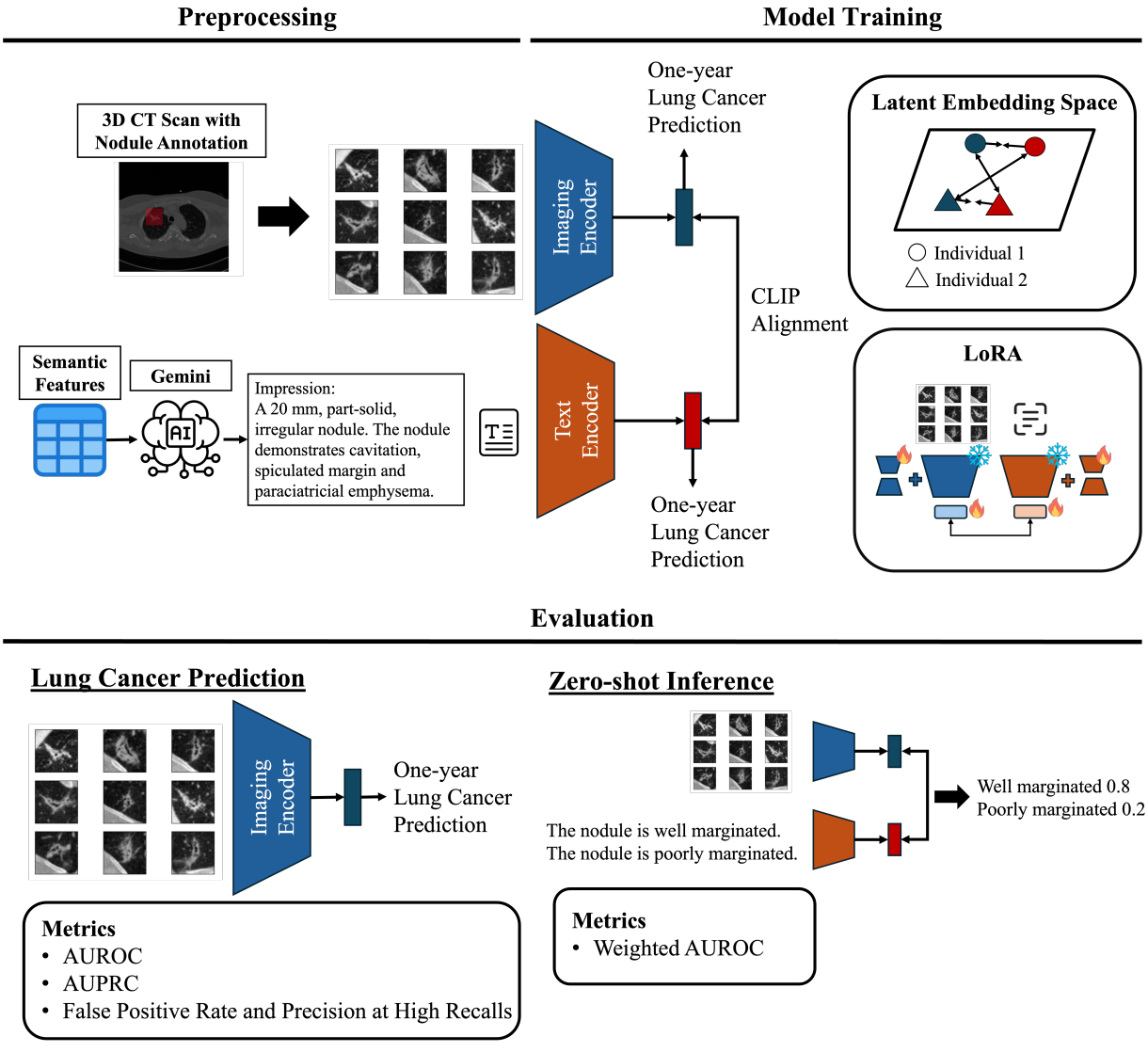
Graphical Abstract

Vision-Language Model-Based Semantic-Guided Imaging Biomarker for Lung Nodule Malignancy Prediction

Luoting Zhuang, Seyed Mohammad Hossein Tabatabaei, Ramin Salehi-Rad, Linh M. Tran, Denise R. Aberle, Ashley E. Prosper, William Hsu



Vision-Language Model-Based Semantic-Guided Imaging Biomarker for Lung Nodule Malignancy Prediction

Luoting Zhuang[a], Seyed Mohammad Hossein Tabatabaei[a], Ramin Salehi-Rad[b], Linh M. Tran[b], Denise R. Aberle[a], Ashley E. Prosper[a], William Hsu[a,*]

[a]*Medical & Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, 90095, CA, USA*
[b]*Department of Medicine, Division of Pulmonology and Critical Care, David Geffen School of Medicine at UCLA, Los Angeles, 90095, CA, USA*

Abstract

Objective: Machine learning models have utilized semantic features, deep features, or both to assess lung nodule malignancy. However, their reliance on manual annotation during inference, limited interpretability, and sensitivity to imaging variations hinder their application in real-world clinical settings. Thus, this research aims to integrate semantic features derived from radiologists' assessments of nodules, guiding the model to learn clinically relevant, robust, and explainable imaging features for predicting lung cancer.

Methods: We obtained 938 low-dose CT scans from the National Lung Screening Trial (NLST) with 1,246 nodules and semantic features. Additionally, the Lung Image Database Consortium dataset contains 1,018 CT scans, with 2,625 lesions annotated for nodule characteristics. Three external datasets were obtained from UCLA Health, the LUNGx Challenge, and the Duke Lung Cancer Screening. For imaging input, we obtained 2D nodule slices from nine directions from $50 \times 50 \times 50$ mm nodule crop. We converted structured semantic features into sentences using Gemini. We fine-tuned a pretrained Contrastive Language-Image Pretraining (CLIP) model with a parameter-efficient fine-tuning approach to align imaging and semantic text features and predict the one-year lung cancer diagnosis.

Results: Our model outperformed state-of-the-art (SOTA) models in the NLST test set with an AUROC of 0.901 and AUPRC of 0.776. It also showed robust results in external datasets. Using CLIP, we also obtained predictions on semantic features through zero-shot inference, such as nodule margin (AUROC: 0.812), nodule consistency (0.812), and pleural attachment (0.840).

Conclusion: Our approach surpasses the SOTA models in predicting lung cancer across datasets collected from diverse clinical settings, providing explainable outputs, aiding clin-

*Corresponding author

icians in comprehending the underlying meaning of model predictions. This approach also prevents the model from learning shortcuts and generalizes across clinical settings. The code is available at `https://github.com/luotingzhuang/CLIP_nodule`.

Keywords: Lung Cancer Early Detection, Computed Tomography, Vision-Language Model, Semantic Features

1. Introduction

Lung cancer remains the leading cause of cancer-related deaths [1, 2]. Computed tomography (CT) has demonstrated effectiveness in detecting lung cancer and decreasing cancer-related mortality during lung cancer screening and also in routine healthcare settings [3, 4]. However, radiologists are experiencing burnout since there is a growing number of detected nodules with the increasing application of CT. Computer-aided diagnostic systems have been proposed to alleviate the workload of radiologists by providing an automatic and accurate lung cancer prediction based on patients' data.

Traditional machine learning models in lung nodule risk assessment focus on semantic features, deep features, or both. Semantic features refer to the descriptive terms radiologists use to characterize regions of interest, such as the shape, margin, and vascularity. Models trained on semantic features are easier to interpret, but they also introduce difficulties when scaled for clinical use, as they require manual annotation from radiologists [5, 6]. Recently, deep learning (DL) has become a powerful tool due to its strong capability of extracting complex features from CT without any manual input [7, 8]. Such imaging-based models still encounter numerous obstacles. For instance, imaging features are sensitive to variations in the acquisition and reconstruction parameters of CT scans, including dose levels, slice thicknesses, and reconstruction kernels [9, 10]. Moreover, research has indicated that DL models learn shortcuts, which are characteristics highly correlated with outcomes but lack clinical significance [11, 12]. This phenomenon could adversely impact the reproducibility of features and the generalization of models. Additionally, the lack of explainability remains one of the major challenges related to deep features. Heatmaps or attribution maps generated by explainability methods, such as GradCAM [13], can only indicate the regions on which the model focuses on but do not reveal the specific features utilized.

Combining these features has become more popular for enhancing lung cancer prediction performance. However, directly merging two features often reduces applicability in clinical settings, as manual annotations are still required. Alternative approaches tackle this problem through co-learning and multi-task learning [14, 15, 16], but these models face challenges regarding explainability or training difficulties when dealing with numerous semantic features.

An emerging approach in vision-language models (VLM), Contrastive Language-Image Pre-training (CLIP) [17], fills the gap by learning the alignment between imaging features and descriptive text. Guided by text, these models allow deep visual features to capture a richer and more robust representation. Therefore, in this study, we have made the following key contributions:

1. We employed CLIP to incorporate the semantic features to direct the model in obtaining clinically significant and robust features for the prediction of lung cancer.
2. We curated a unique dataset comprising CT scans and their corresponding semantic features, which were annotated by radiologists. The semantic features encompass characteristics of both nodules and their surrounding environment.
3. We benchmarked our CLIP model against several state-of-the-art (SOTA) lung cancer prediction models using a comprehensive set of metrics. Our model demonstrated better and more robust results in external datasets that include CT scans collected from different clinical scenarios (e.g., screening and routine clinical care) and different types of CT scans (e.g., low-dose CT, diagnostic CT, and contrast-enhanced CT).
4. We explored the zero-shot inference, a feature of CLIP-based models that generates semantic features despite not being explicitly trained to do so.
5. Several adjustments were introduced to accommodate CLIP fine-tuning on 3D images with limited data. By combining these strategies with extensive domain knowledge, our model shows comparable and better performance than imaging-based models trained on tens of thousands of cases using an order of magnitude fewer cases.

Statement of Significance

Problem or Issue	Imaging-based lung nodule malignancy prediction models lack explainability and tend to learn shortcuts, limiting their generalizability and applicability in diverse clinical settings.
What is Already Known	The vision-language model, CLIP, aligns image and text features in a shared space to integrate visual and textual information effectively.
What this Paper Adds	We utilize CLIP to incorporate both internal and external semantic features to guide the imaging model to learn clinically relevant and reliable imaging biomarkers, aiming to improve lung cancer classification across diverse patient populations and CT acquisition protocols. This model offers explainability through zero-shot inference, helping radiologists and end users understand model predictions with semantic features.

