

Structured Spectral Graph Learning for Anomaly Classification in 3D Chest CT Scans

Theo Di Piazza^{1,2}, Carole Lazarus³, Olivier Nempont³, and Loic Boussel^{1,2}

University of Lyon, INSA Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294
Hospices Civils de Lyon, Lyon, France
Philips Clinical Informatics, Innovation Paris, France
`theo.dipiazza@creatis.insa-lyon.fr`

Abstract. With the increasing number of CT scan examinations, there is a need for automated methods such as organ segmentation, anomaly detection and report generation to assist radiologists in managing their increasing workload. Multi-label classification of 3D CT scans remains a critical yet challenging task due to the complex spatial relationships within volumetric data and the variety of observed anomalies. Existing approaches based on 3D convolutional networks have limited abilities to model long-range dependencies while Vision Transformers suffer from high computational costs and often require extensive pre-training on large-scale datasets from the same domain to achieve competitive performance. In this work, we propose an alternative by introducing a new graph-based approach that models CT scans as structured graphs, leveraging axial slice triplets nodes processed through spectral domain convolution to enhance multi-label anomaly classification performance. Our method exhibits strong cross-dataset generalization, and competitive performance while achieving robustness to z-axis translation. An ablation study evaluates the contribution of each proposed component.

Keywords: 3D Medical Imaging · Chest Computed Tomography · Graph Neural Network · Spectral domain · Multi-label Anomaly Classification.

1 Introduction

Computed Tomography (CT) is a fundamental modality in modern medical imaging, providing radiologists with detailed cross-sectional views of the human body to detect and characterize abnormalities. However, the increasing volume of CT scans has led to an important demand for automated deep learning-based methods to assist radiologists with their growing workload [6]. Deep learning has already demonstrated success in various CT-related tasks [1], including anomaly detection [16], organ segmentation [21], report generation [17], and synthetic volume reconstruction [16] for patient-specific modeling. Among these tasks, multi-label classification of anomalies in 3D CT volumes remains challenging due to the computational complexity of processing volumetric data and the diverse range of pathological patterns. Early deep learning approaches leverage 3D Convolutional Neural Networks (CNNs), effectively capturing local spatial features but

suffering from limited capabilities to model long-ranges dependencies [25]. More recently, Vision Transformers (ViTs) [12], initially designed for natural language processing [31], have been adapted to both 2D [15] and 3D [18] medical imaging. By enabling long-range spatial interactions through self-attention, ViTs have shown promise in various medical imaging tasks [3] through its capabilities to capture global information. However, they remain computationally expensive, requiring large-scale pretraining to generalize effectively [18]. Our work introduces CT-Graph, a new 2.5D GNN-based framework that models 3D chest CT scans as structured graphs, where each node represents a triplet of adjacent axial slices and edges are weighted by inter-slice spacing. This design enables efficient integration of local and global context while preserving spatial structure. Our approach offers the following key advantages:

- CT-Graph demonstrates strong cross-dataset generalization, maintaining consistent performance when trained on a public Turkish 3D chest CT dataset and evaluated on a separate dataset from the United States.
- Our edge weighting strategy based on z-axis distance spacing incorporates spatial awareness with no additional learnable parameters. Ablation studies confirm the effectiveness of GNN modules and graph connectivity patterns.
- By leveraging spectral domain convolution, CT-Graph improves anomaly classification performance and achieves robustness to z-axis translation.

2 Related Work

2.1 3D Visual Encoder

Feature aggregation in 3D medical imaging is crucial for balancing local and long-range dependencies while maintaining global spatial awareness. Early deep learning architectures primarily relied on 3D CNNs [1], which effectively capture local spatial dependencies. These models have been widely applied to tasks such as anomaly detection [20] and segmentation [28]. However, their intrinsic locality limits their ability to model long-range dependencies, which can be crucial for capturing global anatomical structures [25]. The self-attention mechanism [32], initially introduced for natural language processing tasks was rapidly adapted to the visual domain with ViTs [12]. The extension of ViTs [18] and Swin Transformers [34] to 3D tasks has shown promise in applications such as dense image captioning [9] and video processing [24]. In the context of CT imaging, GenerateCT leverages CT-ViT, inspired by ViViT [2], to integrate spatial and causal attention but requires extensive pretraining, limiting its practical applicability [18]. To mitigate computational challenges in 3D volume processing, CT-Net [13] proposes to group triplets of adjacent slices to replicate the three-channel structure of RGB images, extracting features using a pretrained 2D ResNet [19]. While CT-Net subsequently passes these representations through a lightweight 3D CNN for dimensionality reduction, CT-Scroll [11] leverages an alternating global-local attention module to enable feature interactions, effectively reducing the number of parameters while improving classification performance.

