC	F1-score	MCC
Liver tumors	CT	FMCIB	0.8711	0.8082	0.8060	0.6208
		ModelsGenesis	0.6621	0.5695	0.5013	0.1743
		SwinUNETR	0.7249	0.6696	0.6641	0.3406
		VISTA3D	0.6221	0.5962	0.5941	0.2040
		VoCo	0.6053	0.5896	0.5812	0.1802
		TomoGraphView (ours)	**0.9634**	**0.8956**	**0.8959**	**0.7939**
Soft-tissue tumors	MRI	FMCIB	0.6567	0.5882	0.6499	0.2010
		ModelsGenesis	0.5708	0.5432	0.5308	0.0928
		SwinUNETR	0.6926	0.5595	0.6841	0.1583
		VISTA3D	0.6013	0.5380	0.6495	0.0897
		VoCo	0.6651	0.6327	0.7151	0.2737
		TomoGraphView (ours)	**0.8441**	**0.7340**	**0.7957**	**0.4815**
Mean	–	FMCIB	0.7170	0.6518	0.6739	0.3114
		ModelsGenesis	0.6366	0.5518	0.4956	0.1209
		SwinUNETR	0.7057	0.6358	0.6636	0.2703
		VISTA3D	0.5983	0.5466	0.5435	0.0992
		VoCo	0.6241	0.6064	0.6497	0.2173
		TomoGraphView (ours)	**0.8372**	**0.7618**	**0.7756**	**0.5191**

Table F.6
Quantitative results for the finetuning performance of the best-performing frozen backbone, Foundation Model for Cancer Imaging Biomarkers (FMCIB). We employ different warmup durations, along with various batch sizes (BS), and compare their performance with our *TomoGraphView* framework. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Model	Warmup	BS	Params	Brain	Soft-tissue	Breast	Head–neck	Kidney	Liver	Mean
FMCIB (frozen)	n/a	16	100k	0.8985	0.6567	0.6197	0.5733	0.6828	0.8711	0.7170
FMCIB (finetune)	20	256	184M	0.9147	0.7322	0.5049	0.6286	0.5730	0.8708	0.7040
	50	256	184M	0.8981	0.7230	0.5050	0.6308	0.6162	0.8636	0.7061
	100	256	184M	0.9049	0.7013	0.5193	0.6403	0.6589	0.8639	0.7147
	100	16	184M	0.9016	0.6615	0.5540	0.6386	0.6231	0.8811	0.7100
TomoGraphView (ours)	n/a	16	100k	**0.9626**	**0.8485**	**0.7491**	**0.6873**	**0.7887**	**0.9603**	**0.8328**

Table F.7
Quantitative dataset-specific metrics, showing median z-axis spacing, its corresponding z-axis spacing interquartile range (IQR) and z-axis alignment of the target anatomical structure. To quantify anatomical z-axis alignment, we apply principal component analysis (PCA) to the 3D segmentation masks to identify each structure's major axis. The resulting alignment score is computed as the absolute cosine similarity between this major axis and the anatomical z-axis, yielding a value between 0 (no alignment) and 1 (perfect alignment). The unit measuring median slice spacing and IQR slice spacing is given in millimeters (mm).

Metric	Unit	Breast	Brain	Head-neck	Kidney	Liver	Soft-tissue
z-axis spacing (median)	mm	1.00	1.00	2.00	3.00	1.00	4.80
z-axis spacing (IQR)	mm	0.10	0.00	0.00	3.75	0.00	5.86
z-axis alignment score	n/a	0.40	0.32	0.92	0.51	0.10	0.93

Table F.8
Average preprocessing time [s] for different volume slicing strategies. The table reports the average image preprocessing duration (in seconds) across datasets for three slicing strategies: 2D-axial+ slicing, 2.5D+ slicing, and our proposed *omnidirectional* slicing, each using 24 views. Results are shown separately for breast-, brain-, head–neck-, kidney-, liver-, and soft-tissue tumor datasets, along with the overall mean across datasets and volume slicing strategies.

Slicing strategy	Views	Breast	Brain	Head-neck	Kidney	Liver	Soft-tissue	Mean [s]
2D-axial+ slicing	24	2.42	0.53	3.18	3.16	1.95	0.75	1.99
2.5D+ slicing	24	2.50	0.56	3.29	3.13	1.97	0.79	2.04
Omnidirectional slicing (ours)	24	2.67	1.60	3.36	3.74	9.75	8.21	4.90

Table F.9
Comparison of end-to-end average inference time [seconds] and VRAM consumption [Megabyte] between Multi Layer Perceptron (MLP) and Graph Neural Network (GNN) prediction heads across 24 views. Reported time duration and memory consumption is based on a single sample during inference.

