MUGAT-KD: A MULTIMODAL-TO-UNIMODAL KNOWLEDGE DISTILLATION FRAMEWORK FOR PULMONARY NODULE MALIGNANCY CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pulmonary nodule malignancy classification is a vital component of early lung cancer screening. Multimodal models that integrate CT images with structured nodule attributes can derive more comprehensive nodule representations; however, the inference process relies on additional attribute annotations, limiting their practical applicability. Existing distillation methods primarily align model outputs or global representations, making it difficult for a unimodal model to preserve the hierarchical influence of multimodal knowledge on feature representations, attention to critical slices, and local regions of interest. To address this limitation, we propose MUGAT-KD, a multimodal-to-unimodal knowledge distillation framework. The framework employs a graph attention network to model the spatial relationships among CT slices and the associations between slices and nodule attributes. A frozen multimodal teacher then guides the student model at three levels: global features, slice-level attention, and saliency-aware patch masking. We conduct within-dataset classification experiments on LIDC-IDRI and LIDP datasets and evaluate cross-dataset generalization by training on LIDC-IDRI and performing external validation on LIDP. The experimental results demonstrate that MUGAT-KD effectively transfers attribute knowledge from the multimodal teacher to the unimodal student, improving pulmonary nodule malignancy classification without requiring additional nodule attribute annotations and enhancing adaptability to differences in data standards. Ablation studies verify the complementary effects of the distillation modules. Attention visualizations further show that the student model learns the multimodal teacher's attention patterns over critical CT slices and local regions. These results establish MUGAT-KD as an effective solution for pulmonary nodule malignancy classification when nodule attributes are unavailable.

1 INTRODUCTION

Lung cancer is one of the deadliest diseases worldwide (Sung et al., 2021; Bray et al., 2024), and pulmonary nodule malignancy classification is a critical task in early lung cancer screening. In recent years, methods based on 2D or 3D CT images, as well as multimodal methods combining CT images with structured attributes, have substantially improved classification performance. Compared with unimodal models that rely solely on CT images, multimodal models integrating CT images and structured attributes can incorporate clinical prior knowledge. However, these models generally require additional structured attributes during inference, thereby increasing their deployment cost.

In the LIDC-IDRI dataset, which provides both CT images and nodule attributes, each pulmonary nodule is annotated with subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and diameter. Calcification and internal structure are assigned categorical scores, diameter is measured in millimeters, and the remaining attributes are rated on 1–5 ordinal scales (Armato III et al., 2011). As illustrated in Fig. 1, different attributes have corresponding visual manifestations in CT images. This observation suggests that although structured attributes are explicitly annotated by radiologists, some of their corresponding visual representations remain present in CT images.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

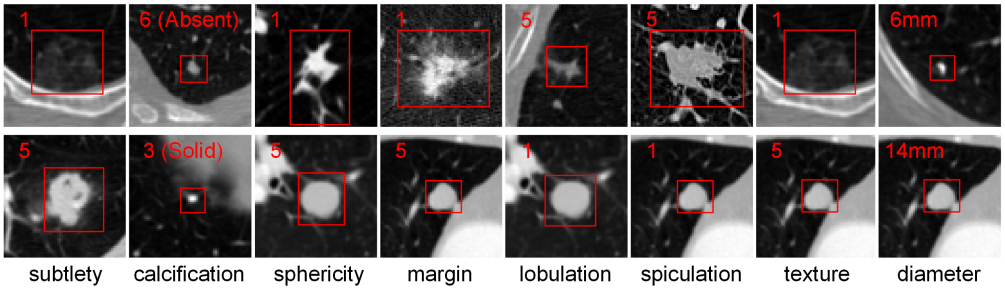

Figure 1: Representative CT appearances of pulmonary nodules with different attribute annotations in LIDC-IDRI. Morphological attributes are rated on 1–5 ordinal scales, calcification is encoded on a 1–6 scale, and diameter is measured continuously in millimeters.

Given the substantial differences in how nodule attributes appear in CT images, we make the following hypothesis: during training, an image encoder can learn attribute-related visual representations and transform this information into features relevant to pulmonary nodule classification, thereby improving the performance of a model that relies solely on CT images during inference. Accordingly, the central research question is how to fully exploit multimodal information available during training to model the spatial relationships among CT slices and their associations with nodule attributes, and how to transfer multimodal knowledge to a unimodal model.

Knowledge distillation (Hinton et al., 2015) provides an effective approach to this problem. However, the multimodal teacher observes both CT images and structured attributes, whereas the student receives only CT images. Therefore, conventional output- or feature-level alignment may cause incomplete knowledge transfer. Moreover, nodule attributes affect diagnosis at multiple levels, including global representations, critical slice selection, and local regions.

To address these challenges, we propose MUGAT-KD, which consists of teacher and student models. The teacher model learns multimodal knowledge by using a graph attention network to model the spatial relationships among CT slices and the associations between slices and structured attributes, whereas the student model constructs a graph using only CT slices. During student training, the frozen multimodal teacher provides supervision at three levels: feature representations, slice-level attention relationships, and local regions within slices.