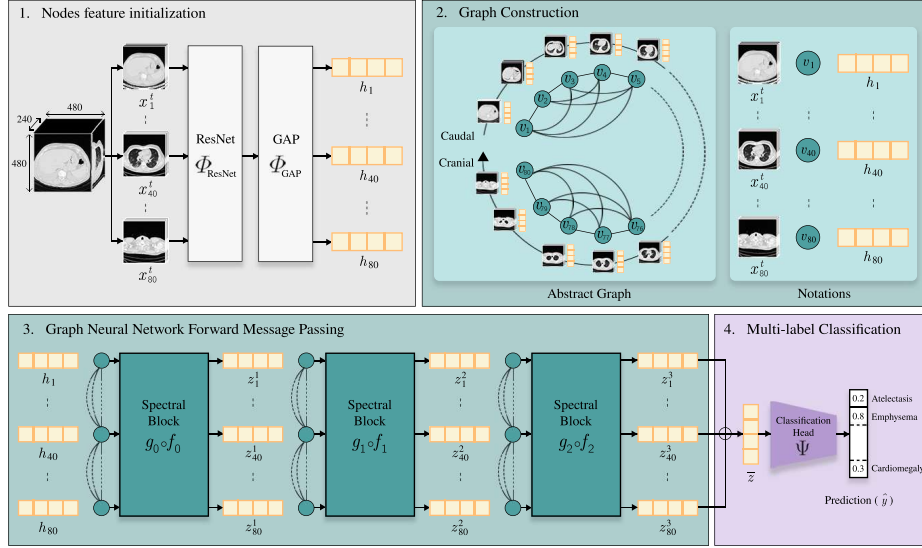


Fig. 1: CT-Graph introduces a structured graph-based architecture, where triplet axial slice features define nodes. Node interactions are modeled through spectral-domain convolutions, enabling contextual aggregation prior before classification.

2.2 Graph Neural Networks

In various application domains such as biology [29] or transportation [26], graphs are a common representation of data found in nature [33]. A graph, denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of edges \mathcal{E} which model the connections between a set of nodes \mathcal{V} . In deep learning, GNNs have become the main approach for tasks involving graph-structured data [4], where each node is associated with a vector representation, which is iteratively updated through neighborhood aggregation during the forward message passing process. Representative models mainly include Convolutional GNNs, which aggregate neighboring node features through graph-based convolutions [10] or Attentional GNNs, which leverage attention mechanisms to weigh the importance of neighbors' contributions [7]. In medical imaging, GNNs have been used in tasks such as medical knowledge integration in radiology report generation [23] and Whole Slide Image analysis [14]. Specifically to 3D medical imaging, recent approaches have explored multi-view modeling, where each node encodes a triplet of orthogonal slices with axial, coronal, and sagittal views to capture complementary anatomical information [22].

3 Method

As shown in Figure 1, CT-Graph models the 3D CT scan as a graph of *triplet axial CT slices* connected by their *physical z -axis distance*. Each node corresponds to a triplet of axial slices connected by neighborhood nodes with an edge

weighted by their physical distance. Node features interact through a GNN module before being summed and given to a classification head.

