Explainable Multi-Scale Multiple Instance Learning for Lung Nodule Malignancy Prediction

Changhyun Park[1] and Juho Jung[1]

VUNO Inc, South Korea
changhyun.park@vuno.co
juho.jung@vuno.co

Abstract. The LUNA25 challenge focuses on the binary classification of lung nodules as malignant or benign from chest CT scans. This task is critical for early lung cancer detection and treatment planning. Our approach builds upon a multi-scale multiple instance learning (MIL) framework, wherein each volume of interest (VOI) is treated as a bag of instances to effectively account for spatial ambiguity in both the localization of pathological structures and the distribution of malignancy-associated evidence. This formulation not only enables robust learning under weak supervision but also enhances interpretability by providing spatially resolved malignancy evidence across the VOI. To improve generalization, we adopt a curriculum learning strategy: pretraining on LIDC-IDRI dataset with multi-attribute nodule analysis, followed by fine-tuning on LUNA25 for malignancy prediction. Our final ensemble achieved an AUROC of 0.9366 on the Open Development Phase Leaderboard. Code is available at: https://github.com/chpark-ML/luna25-challenge.

Keywords: LUNA25 · 3D Medical Imaging · Chest CT Volume · Malignancy

1 Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, and early detection plays a critical role in improving patient outcomes. Computed tomography (CT) of the chest has become a cornerstone in the screening and diagnosis of pulmonary nodules, which may serve as early indicators of malignancy. However, accurately distinguishing malignant nodules from benign ones remains a challenging task due to the high variability in nodule appearance and overlapping radiographic features.

The LUNA25 challenge [1] addresses this problem by focusing on the binary classification of lung nodules—determining whether a given nodule is malignant or benign—using volumetric chest CT scans. Specifically, each case consists of a chest CT volume along with the 3D coordinates of a nodule center. Based

[1] https://luna25.grand-challenge.org/

on this information, a volume of interest (VOI) is extracted around the specified location, and the objective is to predict the malignancy of the nodule contained within the VOI. The dataset provided for this challenge, the LUNA25 dataset, comprises expertly annotated CT scans with rich spatial and semantic information about lung nodules, offering a valuable benchmark for algorithmic development.

Our approach is motivated by the clinical observation that radiologists do not rely solely on the appearance of the nodule itself, but also incorporate surrounding anatomical context and integrate both coarse and subtle radiographic cues when making malignancy assessments. To emulate this multi-faceted diagnostic process, we adopt a multi-scale multiple instance learning (MIL) framework that models each VOI as a bag of instances sampled at different spatial scales. This formulation not only enables the model to capture both localized and contextual malignancy evidence, but also facilitates interpretability by providing spatially resolved predictions that highlight diagnostically relevant regions across the volume.

In addition, we incorporate transfer learning from a model pretrained on the LIDC-IDRI dataset through auxiliary tasks such as nodule attribute classification and segmentation. This pretraining enables the model to learn rich morphological representations of physical nodules, including shape, margin, texture, and spatial context, thereby providing a strong inductive prior for downstream malignancy classification. The pretrained model is then fine-tuned on the LUNA25 dataset through a curriculum-based learning strategy, allowing for gradual adaptation to the malignancy prediction task.

Model classification performance is primarily evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC), a widely adopted metric for binary classification, particularly in medical imaging contexts where both sensitivity and specificity are critical. Our method demonstrates strong predictive performance, highlighting its potential for aiding in early lung cancer detection.

Problem Statement

Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ denote a set of input samples, where each x_i corresponds to a 3D VOI extracted from a chest CT scan. In typical clinical and algorithmic settings, a full CT image volume along with the spatial coordinates of a lung nodule are provided. Given this information, a VOI of fixed size—centered at the annotated nodule location—is extracted and used as the input x_i. Each x_i contains a single annotated nodule and is associated with a binary label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates a malignant nodule and $y_i = 0$ indicates a benign one. The goal is to learn a predictive function:

$$f_\theta : \mathcal{X} \to [0, 1],$$

parameterized by θ, that assigns to each input volume $x_i \in \mathcal{X}$ a malignancy probability $\hat{y}_i = f_\theta(x_i)$, quantifying the likelihood that the nodule present in x_i is malignant.