2. Related Work

2.1. Imaging-based Lung Cancer Prediction Model

DL models trained on CT scans have been made available online for potential implementation in real-world clinical settings. Sybil [7] employs ResNet18 to predict lung cancer risk scores for up to six years using low-dose CT (LDCT) scans. Sybil was trained utilizing the National Lung Screening Trial (NLST) dataset, with around 15,000 CT scans, and has demonstrated consistently robust performance during external validation using screening scans from two distinct sites. In another study, Venkadesh et al. [8] developed a model to predict the malignancy risk of pulmonary nodules by integrating ResNet18 for 2D nodule crops and Inception-V1 for 3D nodule crops. This model was trained using 16,077 nodules from the NLST dataset and has shown robust performance in the external cancer screening datasets.

Although these models have shown effectiveness in predicting lung cancer, they still lack explainability. It is unclear what features have been captured from the image. Moreover, as these models were trained on NLST data and only evaluated on screening cohorts, the models can overfit to LDCT scans and have a limited scope of clinical application, such as in diagnostic CT and contrast-enhanced CT. Studies have also shown that the risk scores generated from Sybil are sensitive to the CT reconstruction parameters, such as slice thickness and kernel [9].

2.2. Imaging Feature Learning Guided by Semantic Features

Many studies have also investigated incorporating clinical or semantic features to improve the accuracy of lung cancer prediction. These models can capture the relationship between imaging and semantic data, making the resulting embeddings more clinically meaningful and interpretable. However, unlike medical imaging, which is more readily available, semantic features often require labor-intensive annotation by radiologists. The existing approach of directly fusing two modalities can reduce the model's practical value in clinical settings and increase radiologists' workload. Several studies attempted to address the issue. For example, DeepIPN [15] is a co-learning model trained on 1,284 in-house data, integrating both imaging features and clinical features, which consist of demographics and nodule characteristics. Imaging features were obtained from the top five most likely nodules based on a nodule detection algorithm. They implemented three prediction branches, each for the imaging feature, clinical tabular feature, and the fusion of the two. However, the model cannot generate explanations of its outputs, despite jointly modeling information across two modalities.

Moreover, a hierarchical semantic convolutional neural network [14] was developed using a multi-task learning framework. It simultaneously performs a low-level task of predicting

semantic features and a high-level task of assessing nodule malignancy. The model combines embeddings from the low-level semantic predictions with global convolutional representations to generate the final malignancy assessment. This architecture is inherently interpretable, as it provides explicit predictions for semantic features through its low-level branches. However, challenges arise when many semantic features are involved. Conflicting tasks can lead to inconsistent gradient updates in the shared layers, ultimately degrading overall model performance.

2.3. Vision-Language Model

VLMs have recently gained popularity, employing DL to simultaneously understand visual and linguistic information. Generally, joint representations are learned by mapping features from different modalities into a shared latent space. One of the VLMs is CLIP [17], which aligns natural images with their corresponding captions. It has exhibited exceptional performance across a variety of downstream tasks. This framework has also been successfully applied in the medical field, as clinicians generally write reports linked to various medical modalities, such as echocardiograms [18], chest X-rays [19], histopathology images [20], and CT scans [21].

The CLIP framework can also be suitable for predicting lung cancer by aligning nodule-specific imaging features with semantic features. By projecting both feature types from the same patient into a shared embedding space, the model enables image features to incorporate semantic context. Additionally, the model offers greater explainability, as it can perform zero-shot inference to evaluate how closely an image aligns with a given semantic feature. However, several questions arise when training such a contrastive learning model on medical imaging and semantic features. (1) The success of CLIP relies on a large dataset. Although medical imaging data and reports can be collected with relative ease, it is challenging to obtain detailed characterizations for specific regions like nodules. Therefore, training a VLM model from scratch is impractical, and it is necessary to use a pretrained CLIP model. However, issues emerge when 3D medical images are input into a pretrained CLIP model, which is only compatible with 2D images. (2) Semantic features are typically presented in tabular format and may contain missing values. Tabular data limits the use of the pretrained text encoder from the CLIP model, which already maintains some degree of alignment between images and text.

In a previous study, CLIP-Lung [22] was introduced to utilize CLIP for learning generalized visual representations from semantic features using the Lung Image Database Consortium (LIDC) dataset. However, one of the critical limitations of the study is the lack of investigation into one of CLIP's most distinctive capabilities, zero-shot inference. While

Table 1: **Datasets.** Datasets utilized in this study are summarized with respect to cohort size, proportion of lung cancer cases, imaging modality, and data source. It is important to note that the biopsy-confirmed diagnoses for lung cancer in the LIDC dataset are incomplete.

Dataset	# Cases (% positive)	# Nodules	CT	Patient Cohort
Training Datasets				
NLST	938 (23%)	1,261	Low-dose	Screening
LIDC	1,018 (9%)	2,625	Low-dose & Diagnostic	Screening & Incidental
External Datasets				
LUNGx Challenge	70 (50%)	83	Diagnostic & Contrast-enhanced	Incidental
UCLA Health	51 (55%)	52	Low-dose & Diagnostic	Incidental; Neversmoker
DLCS	856 (11%)	1,388	Low-dose	Screening

CLIP is not trained to predict semantic features, zero-shot inference identifies semantic features most similar to a given imaging feature embedding. Leveraging this could improve model explainability and enhance radiologists' trust in the system. Second, the authors did not perform a benchmark to evaluate the model against SOTA lung cancer risk prediction models across datasets with different properties. Additionally, the model made predictions of risk score estimations from radiologists instead of the biopsy-confirmed results. Furthermore, the model was trained on a limited number of semantic features that do not fully characterize the properties of the nodule and the surrounding areas.

3. Methods

3.1. Datasets

3.1.1. Training and Test Data

The training data were collected from two publicly available datasets (Table 1). First, we obtained 938 LDCT scans containing at least one nodule from the NLST [3, 23]. This is a unique cohort in which thorough annotations were performed by fellowship-trained thoracic radiologists at UCLA Health. They annotated the nodule's location and curated a holistic set of 19 semantic features, listed in Table A1, for a total of 1,261 nodules [24]. These features encompass general features (e.g., shape, margin, and consistency), internal characteristics (e.g., necrosis and cyst-like spaces), and external features (e.g., vascular convergence and emphysema). Additionally, the LIDC dataset contains both diagnostic and screening CT scans from 1,018 cases, with 2,625 lesions annotated for nodule characteristics, including sphericity, lobulation, consistency, internal structure, margin, and spiculation [25, 26]. Since multiple radiologists annotated the nodules in the study, and their annotations varied, we used a majority vote at the pixel level for the nodule annotations and took the median of the semantic feature scores.

3.1.2. External Data

We collected three external datasets, each representing distinct patient cohorts and CT imaging types. First, we obtained 51 diagnostic chest CT scans from a subset of patients at UCLA Health who underwent percutaneous CT-guided lung biopsies. These patients self-identified as never-smokers and were not eligible for lung cancer screening. Despite this, they are considered higher risk due to the incidental detection of nodules that warranted biopsy. Fellowship-trained radiologists annotated the locations of the most suspicious nodules. Second, we utilized the publicly available LUNGx Challenge dataset, which includes 70 diagnostic and contrast-enhanced CT scans with 83 annotated nodules [27, 28]. Lastly, the Duke Lung Cancer Screening (DLCS) dataset comprises 1,613 low-dose CT (LDCT) scans with 2,487 nodules, collected as part of a lung cancer screening program. Nodule detection was performed using a DL-based algorithm, and some nodules were further reviewed by medical students and radiologists [29]. We excluded individuals with less than one year of follow-up, resulting in a final dataset of 856 patients and 1,388 nodules.

These three datasets also differ in their lung cancer prevalence. The LUNGx and UCLA datasets have relatively balanced distributions, with 50% (35 cases) and 55% (28 cases) of nodules diagnosed as malignant tumors, respectively. In contrast, the DLCS dataset has a lower malignancy rate, with only 11% (94 cases) confirmed as lung cancer, reflective of a screening population.

3.2. Data Preprocessing

3.2.1. Semantic feature preprocessing

There are several challenges when directly aligning tabular semantic and imaging features using the CLIP framework. First, we have limited training data compared to the quantity usually used in contrastive learning, so it is beneficial to use the pretrained CLIP model, which already exhibits a certain degree of alignment between image and text features. Second, tabular data contains numerous missing values. For example, radiologists at UCLA overlooked certain semantic features during annotation. Specifically, 71% are missing in airway cutoff, and 1-2% are missing in nodule consistency, shape, and eccentric calcification. (Table 4). Furthermore, the NLST and LIDC datasets contain distinct sets of semantic features, resulting in the missingness of a majority of these features during the harmonization process. Therefore, we proposed a workaround to convert semantic features into text to resolve the challenge above. First, we consulted with radiologists to standardize the semantic features by matching the terms between the two datasets as closely as possible. The harmonization of semantic features from LIDC and NLST data is shown in Appendix C. We treated features that appear in NLST but not in LIDC as missing values. Second, we

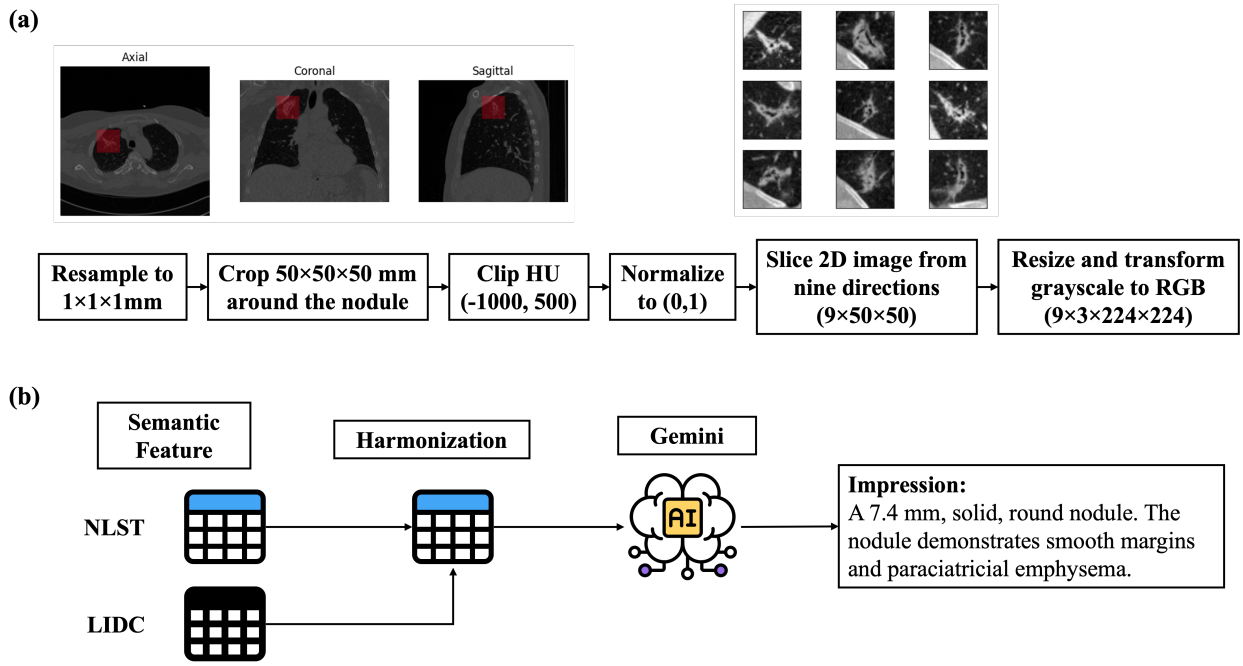


Figure 1: **CT and Semantic Features Preprocessing Steps.** (a) For the CT scan, we resampled it to a spacing of $1 \times 1 \times 1$ mm and cropped it with a bounding box of size 50 mm. The Hounsfield Unit values were clipped between -1000 and 500, then normalized to a scale of 0 to 1. We sliced the nodule from nine different angles to accommodate the CLIP architecture while preserving most of the nodule's structures. Each of the 2D images was then resized and transformed from grayscale to RGB format to meet the requirement of the CLIP input. (b) For semantic features, we first harmonized LIDC and NLST semantic features, and then input the features into Gemini with a prompt designed to generate text similar to a radiology report.