We evaluate MUGAT-KD on two datasets, LIDC-IDRI (Armato III et al., 2011) and LIDP (Shao et al., 2022). In addition to within-dataset five-fold cross-validation, we train the model on LIDC-IDRI and perform external validation on LIDP. Because the two datasets differ in attribute types, label definitions, and data distributions, this setting more closely reflects missing attributes and inconsistent cross-center standards encountered in clinical practice. The experimental results show that MUGAT-KD outperforms comparison methods that rely solely on images during inference on both datasets and exhibits favorable generalization in cross-dataset external validation.

Our contributions are summarized as follows:

- We propose MUGAT-KD, which formulates multimodal-to-unimodal knowledge transfer as a hierarchical knowledge distillation problem from a multimodal heterogeneous graph teacher to a unimodal sliced graph student.
- We design a three-level distillation mechanism comprising feature-level, slice-level attention, and region-level saliency-aware patch mask distillation, which transfer the multimodal teacher's knowledge from multiple levels.
- We conduct external validation across two datasets with different nodule attribute annotations and demonstrate the effectiveness of the model under a realistic missing-attribute setting.

2 Related Work

2.1 Pulmonary Nodule Classification Using CT Images and Structured Attributes

CT-based pulmonary nodule malignancy classification methods mainly employ 2D slices, multiple views, or 3D volumetric data. Early methods typically selected the central nodule slice or multiple 2D projection views to extract visual features that capture nodule morphology and internal heterogeneity.

MT-CNN (Zhai et al., 2020) selected nine 2D projection views from different directions, including axial, coronal, sagittal, and multiple oblique views, to capture internal nodule relationships (Setio et al., 2016; Kang et al., 2017; Xie et al., 2018). Compared with single-slice methods, 3D convolutional neural networks can further exploit spatial information across different slices or orientations. MVCNN (Yang et al., 2023) extracts 3D information from multiple directions and introduces the Squeeze-and-Excitation (SE) attention mechanism to weight features from different views. However, these methods have lower performance due to the lack of attribute information guidance.

Existing multimodal methods for pulmonary nodule classification generally employ decision-level fusion, feature concatenation, or graph-based modeling to fuse cross-modal features. Lin et al. (2024) extract deep features from CT images and encode them for nodule prediction, while structured attributes were processed using logistic regression for binary prediction. The outputs are combined during prediction, however, this approach provides insufficient interaction between modalities during feature learning. Tang et al. (2022) concatenate the image encoding and attribute representations. Feature concatenation directly combines image encodings with attribute representations and enables joint optimization, but it cannot explicitly characterize complex relationships between heterogeneous modalities. MIE-GAT (Li et al., 2025) uses a graph neural network to construct slice and attribute subgraphs. During training, the model integrates spatial CT information with attribute information through inter-slice and inter-attribute connections. This process employs GNN message passing (Kipf & Welling, 2017) and attention mechanisms to enable cross-modal information exchange, allowing the image modality to acquire complementary information from nodule attributes instead of making independent decisions as in feature concatenation or decision-level fusion. Although these graph-based methods effectively exploit structured attributes, they typically require both CT images and attribute data during inference, limiting their applicability when attributes are unavailable.

2.2 Multimodal-to-Unimodal Knowledge Transfer

To eliminate dependence on additional modalities during inference, some studies explored transferring multimodal knowledge to single-modal models. Inspired by vision-language pretraining models (Radford et al., 2021; Wang et al., 2022; Jia et al., 2021; Huang et al., 2021), some studies have transferred medical textual knowledge to image encoders. CLIP-Lung (Lei et al., 2023) maps pulmonary nodule CT images, malignancy labels, and nodule attributes text into a shared semantic space. Through contrastive learning, it aligns image and text feature representations to enhance the ability of image features to express the semantics of pulmonary nodule malignancy using attribute knowledge. Nodule-CLIP (Sun et al., 2024) further improves the image encoding process by using nodule segmentation to achieve finer-grained image-text semantic alignment.

At the early stage of training, class text, attribute text, and image representations have not yet formed a stable and consistent semantic space. Directly imposing strong image-attribute alignment constraints may therefore produce supervision inconsistent with the primary classification objective. Furthermore, these methods primarily perform image-text alignment in the global feature space and do not explicitly model the hierarchical influence of structured attributes on different CT slices and local regions within individual slices.

In contrast, MUGAT-KD employs a frozen multimodal graph teacher to guide a CT-only graph student and transfers attribute knowledge at three levels: global feature representation, slice-selection attention, and saliency-aware local regions. The student's image encoder is consistently supervised by the teacher parameters that optimally capture classification-relevant attribute knowledge.