Triplet Slices Feature Extraction. Following a strategy similar to CT-Net [13], we partition the input volume $x \in \mathbb{R}^{240 \times 480 \times 480}$ into non-overlapping triplets of slices, noted $\{x_i^t\}_{i=1}^{80}$ forming a tensor of dimension $80 \times 3 \times 480 \times 480$. Each triplet is processed by a ResNet [19] Φ_{ResNet} pretrained on ImageNet [30] to extract a corresponding feature map. The feature maps are then processed independently, with each one being passed through a Global Average Pooling (GAP) layer [11] Φ_{GAP} to obtain a compact vector representation for each triplet, noted $h_i \in \mathbb{R}^{512}$ ($i \in \{1, \dots, 80\}$), such that:

$$h_i = (\Phi_{\text{GAP}} \circ \Phi_{\text{ResNet}})(x_i^t), \quad \forall i \in \{1, \dots, 80\}. \quad (1)$$

Graph Construction. We define the volumetric representation as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, H, A)$, where:

- $\mathcal{V} = \{v_i\}_{i=1}^N$ is the set of nodes, where each node v_i represents a triplet of consecutive slices. Hence, the number of nodes is $N = 80$.
- $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, where an edge $(v_i, v_j) \in \mathcal{E}$ is weighted based on a function of inter-triplet distance and z-axis spacing. An undirected edge $(v_i, v_j) \in \mathcal{E}$ is established if and only if the corresponding triplet slices are separated by at most $q \in \mathbb{N}^+$ other triplet slices in the sequence, such that:

$$\mathcal{E} = \{(v_i, v_j) \mid |i - j| \leq q\}. \quad (2)$$

- $H = \{h_1, \dots, h_N\} \in \mathbb{R}^{N \times d}$ is the node feature matrix, where $\mathbf{h}_i \in \mathbb{R}^d$ denotes the feature embedding of node v_i ($\forall i \in \{1, \dots, N\}$). We set $d = 512$.
- $A \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where $A_{ij} = w_{i,j} \in \mathbb{R}^+$ encodes the connectivity and spatial relationship between triplets, $w_{i,j}$ being the edge weight such that:

$$A_{ij} = \begin{cases} w_{ij}, & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Graph Neural Network module. A key challenge in this formulation is the variability in anatomical positioning across patients due to differences in scan length and body proportions. Traditional spatial graph convolutions, such as GraphConv [27], aggregate information from fixed local neighborhoods, which can be suboptimal in this context as anatomical structures do not consistently align across scans. Instead, we leverage Chebyshev convolutions [10] to define graph convolutions in the spectral domain, each followed by a feedforward neural network. Unlike spatial approaches, which struggle with non-uniform neighborhood structures [8], ChebConv utilizes polynomial approximations of the graph Laplacian [5] to capture hierarchical feature representations while preserving spatial localization. This allows the model to adapt to variations in caudal-cranial slice positioning and effectively learn long-range anatomical relationships, making it more robust to inter-patient variability. Our GNN module, denoted as Φ_{GNN} , consists of 3 Chebyshev Convolutional Layers [10], each

noted f_n ($n \in \{0, 1, 2\}$) and followed by a feedforward neural network consisting of a linear layer followed by a ReLU, denoted as g_n , matching the depth of CT-Scroll [11] for fair comparison. For each layer, the scaled and normalized Laplacian \hat{L} is defined as:

$$\hat{L} = \frac{2}{\lambda_{\max}}(D - A) - I, \quad (4)$$

where λ_{\max} is the largest eigenvalue of the graph Laplacian $L = D - A$. The degree matrix D is a diagonal matrix where $D_{i,i} = \sum_{j=1}^N w_{i,j}$. $w_{i,j}$ denotes the edge weight from source node i to target node j , defined such that:

$$w_{i,j} = 1 + \frac{1}{1 + \text{dist}(i, j)} = 1 + \frac{1}{1 + 3 \times |i - j| \times s_z}, \quad (5)$$

where s_z is the spacing along the z-axis in decimetre. The convolution operation is parameterized using Chebyshev polynomials $T_j(\hat{L}) \in \mathbb{R}^{N \times N}$, resulting in a recurrence relation for the transformation of the node feature matrix. Let $Z^0 = H$ be the initial node feature matrix, $\theta_k \in \mathbb{R}^{d \times d}$ be the learnable parameters, and K be the Chebyshev filter size fixed to 3 for all experiments, to align with common practice [10]. The recurrence relation is given by:

$$Z^{n+1} = (g_n \circ f_n)(Z^n) = g_n\left(\sum_{k=0}^{K-1} T_k(\hat{L})Z^n\theta_k\right), \quad \forall n \in \{0, 1, 2\}. \quad (6)$$