Pred. Head	Views	Mean inference time [s]	Mean VRAM Pred. Head [MB]	Mean VRAM encoder + Pred. Head [MB]
MLP	24	2.588×10^{-3}	4.80	4111.25 (4106.45 + 4.80)
GNN	24	2.524×10^{-3}	5.04	4111.49 (4106.45 + 5.04)

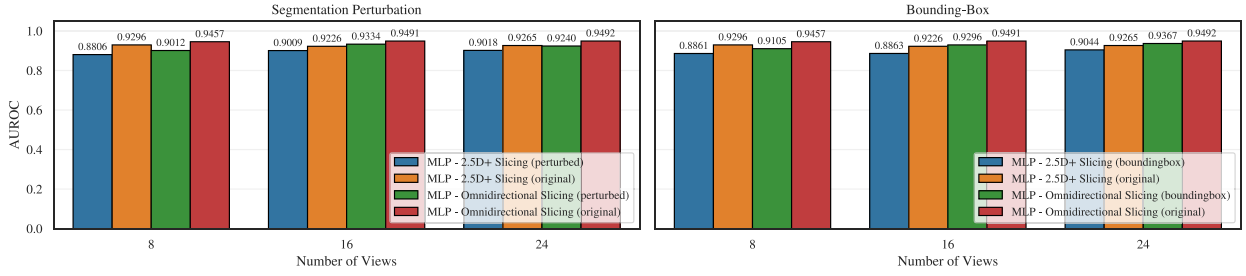

Fig. F.10. Effect of slicing strategy and input slice selection on classification performance. AUROC values obtained using multi-layer perceptron (MLP) classifiers trained on features extracted with 2.5D+ slicing and *omnidirectional* slicing, evaluated with 8, 16, and 24 views. The left panel shows results when using perturbed segmentation's, while the right panel shows results when using bounding-box and its corresponding middle slice instead of biggest segmentation lesion area.