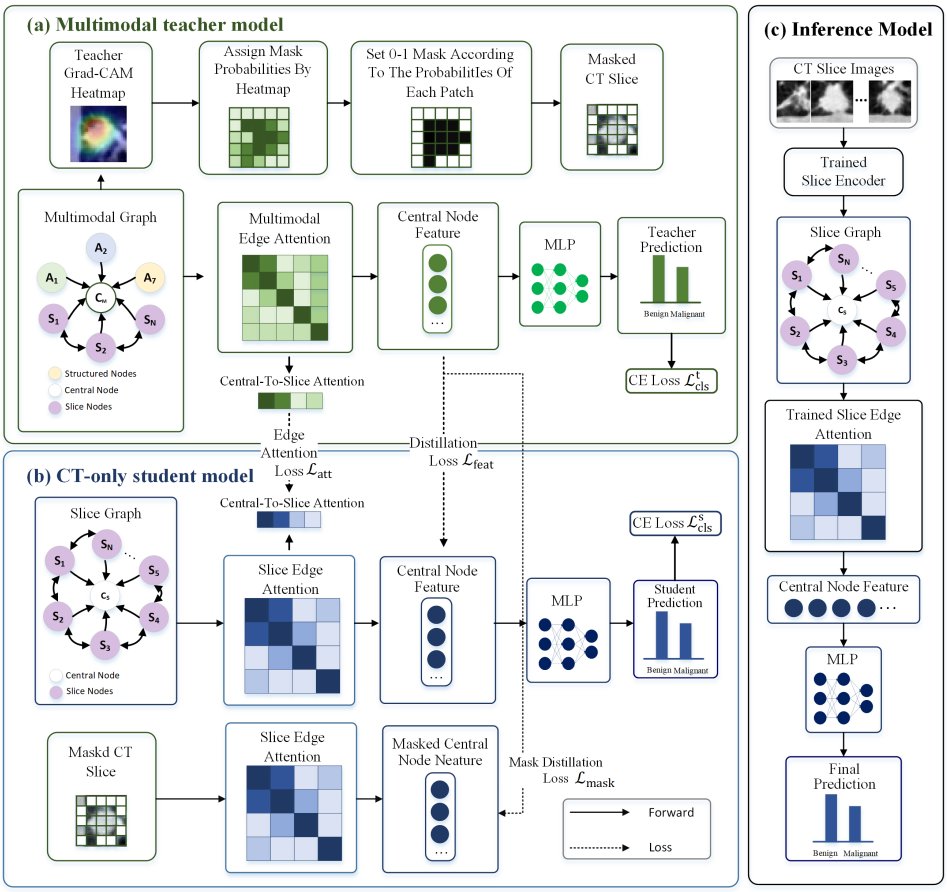

Figure 2: Overall framework of the proposed MUGAT-KD.

3 METHOD

3.1 OVERALL FRAMEWORK

As illustrated in Fig. 2, MUGAT-KD is a multimodal-to-unimodal knowledge distillation framework consisting of multimodal teacher pretraining, student distillation training, and image-only inference. In the first stage, a multimodal graph teacher is trained using both CT images and structured nodule attributes. In the second stage, the best-performing teacher is frozen and guides a CT-only graph student through feature-level, slice-level attention, and saliency-aware patch mask distillation. During inference, only the student model is retained, enabling pulmonary nodule malignancy classification without additional attribute annotations.

3.2 MULTIMODAL TEACHER AND CT-ONLY STUDENT

Given a CT volume $\mathbf{X} \in \mathbb{R}^{B\times S\times H\times W}$ and its structured attributes $\mathbf{F} \in \mathbb{R}^{B\times T}$, the student and teacher image encoders extract slice-wise representations as

$$\mathbf{Z}^s = \phi_s(\mathbf{X}) = \{\mathbf{z}_i^s\}_{i=1}^S, \qquad \mathbf{Z}^t = \phi_t(\mathbf{X}) = \{\mathbf{z}_i^t\}_{i=1}^S, \tag{1}$$

where each slice feature belongs to $\mathbb{R}^d$. The teacher further maps the structured attributes into attribute tokens:

$$\mathbf{Z}^a = \psi(\mathbf{F}) = \{\mathbf{z}_j^a\}_{j=1}^T, \qquad \mathbf{z}_j^a \in \mathbb{R}^d. \tag{2}$$

As illustrated in Fig. 3, the CT-only student graph and multimodal teacher graph are respectively defined as

$$\mathcal{G}_s = \{\mathbf{c}_s, \mathbf{Z}^s\}, \qquad \mathcal{G}_t = \{\mathbf{c}_t, \mathbf{Z}^t, \mathbf{Z}^a\}, \tag{3}$$