The GNN module Φ_{GNN} produces the final output vector representation, which we denote as $Z = Z^3 \in \mathbb{R}^{N \times d}$ and which is defined as:

$$Z = \{z_1^3, \dots, z_N^3\} = \Phi_{\text{GNN}}(H). \quad (7)$$

Feature aggregation. The obtained vector representations are aggregated through summation to derive a vector representation, denoted as $\bar{z} \in \mathbb{R}^d$, which is subsequently passed to a classification head Ψ implemented as a lightweight multilayer perceptron. Ψ predicts the logit vector $\hat{y} \in \mathbb{R}^{18}$. The model is trained on a multi-label classification task using Binary Cross-Entropy as the loss function.

4 Experimental results

4.1 Dataset preparation

We train and evaluate our methods on the public CT-RATE dataset [16], which consists of non-contrast chest CT scans with 18 annotated anomalies extracted from radiology reports. The training set includes 17,799 unique patients, while the validation and test sets both contain 1,314 unique patients. Additionally, we extend our evaluation on the publicly available Rad-ChestCT dataset [13], comprising non-contrast chest CT scans from 1,344 unique patients, focusing on the 16 anomalies shared with CT-RATE [16]. Consistent with prior work [17, 11], volumes for both datasets are center-cropped or padded to a resolution of $240 \times 480 \times 480$, with a spacing of 0.75 mm on the x and y and 1.5 mm on the z axis. Hounsfield Unit values are clipped to the range $[-1000, 200]$, reflecting practical diagnostic limits [17].

Table 1: Quantitative evaluation on the CT-RATE and Rad-ChestCT test sets. Reported mean and standard deviation metrics were computed over 5 independent runs. **Best** results are in bold, second best are underlined.

Dataset	Method	F1	Recall	AUROC	Accuracy
CT-RATE	Random Pred.	27.78±0.51	50.42±1.05	49.88±0.62	49.89±0.31
	ViViT [2]	49.91±0.28	66.39±1.48	79.19±0.28	75.95±0.71
	Swin3D [24]	50.64±0.25	<u>67.96±0.58</u>	79.94±0.15	75.95±0.25
	CT-Net [13]	51.39±0.50	66.42±1.99	79.37±0.27	77.37±0.40
	CNN3D [1]	52.92±1.08	67.60±1.01	81.47±0.78	77.80±0.37
	CT-Scroll [11]	<u>53.97±0.21</u>	65.36±1.91	<u>81.80±0.22</u>	79.49±0.45
	CT-Graph	54.59±0.17	68.77±0.92	82.44±0.14	<u>78.66±0.36</u>
Rad-ChestCT	Random Pred.	35.91±0.41	51.51±0.75	49.68±0.55	50.40±0.32
	ViViT [2]	48.59±0.97	69.27±1.64	67.83±0.38	60.22±1.15
	Swin3D [24]	47.98±0.41	66.76±0.63	67.29±0.23	60.67±0.60
	CT-Net [13]	47.53±0.93	68.45±1.18	67.71±0.83	60.05±1.93
	CNN3D [1]	<u>49.28±0.93</u>	70.47±0.73	71.13±0.62	61.08±0.60
	CT-Scroll [11]	48.55±0.54	66.63±1.49	<u>71.21±0.37</u>	63.02±0.93
	CT-Graph	49.52±0.76	<u>69.30±1.48</u>	72.18±0.29	<u>62.60±0.52</u>

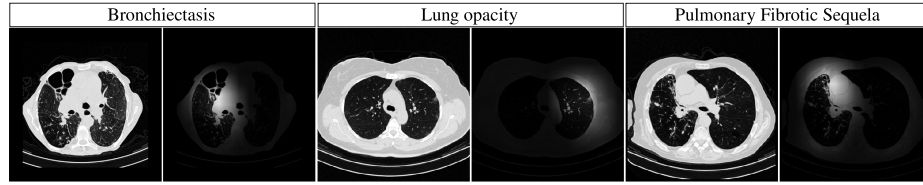


Fig. 2: GradCAM activation maps extracted from the 2D ResNet module.