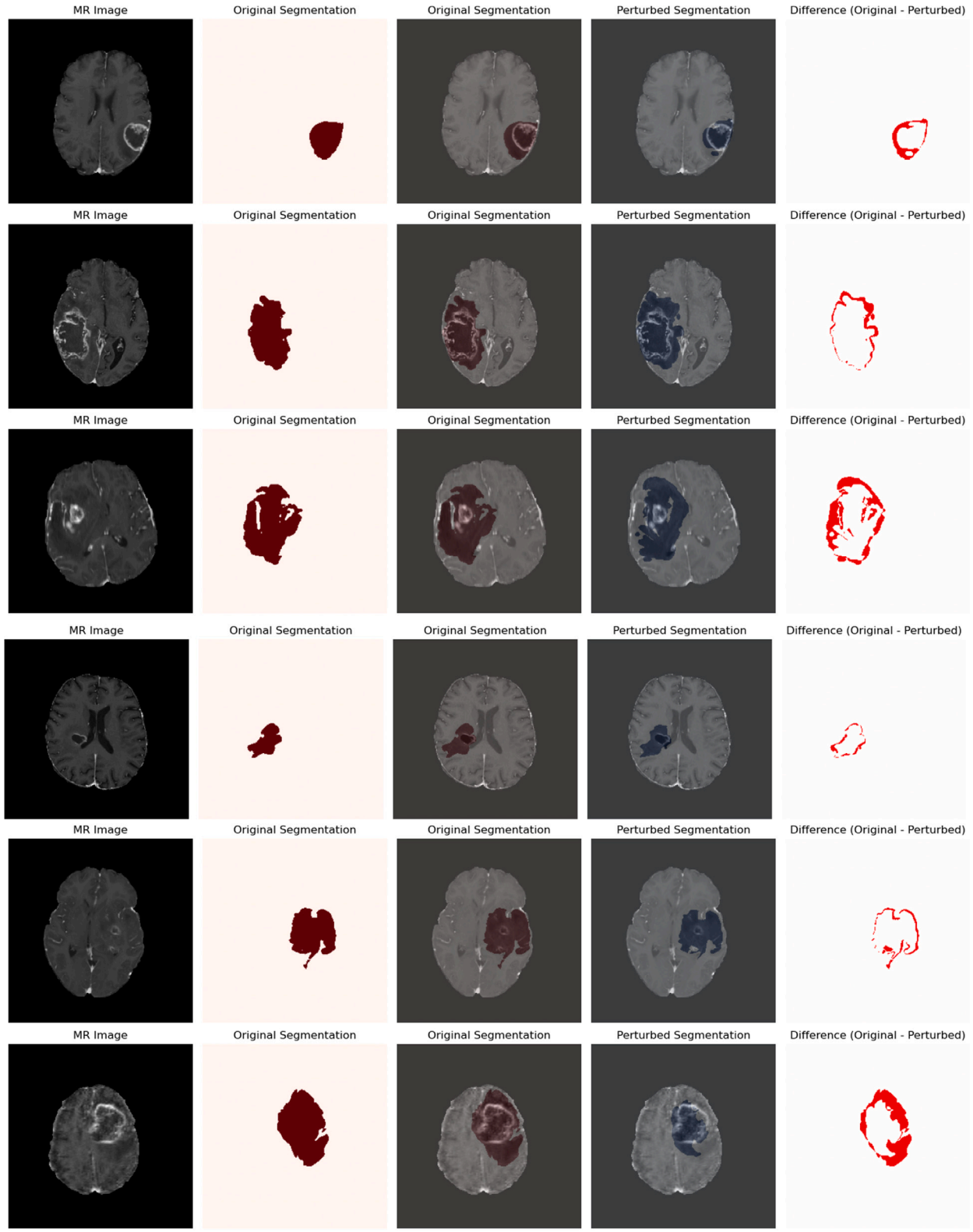

Fig. F.11. Segmentation perturbation of the original lesion mask. Each row shows a representative axial MR slice with the corresponding segmentation and its perturbation. From left to right: (1) input brain MR image, (2) original lesion segmentation mask, (3) overlay of the original segmentation on the MR image, (4) overlay of the segmentation obtained after perturbation, and (5) spatial difference between the original and perturbed segmentation, where the difference map highlight the local regions where the lesion boundaries change under the applied inter-observer variability patterns (Weltens et al., 2001; Gillett et al., 2025).