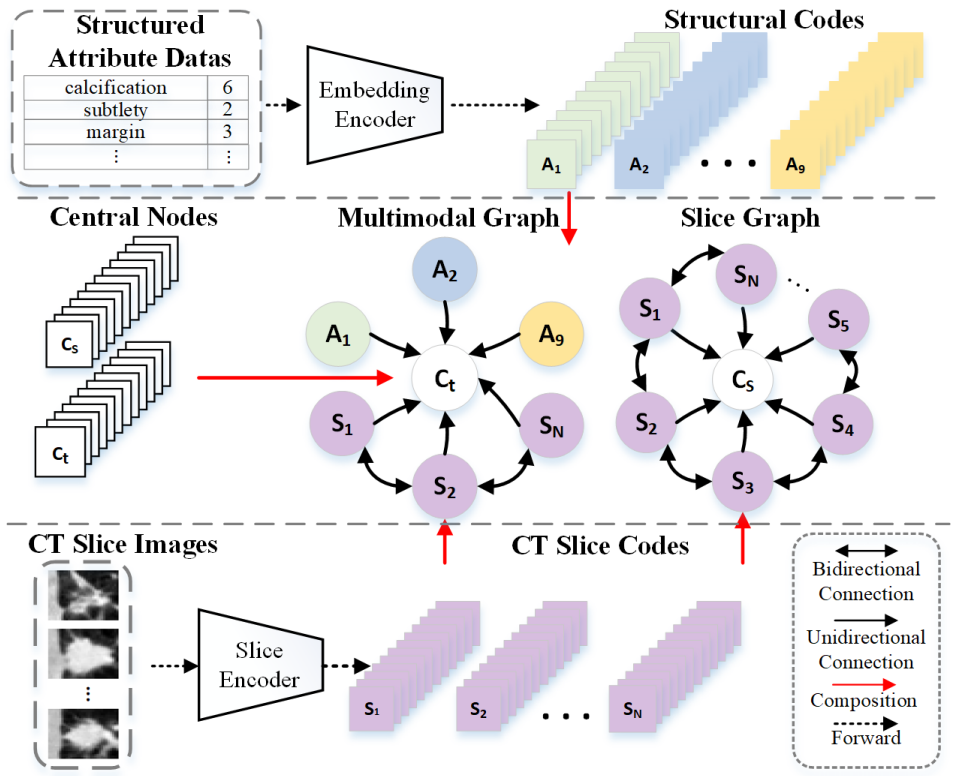

Figure 3: Architectures of the multimodal graph and slice graph.

where $\mathbf{c}_s$ and $\mathbf{c}_t$ denote the center nodes. The center nodes are initialized as zero vectors to avoid introducing additional prior bias. The student center node aggregates information from CT slice nodes, whereas the teacher center node integrates both slice and attribute information. Given the predefined adjacency matrices $\mathbf{A}_s$ and $\mathbf{A}_t$, the two graphs are processed by graph attention networks:

$$(\mathbf{H}_q, \mathbf{A}_q^{\text{att}}) = \text{GAT}_q(\mathcal{G}_q, \mathbf{A}_q), \qquad q \in \{s, t\}. \tag{4}$$

The center-node representations $\mathbf{h}_q = \mathbf{H}_q[:, 0, :]$ are used for classification:

$$\hat{\mathbf{y}}_q = \text{softmax}\big(\text{MLP}_q(\mathbf{h}_q)\big), \qquad q \in \{s, t\}. \tag{5}$$

The teacher is pretrained and student is trained using

$$\mathcal{L}_{\text{cls}}^q = \text{CE}(\hat{\mathbf{y}}_q, \mathbf{y}), \qquad q \in \{s, t\}. \tag{6}$$

After pretraining, the optimal teacher parameters are frozen and used to provide feature representations, slice-attention distributions, and Grad-CAM maps for student training.

3.3 Hierarchical Knowledge Distillation

Feature-Level Distillation. To transfer the multimodal teacher's global representation, the student center-node feature is aligned with its teacher counterpart:

$$\mathcal{L}_{\text{feat}} = \text{MSE}\left(\mathbf{h}_s, \text{sg}(\mathbf{h}_t)\right), \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation.

Slice-Level Attention Distillation. The attention weights connecting the center node to the CT slice nodes are extracted from the final graph attention layer:

$$\mathbf{a}_s = \mathbf{A}_s^{\text{att}}[:, 0, 1 : S+1], \qquad \mathbf{a}_t = \mathbf{A}_t^{\text{att}}[:, 0, 1 : S+1]. \tag{8}$$

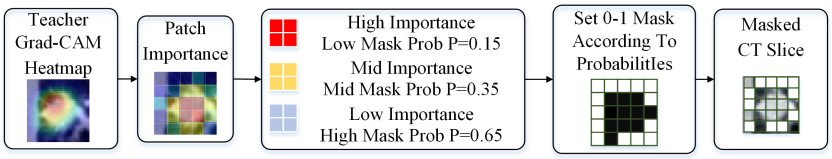

Figure 4: Generation of 2D patch-masked slices.

Because the teacher additionally attends to attribute nodes, the student and teacher slice-attention vectors are normalized over slice nodes:

$$\bar{\mathbf{a}}_s = \frac{\mathbf{a}_s}{\sum_{i=1}^{S} a_{s,i} + \epsilon}, \qquad \bar{\mathbf{a}}_t = \frac{\mathbf{a}_t}{\sum_{i=1}^{S} a_{t,i} + \epsilon}. \tag{9}$$

The attention distillation loss is then defined as

$$\mathcal{L}_{\text{att}} = \text{MSE}\left(\bar{\mathbf{a}}_s, \text{sg}(\bar{\mathbf{a}}_t)\right). \tag{10}$$

encouraging the student to reproduce the teacher's selection of diagnostically important CT slices.

Saliency-Aware Patch Mask Distillation. As shown in Fig. 4, inspired by attention-guided masked distillation (Kakogeorgiou et al., 2022), to transfer local spatial attention, we generate a normalized Grad-CAM map $\mathbf{M}$ from the last layer of frozen teacher image encoder and combine it with the teacher's slice-level attention:

$$\widetilde{\mathbf{M}}_s = \mathbf{M}_s \frac{a_{t,s}}{\max_k a_{t,k} + \epsilon}. \tag{11}$$

Thus, $\widetilde{\mathbf{M}}$ simultaneously reflects the importance of each CT slice and the regions within that slice.