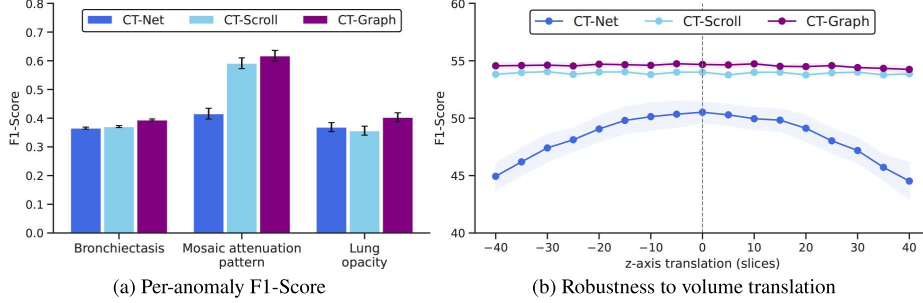
4.2 Implementation Details

CT-Graph and baseline methods are trained with a batch size of 4 using the AdamW optimizer with $(\beta_1, \beta_2) = (0.9, 0.99)$ and a weight decay of 0.01. The learning schedule follows a cosine decay with a warm-up phase of 20,000 steps, a maximum learning rate of 0.0001, and training runs for 200,000 iterations.

4.3 Quantitative results

For each method and each label, we select the threshold that maximizes F1-Score on the validation set and report all metrics on the test set. We compare our method against a 3D CNN, ViViT [2], a video-adapted Vision Transformer which also forms the architectural basis for CT-ViT, and Swin3D [34], an extension of Swin Transformer for volumetric data. We also include CT-Net [13] and CT-Scroll [11], two 2.5D approaches that employ CNN-based feature extractors.

Fig. 3: (a) Per-anomaly F1-Score comparison for the 3 anomalies with highest improvement over baselines. (b) Model robustness to z-axis volume shift. F1 are reported for volumes translated along the z-axis with minimum-value padding.



CT-Net relies on convolutional layers for feature aggregation and dimensionality reduction, whereas CT-Scroll leverages an alternating attention mechanism to capture cross-slice dependencies. ResNet-based models used ImageNet pre-trained weights; others were initialized via weight inflation [35] for comparability. Table 1 shows that CT-Graph consistently outperforms all baselines across AUROC, F1-Score and Recall. On the CT-RATE test set, our method achieves an F1-Score of 54.59, representing a $+ \Delta 1.15\%$ improvement over CT-Scroll [11] and $+ \Delta 5.93\%$ over CT-Net [13]. For the F1-Score, a paired t-test comparing the performance of CT-Graph against each baseline consistently yields a p-value < 0.01 , demonstrating statistical significance. As shown in Fig. 3.a, CT-Graph yields the largest improvements on diffuse anomalies such as bronchiectasis, mosaic attenuation, and lung opacity. Referring to Fig. 3.b, both attention and spectral convolution demonstrate robustness to z-axis translations, whereas standard convolution is sensitive to such shifts. To evaluate this property, we simulate patient body shifts by applying controlled translations along the z-axis with appropriate padding. Fig. 2 illustrates CT-Graph’s ability to classify anomalies from relevant regions.

4.4 Ablation study

Comparison of representative GNNs. Table 2 highlights the performance gains achieved by incorporating Chebyshev Convolutions [10] in our GNN module. Compared to a direct neighborhood aggregation approach [27], ChebConv improves AUROC by $+ \Delta 0.42\%$ and F1-Score by $+ \Delta 1.25\%$, suggesting that spectral-domain convolutions may enhance feature aggregation while demonstrating robustness to variations in cranial-caudal slice positioning (Fig. 3). Inference time takes approximately 70 milliseconds for all GNN variants.

Graph construction. Table 2 and Table 3 demonstrate that neighborhood graph construction consistently improves AUROC and F1-score across all GNN

Table 2: Comparison of graph connectivity schemes and GNN modules, evaluated on the CT-RATE test set. The neighborhood size is fixed to 16 for these runs.