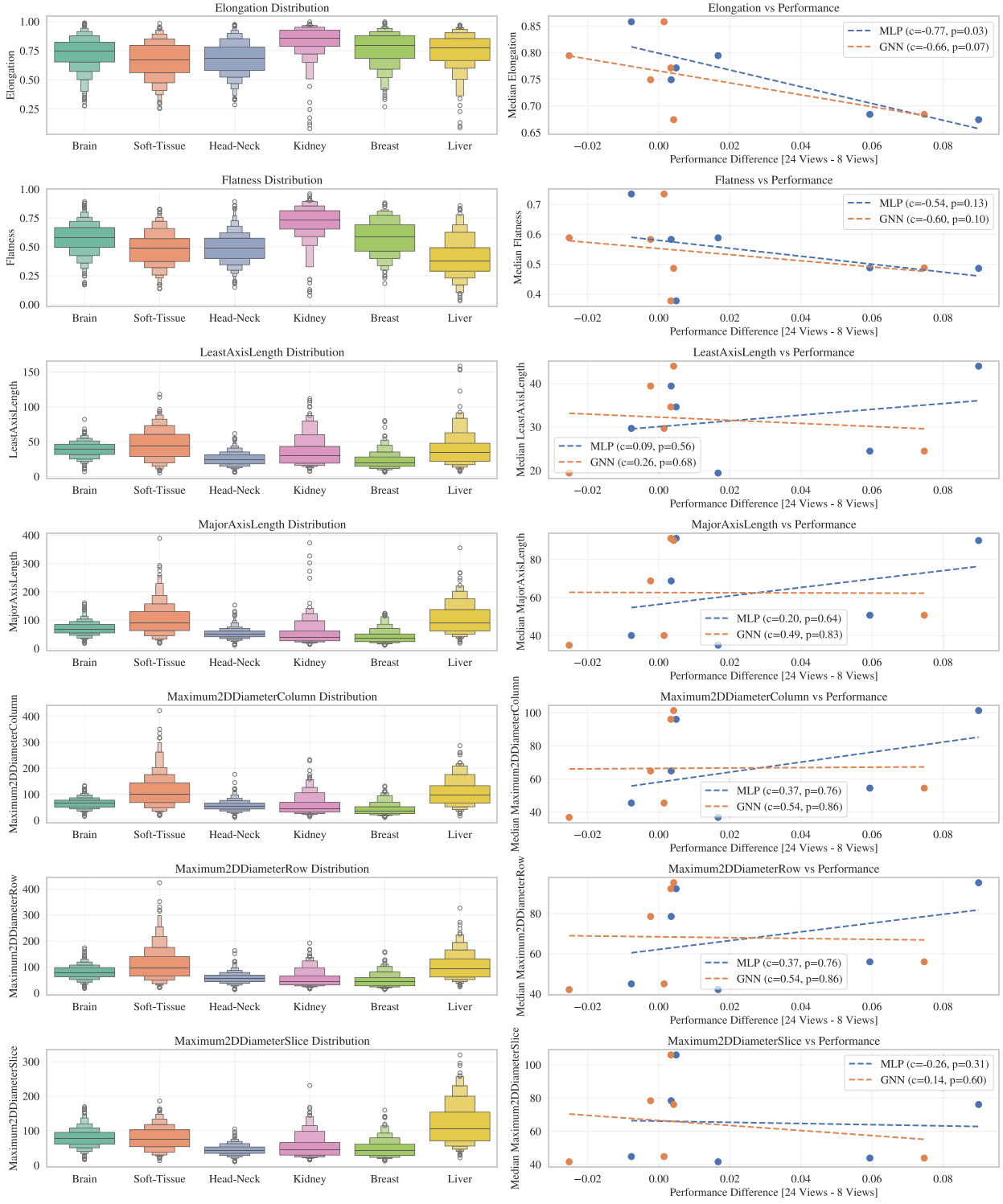

Fig. F.12. (Part 1): Radiomics shape feature distributions and their relationship with model performance across datasets. Left column: Boxplots showing the distribution of extracted shape descriptors across brain, soft tissue, head–neck, kidney, breast, and liver datasets. Right column: Scatter plots depicting the association between median shape feature values (y-axis) and the performance difference (24 views minus 8 views) for both backbones (MLP and GNN) and *omnidirectional* slicing (x-axis). Dashed lines indicate linear regression fits, with corresponding Spearman correlation coefficients (c) and p-values (p) reported in the legends.