The weighted heatmap is divided into patches, and the importance of each patch is calculated as

$$r_{ij} = \text{AvgPool}(\widetilde{\mathbf{M}})_{ij}. \tag{12}$$

The masking probability is assigned according to patch importance:

$$P_{\text{mask}}(r_{ij}) = \begin{cases} p_{\text{high}}, & r_{ij} \geq \tau_h, \\ p_{\text{mid}}, & \tau_l < r_{ij} < \tau_h, \\ p_{\text{low}}, & r_{ij} \leq \tau_l, \end{cases} \tag{13}$$

Consequently, highly discriminative regions are more likely to be retained, whereas less informative regions are masked with higher probability.

A binary patch mask is sampled as $\mathbf{R}_{ij} \sim \text{Bernoulli}(P_{\text{mask}}(r_{ij}))$ and upsampled to the original image resolution. The masked CT input is given by

$$\mathbf{X}_{\text{mask}} = \mathbf{X} \odot (1 - \mathbf{R}). \tag{14}$$

It is processed by the student to obtain the masked center-node representation $\mathbf{h}_s^{\text{mask}}$, which is constrained by

$$\mathcal{L}_{\text{mask}} = \text{MSE}\left(\mathbf{h}_s^{\text{mask}}, \text{sg}(\mathbf{h}_t)\right). \tag{15}$$

3.4 Training Objective

Experiments show the multimodal teacher model stabilizes at around $e = 75$. Its optimal parameters are subsequently frozen, and the student is progressively optimized using the three distillation objectives, loss coefficients and transition thresholds for distillation are selected via validation:

$$\mathcal{L}(e) = \begin{cases} \mathcal{L}_{\text{cls}}^t, & 0 \leq e < 75, \\ 0.9\mathcal{L}_{\text{cls}}^s + 0.1\mathcal{L}_{\text{feat}}, & 75 \leq e < 100, \\ 0.8\mathcal{L}_{\text{cls}}^s + 0.1\mathcal{L}_{\text{feat}} + 0.1\mathcal{L}_{\text{att}}, & 100 \leq e < 150, \\ 0.7\mathcal{L}_{\text{cls}}^s + 0.1\mathcal{L}_{\text{feat}} + 0.1\mathcal{L}_{\text{att}} + 0.1\mathcal{L}_{\text{mask}}, & e \geq 150, \end{cases} \tag{16}$$

Table 1: Comparison with existing methods on LIDC-IDRI dataset.

Method	AUC (%)	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)
2D CNN-Based Methods						
Multichannel CNN (Zheng et al., 2024)	95.66 ± 0.17	90.11 ± 0.24	–	–	–	–
DTCNN-ELM (Huang et al., 2020)	94.94	$\mathbf{94.57}$	$\mathbf{93.69}$	95.15	–	–
MV-KBC (Xie et al., 2018)	95.70 ± 0.24	91.60 ± 0.15	86.52 ± 0.25	94.00 ± 0.30	87.75 ± 0.52	87.13 ± 0.16
3D CNN-Based Methods						
STLF-VA (Wu et al., 2023)	97.17 ± 0.32	92.36 ± 0.18	91.62 ± 0.29	93.08 ± 0.48	92.99 ± 0.38	92.55 ± 0.17
Multiside GNN (Saihood et al., 2024)	95.00 ± 1.22	87.17 ± 0.84	88.60 ± 1.42	–	91.01 ± 1.14	89.07 ± 0.87
CSEGAT (Tong et al., 2025)	95.60	90.85	86.05	$\mathbf{96.46}$	–	–
Multimodal-to-Unimodal Methods						
Nodule-CLIP (Sun et al., 2024)	–	90.60	92.81	92.60	–	–
CLIP-Lung (Lei et al., 2023)	95.92 ± 1.33	87.59 ± 1.81	92.93 ± 1.90	81.69 ± 3.27	82.66 ± 3.02	87.94 ± 2.52
Our Method						
MUGAT-KD	$\mathbf{98.15 \pm 1.63}$	93.28 ± 1.20	93.03 ± 2.75	93.52 ± 3.81	$\mathbf{93.22 \pm 3.74}$	$\mathbf{93.04 \pm 1.11}$

Table 2: Comparison with existing methods on LIDP dataset.

Method	AUC (%)	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)
2D CNN-Based Methods						
MV-KBC (Xie et al., 2018)	72.51 ± 2.92	66.50 ± 2.23	86.44 ± 6.29	39.21 ± 9.36	66.39 ± 2.64	74.77 ± 2.11
DTCNN-ELM (Huang et al., 2020)	67.49 ± 3.33	67.50 ± 1.12	93.42 ± 0.96	31.07 ± 3.15	65.61 ± 0.98	77.07 ± 0.60
3D CNN-Based Methods						
MC-CNN (Shen et al., 2017)	73.37 ± 3.04	67.89 ± 2.00	79.33 ± 4.57	52.24 ± 4.74	69.50 ± 1.56	73.99 ± 2.13
NAS-Lung (Jiang et al., 2021)	70.42 ± 2.43	63.50 ± 2.02	77.69 ± 7.13	44.08 ± 13.66	66.38 ± 3.58	70.98 ± 1.21
Multimodal-to-Unimodal Methods						
CLIP-Lung (Lei et al., 2023)	79.33 ± 2.01	73.73 ± 3.91	92.68 ± 7.71	$\mathbf{58.53 \pm 8.12}$	72.61 ± 6.38	80.01 ± 7.37
Our Method						
MUGAT-KD	$\mathbf{81.00 \pm 3.10}$	$\mathbf{74.44 \pm 4.11}$	$\mathbf{96.15 \pm 8.73}$	44.73 ± 9.82	$\mathbf{75.56 \pm 4.86}$	$\mathbf{84.62 \pm 5.10}$

Table 3: Ablation study on LIDC-IDRI and LIDP datasets.