converted semantic features into radiology report-like text using Gemini 1.5 Flash [30, 31], a large language model developed by Google DeepMind (Fig. 1b). Specifically, we input a list of semantic features with a prompt. The generative model was instructed to produce a report including findings and impression sections, emulating a radiology report. The findings section listed all semantic features, whereas the impression section provided a high-level summary of these features in a few sentences. The absence of features was only mentioned within the findings section. To prevent the generation of irrelevant information, explicit instructions were given to exclude any content beyond the provided data. For missing values, while imputation methods typically add noise and bias into data, the textual format allows skipping the missing values, and the downstream language models can generate structured feature embeddings. The exact prompt utilized and an example report generated are shown in Appendix B. Furthermore, we implemented natural language preprocessing augmentation to substitute the text with synonyms and randomly crop the sentence using the nlpaug Python package [32], allowing for greater variations in text during training. During training, we also randomly selected text from the findings and impressions parts.

3.2.2. CT images preprocessing

Fig. 1a illustrates the CT preprocessing pipeline. First, we standardized and resampled 3D CT images to a spatial resolution of $1 \times 1 \times 1$ mm and placed a $50 \times 50 \times 50$ mm bounding box around the nodule. For nodules located at the boundaries, padding was applied to the cropped image to ensure a consistent input size. We chose the size of 50 mm because our study focuses on pulmonary nodules, which are typically smaller than 30 mm. This size ensures that the cropped image includes the entire nodule and its perinodular region, while still preserving sufficient detail to capture small nodules. Then, we clipped the Hounsfield Unit values within the range of -1000 to 500 and normalized them to the range of 0 to 1. Since we aimed to use the CLIP model, we converted the 3D nodule crops into 2D images obtained from nine different planes, all passing through the nodule centroid. This approach is referred to as 2.5D, where multiple 2D images are used to approximate 3D information. This is especially important for characterizing nodules, as different areas of the nodule can exhibit varying attributes. Furthermore, since CT scans are grayscale, we repeated the 2D image three times and stacked them together to mimic three RGB channels. Then, we normalized each 2D image using CLIP preprocessing and resized it to be 224×224 to maintain consistency with the pretrained model's input requirements.

During training, we applied a random jitter of up to 5 mm to the nodule centroid. Using the TorchIO Python package [33], we performed random flipping (with 50% probability), affine transformations (with up to 10 degrees of rotation), the addition of Gaussian noise (mean = 0, standard deviation = 0.02), and contrast adjustments by raising voxel intensities to powers between -0.02 and 0.02.

3.3. Experimental Setup

3.3.1. Model Architecture and Loss Functions

The architecture of the model is illustrated in Figure 2a. The CLIP model was initialized using the weights from OpenAI's CLIP (ViT-B32) and was subsequently fine-tuned to align imaging and semantic features. The 2.5D images were processed through the image encoder, and an attention-based multiple instance learning model was employed to generate the attention scores and aggregate features across nine 2D images [34]. Simultaneously, we provided Gemini-generated radiology report-like text to the text encoder to obtain feature embeddings. Two projection heads were attached after each encoder to transform the features into 256-dimensional feature embeddings, which were aligned with two InfoNCE losses [35]:

$$\mathcal{L}_{\text{InfoNCE-Image}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp\left(\text{sim}(I_i, S_i)/\tau\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}(I_i, S_j)/\tau\right)}\right) \tag{1}$$

$$\mathcal{L}_{\text{InfoNCE-semantic}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp\left(\text{sim}(I_i, S_i)/\tau\right)}{\sum_{k=1}^{B}\exp\left(\text{sim}(I_k, S_i)/\tau\right)}\right) \tag{2}$$

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{\text{InfoNCE-Image}} + \mathcal{L}_{\text{InfoNCE-semantic}}) \tag{3}$$

where I_i and S_i represent the image deep features and the corresponding semantic features from nodule i, which we consider a positive pair. The negative pairs are imaging I_i and all other semantic features in the batches of size B, S_j, where $j \neq i$. In addition, negative pairs also include semantic features S_i and all other imaging I_k, where $k \neq i$. The cosine similarity (sim) is computed for both positive and negative pairs, and a softmax function is used to transform the similarity into probability, which indicates how likely each imaging and semantic feature is in the positive pair. The temperature, τ, controls the smoothness of the output probability. Rather than being manually tuned, the temperature is set as a learnable parameter and is optimized jointly with the model parameters during training.

This framework allows the image encoder to learn clinically meaningful features by incorporating semantic information, resulting in more robust imaging representations that focus on relevant regions rather than spurious patterns. However, limited training data can hinder effective alignment between modalities. To address this, we introduced two prediction branches, one for imaging features and one for semantic text features, to predict one-year lung cancer diagnoses, encouraging the model to learn diagnosis-relevant features while aligning both modalities. Each branch is trained using cross-entropy loss. Since our data is highly imbalanced, during training, class weighting was applied in the cross-entropy loss functions to assign a higher weight to malignant samples. In addition to the prediction branches, we utilized Low-Rank Adaptation (LoRA) , a parameter-efficient technique that resulted in only 0.4% of the trainable parameters, effectively preventing overfitting [36]. Specifically, low-rank matrices were incorporated into each query, key, and value layer of the vision and text transformer. The final weights were obtained by summing the pretrained weights with the low-rank matrices. The pretrained weights were kept frozen throughout training, and only the low-rank matrices were updated.

3.3.2. Model Training and Hyperparameter Settings

We selected 20% of cases (N = 188) from NLST as the held-out test set. The remaining nodules from NLST were combined with LIDC data for training. We performed 5-fold cross-validation with 80% and 20% split on the patient level. We selected the best hyperparameters based on the average Area Under the Receiver Operating Characteristic (AUROC) score across five validation sets. The best models were trained using three loss functions, including

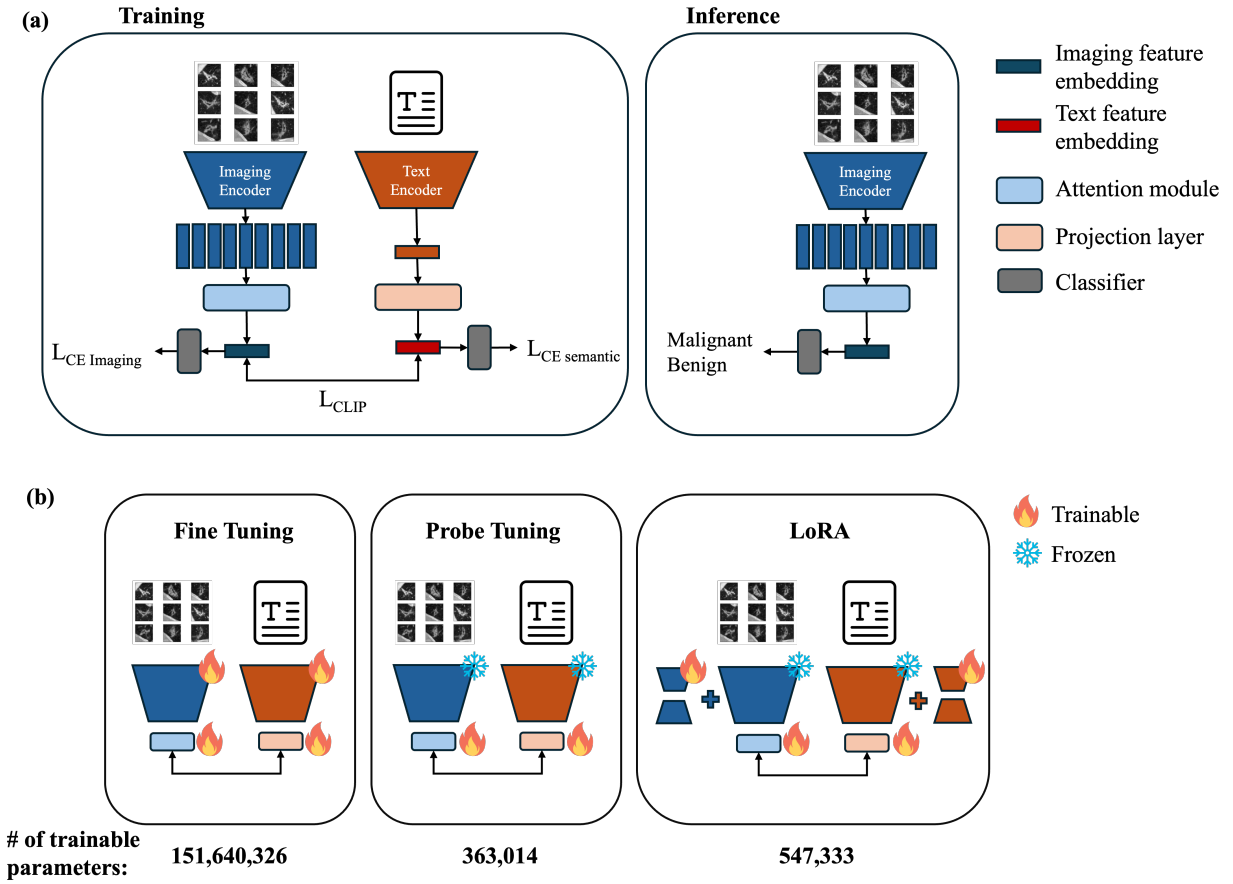


Figure 2: **CLIP Model Architecture and Fine-tuning Methods.** (a) The pretrained CLIP was fine-tuned to pull the paired imaging and semantic features close together, which allowed the model to learn meaningful relationships between imaging features and semantic features. During training, nodule images sliced from nine different directions of 3D nodule crops were passed into the vision transformer image encoder. The attention-based multiple-instance learning module aggregated the output imaging features to obtain a single embedding. The sentence containing the semantic features was passed into the text encoder to generate the text feature embedding. The visual and text features were then aligned using CLIP loss. Two prediction heads were independently attached after the encoders to predict the one-year lung cancer risk. During inference, only the imaging feature was required, allowing it to be applied universally without the need for a radiologist's evaluation on CT. (b) We examined the training of the CLIP model using three distinct tuning methods as part of our ablation study. All parameters were fine-tuned (left). Only two projection layers were fine-tuned (middle). A parameter-efficient fine-tuning method called LoRA, which involves inserting trainable low-rank matrices into each layer of the vision and text encoders, was used in our final model (right). While keeping the pretrained weights frozen, we updated the low-rank matrices. We also allowed the projection layers to be fine-tuned.