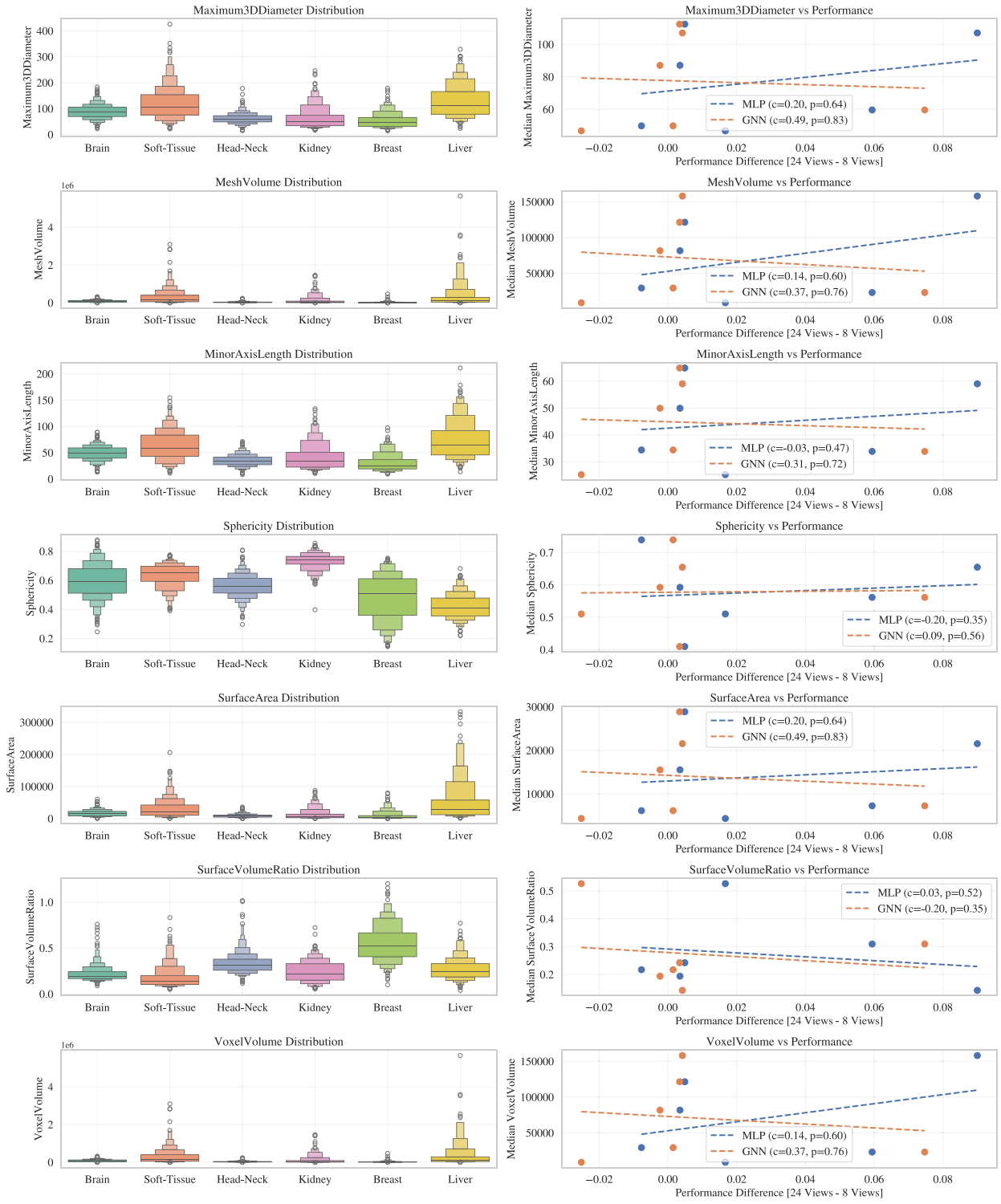

Fig. F.13. (Part 2): Radiomics shape feature distributions and their relationship with model performance across datasets. Left column: Boxplots showing the distribution of extracted shape descriptors across brain, soft tissue, head–neck, kidney, breast, and liver datasets. Right column: Scatter plots depicting the association between median shape feature values (y-axis) and the performance difference (24 views minus 8 views) for both backbones (MLP and GNN) and *omnidirectional* slicing (x-axis). Dashed lines indicate linear regression fits, with corresponding Spearman correlation coefficients (c) and p-values (p) reported in the legends.

Data availability

All data, except the in-house soft-tissue sarcoma dataset, used in this study, are publicly available. We provide detailed descriptions and proper citations for each dataset used to ensure full transparency and reproducibility. Additionally, all code for data preprocessing and preparation is publicly released and can be used to fully reproduce our results. The code is available at http://github.com/compai-lab/2025-MedIA-kiechle. The in-house soft-tissue sarcoma data may be obtained from the corresponding author upon reasonable request and may be published online in the future after approval by the institutional ethics review boards. Furthermore, we provide a user-friendly library for *omnidirectional* volume slicing at https://pypi.org/project/OmniSlicer.

References

Aerts, H., Pai, S., Hadzic, I., Fedorov, A., Mak, R., 2025. Foundation model embeddings for quantitative tumor imaging biomarkers. Res. Sq. rs–3.

Ahamed, S., Xu, Y., Bloise, I., O, J.H., Uribe, C.F., Dodhia, R., Ferres, J.L., Rahmim, A., 2023. A slice classification neural network for automated classification of axial PET/CT slices from a multi-centric lymphoma dataset. In: Colliot, O., Isgum, I. (Eds.), Medical Imaging 2023: Image Processing, San Diego, CA, USA, February 19-23, 2023. In: SPIE Proceedings, vol. 12464, SPIE.

Antropova, N., Abe, H., Giger, M.L., 2018. Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. J. Med. Imaging 5 (1), 014503.

Ayzenberg, L., Giryes, R., Greenspan, H., 2024. DINOv2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D., 2023. Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689.

Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., Peng, W., 2023. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks. arXiv preprint arXiv:2312.02366.

Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Process. Mag. 34 (4), 18–42.

Buchner, J.A., Kofler, F., Etzel, L., Mayinger, M., Christ, S.M., Brunner, T.B., Wittig, A., Menze, B., Zimmer, C., Meyer, B., et al., 2023. Development and external validation of an MRI-based neural network for brain metastasis segmentation in the AURORA multicenter study. Radiother. Oncol. 178, 109425.

Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P., 2022. The University of California San Francisco preoperative diffuse glioma MRI dataset. Radiol.: Artif. Intell. 4 (6), e220058.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al., 2024. Towards a general-purpose foundation model for computational pathology. Nature Med. 30 (3), 850–862.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. Simclr: A simple framework for contrastive learning of visual representations. In: International Conference on Learning Representations. Proceedings of Machine Learning Research, New York, NY, USA.

Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P., 2020. Principal neighbourhood aggregation for graph nets. Adv. Neural Inf. Process. Syst. 33, 13260–13271.

Dadoun, H., Rousseau, A.-L., de Kerviler, E., Correas, J.-M., Tissier, A.-M., Joujou, F., Bodard, S., Khezzane, K., de Margerie-Mellon, C., Delingette, H., et al., 2022. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. Radiol.: Artif. Intell. 4 (3), e210110.

Delaunay, B., 1934. Sur la sphère vide. Izv. Akad. Nauk SSSR Otd. Mat. Estestv. Nauk 7, 793–800.

Di Piazza, T., Lazarus, C., Nempont, O., Boussel, L., 2025. Structured spectral graph learning for anomaly classification in 3D chest CT scans. arXiv preprint arXiv:2508.01045.

Dijkstra, E.W., 2022. A note on two problems in connexion with graphs. In: Edsger Wybe Dijkstra: His Life, Work, and Legacy. pp. 287–290.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2625–2634.

Erdur, A.C., Rusche, D., Scholz, D., Kiechle, J., Fischer, S., Llorian-Salvador, O., Buchner, J.A., Nguyen, M.Q., Etzel, L., Weidner, J., et al., 2025. Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. Strahlenther. Onkol. 201 (3), 236–254.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2024. Eva-02: A visual representation for neon genesis. Image Vis. Comput. 149, 105171.

Filiot, A., Jacob, P., Mac Kain, A., Saillard, C., 2024. Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173.

Gillett, H., Stanley, E.A., Souza, R., Wilms, M., Forkert, N.D., 2025. Simulating inter-observer variability across clinical experience levels for brain tumour segmentation. In: International Workshop on Human-AI Collaboration. Springer, pp. 81–90.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 1263–1272.

Grossberg, A.J., Mohamed, A.S., Elhalawani, H., Bennett, W.C., Smith, K.E., Nolan, T.S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M.E., et al., 2018. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci. Data 5 (1), 1–10.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. 30.

He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al., 2025. VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20863–20873.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The KITS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445.

Herent, P., Schmauch, B., Jehanno, P., Dehaene, O., Saillard, C., Balleyguier, C., Arfi-Rouche, J., Jégou, S., 2019. Detection and characterization of MRI breast lesions using deep learning. Diagn. Interv. Imaging 100 (4), 219–225.

Hu, Q., Whitney, H.M., Li, H., Ji, Y., Liu, P., Giger, M.L., 2021. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection MRI in breast cancer diagnosis using dynamic contrast-enhanced MRI. Radiol.: Artif. Intell. 3 (3), e200159.

Huang, Y., Zou, J., Meng, L., Yue, X., Zhao, Q., Li, J., Song, C., Jimenez, G., Li, S., Fu, G., 2024. Comparative analysis of imagenet pre-trained deep learning models and DINOv2 in medical imaging classification. In: 2024 IEEE 48th Annual Computers, Software, and Applications Conference. COMPSAC, IEEE, pp. 297–305.

Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K., 2024. Are natural domain foundation models useful for medical image classification? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7634–7643.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jang, J., Hwang, D., 2022. 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. pp. 20718–20729.

Keriven, N., 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. Adv. Neural Inf. Process. Syst. 35, 2268–2281.

Kiechle, J., Lang, D.M., Fischer, S.M., Felsner, L., Peeken, J.C., Schnabel, J.A., 2024. Graph neural networks: A suitable alternative to MLPs in latent 3D medical image classification? In: International Workshop on Graphs in Biomedical Image Analysis. Springer, pp. 12–22.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C., 2024. DinoBloom: a foundation model for generalizable cell embeddings in hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 520–530.

Kwan, J., Su, J., Huang, S., et al., 2019. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma. The Cancer Imaging Archive, 2019. In: Book Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma. The Cancer Imaging Archive.

Lilhore, U.K., Sharma, Y.K., Shukla, B.K., Vadlamudi, M.N., Simaiya, S., Alroobaea, R., Alsafyani, M., Baqasah, A.M., 2025. Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification. Sci. Rep. 15 (1), 12082.

Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., Bercea, C.I., Ouyang, C., Chen, C., Xiong, Z., Wiestler, B., Wachinger, C., Rueckert, D., Bai, W., Arcucci, R., 2025. Does DINOv3 set a new medical vision standard? arXiv:2509.06467.