No.	Img	Feat.	Slice	Mask	LIDC-IDRI Dataset				LIDP Dataset			
					AUC	Accuracy	Recall	Specificity	AUC	Accuracy	Recall	Specificity
0	Teacher Model				98.70±1.2	94.87±2.8	97.08±7.5	93.12±3.8	83.05±1.1	85.35±3.2	95.50±8.3	71.09±7.8
1	✓	×	×	×	93.73±1.3	86.29±2.7	87.00±8.7	85.43±6.8	74.31±2.4	69.44±5.3	73.07±4.3	64.71±3.3
2	✓	✓	×	×	97.14±2.4	91.67±1.7	90.60±5.6	92.57±4.6	78.64±3.1	70.55±8.1	80.76±9.9	56.67±9.3
3	✓	×	✓	×	96.10±1.2	86.71±1.8	90.66±8.3	83.18±2.3	76.46±3.2	**79.29**±4.9	87.26±2.2	**68.48±7.2**
4	✓	×	×	✓	96.23±1.9	82.65±1.3	93.66±7.3	72.32±4.1	78.17±3.0	76.76±4.0	95.26±5.5	52.42±7.5
5	✓	✓	✓	×	97.30±1.3	87.61±2.3	**98.78±5.2**	77.33±2.8	80.77±2.8	74.24±3.9	93.89±4.7	60.44±6.9
6	✓	✓	×	✓	97.52±1.7	93.13±1.4	91.51±5.4	**94.69±6.4**	79.06±2.2	74.44±5.5	82.69±6.9	63.05±7.1
7	✓	×	✓	✓	96.89±1.5	92.11±1.5	91.21±7.7	92.89±3.2	78.30±2.1	67.22±4.4	82.30±3.4	47.47±9.4
8	✓	✓	✓	✓	**98.15±1.6**	**93.28±1.2**	93.03±2.7	93.52±3.8	**81.00±3.1**	74.44±4.1	**96.15±8.7**	44.73±9.8

4 EXPERIMENTS

4.1 DATASETS AND SETTINGS

LIDC-IDRI is a public dataset comprising 1,018 thoracic CT scans. Pulmonary nodules were annotated by four radiologists based on CT images. Nodules with an average malignant score less than 3 and greater than 3 are labeled as benign andmalignant,respectively. Exclusion with a score equals 3. Each nodule was annotated with nine attributes. LIDP comprises 990 pulmonary nodules, whose labels were determined through pathological examination. LIDP also provides seven structured attributes describing nodule attributes, which are not entirely consistent with those in LIDC-IDRI. These attributes include spiculation, lobulation, calcification, pleural retraction, texture, location, and long-axis diameter.

Each CT volume was cropped into a $32 \times 32 \times 32$ cube centered on the nodule and augmented by flipping, cropping, rotation, and axis permutation. To avoid data leakage, we employed patient-level five-fold cross-validation, ensuring that all nodules from the same patient were assigned to the same fold. The mean and standard deviation across the five test folds were reported. Training used SGD with an initial learning rate of 0.01 on two NVIDIA Quadro RTX 8000 GPUs.

Table 4: External validation using the complete LIDC-IDRI dataset for training and the complete LIDP dataset for testing.

Setting				LIDC Train–LIDP Predict			
Img	Feature	Slice	Mask	AUC	Accuracy	Recall	Specificity
✓	×	×	×	68.84	73.66	80.12	49.69
✓	✓	×	×	73.35	74.16	79.39	54.76
✓	×	✓	×	71.11	74.67	82.16	46.88
✓	×	×	✓	69.87	70.03	84.07	17.94
✓	×	✓	✓	71.56	75.17	82.69	47.27
✓	✓	✓	×	71.79	73.23	83.20	36.24
✓	✓	×	✓	75.03	76.76	**85.98**	42.55
✓	✓	✓	✓	**75.28**	**77.77**	83.33	**57.14**

The slice-wise image encoder contains four convolutional blocks, each comprising a 3×3 convolution, batch normalization, max pooling, and LeakyReLU. Each slice is mapped to a 64-dimensional feature vector and divided into 8×8 patches. Patches with mean Grad-CAM responses ≥ 0.7, ≤ 0.3, and otherwise are treated as high-, low-, and medium-importance regions, with masking probabilities of 0.15, 0.65, and 0.35, respectively. These hyperparameters were selected using only the training and validation sets.

4.2 Comparison with SOTA Methods

We compare MUGAT-KD with 2D CNN-, 3D CNN-, and multimodal-to-unimodal methods that require only CT images during inference. Publicly available methods were reproduced under the same experimental protocol, while others used reported results. As shown in Tab. 1 and 2, MUGAT-KD achieves the best AUC and F1 scores on LIDC-IDRI, reaching 98.15% and 93.04%, respectively. On LIDP, it also obtains the highest AUC, recall, and F1 scores of 81.00%, 96.15%, and 84.62%. Although its specificity is relatively low on LIDP, the high recall indicates that the model identifies most malignant nodules under the more challenging pathology-defined labels.