the CLIP loss in Equation 3 and two cross-entropy losses, all with equal weighting. Although the temperature in CLIP loss was learnable during training, we initially set it to 0.03. Optimization was performed using AdamW [37] with a learning rate of 0.0001 and a weight decay of 0.1. The model was trained with a batch size of 16. To diversify the nodules within a batch, we selected samples based on the frequency of semantic features, upsampling nodules

with rare features. In LoRA, we set the rank of the inserted matrices to 2, the scale factor to 1, and the dropout rate to 0.25.

Since each individual can have multiple nodules, we took the maximum of predicted probabilities to represent the patient-level risk score. We performed Beta calibration [38] to prevent the model from being overconfident and ensure that the probabilities reflect the true likelihoods. Uncalibrated and overconfident probabilities may lead to confusion in clinical interpretations and could potentially result in misleading clinical decisions [39]. Similar to Sybil, we adopted an ensemble approach for inference by averaging the calibrated predictions from all five trained models.

3.3.3. Hardware and Computation

Model training and inference were performed on an NVIDIA Quadro RTX 8000 GPU (48 GB memory). Inference averaged 0.011 ± 0.001 seconds per sample, allowing models from five folds to process a nodule in under one second. CPU inference was also efficient, averaging 0.095 ± 0.068 seconds per nodule, using an AMD Ryzen Threadripper 3970X (32 cores, 64 threads) with 256 GB RAM.

3.4. Evaluation

We compared our model to three SOTA lung cancer risk prediction models, Sybil, Venkadesh et al., and DeepIPN, which were previously introduced in Section 2. We evaluated the model's performance in predicting lung cancer risk within one year using the AUROC and the Area Under the Precision-Recall Curve (AUPRC). While AUROC reflects a model's overall ability to distinguish between benign and malignant nodules, AUPRC focuses on the precision-recall trade-off and offers a more informative evaluation for imbalanced datasets. This is particularly valuable in cancer screening, where correctly identifying the relatively few malignant cases among many benign ones is critical. In addition to point estimates, we computed 95% confidence intervals (CIs) by bootstrapping predicted probabilities 10,000 times. Given the importance of high sensitivity in lung cancer screening, we also evaluated model performance at higher recall levels. Since AUROC and AUPRC summarize performance across all thresholds, including those with low recall, we additionally reported false positive rate (FPR) and precision at recall levels closest to 0.6, 0.7, 0.8, and 0.9. This helps assess how well each model minimizes false discoveries while capturing more cancer cases.

Using CLIP, we also obtained predictions on semantic features for the NLST test set using zero-shot inference. For categorical semantic features, we inserted all possible words into the sentence "This nodule [margin/shape/consistency/...] is [...]". For binary semantic features, we utilized "There is [pleural retraction/cyst-like spaces/...]." to indicate presence and "No findings." to suggest absence. We calculated the cosine similarity between the features

generated from nodule images and each sentence. The softmax function was applied to the cosine similarity scores. Subsequently, we computed the weighted AUROC score to assess the effectiveness of the zero-shot inference for semantic features with multiple categories and the standard AUROC for those with binary categories.

3.5. Ablation Studies

We investigated three different tuning methods to train the CLIP model (Fig.2b). First, fine-tuning involved optimizing all parameters, but this approach can lead to overfitting. Second, in probe-tuning, we froze the image and text encoders while training only the projection layers. This can decrease the risk of overfitting, but it is less effective when there are significant domain shifts. For instance, we adapted a model originally trained on RGB natural images and captions to work with grayscale medical images and text similar to radiology reports. Third, with the original weights frozen, LoRA, described in Section 3.3.1, preserves the knowledge from the pretrained model while allowing adaptation to new domains using a small number of trainable parameters.

To compare the model performance with different modalities, we trained a logistic regression model with L1 regularization using the semantic features from the NLST training data and assessed its performance on the NLST test set. We also trained the model using only the CLIP vision encoder on the imaging data and evaluated the outcomes on the NLST test set and three external datasets.

We calculated the mean and standard deviation of the AUROC and AUPRC scores for each model across five folds. The 95% CIs were computed using the t-distribution based on the standard error across folds. To evaluate the significance of differences in performance among the three models, we conducted statistical testing utilizing the Friedman Chi-Square test. A post-hoc Nemenyi test was used to obtain the pairwise p-value. To compare two models, we applied the Wilcoxon signed-rank test at a significance level of 0.05.

4. Results

4.1. Lung Cancer Risk Prediction

We present the model performance of one-year lung cancer prediction with AUROC and AUPRC scores in Table 2. FPR and precision at recall levels approximating 0.6, 0.7, 0.8, and 0.9 are shown in Table 3. Since DeepIPN failed to generate predictions for some cases, we also report performance metrics excluding those cases to ensure a fairer comparison in Appendix D.

In the NLST test set, our model surpasses DeepIPN (AUROC: 0.862; AUPRC: 0.679), achieving an AUROC of 0.901 (95% CI: 0.843, 0.950) and an AUPRC of 0.776 (95% CI: 0.642,

Table 2: **Model Performance on Lung Cancer Prediction Within One Year.** We conducted a comparative analysis of our model against three SOTA lung cancer risk assessment models, evaluating performance through AUROC and AUPRC scores on a held-out test set, as well as on three external datasets, each encompassing various types of CT scans and patient cohorts. We also reported the 95% confidence interval (CI) computed by bootstrapping. Merics are displayed as a point estimate in the first row, followed by [95%CI lower bound, 95%CI upper bound] in the second row in each cell. Sybil and Venkadesh et al. models were not assessed on the NLST test set, as both models were trained on NLST. The performance of Venkadesh et al. on DLCS and UCLA could not be obtained due to issues with data privacy and platform credits. Since DeepIPN failed to process some cases across the datasets, the comparison here is not strictly equivalent. In Table A3, these cases were excluded, and the corresponding results are reported.

Models	**NLST Test Set** **NLST** **N=188 (43)**		**External Datasets** **LUNGx** **70 (35)**		**DLCS** **856 (94)**		**UCLA** **N = 51 (28)**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Imaging Only Models								
Sybil	-	-	0.662 [0.525,0.795]	0.669 [0.502,0.850]	0.797 [0.742,0.848]	0.468 [0.361,0.573]	0.734 [0.585,0.865]	0.780 [0.610,0.906]
Venkadesh et al.	-	-	0.684 [0.554,0.803]	0.702 [0.537,0.838]	-	-	-	-
Imaging + Semantic Features Models								
DeepIPN	0.862 [0.794,0.920]	0.679 [0.522,0.815]	0.709 [0.577,0.835]	0.658 [0.496,0.856]	0.851 [0.803,0.894]	**0.543** [0.435,0.645]	0.624 [0.458,0.779]	0.601 [0.435,0.821]
CLIP	**0.901** [0.843,0.950]	**0.776** [0.642,0.880]	**0.771** [0.652,0.880]	**0.813** [0.679,0.914]	**0.861** [0.820,0.897]	0.489 [0.382,0.595]	**0.778** [0.643,0.898]	**0.823** [0.661,0.935]

0.880). The CLIP model consistently shows favorable FPR and precision scores across four distinct recall levels.

When evaluated externally on the LUNGx, our model exhibits superior performance with an AUROC of 0.771 (95% CI: 0.652, 0.880) and an AUPRC of 0.813 (95% CI: 0.679, 0.914). We also observe lower FPR scores and higher precision scores at a recall level of 0.6, 0.7, and 0.9 for our model. In contrast, DeepIPN performs better at a recall of 0.8. The AUROC of our model on the nodule level is 0.769, which outperforms all the reported machine learning-based models in the challenge [27] (mean AUROC of 0.620, with a range of 0.500 to 0.680). The AUC values for the six radiologists range from 0.700 to 0.850 in the observer study. Our model outperforms two of the radiologists in the study but underperforms when compared to the other four.

In the UCLA never-smoker cohort, our model also achieves better performance with an AUROC of 0.778 (95% CI: 0.643, 0.898) and an AUPRC of 0.823 (95% CI: 0.661, 0.935). However, DeepIPN struggles with the UCLA dataset, with the lowest AUROC of 0.624 (95% CI: 0.458, 0.779) and AUPRC of 0.601 (95% CI: 0.435, 0.821), alongside the highest FPR and the lowest precision across all recall levels. Sybil demonstrates fairly strong performance, achieving the highest precision at recall levels of 0.6 and 0.9, while our model presents consistently better FPR and precision at recall levels of 0.7 and 0.8.

In the DLCS, our model achieves the highest AUROC of 0.861 (95% CI: 0.820, 0.897), although it has a less favorable AUPRC of 0.489 (95% CI: 0.382, 0.595). DeepIPN achieves

Table 3: **False Positive Rate and Precision at a Given Recall.** High recall is important in lung cancer detection, especially in screening programs, to avoid missing cancer cases. Metrics such as AUROC and AUPRC scores provide an overall evaluation across all thresholds but may not accurately reflect how the model performs in clinically relevant scenarios. This table presents the false positive rate (FPR) and precision at high recall, ranging from 0.6 to 0.9. Lower FPR and higher precision are more optimal. Since DeepIPN failed to process some cases across the datasets, the comparison here is not strictly equivalent. In Table A4, these cases were excluded, and the corresponding results are reported.