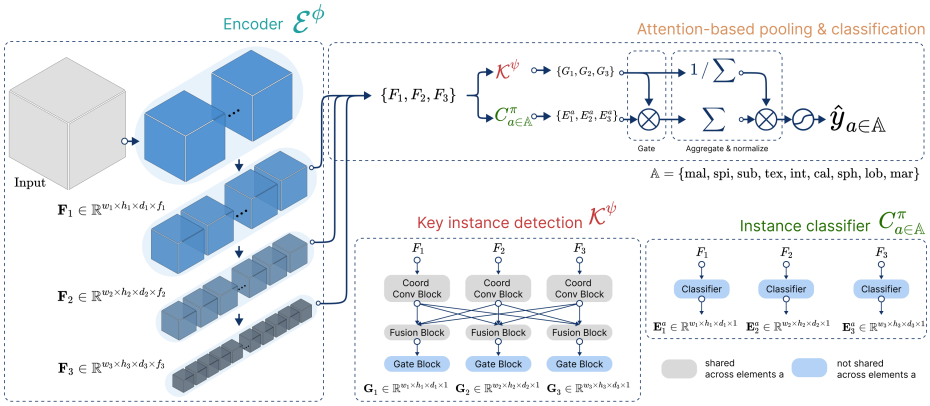

Fig. 1. Overall architecture of our proposed method.

2 Method

2.1 Multi-scale Multiple Instance Learning

Bag-of-Local-Features The proposed architecture adopts a MIL formulation, where each VOI is treated as a bag of spatially localized feature vectors. As illustrated in Figure 1, our model is built upon a 3D U-Net backbone that extracts multi-level feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$ from different stages of the encoder. These feature maps correspond to increasingly larger receptive fields, thereby capturing local information at fine, intermediate, and coarse spatial resolutions.

Each feature map $\mathbf{F}_l$ ($l \in \{1, 2, 3\}$) consists of a grid of feature vectors that represent semantic embeddings of 3D patches across the spatial extent of the VOI. Specifically, $\mathbf{F}_1$ encodes texture-level and edge-level details from small receptive fields, whereas $\mathbf{F}_2$ and $\mathbf{F}_3$ progressively capture more abstract and contextual representations. These spatially distributed vectors are treated as individual instances within the MIL framework, allowing the entire VOI to be modeled as a bag of heterogeneous, scale-specific local descriptors.

This formulation enables the network to learn fine-grained and context-aware malignancy patterns without requiring pixel-wise or voxel-level annotations. It also supports learning under weak supervision, leveraging the assumption that at least one instance within the bag (i.e., the VOI) may contain diagnostic evidence of malignancy.

Attention Gate Modules To enhance spatial awareness and selective feature encoding, we incorporate multi-scale attention gating mechanisms guided by anatomical priors. Given that the nodule center is known and lies near the center of the VOI, absolute positional information is a valuable cue for malignancy assessment. Therefore, we explicitly encode spatial coordinates into the attention module, allowing the model to attend to regions that are both semantically relevant and anatomically plausible.

Feature extraction is conducted via multi-scale 3D convolutions, producing hierarchical feature maps that are fused through a Feature Pyramid Network (FPN). The FPN aggregates semantic information from coarse to fine resolutions, enabling input-dependent spatial discrimination. At each resolution level l, an attention gating module generates a spatial weight map $\mathbf{G}_l$, which reflects the estimated discriminativeness of each local region within the feature map $\mathbf{F}_l$.

The gated representations $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$ serve as instance-level importance scores across different spatial scales, forming the basis for stochastic instance selection in the pooling stage.

Attention-Gated Pooling Final malignancy prediction is derived by aggregating instance-level predictions across all spatial locations in a weighted manner guided by attention. Each local feature vector in $\mathbf{F}_l$ is independently passed through a shared lightweight classifier to produce a logit value, resulting in a dense logit map over the VOI.