Luo, J., Wan, X., Du, J., Liu, L., Zhao, L., Peng, X., Wu, M., Huang, S., Nie, X., 2025. Comprehensive multi-phase 3D contrast-enhanced CT imaging for primary liver cancer. Sci. Data 12 (1), 768.

Meneghetti, A.R., Hernández, M.L., Kuehn, J.-P., Löck, S., Carrero, Z.I., Perez-Lopez, R., Bressem, K.K., Brinker, T.J., Pearson, A.T., Truhn, D., et al., 2025. End-to-end prediction of clinical outcomes in head and neck squamous cell carcinoma with foundation model-based multiple instance learning. BMC Artif. Intell. 1 (1), 3.

Miron, R., Moisii, C., Dinu, S., Breaban, M.E., 2021. Evaluating volumetric and slice-based approaches for COVID-19 detection in chest CTs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 529–536.

Müller, T.T., Starck, S., Dima, A., Wunderlich, S., Bintsi, K.-M., Zaripova, K., Braren, R., Rueckert, D., Kazi, A., Kaissis, G., 2024. A survey on graph construction for geometric deep learning in medicine: Methods and recommendations. Trans. Mach. Learn. Res.

Müller-Franzes, G., Khader, F., Siepmann, R., Han, T., Kather, J.N., Nebelung, S., Truhn, D., 2025. Medical slice transformer for improved diagnosis and explainability on 3D medical images with DINOv2. Sci. Rep. 15 (1), 23979.

Naser, M.A., Deen, M.J., 2020. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Comput. Biol. Med. 121, 103758.

Navarro, F., Dapper, H., Asadpour, R., Knebel, C., Spraker, M.B., Schwarze, V., Schaub, S.K., Mayr, N.A., Specht, K., Woodruff, H.C., et al., 2021. Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. Cancers 13 (12), 2866.

Nguyen, N.T., Tran, D.Q., Nguyen, N.T., Nguyen, H.Q., 2020. A CNN-LSTM architecture for detection of intracranial hemorrhage on CT scans. MedRxiv 2020–04.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., et al., 2024. Foundation model for cancer imaging biomarkers. Nat. Mach. Intell. 6 (3), 354–367.

Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A., 2025. Foundation models in radiology: What, how, why, and why not. Radiology 314 (2), e240597.

Piazza, T.D., Lazarus, C., Nempont, O., Boussel, L., 2025. Structured spectral graph representation learning for multi-label abnormality analysis from 3D CT scans. arXiv:2510.10779.

Prakash, S.S., Visakha, K., 2020. Breast cancer malignancy prediction using deep learning neural networks. In: 2020 Second International Conference on Inventive Research in Computing Applications. ICIRCA, IEEE, pp. 88–92.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 8748–8763.

Rana, M., Bhushan, M., 2023. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools Appl. 82 (17), 26731–26769.

Rasuli, F., 2025. Deep learning in radiology: Challenges and opportunities for improved diagnosis. In: 2025 International Conference on Frontier Technologies and Solutions. ICFTS, IEEE, pp. 1–10.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Saha, A., Harowicz, M., Grimm, L., Weng, J., Cain, E., Kim, C., Ghate, S., Walsh, R., Mazurowski, M., 2021. Dynamic Contrast-Enhanced Magnetic Resonance Images of Breast Cancer Patients with Tumor Locations. The Cancer Imaging Archive, Bethesda, MD, USA.

Saint-Esteven, A.L.G., Bogowicz, M., Konukoglu, E., Riesterer, O., Balermpas, P., Guckenberger, M., Tanadini-Lang, S., van Timmeren, J.E., 2022. A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer. Comput. Biol. Med. 142, 105215.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE Trans. Neural Netw. 20 (1), 61–80.

Schulz, S., Woerl, A.-C., Jungmann, F., Glasner, C., Stenzel, P., Strobl, S., Fernandez, A., Wagner, D.-C., Haferkamp, A., Mildenberger, P., et al., 2021. Multimodal deep learning for prognosis prediction in renal cancer. Front. Oncol. 11, 788740.

Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al., 2025. DINOv3. arXiv preprint arXiv:2508.10104.

Song, X., Xu, X., Yan, P., 2024. DINO-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 608–617.