Models	**Recall ≈ 0.6**		**Recall ≈ 0.7**		**Recall ≈ 0.8**		**Recall ≈ 0.9**	
	FPR↓	**Precision↑**	**FPR↓**	**Precision↑**	**FPR↓**	**Precision↑**	**FPR↓**	**Precision↑**
NLST Test Set								
DeepIPN	0.070	0.667	0.148	0.588	0.190	0.486	0.430	0.386
Ours (CLIP)	**0.034**	**0.812**	**0.090**	**0.638**	**0.145**	**0.618**	**0.338**	**0.406**
LUNGx								
Sybil	0.229	0.724	0.543	0.568	0.743	0.519	0.914	0.500
Venkadesh et al.	0.371	0.618	0.571	0.568	0.600	0.571	0.771	**0.542**
DeepIPN	0.206	0.724	0.265	0.706	**0.471**	**0.596**	0.882	0.492
Ours (CLIP)	**0.086**	**0.875**	**0.171**	**0.774**	0.486	0.560	**0.743**	**0.542**
DLCS								
Sybil	0.136	0.354	0.227	0.273	0.444	0.187	0.580	0.153
DeepIPN	**0.097**	**0.424**	**0.145**	**0.361**	**0.224**	**0.301**	0.477	0.193
Ours (CLIP)	0.110	0.400	0.154	0.353	0.234	0.295	**0.379**	**0.203**
UCLA								
Sybil	**0.217**	**0.773**	0.391	0.690	0.609	0.622	0.652	**0.634**
DeepIPN	0.304	0.586	0.522	0.588	0.609	0.611	0.696	0.581
Ours (CLIP)	**0.217**	0.739	**0.261**	**0.769**	**0.261**	**0.733**	**0.565**	0.568

results comparable to those of our model, with a lower AUROC of 0.851 (95% CI: 0.803, 0.894) but a higher AUPRC of 0.543 (95% CI: 0.435, 0.645). In a more equitable comparison shown in Table A4, while DeepIPN has a lower FPR and higher precision at recall levels of 0.6 and 0.7, our model outperforms it at a recall level of 0.9.

4.2. Model Explainability

We employed the inherent feature of CLIP, zero-shot inference, to predict semantic features. Table 4 presents the performance metrics for general, internal, and external features. Our model demonstrates robust AUROC scores for general nodule features, including nodule margin (0.812), margin conspicuity (0.859), and consistency (0.812). Performance in predicting external features is fairly strong, with an AUROC of 0.747 for vascular convergence, 0.840 for pleural attachment, and 0.756 for paracicatrial emphysema. However, the model struggles to predict most internal features, achieving notable results only for cyst-like spaces (0.731) and eccentric calcification (0.794).

Table 4: **Semantic Features Prediction Through Zero-shot Inference.**We presented a subset of the semantic features and their respective classes. We assessed the performance of predicting semantic features by using weighted AUROC for cases with multiple classes and standard AUROC for those with binary elements. The total percentage of some features may not add up to 100% because of missing values. There can be more than one nodule margin annotated for one nodule. While the prediction shows solid performance in both general and external features, the internal features fail to provide satisfactory zero-shot inference results. This can be attributed to two main factors. First, internal features show a significant imbalance in their presence and absence. Second, the alignment between two modalities may be dominated by semantic features indicating malignancy.

Semantic Features	**Classes (%)**	**AUROC**
General Features		
Nodule Margin	Smooth (67); Lobulated (22); Spiculated (24); Ill-defined (19)	0.812
Nodule Consistency	Peri-cystic (2); Solid (73); Pure ground glass (9); Semiconsolidation (8); Part-solid (8)	0.812
Nodule Shape	Irregular (31); Ovoid (35); Polygonal (14); Round (19)	0.670
Nodule Margin Conspicuity	Well marginated (84); Poorly marginated (16)	0.859
Internal Features		
Nodule Reticulation	Present (88); Absent (12)	0.411
Cyst-like Spaces	Present (14); Absent (86)	0.731
Necrosis	Present (0.4); Absent (99.6)	0.112
Eccentric Calcification	Present (3); Absent (96)	0.794
Cavitation	Present (0.4); Absent (99.6)	0.620
Intra-nodular bronchiectasis	Present (3); Absent (97)	0.425
Airway Cutoff	Present (2); Absent (27)	0.387
External Features		
Vascular Convergence	Present (10); Absent (89)	0.747
Pleural Retraction	Present (80) - Mild and Obvious Dimpling; Absent (20)	0.689
Pleural Attachment	Present (53); Absent (47)	0.840
Paracicatricial Emphysema	Present (12); Absent (88)	0.756
Septal Stretching	Present (69); Absent (30)	0.670

4.3. Ablation Studies

In Table 5, we present the mean and standard deviation of AUROC and AUPRC across five folds for models trained with various tuning methods and modalities. The CLIP model trained with LoRA performs better in predicting lung cancer risk than one trained with fine-tuning and probe-tuning. Furthermore, in the NLST test set, with only the semantic features, the model achieves a mean AUROC of 0.888 (95% CI: 0.886, 0.890), which is comparable to our model, and a mean AUPRC of 0.889 (95% CI: 0.876, 0.902), which is the highest among the three trained models. The CLIP model, trained with both semantic and imaging features, outperforms the CLIP vision encoder trained with imaging data only across all datasets.

Table 5: **Performance of Ablation Studies.** The first three rows display the outcome of the CLIP model, which was trained using various tuning methods. The last three rows display the performance of models trained on different modalities. In each cell, mean and standard deviation of AUROC scores across five-fold models were presented in the first row. In the second row, the 95% confidence intervals were computed using the t-distribution based on the standard error across folds. The * superscript denotes a statistically significant difference in performance compared to the best-performing model. P-values are shown in Table A5. Abbreviation: FT - fine-tuning; PT - probe-tuning; LORA - low-rank adaptation; S - logistic regression model trained on semantic features; V - CLIP vision encoder trained with imaging only; CLIP - contrastive language–image pre-training.

	NLST Test Set **NLST** **N=188 (43)**		**External Datasets** **LUNGx** **70 (35)**		**DLCS** **856 (94)**		**UCLA** **N = 51 (28)**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Training With Different Tuning Methods								
FT	0.809 ± 0.012* [0.793,0.825]	0.533 ± 0.050* [0.463,0.603]	0.656 ± 0.040* [0.600,0.712]	0.638 ± 0.060* [0.555,0.722]	0.798 ± 0.014* [0.778,0.818]	0.328 ± 0.018* [0.302,0.353]	0.576 ± 0.038* [0.523,0.629]	0.602 ± 0.052* [0.529,0.675]
PT	0.839 ± 0.018 [0.814,0.863]	0.634 ± 0.026 [0.599,0.670]	0.733 ± 0.016 [0.710,0.755]	0.728 ± 0.031 [0.684,0.771]	0.836 ± 0.002 [0.832,0.839]	0.423 ± 0.009 [0.410,0.436]	0.722 ± 0.017 [0.698,0.746]	0.753 ± 0.007 [0.743,0.762]
LoRA	**0.889** ± 0.009 [0.876,0.902]	**0.757** ± 0.017 [0.733,0.781]	**0.763** ± 0.013 [0.745,0.782]	**0.801** ± 0.017 [0.778,0.825]	**0.852** ± 0.010 [0.838,0.866]	**0.461** ± 0.041 [0.404,0.517]	**0.760** ± 0.041 [0.702,0.817]	**0.805** ± 0.035 [0.756,0.854]
Training With Different Modalities								
S	0.888 ± 0.002 [0.886,0.890]	**0.780** ± 0.009 [0.772,0.795]	-	-	-	-	-	-
V	0.812 ± 0.032* [0.768,0.857]	0.587 ± 0.046* [0.523,0.651]	0.685 ± 0.036 [0.635,0.735]	0.718 ± 0.034 [0.671,0.765]	0.798 ± 0.014 [0.778,0.817]	0.385 ± 0.036 [0.335,0.436]	0.685 ± 0.046 [0.621,0.750]	0.709 ± 0.037 [0.657,0.760]
CLIP	**0.889** ± 0.009 [0.876,0.902]	0.757 ± 0.017 [0.733,0.781]	**0.763** ± 0.013 [0.745,0.782]	**0.801** ± 0.017 [0.778,0.825]	**0.852** ± 0.010 [0.838,0.866]	**0.461** ± 0.041 [0.404,0.517]	**0.760** ± 0.041 [0.702,0.817]	**0.805** ± 0.035 [0.756,0.854]

4.4. Error Analysis

In Figure 3, we present a CT scan from a lung cancer patient and a non-lung cancer patient for each dataset. More examples are shown in Figure A1. It is important to note that both Sybil and DeepIPN use only the CT scan as input, whereas our model and that of Venkadesh et al. incorporate both the CT scan and the annotated nodule location as inputs. Although the pipeline in DeepIPN involves a nodule detection step, the detected nodules may not correspond to the annotated nodules shown in the figure.

In Figure 3a and c, all models fail to identify the patient as high-risk even though they are diagnosed with lung cancer within a year. In Figure 3a, the nodule crop shows a poorly marginated, pure ground glass nodule with an irregular margin. The malignant nodule in Figure 3c is spiculated with complex shape, septal stretching, and vascular convergence. Our model outputs relatively higher risk scores of 0.306 and 0.443, in contrast to the extremely low scores given by other models. Figure 3e illustrates a challenging and uncommon case where the nodule is within the right main stem bronchus. All models assign a relatively lower score while the nodule is malignant. In Figure 3g, a poorly marginated pure ground glass nodule is presented. Among the models, only DeepIPN successfully assigns a high risk score to this malignant case, while the others fail.