4.3 Ablation Studies

Tab. 3 presents the ablation results of our model on LIDC-IDRI and LIDP datasets. Configuration 0 is the performance of the teacher model. Configuration 1 uses only CT images for training and testing. Configurations 2–8 evaluate different combinations of feature-level distillation, slice-level attention distillation, and saliency-aware patch mask distillation.

Configuration 8 is the complete model, reaching 98.15% on LIDC-IDRI and 81.00% on LIDP. Compared with the image-only baseline, these results represent absolute AUC improvements of 4.42 and 6.69 percentage points, respectively.

However, the complete model exhibits reduced specificity on LIDP, suggesting that it favors higher sensitivity and reflects a trade-off between sensitivity and specificity. This trade-off may be further adjusted through classification-threshold calibration.

4.4 External Validation

To evaluate cross-dataset generalization, we train the model on LIDC-IDRI and externally validate it on LIDP, which differs in attribute availability, label definitions, and data distribution. As shown in Tab. 4, we train the model using all LIDC data and test it with all LIDP data, the full model improves the AUC from 68.84% to 75.28% over the image-only baseline and achieves an accuracy, recall, and specificity of 77.77%, 83.33%, and 57.14%, respectively.

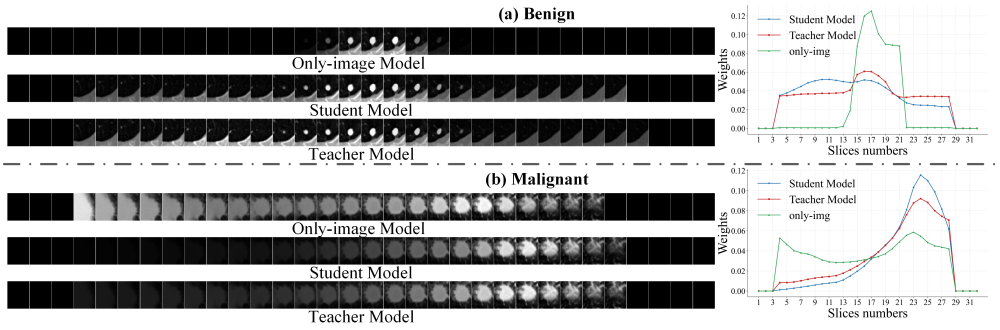

Figure 5: Visualization of slice-level attention for benign and malignant pulmonary nodules.

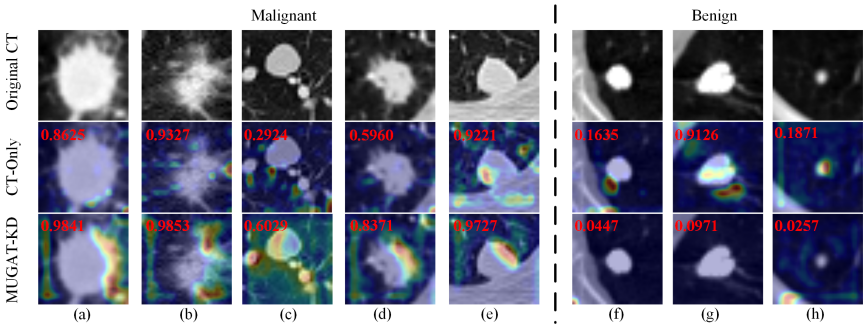

Figure 6: Grad-CAM visualizations for malignant (a–e) and benign (f–h) nodules.

4.5 Slice Attention and 2D Attention Maps

To examine how nodule attributes affect visual attention, we visualize both slice-level and 2D spatial attention. Slice attention is obtained from the center-node attention coefficients to the 32 CT slices in the final graph attention layer and renormalized across slices.

As shown in Fig. 5, the image-only model either over-focuses on a few central slices or assigns attention to less informative slices. After distillation, the student's attention distribution becomes closer to that of the multimodal teacher for both benign and malignant nodules, indicating that slice-level attention distillation successfully transfers slice-selection knowledge.

Fig. 6 presents the corresponding Grad-CAM visualizations. Specifically, (a) large diameter, (b) blurred-margin, (c) low confidence, (d) false negative corrected, (e) attending to opposite sides of the nodule, (f) well-circumscribed benign nodule, (g) false positive corrected, (h) small benign nodule. Red values indicate predicted malignancy probabilities. For malignant nodules, the distilled student focuses more on nodule boundaries, branches, and surrounding structures; for benign nodules, it suppresses responses to nonspecific tissues and adjacent vessels. Combined with probability outputs, these results indicate more reliable slice-level and regions attention than the CT-only model.

5 Conclusion

We propose MUGAT-KD, a multimodal-to-unimodal distillation framework that transfers feature-, slice-, and region-level knowledge from an attribute-assisted teacher to a CT-only student. It achieves AUCs of 98.15% and 81.00% on LIDC-IDRI and LIDP, respectively, and 75.28% under external validation. Ablation and visualization results confirm the complementary effects of the distillation modules, while CT-only inference enables deployment without nodule attributes. The results show that MUGAT-KD successfully transfers attribute knowledge to the unimodal student, improving classification adaptability and robustness without requiring attributes at inference.