Connectivity	Module	F1	AUROC	Accuracy
<i>Fully connected</i>	GATv2Conv [7]	53.72 \pm 0.34	81.56 \pm 0.03	78.04 \pm 0.31
	GraphConv [27]	53.73 \pm 0.36	81.99 \pm 0.40	78.15 \pm 0.31
	ChebConv [10]	54.40 \pm 0.15	82.34 \pm 0.12	79.01 \pm 0.55
<i>Neighbourhood</i>	GATv2Conv [7]	54.06 \pm 0.19	82.22 \pm 0.05	78.59 \pm 0.25
	GraphConv [27]	54.16 \pm 0.24	82.33 \pm 0.18	78.68 \pm 0.52
	ChebConv [10]	54.41\pm0.12	82.47\pm0.26	79.12\pm0.53

Neighbourhood size	F1 Score	Recall	Precision	AUROC	Accuracy
4	53.76 \pm 0.24	66.02 \pm 0.92	47.84\pm0.22	82.22 \pm 0.05	78.97\pm0.58
16	54.14\pm0.24	67.99 \pm 0.75	47.34 \pm 0.30	82.33\pm0.18	78.68 \pm 0.52
80 (Fully connected)	53.73 \pm 0.36	69.34\pm0.91	45.80 \pm 0.59	81.99 \pm 0.40	78.15 \pm 0.31

Table 3: Impact of the neighbourhood size, using GraphConv. Neighborhood size, noted as q , refers to the number of nodes each node is connected to.

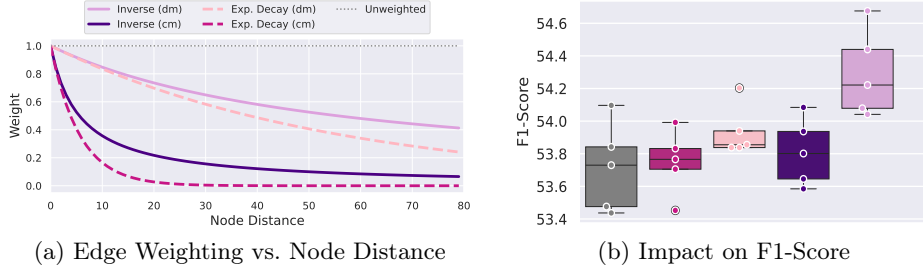


Fig. 4: Impact of the edge weighting functions, on the CT-RATE test set. We use a GraphConv module and a fully connected graph for all experiments.

variants, with particularly pronounced gains for GATv2Conv and GraphConv, with ChebConv showing marginal gains.

Impact of the weight function. Among the evaluated edge weighting functions, the inverse function (see Eq. 5) with z-axis spacing measured in decimeters (dm) yields the best classification performance, as illustrated in Figure 4.

5 Discussion and Conclusion

In this work, we introduced CT-Graph, a new graph-based approach for multi-label anomaly classification from 3D Chest CT volumes. Each scan is represented as a structured graph, where nodes correspond to triplets of adjacent axial slices. To enable effective feature aggregation across this graph, we leverage a spectral approach based on Chebyshev convolution, which captures both short-range and long-range dependencies along the axial direction. Additionally, we show that incorporating spatially-aware graph structures, through both weighted edges and constrained neighborhood connectivity, enhances performance across multiple Graph Neural Network variants. CT-Graph demonstrates robustness to variations in patient body positioning along the z-axis and provides a flexible framework for modeling volumetric data. Future work may include anatomical segmentation-driven graph construction, transformer-based hybridization with patch representations, multi-view modeling extension, and exploration of architectural factors such as convolution depth and Chebyshev filter size.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anaya-Isaza, A., Mera-Jiménez, L., Zequera-Diaz, M.: An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked* **26**, 100723 (Jan 2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: ViViT: A Video Vision Transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6816–6826. IEEE, Montreal, QC, Canada (Oct 2021)
3. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision Transformers: A comprehensive review. *Medical Image Analysis* **91**, 103000 (Jan 2024)
4. Bechler-Speicher, M., Globerson, A., Gilad-Bachrach, R.: The Intelligible and Effective Graph Neural Additive Networks (Dec 2024), arXiv:2406.01317 [cs]
5. Belkin, M., Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: *Advances in Neural Information Processing Systems*. vol. 14. MIT Press (2001)
6. Broder, J., Warshauer, D.M.: Increasing utilization of computed tomography in the adult emergency department, 2000-2005. *Emergency Radiology* **13**(1), 25–30 (Oct 2006)
7. Brody, S., Alon, U., Yahav, E.: How Attentive are Graph Attention Networks? (Jan 2022), arXiv:2105.14491 [cs]
8. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral Networks and Locally Connected Networks on Graphs (May 2014), arXiv:1312.6203 [cs]
9. Chen, D.Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding (Dec 2022), arXiv:2212.00836 [cs]
10. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering (Feb 2017), arXiv:1606.09375 [cs]