Figure 3b shows a non-cancerous case with a lobulated and serrated nodule exhibiting complex shape and pleural retraction. While DeepIPN correctly identifies the case as low

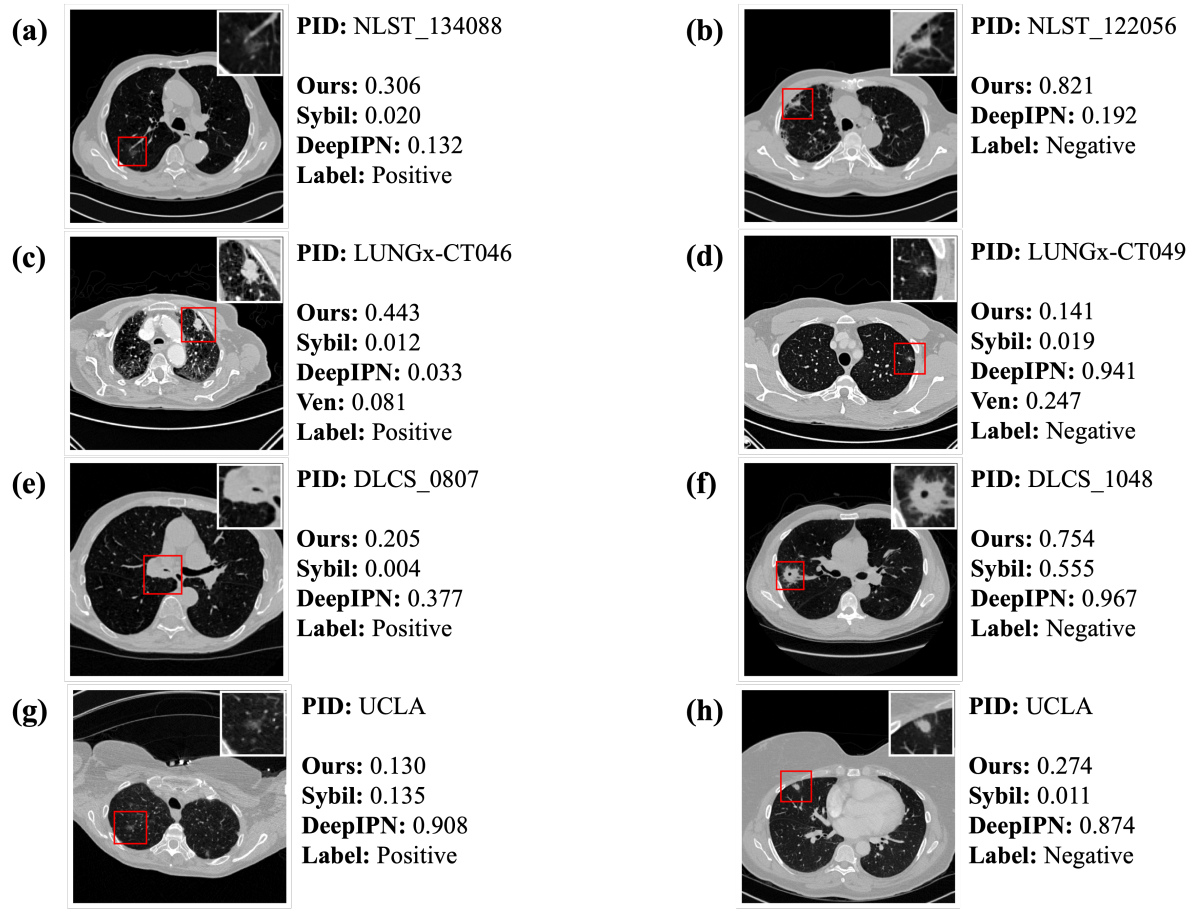


Figure 3: **Error Analysis.** One lung cancer and one non-lung cancer case from each dataset are presented. We present the CT scan slice corresponding to the middle of the nodule, highlighted with a red bounding box. In the top right corner, we place the magnified view of the nodule for clearer visualization. Certain features may not be fully appreciable in the single slice but are visible when viewing the whole series. Patient-level risk score from each model is shown beside each image. The scale and range of the predicted probability across different models can vary significantly. The probabilities from Sybil, Venkadesh et al., and our models have been calibrated. Therefore, the predicted probability from these three models matches the observed frequency. Sybil risk scores for NLST are listed only if the case is included in Sybil's test data.

risk, our model assigns a high risk score of 0.821. In contrast, Figures 3d and h illustrate two negative cases where DeepIPN assigns unusually high risk scores, whereas our model and others provide more conservative estimates. This discrepancy may stem from DeepIPN's miscalibration, leading to overconfident predictions. Additionally, visual inspection revealed that DeepIPN sometimes detects non-nodule regions, which may contribute to inflated risk scores. In Figure 3f, although the biopsy confirmed a benign finding, the nodule appears highly suspicious due to its large size, irregular shape, and spiculated margins. All models, including ours, assign high risk scores to this case. Radiologists also classified it as Lung-RADS 4X, indicating a high suspicion of malignancy with additional concerning imaging features [40].

5. Discussion

In this study, we trained a CLIP model to align CT imaging features with semantic features to facilitate better lung cancer prediction. We preprocessed both image and semantic features to ensure alignment with the pretrained CLIP format. To tackle the challenge of fine-tuning with a small dataset, we integrated LoRA and supervised branches into CLIP. Our model outperforms SOTA models, showing consistent robustness across external datasets.

In the test and external datasets, we include a variety of CT scans collected in different scenarios. The performance of Sybil and Venkadesh et al. noticeably decreases when applied to the LUNGx dataset, which includes diagnostic and contrast-enhanced CT scans. This decrease can be attributed to the fact that both models were trained on the NLST dataset, which only contains LDCT, and the resulting distribution shift adversely impacted their performance. However, our model remains robust even though we did not explicitly train on contrast-enhanced CT. Additionally, our model exhibits superior performance on the UCLA dataset, which comprises incidental nodules from non-smoking patients. Nevertheless, the sample size in the external dataset may be limited, and the range of the 95% CI is quite large. Our model achieves the highest performance in terms of AUROC on the DLCS dataset, but yields a lower overall AUPRC. One potential cause of suboptimal results of our model in DLCS is that, although radiologists review the detected nodules from the algorithm, some still do not correspond to true nodules. As our model was trained exclusively on radiologists-verified nodules, its performance can be unstable when evaluated on regions that do not contain confirmed nodules.

There are several benefits of incorporating semantic features into training. First, the use of semantic features during training guides imaging-based models to learn clinically meaningful patterns of nodules, rather than relying on spurious correlations. In a previous study [41], we investigated the characteristics and robustness of various imaging features by selecting prompts based on feature similarity for nodule segmentation in generalist foundation models. We extracted imaging features from our CLIP model and compared them against pixel intensity, radiomics features, and imaging biomarker foundation features for prompt selection. The results demonstrated that CLIP imaging features are more effective in focusing on relevant nodule characteristics, which led to superior segmentation accuracy and robustness on external datasets using generalist segmentation models. Second, even though the model trained exclusively on semantic features yields comparable or surpassing results to CLIP, using semantic features for prediction is not suitable for clinical use due to subjectivity and variability in semantic feature annotations among radiologists, and the added burden of requiring detailed semantic assessments from radiologists. On the other hand, even though the semantic features were utilized during training in CLIP, our CLIP model only requires

CT images as input data at the inference stage, with potential applications across more diverse clinical settings. Third, incorporating domain knowledge about nodule characteristics makes the training process more data-efficient. Our model achieves performance comparable to or surpassing imaging-based SOTA models despite being trained on limited data. Our model utilized approximately 2,000 cases, whereas Sybil and Venkadesh et al. were trained on approximately 15,000 and 10,000 cases, respectively.

The zero-shot inference capability of the CLIP model improves model explainability, allowing our model to predict semantic features without requiring explicit training. This capability can help clinicians understand the underlying meaning of model predictions. Our model achieves high performance in predicting various general nodule characteristics and external features, but predictions for other features, like nodule shape and most internal features, did not perform well. First, as shown in Table 4, most of the internal features exhibit a significant imbalance in the presence and absence cases. For example, only 0.4% of the nodules present with necrosis, and 2% of nodules have an airway cutoff. Despite upsampling cases with rare features, the limited number of such instances may result in underrepresentation for certain subgroups and limit the model's effectiveness in differentiating such features. Second, not all features contribute equally to predicting nodule malignancy. It is possible that general and external features are more informative than internal features. In addition, a well-known limitation of CLIP, its lack of fine-grained alignment, could cause the variability in zero-shot semantic feature prediction performance. Imaging and semantic features are converted into a single embedding and subsequently aligned. Therefore, some semantic features can overshadow others during the training process. In the future, our goal is to develop and validate the model further to facilitate more granular alignment.

While we used Gemini to convert tabular semantic features into radiology report–like text, we did not validate the generated outputs against actual radiology reports. However, our primary goal is not to replicate exact clinical language, but to show the promise of using the CLIP model to effectively align text describing nodule characteristics with nodule images for improved lung cancer diagnosis. Future work could incorporate millions of institution-specific radiology reports, containing detailed descriptions of nodule characteristics, to further enhance the training of CLIP-based foundation models. In addition, while we extracted nodules from nine distinct directions to preserve 3D characteristics to the greatest extent possible, certain features could still be overlooked during this process. Moreover, since our model focused exclusively on the nodule region, we did not include broader lung features like fibrosis and emphysema, which could also serve as risk factors for lung cancer. Furthermore, we did not evaluate the model developed by Venkadesh et al. on the DLCS and UCLA datasets due to limitations of running the model and sharing local data in a

public cloud environment.

6. Conclusion

We presented the benefits of integrating semantic features using the CLIP model, which allows the imaging encoder to learn more clinically significant and robust features. This model has achieved outstanding results in predicting lung cancer risk across external datasets with various patient cohorts and CT scan types. Additionally, our model provides explainability regarding nodule characteristics through zero-shot inference, equipping clinicians with better insights into how the model makes its predictions. Although the model was trained with semantic features, our model does not require manual annotation from radiologists, which enhances its scalability in diverse clinical settings.

CRediT authorship contribution statement

Luoting Zhuang: Writing - Original Draft, Writing - Review & Editing, Data Curation, Software, Methodology, Formal analysis, Visualization, Validation, Investigation, Conceptualization. **Seyed Mohammad Hossein Tabatabaei:** Data Curation, Writing – review & editing. **Ramin Salehi-Rad:** Data Curation, Writing – review & editing. **Linh M. Tran:** Data Curation, Writing – review & editing. **Denise Aberle:** Data Curation, Writing – review & editing, Funding acquisition, Conceptualization. **Ashley Prosper:** Data Curation, Writing – review & editing, Conceptualization. **William Hsu:** Data Curation, Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization, Resources.

Declaration of competing interest

William Hsu reports funding support from the National Institutes of Health, Agency for Healthcare Research and Quality, Early Diagnostics Inc, personal fees from the Radiological Society of North America related to editorial board work, and consulting fees from LungLife AI, Inc. Ashley E. Prosper report funding support from the National Institutes of Health. Linh M. Tran reports funding support from the Department of Veterans Affairs Merit Review. Denise R. Aberle reports funding support from the National Institutes of Health. If there are other authors, they 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

We gratefully acknowledge the support of NIH/National Cancer Institute U2C CA271898 (to RSR, LMT, DA, AP, WH), U01 CA233370 (to LZ, DA, AP, WH), the V Foundation (to DA, WH), and the Department of Veterans Affairs Merit Review I01BX005721 (to LMT). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study. LUNGx data used in this research were obtained from The Cancer Imaging Archive (TCIA) sponsored by the SPIE, NCI/NIH, AAPM and The University of Chicago.