Spohn, S.K., Schmidt-Hegemann, N.-S., Ruf, J., Mix, M., Benndorf, M., Bamberg, F., Makowski, M.R., Kirste, S., Rühle, A., Nouvel, J., et al., 2023. Development of PSMA-PET-guided CT-based radiomic signature to predict biochemical recurrence after salvage radiotherapy. Eur. J. Nucl. Med. Mol. Imaging 50 (8), 2537–2547.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.

Takahashi, K., Fujioka, T., Oyama, J., Mori, M., Yamaga, E., Yashima, Y., Imokawa, T., Hayashi, A., Kujiraoka, Y., Tsuchiya, J., et al., 2022. Deep learning using multiple degrees of maximum-intensity projection for PET/CT image classification in breast cancer. Tomography 8 (1), 131–141.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. 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. pp. 20730–20740.

Thomson, J.J., 1904. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. Lond. Edinb. Dublin Philos. Mag. J. Sci. 7 (39), 237–265.

Thrall, J.H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., Brink, J., 2018. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J. Am. Coll. Radiol. 15 (3), 504–508.

Truong, T., Mohammadi, S., Lenga, M., 2021. How transferable are self-supervised features in medical image classification tasks? In: Machine Learning for Health. Proceedings of Machine Learning Research, pp. 54–74.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al., 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786.

Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107.

Veasey, B.P., Amini, A.A., 2024. Parameter-efficient fine-tuning of DINOv2 vision transformers for lung nodule classification. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

van Veldhuizen, V., Botha, V., Lu, C., Cesur, M.E., Lipman, K.G., de Jong, E.D., Horlings, H., Sanchez, C.I., Snoek, C.G., Wessels, L., et al., 2025. Foundation models in medical imaging–A review and outlook. arXiv preprint arXiv:2506.09095.

Venkataramanan, S., Pariza, V., Salehi, M., Knobel, L., Gidaris, S., Ramzi, E., Bursuc, A., Asano, Y.M., 2025. Franca: Nested matryoshka clustering for scalable visual representation learning. arXiv preprint arXiv:2507.14137.

Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Köhler, G., Maier-Hein, K.H., 2024. An OpenMind for 3D medical vision self-supervised learning. arXiv preprint arXiv:2412.17041.

Wang, R., Gao, L., Zhang, X., Han, J., 2024. SVFR: A novel slice-to-volume feature representation framework using deep neural networks and a clustering model for the diagnosis of Alzheimer's disease. Heliyon 10 (1).

Wee, L., Dekker, A., 2019. Data from Head-Neck-Radiomics-HN1. The Cancer Imaging Archive.

Wei, X., Yu, R., Sun, J., 2020. View-GCN: View-based graph convolutional network for 3D shape analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Wei, X., Yu, R., Sun, J., 2023. Learning view-based graph convolutional network for multi-view 3D shape analysis. IEEE Trans. Pattern Anal. Mach. Intell. 45 (6), 7525–7541. http://dx.doi.org/10.1109/TPAMI.2022.3221785.

Weltens, C., Menten, J., Feron, M., Bellon, E., Demaerel, P., Maes, F., Van den Bogaert, W., van der Schueren, E., 2001. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. Radiother. Oncol. 60 (1), 49–59.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2020. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32 (1), 4–24.

Wu, L., Zhuang, J., Chen, H., 2024. VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22873–22882.

Xia, K., Wang, J., 2023. Recent advances of transformers in medical image analysis: a comprehensive review. MedComm–Future Med. 2 (1), e38.

Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2018. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.

Yan, K., Wang, X., Lu, L., Summers, R.M., 2018. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging 5 (3), 036501.

Ye, J., Tang, H., 2025. Multimodal large language models for medicine: A comprehensive survey. arXiv preprint arXiv:2504.21051.

Yin, Y., Yakar, D., Slangen, J.J., Hoogwater, F.J., Kwee, T.C., de Haas, R.J., 2023. The value of deep learning in gallbladder lesion characterization. Diagnostics 13 (4), 704.

Zhang, H., Zhang, J., Zhang, Q., Kim, J., Zhang, S., Gauthier, S.A., Spincemaille, P., Nguyen, T.D., Sabuncu, M., Wang, Y., 2019. RSANet: recurrent slice-wise attention network for multiple sclerosis lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 411–419.

Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models Genesis: Generic autodidactic models for 3D medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 384–393.

Zhu, W., Xie, L., Han, J., Guo, X., 2020. The application of deep learning in cancer prognosis prediction. Cancers 12 (3), 603.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J., 2023. Segment everything everywhere all at once. Adv. Neural Inf. Process. Syst. 36, 19769–19782.