11. Di Piazza, T., Lazarus, C., Nempont, O., Boussel, L.: Imitating Radiological Scrolling: A Global-Local Attention Model for 3D Chest CT Volumes Multi-Label Anomaly Classification (Jan 2025)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021), arXiv:2010.11929 [cs]
13. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis* **67**, 101857 (Jan 2021)
14. Guo, Z., Zhao, W., Wang, S., Yu, L.: HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis (Sep 2023), arXiv:2309.07400 [cs]
15. Halder, A., Gharami, S., Sadhu, P., Singh, P.K., Woźniak, M., Ijaz, M.F.: Implementing vision transformer for classifying 2D biomedical images. *Scientific Reports* **14**(1), 12567 (May 2024), publisher: Nature Publishing Group
16. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Ozdemir, M.K., Menze, B.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities (Mar 2024), arXiv:2403.17834 [cs]
17. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging (Mar 2024), arXiv:2403.06801 [cs, eess]
18. Hamamci, I.E., Er, S., Simsar, E., Sekuboyina, A., Prabhakar, C., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Doğan, I., Dasdelen, M.F., Reynaud, H., Pati, S., Bluethgen, C., Ozdemir, M.K., Menze, B.: GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes (Nov 2023), arXiv:2305.16037 [cs]
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015), arXiv:1512.03385 [cs]
20. Ibrahim, D.M., Elshennawy, N.M., Sarhan, A.M.: Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Computers in Biology and Medicine* **132**, 104348 (May 2021)
21. Ilesanmi, A.E., Ilesanmi, T.O., Ajayi, B.O.: Reviewing 3D convolutional neural network approaches for medical image segmentation. *Heliyon* **10**(6), e27398 (Mar 2024)
22. Kiechle, J., Lang, D.M., Fischer, S.M., Felsner, L., Peeken, J.C., Schnabel, J.A.: Graph Neural Networks: A suitable Alternative to MLPs in Latent 3D Medical Image Classification? (Jul 2024), arXiv:2407.17219 [cs]
23. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation (Jun 2021), arXiv:2106.06963 [cs]
24. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer (Jun 2021), arXiv:2106.13230 [cs]
25. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation (Jan 2024), arXiv:2401.04722 [eess]
26. Makarov, N., Narayanan, S., Antoniou, C.: Graph neural network surrogate for strategic transport planning (Aug 2024), arXiv:2408.07726 [cs]
27. Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M.: Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks (Nov 2021), arXiv:1810.02244 [cs]

28. Rayed, M.E., Islam, S.M.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked* **47**, 101504 (Jan 2024)
29. Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., Friederich, P.: Graph neural networks for materials science and chemistry. *Communications Materials* **3**(1), 1–18 (Nov 2022), publisher: Nature Publishing Group
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge (Jan 2015), arXiv:1409.0575 [cs]
31. Tucudean, G., Bucos, M., Dragulescu, B., Căleanu, C.D.: Natural language processing with transformers: a review. *PeerJ. Computer Science* **10**, e2222 (2024)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (Aug 2023), arXiv:1706.03762 [cs]
33. Veličković, P.: Everything is Connected: Graph Neural Networks. *Current Opinion in Structural Biology* **79**, 102538 (Apr 2023), arXiv:2301.08210 [cs]
34. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding (Aug 2023), arXiv:2304.06906 [cs]
35. Zhang, Y., Huang, S.C., Zhou, Z., Lungren, M.P., Yeung, S.: Adapting Pre-trained Vision Transformers from 2D to 3D through Weight Inflation Improves Medical Image Segmentation (Feb 2023), arXiv:2302.04303 [cs]