Appendix A. Semantic Features Annotated

Table A1: Semantic Features Annotated by Radiologists for NLST Data.

Feature Type	**Feature**	**Elements Contained**
General features	longest axial diameter	3.7-62 mm
General features	short diameter	2-54 mm
General features	nodule consistency	Solid; Pure ground glass; Semiconsolidation; Part-solid; Peri-cystic
General features	nodule margin conspicuity	Well marginated; Poorly marginated
General features	nodule margins	Spiculated/Serrated; Smooth; Lobulated; Ill-defined; Notched
General features	nodule shape	Ovoid; Round; Complex/Irregular; Polygonal
Internal features	nodule reticulation	Present; Absent
Internal features	cyst-like spaces	Present; Absent
Internal features	intra-nodular bronchiectasis	Present; Absent
Internal features	necrosis	Present; Absent
Internal features	cavitation	Present; Absent
Internal features	eccentric calcification	Present; Absent
Internal features	airway cut-off	Present; Absent
External features	pleural attachment	Present; Absent
External features	pleural retraction	Absent; Mild dimpling; Obvious dimpling
External features	vascular convergence	Present; Absent
External features	septal stretching	Present; Absent
External features	paracicatricial emphysema	Present; Absent
General Assessment	level of suspicion of lung cancer	Very Low; Moderately Low; Intermediate; Moderately High; High

Appendix B. Text Generation from Tabular Data Using Gemini

Prompt:

> Think as if you are a radiologist, you have a table/dictionary for lung nodule evaluation, and you want to transform it into radiology reports. Here is the dictionary . Just show me the report. Do not add any additional information. In English! You should have an impression and findings part. Findings are bullet points listing each of the features separately. These features should be in random order. Combine nodule margins and additional nodule margins. If there is a missing value, just ignore it. For impression, do a quick summary of the findings in sentence and then state the suspiciousness of lung cancer. You can mention the absences in the findings but do not mention those in the impression.

Example:

Findings:
- Axial location: Central
- Longest axial diameter (mm): 17.0
- Short diameter (mm): 5.1
- Nodule margins: Ill defined, Spiculated, Serrated
- Nodule shape: Complex, Irregular
- Nodule consistency: Part solid
- Nodule reticulation: Present
- Cyst like spaces: Absent
- Intra nodular bronchiectasis: Absent
- Necrosis: Absent
- Cavitation: Absent
- Eccentric calcification: Absent
- Airway cut off: Absent
- Pleural attachment: Present
- Pleural retraction: Mild dimpling
- Vascular convergence: Present
- Septal stretching: Present
- Paracicatricial emphysema: Absent

Impression:
A 17.0 x 5.1 mm, central, part solid nodule is identified. The nodule demonstrates complex, spiculated margins and is associated with pleural attachment, mild dimpling, vascular convergence, and septal stretching. The level of suspicion for lung cancer is moderately high.

Appendix C. LIDC and NLST Semantic Features Harmonization

Table A2: **Harmonization of LIDC and NLST semantic features.** We consulted radiologists at our institution to map the semantic features in LIDC to the corresponding semantic features in our internal NLST database. Semantic features present in our in-house data but absent in LIDC are considered missing and were excluded during text generation.

LIDC Semantic Features	**Radiologists' Entry**	**Mapped to NLST Semantic Features**
Internal Structure	"Air"	Cyst-like spaces = "Present"
	All Others	Cyst-like spaces = "Absent"
Calcification	"Non central appearance"	Eccentric Calcification == "Present"
	All Others	Eccentric Calcification == "Absent"
Sphericity	> 3	Nodule Shape = "Round"
	≤ 3 4	Nodule Shape = "Ovoid"
Margin	≥ 3	Nodule Margin Consipicuity = "Well marginated"
	< 3 4	Nodule Margin Consipicuity = "Poorly marginated"
Lobulation	≥ 3	Nodule Margins = "Lobulated"
Spiculation	≥ 3	Nodule Margins = "Spiculated"
Texture	> 4	Nodule Consistency = "Solid"
	$= 2, 3, 4$	Nodule Consistency = "Part-solid"
	< 2	Nodule Consistency = "Pure ground glass"

Appendix D. Model Performance Removing Cases Failed in DeepIPN

To ensure fair comparison with DeepIPN, which failed to predict lung cancer risk for certain cases, we removed the failed cases and reported AUROC and AUPRC scores in Table A3 and FDR and Precision at different recall levels in Table A4.

Table A3: **Model Performance on Lung Cancer Prediction Within One Year.**

Models	NLST Test Set NLST N=186 (43)		External Datasets LUNGx 69 (35)		DLCS 838 (94)		UCLA N = 51 (28)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Imaging Only Models								
Sybil	-	-	0.660 [0.523,0.790]	0.670 [0.503,0.850]	0.798 [0.743,0.849]	0.472 [0.365,0.576]	0.734 [0.585,0.865]	0.780 [0.610,0.906]
Venkadesh et al.	-	-	0.675 [0.543,0.800]	0.703 [0.538,0.840]	-	-	-	-
Imaging + Semantic Features Models								
DeepIPN	0.862 [0.794,0.920]	0.679 [0.522,0.815]	0.709 [0.577,0.835]	0.658 [0.496,0.856]	0.851 [0.803,0.894]	**0.543** [0.436,0.645]	0.624 [0.458,0.779]	0.601 [0.435,0.821]
CLIP	**0.903** [0.845,0.951]	**0.779** [0.638,0.884]	**0.764** [0.641,0.876]	**0.814** [0.680,0.912]	**0.862** [0.821,0.899]	0.498 [0.389,0.604]	**0.778** [0.643,0.898]	**0.823** [0.661,0.935]

Table A4: **False Positive Rate and Precision at a Given Recall.**

Models	Recall ≈ 0.6		Recall ≈ 0.7		Recall ≈ 0.8		Recall ≈ 0.9	
	FPR↓	Precision↑	FPR↓	Precision↑	FPR↓	Precision↑	FPR↓	Precision↑
NLST Test Set								
DeepIPN	0.070	0.667	0.148	0.588	0.190	0.486	0.430	0.386
Ours (CLIP)	**0.035**	**0.812**	**0.092**	**0.638**	**0.141**	**0.630**	**0.331**	**0.419**
LUNGx								
Sybil	0.235	0.724	0.294	0.615	0.735	0.483	0.912	0.500
Venkadesh et al.	0.382	0.600	0.441	0.571	0.618	0.560	**0.676**	**0.542**
DeepIPN	0.206	0.724	0.265	0.706	**0.471**	**0.596**	0.882	0.492
Ours (CLIP)	**0.088**	**0.875**	**0.176**	**0.774**	0.500	0.560	0.765	**0.542**
DLCS								
Sybil	0.132	0.368	0.223	0.281	0.438	0.193	0.578	0.157
DeepIPN	**0.097**	**0.424**	**0.145**	0.361	**0.224**	0.301	0.477	0.193
Ours (CLIP)	0.108	0.412	0.151	**0.363**	0.228	**0.305**	**0.378**	**0.208**
UCLA								
Sybil	**0.217**	**0.773**	0.391	0.690	0.609	0.622	0.652	**0.634**
DeepIPN	0.304	0.586	0.522	0.588	0.609	0.611	0.696	0.581
Ours (CLIP)	**0.217**	0.739	**0.261**	**0.769**	**0.261**	**0.733**	**0.565**	0.568

Appendix E. Result of Ablation Studies: P-values

Table A5: **Hypothesis Testing P-values in Ablation Studies.** This table shows the p-values for hypothesis testing performed in ablation studies. The * superscript denotes a statistically significant difference at a significance level of 0.05. To compare more than three models. We used the Friedman Chi-Square test with a post-hoc Nemenyi test. For comparing two models, we conducted the Wilcoxon signed-rank test. It is important to note that since we only have five samples, one from each fold, the lowest p-value for the Wilcoxon signed-rank test is 0.063, so it will never yield a significant result. Abbreviation: F - fine-tuning; P - probe-tuning; L - low-rank adaptation; S - logistic regression model trained on semantic features; V - CLIP vision encoder trained with imaging only; CLIP - contrastive language–image pre-training.

	NLST Test Set NLST N=188 (43)		**LUNGx 70 (35)**		**External Datasets DLCS 856 (94)**		**UCLA N = 51 (28)**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
CLIP Training With Different Tuning Methods								
F vs P	0.609	0.254	0.139	0.254	0.139	0.139	0.139	0.139
F vs L	0.012*	0.004*	0.012*	0.004*	0.012*	0.012*	0.012*	0.012*
P vs L	0.139	0.254	0.609	0.254	0.609	0.609	0.609	0.609
Training With Different Modalities								
S vs V	0.069	0.012*	-	-	-	-	-	-
S vs CLIP	0.946	0.609	-	-	-	-	-	-
V vs CLIP	0.031*	0.139	0.063	0.063	0.063	0.125	0.063	0.063