AI USE STATEMENT

Generative AI tools were used only for translation and language polishing. The authors reviewed the output and take full responsibility for the final manuscript.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This study uses existing LIDC-IDRI and LIDP CT data and nodule annotations for retrospective computational evaluation; no patients were recruited or affected by the model's predictions. We report aggregate results and follow the datasets' applicable access and use conditions.

REPRODUCIBILITY STATEMENT

We provide the necessary information to reproduce the reported results. Sections 3.2–3.4 specify the MUGAT-KD architecture, including the multimodal graph teacher, the CT-only graph student, and the three-level knowledge-distillation method based on global feature alignment, slice-attention alignment, and saliency-aware patch masking. Section 4.1 describes the LIDC-IDRI and LIDP datasets, their malignancy-label definitions and structured nodule attributes, the $32 \times 32 \times 32$ nodule inputs, and the data-augmentation operations, including flipping, random cropping, rotation, and axis permutation. It also specifies the five-fold evaluation protocol, SGD optimization, and computational resources. All experimental details are configured in the source code in the supplementary material and can be used to reproduce the experiments.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74 (3):229–263, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3942–3951, 2021.

Xufeng Huang, Qiang Lei, Tingli Xie, Yahui Zhang, Zhen Hu, and Qi Zhou. Deep transfer convolutional neural network and extreme learning machine for lung nodule diagnosis on ct images. *Knowledge-Based Systems*, 204:106230, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Hanliang Jiang, Fuhao Shen, Fei Gao, and Weidong Han. Learning efficient, explainable and discriminative representations for pulmonary nodules classification. *Pattern Recognition*, 113: 107825, 2021.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2022.

Guixia Kang, Kui Liu, Beibei Hou, and Ningbo Zhang. 3d multi-view convolutional neural networks for lung nodule classification. *PloS one*, 12(11):e0188290, 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Yiming Lei, Zilong Li, Yan Shen, Junping Zhang, and Hongming Shan. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 403–412. Springer, 2023.

Yuan Li, Yinjian Zhao, Minghao Wang, Xia Feng, Qian Peng, Airu Yin, and Hua Ji. Mie-gat: Multi-perspective information enhancement for slice-based image retrieval in multi-modal medical diagnosis. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 796–805, 2025.

Chia-Ying Lin, Shu-Mei Guo, Jenn-Jier James Lien, Wen-Tsen Lin, Yi-Sheng Liu, Chao-Han Lai, I-Lin Hsu, Chao-Chun Chang, and Yau-Lin Tseng. Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest ct. *La radiologia medica*, 129(1): 56–69, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ahmed Ali Saihood, Mustafa Asaad Hasan, Mohammed A Fadhel, Laith Alzubaid, Ashish Gupta, Yuantong Gu, et al. Multiside graph neural network-based attention for local co-occurrence features fusion in lung nodule classification. *Expert Systems with Applications*, 252:124149, 2024.

Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram Van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

Yanbo Shao, Minghao Wang, Juanyun Mai, Xinliang Fu, Mei Li, Jiayin Zheng, Zhaoqi Diao, Airu Yin, Yulong Chen, Jianyu Xiao, et al. Lidp: A lung image dataset with pathological information for lung cancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 770–779. Springer, 2022.

Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017.

Lijing Sun, Mengyi Zhang, Yu Lu, Wenjun Zhu, Yang Yi, and Fei Yan. Nodule-clip: Lung nodule classification based on multi-modal contrastive learning. *Computers in Biology and Medicine*, 175:108505, 2024.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71 (3):209–249, 2021.

Ning Tang, Rui Zhang, Zeliang Wei, Xicheng Chen, Gaoming Li, Qiuyue Song, Dong Yi, and Yazhou Wu. Improving the performance of lung nodule classification by fusing structured and unstructured data. *Information Fusion*, 88:161–174, 2022.

Guiyuan Tong, Zhaofeng Xue, Tongyang Dang, Tiancheng Cao, and XinZhuo Zhao. Lung nodule malignancy classification on 3d ct images using a cosine similarity-enhanced graph attention network. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–6. IEEE, 2025.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.

Ruoyu Wu, Changyu Liang, Yuan Li, Xu Shi, Jiuquan Zhang, and Hong Huang. Self-supervised transfer learning framework driven by visual attention for benign–malignant lung nodule classification on chest ct. *Expert Systems with Applications*, 215:119339, 2023.

Yutong Xie, Yong Xia, Jianpeng Zhang, Yang Song, Dagan Feng, Michael Fulham, and Weidong Cai. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. *IEEE transactions on medical imaging*, 38(4):991–1004, 2018.

Yang Yang, Xiaoqin Li, Jipeng Fu, Zhenbo Han, and Bin Gao. 3d multi-view squeeze-and-excitation convolutional neural network for lung nodule classification. *Medical Physics*, 50(3):1905–1916, 2023.

Penghua Zhai, Yaling Tao, Hao Chen, Ting Cai, and Jinpeng Li. Multi-task learning for lung nodule classification on chest ct. *IEEE Access*, 8:180317–180327, 2020. doi: 10.1109/ACCESS.2020.3027812.

Rong Zheng, Hongqiao Wen, Feng Zhu, and Weishun Lan. Attention-guided deep neural network with a multichannel architecture for lung nodule classification. *Heliyon*, 10(1), 2024.