These logits are then aggregated via attention-gated pooling, where each logit is weighted by its corresponding gating score $\mathbf{G}_l$. Since the final malignancy score is computed as a linear combination of spatially localized logits and attention weights, the model's decision can be explicitly attributed to specific regions within the VOI.

This formulation enables spatial interpretability: the combination of $\mathbf{F}_l$ and $\mathbf{G}_l$ directly defines the contribution of each spatial location to the final prediction, effectively exposing localized malignancy evidence. By performing this weighted aggregation across multiple scales, the model integrates complementary diagnostic cues while preserving spatial coherence. The resulting scalar output represents a soft, interpretable fusion of local predictions under the MIL paradigm.

2.2 Three-Phase Curriculum Learning Framework

To effectively address the challenges posed by limited supervisory signals and inter-scan heterogeneity in pulmonary nodule malignancy classification, we propose a structured three-phase curriculum learning framework. This paradigm organizes the training process from fundamental morphological attribute learning to higher-order diagnostic reasoning, thereby enabling the network to progressively acquire transferable inductive priors before specializing on the final malignancy prediction task.

Phase 1: Morphological Prior Induction through Multi-Attribute Prediction and Segmentation Supervision In the initial phase, we pretrain the model using the LIDC-IDRI dataset [1], which provides expert-annotated descriptors of pulmonary nodule morphology. The pretraining is formulated as a multi-task learning problem encompassing nine clinically relevant attributes: *malignancy*, *subtlety*, *sphericity*, *lobulation*, *spiculation*, *margin*, *texture*, *calcification*, and *internal structure*. To encourage the emergence of disentangled,

attribute-specific representations, distinct classification heads are attached to the backbone for each attribute. Concurrently, an auxiliary segmentation branch is jointly optimized using binary nodule masks, compelling the network to capture boundary-aware and spatially localized features. This stage facilitates the acquisition of robust semantic priors that are fundamental for subsequent malignancy reasoning.

Phase 2: Cross-Cohort Joint Learning for Malignancy Discrimination Leveraging the pretrained representations from Phase 1, the model undergoes a joint learning procedure utilizing both the LIDC-IDRI and LUNA25 datasets. This phase is designed to enhance the network's ability to discriminate between benign and malignant nodules while improving robustness to domain shifts caused by heterogeneous imaging conditions, such as variations in scanner hardware, reconstruction protocols, and labeling policies. The joint optimization is formulated as an end-to-end binary malignancy classification task, where samples from both datasets are simultaneously integrated during training. This multi-cohort learning paradigm enables the acquisition of domain-invariant yet highly discriminative feature representations, thereby improving generalization across diverse clinical settings and preparing the model for task-specific fine-tuning in Phase 3.

Phase 3: Concept-Level Fine-Tuning on LUNA25 In the final stage, the model undergoes concept-level fine-tuning on the LUNA25 dataset, whose annotation schema is closely aligned with the target malignancy classification task. During this phase, all network parameters remain trainable, enabling the model to holistically adapt both low-level morphological features and high-level diagnostic abstractions. This comprehensive optimization ensures that the final predictor is fully aligned with the labeling semantics and evaluation metrics defined by the LUNA25 benchmark, thereby achieving task-specific performance fidelity.

3 Experiments

3.1 Datasets and Preprocessing

To standardize input representations and emphasize diagnostically salient regions, we implemented a structured preprocessing pipeline for all CT volumes. For each annotated nodule, a lesion-centered 3D patch was extracted to ensure nodule-focused modeling. To mitigate inter-scan resolution heterogeneity, all CT volumes were resampled to a standardized anisotropic voxel spacing of 1.0mm (axial slice thickness) and 0.67mm (in-plane pixel size along both y- and x-axes). Intensity normalization was performed via DICOM windowing to enhance soft-tissue contrast and highlight pulmonary structures.

For training and evaluation, we adopted a 7-fold stratified cross-validation protocol to ensure robust generalization while preserving class balance across

folds. An independent model was trained on each fold. Final predictions were obtained via an ensemble in which fold-wise outputs were combined using a average.