Appendix F. Error Analysis: More Examples

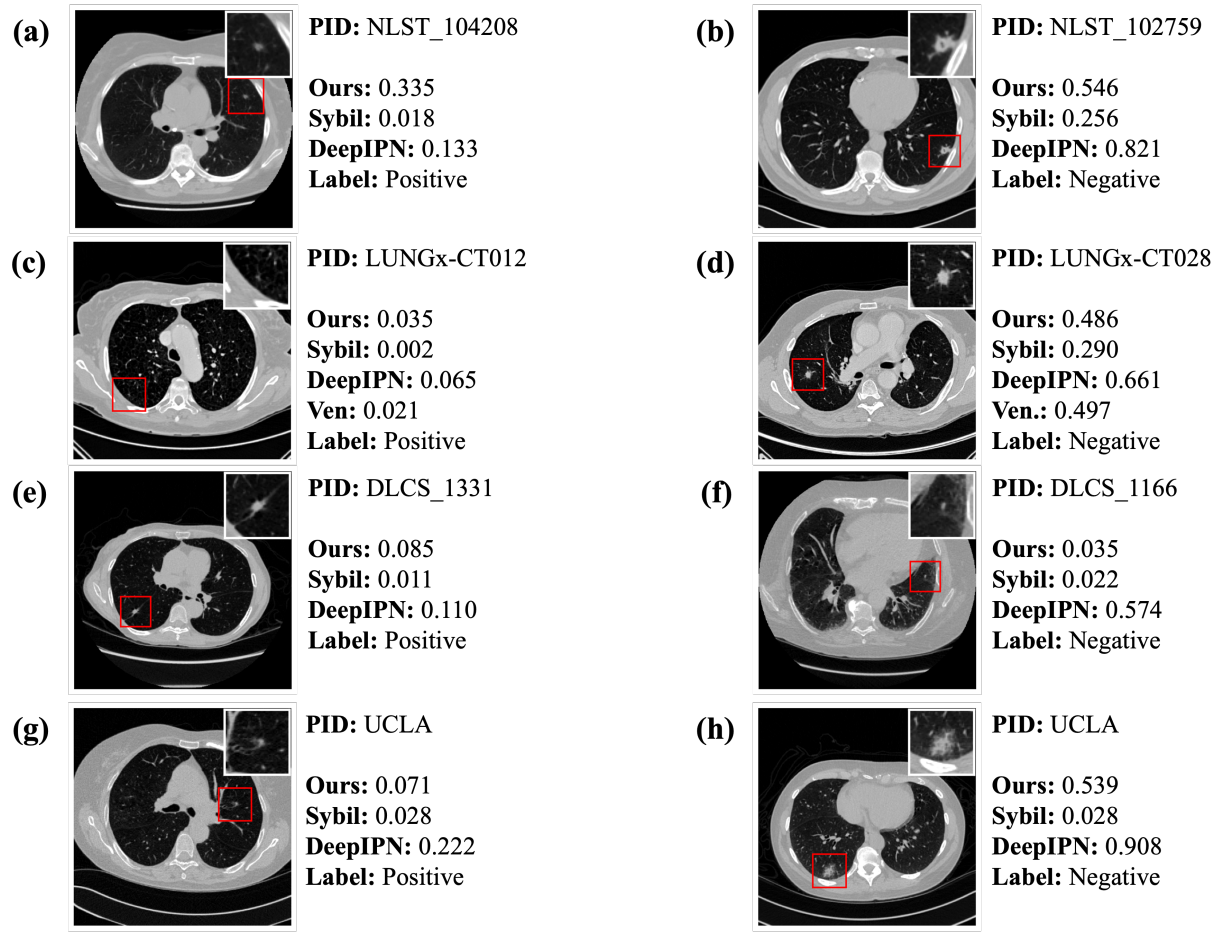


Figure A1: **Error Analysis (Extended).** One lung cancer and one non-lung cancer case from each dataset are presented. We present the CT scan slice corresponding to the middle of the nodule, highlighted with a red bounding box and magnified in the top right corner. Certain features may not be fully appreciable in the single slice but are visible when viewing the whole series. Patient-level risk scores are shown beside images.

Figure A1c, e, and g depict nodules diagnosed as lung cancer, but our model underestimates the risk. Although these nodules are relatively small, they exhibit irregular shapes (c), spiculated margins (e), mixed densities (g), and pleural attachment (c, e, g). Interestingly, all other models also predict low scores for these nodules. In Figure A1a, a lobulated, solid, well-marginated nodule is shown with few suspicious features. It is reasonable that our model predicts a 33% likelihood of malignancy, but all other models assign extremely low scores to this cancer patient.

In Figure A1b, the nodule is solid with spiculated and lobulated margins and an irregular shape. Concerning characteristics include cyst-like spaces and pleural retraction. In Figure A1d, the nodule appears to be solid with spiculated margins, septal stretching, and vascular convergence. In Figure A1h, a large, part-solid nodule with poorly defined margins is

observed. Given these suspicious features, most models assign relatively higher risk scores, despite the patient not being diagnosed with lung cancer. However, in Figure A1h, Sybil accurately identifies it as having a lower score. In contrast, Figure A1f shows a very small nodule with no apparent suspicious features, but DeepIPN assigns a comparatively higher risk. We found that the nodules detected by DeepIPN appear to be fibrosis and vessels.

References

[1] J. A. Barta, C. A. Powell, J. P. Wisnivesky, Global epidemiology of lung cancer, Annals of global health 85 (1) (2019) 8.

[2] C. S. D. Cruz, L. T. Tanoue, R. A. Matthay, Lung cancer: epidemiology, etiology, and prevention, Clinics in chest medicine 32 (4) (2011) 10–1016.

[3] N. L. S. T. R. Team, Reduced lung-cancer mortality with low-dose computed tomographic screening, New England Journal of Medicine 365 (5) (2011) 395–409.

[4] H. J. de Koning, C. M. van Der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, et al., Reduced lung-cancer mortality with volume ct screening in a randomized trial, New England journal of medicine 382 (6) (2020) 503–513.

[5] W. Wu, L. A. Pierce, Y. Zhang, S. N. Pipavath, T. W. Randolph, K. J. Lastwika, P. D. Lampe, A. M. Houghton, H. Liu, L. Xia, et al., Comparison of prediction models with radiological semantic features and radiomics in lung cancer diagnosis of the pulmonary nodules: a case-control study, European radiology 29 (2019) 6100–6108.

[6] U. Bashir, B. Kawa, M. Siddique, S. M. Mak, A. Nair, E. Mclean, A. Bille, V. Goh, G. Cook, Non-invasive classification of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features, The British journal of radiology 92 (1099) (2019) 20190159.

[7] P. G. Mikhael, J. Wohlwend, A. Yala, L. Karstens, J. Xiang, A. K. Takigami, P. P. Bourgouin, P. Chan, S. Mrah, W. Amayri, et al., Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography, Journal of Clinical Oncology 41 (12) (2023) 2191–2200.

[8] K. V. Venkadesh, A. A. Setio, A. Schreuder, E. T. Scholten, K. Chung, M. M. W. Wille, Z. Saghir, B. van Ginneken, M. Prokop, C. Jacobs, Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening ct, Radiology 300 (2) (2021) 438–447.

[9] L. Zhuang, A. Yadav, G. H. Kim, S. M. H. Tabatabaei, A. Prosper, W. Hsu, Exploring the impact of acquisition and reconstruction parameters on an imaging-based lung cancer risk model, in: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2024, pp. 1–5.

[10] N. Emaminejad, M. W. Wahi-Anwar, G. H. J. Kim, W. Hsu, M. Brown, M. McNitt-Gray, Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between ct acquisition and reconstruction parameters, Medical physics 48 (6) (2021) 2906–2919.

[11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (11) (2020) 665–673.

[12] A. J. DeGrave, J. D. Janizek, S.-I. Lee, Ai for radiographic covid-19 detection selects shortcuts over signal, Nature Machine Intelligence 3 (7) (2021) 610–619.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[14] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, W. Hsu, Explainable hierarchical semantic convolutional neural network for lung cancer diagnosis., in: CVPR workshops, 2019, pp. 63–66.

[15] R. Gao, T. Li, Y. Tang, K. Xu, M. Khan, M. Kammer, S. L. Antic, S. Deppen, Y. Huo, T. A. Lasko, et al., Reducing uncertainty in cancer risk estimation for patients with indeterminate pulmonary nodules using an integrated deep learning model, Computers in biology and medicine 150 (2022) 106113.

[16] L. Liu, Q. Dou, H. Chen, J. Qin, P.-A. Heng, Multi-task deep model with margin ranking loss for lung nodule analysis, IEEE transactions on medical imaging 39 (3) (2019) 718–728.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[18] M. Christensen, M. Vukadinovic, N. Yuan, D. Ouyang, Vision–language foundation model for echocardiogram interpretation, Nature Medicine 30 (5) (2024) 1481–1488.

[19] X. Zhang, C. Wu, Y. Zhang, W. Xie, Y. Wang, Knowledge-enhanced visual-language pre-training on chest radiology images, Nature Communications 14 (1) (2023) 4542.

[20] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature medicine 29 (9) (2023) 2307–2316.

[21] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, O. F. Durugol, B. Wittmann, T. Amiranashvili, et al., Developing generalist foundation models from a multimodal dataset for 3d computed tomography, arXiv preprint arXiv:2403.17834 (2024).

[22] Y. Lei, Z. Li, Y. Shen, J. Zhang, H. Shan, Clip-lung: Textual knowledge-guided lung nodule malignancy prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 403–412.

[23] N. L. S. T. R. Team, Data from the national lung screening trial (nlst) (2013). doi:10.7937/TCIA.HMQ8-J677.
URL https://doi.org/10.7937/TCIA.HMQ8-J677

[24] D. Aberle, Challenges in the semantic annotation of indeterminate nodules, in: Society of Thoracic Radiology Annual Meeting, San Diego, CA, 2025, keynote Speech.

[25] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. 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) (2011) 915–931.

[26] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Van Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, L. P. Clarke, Data from lidc-idri (2015). doi:10.7937/K9/TCIA.2015.LO9QL9SX.
URL https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

[27] S. G. Armato III, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby, et al., Lungx challenge for computerized lung nodule classification, Journal of Medical Imaging 3 (4) (2016) 044506–044506.

[28] S. G. Armato III, L. Hadjiiski, G. D. Tourassi, K. Drukker, M. L. Giger, F. Li, G. Redmond, K. Farahani, J. S. Kirby, L. P. Clarke, Spie-aapm-nci lung nodule classification challenge dataset (2015). doi:10.7937/K9/TCIA.2015.UZLSU3FL.
URL https://doi.org/10.7937/K9/TCIA.2015.UZLSU3FL

[29] A. J. Wang, F. I. Tushar, M. R. Harowicz, B. C. Tong, K. J. Lafata, T. D. Tailor, J. Y. Lo, The duke lung cancer screening (dlcs) dataset: a reference dataset of annotated low-dose screening thoracic ct, Radiology: Artificial Intelligence 7 (4) (2025) e240248.

[30] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[31] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint arXiv:2403.05530 (2024).

[32] D. Joshi, A. Shinde, S. Das, O. Deokar, D. Shetiya, S. Jagtap, Text data augmentation, in: 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), IEEE, 2023, pp. 392–396.

[33] F. Pérez-García, R. Sparks, S. Ourselin, Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning, Computer methods and programs in biomedicine 208 (2021) 106236.

[34] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International conference on machine learning, PMLR, 2018, pp. 2127–2136.

[35] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2) (2022) 3.

[37] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[38] M. Kull, T. Silva Filho, P. Flach, Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers, in: Artificial intelligence and statistics, PMLR, 2017, pp. 623–631.

[39] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.

[40] A. C. of Radiology, et al., Lung ct screening reporting and data system (lung-rads) (2014).

[41] L. Zhuang, S. M. H. Tabatabaei, D. R. Aberle, A. E. Prosper, W. Hsu, Comparing the characteristics and robustness of imaging features via prompt selection in generalist segmentation models, in: Proc. of SPIE Vol, Vol. 13411, 2025, pp. 134110Z–1.