3.2 Implementation Details

Phase 1: Morphological Prior Induction. The network was pretrained on the LIDC-IDRI dataset, where supervision was provided through nine attribute-classification tasks (*malignancy*, *subtlety*, *sphericity*, *lobulation*, *spiculation*, *margin*, *texture*, *calcification*, and *internal structure*) alongside an auxiliary nodule segmentation task. For this stage, only CT scans with a slice thickness of 3.5mm or less were utilized, and training and validation were restricted to nodules annotated by at least two independent radiologists. Each attribute score was mapped to a continuous scale of 0–1, and the averaged score across the annotating radiologists was used as the annotation. Segmentation masks were prepared according to a 50% consensus rule, where a voxel was included in the mask if at least half of the annotating radiologists marked it as part of the nodule. This setup encouraged the development of disentangled, attribute-specific representations while capturing boundary-aware features essential for semantic understanding.

Phase 2: Cross-Cohort Joint Learning. Building on the pretrained representations, the model was jointly trained on both the LIDC-IDRI and LUNA25 datasets to enhance binary malignancy discrimination. For the LIDC-IDRI subset, only nodules annotated by at least three radiologists were included to ensure higher annotation reliability. The malignancy score for each nodule was obtained by mapping individual radiologists' ratings to a 0–1 scale and averaging them to form the final annotation. This phase aimed to mitigate inter-dataset domain shifts caused by scanner variability and differing annotation policies. Samples from both datasets were jointly optimized in an end-to-end classification task, enabling the model to learn domain-invariant yet discriminative features for malignancy prediction.

Phase 3: Concept-Level Fine-Tuning. Finally, the model was fine-tuned on the LUNA25 dataset, aligning the network with the target malignancy classification objectives. During this phase, all layers were unfrozen to permit full adaptation of both low-level morphological features and high-level diagnostic reasoning.

Optimization We employed the AdamW optimizer with a weight decay of 1×10^{-3} and utilized a one-cycle learning rate policy with cosine annealing, where the learning rate was increased during the first 20% of training and decayed thereafter. The batch size was fixed at 32, and gradient norms were clipped to a maximum of 1.0 to maintain training stability. An exponential moving average of model parameters was maintained throughout training to stabilize convergence and improve generalization.

To ensure stable training of the gate network and to encourage effective discovery of class evidence across multiple feature levels, an auxiliary loss was introduced and jointly optimized with the main objective.

For each training phase, the maximum learning rate and training duration were configured as follows:

Phase 1: 200 epochs with a maximum learning rate of 1×10^{-3}.
Phase 2: 100 epochs with a maximum learning rate of 1×10^{-4}.
Phase 3: 50 epochs with a maximum learning rate of 1×10^{-5}.

4 Conclusion

In this work, we proposed a robust and interpretable framework for lung nodule malignancy classification, developed for the LUNA25 challenge. Our approach uses a multi-scale MIL architecture with 3D convolutional backbones and attention-based instance selection to capture spatially distributed malignancy evidence without dense annotations.

A key component of our approach is the incorporation of a three-phase curriculum transfer learning strategy. In the first phase, the model undergoes multi-attribute pretraining on the LIDC-IDRI dataset with auxiliary segmentation supervision, enabling the acquisition of robust morphological priors. In the second phase, a cross-cohort joint learning stage is performed on both the LIDC-IDRI and LUNA25 datasets, enhancing malignancy discrimination and improving robustness to inter-dataset domain shifts. Finally, in the third phase, the model is fine-tuned on the LUNA25 dataset to fully align with the target malignancy classification task. This progressive training paradigm facilitates improved generalization, data efficiency, and convergence while enabling the network to gradually evolve from low-level morphological understanding to high-level diagnostic reasoning.

Overall, our findings emphasize the importance of combining domain-specific knowledge with structured learning strategies for interpretable and accurate pulmonary nodule assessment. Our final ensemble model achieved an AUROC of 0.9366 on the Open Development Phase Leaderboard, demonstrating the effectiveness of the proposed framework.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., 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)