Multimodal Fusion of 3D CNN Features and Radiologist Attributes for Lung Nodule Malignancy Classification

B.-A. Banescu, L. Ichim, *Member, IEEE*, and D. Popescu, *Member, IEEE*

Abstract— **Lung cancer remains one of the leading causes of cancer-related mortality worldwide, making early detection of malignant pulmonary nodules essential for improving patient outcomes. Recent advances in deep learning have enabled automated analysis of computed tomography (CT) scans, but many approaches rely solely on imaging data and overlook valuable diagnostic information provided by radiologists. In this work, we propose a multimodal framework for lung nodule malignancy classification that integrates CT image features with radiologist-annotated semantic attributes from LIDC-IDRI dataset. A 3D convolutional neural network (CNN) is used to extract spatial features from CT patches, while a multilayer perceptron (MLP) processes six radiological descriptors describing nodule characteristics. The two feature representations are fused through feature concatenation and classified using a fully connected network. Experiments performed using a patient-level data split demonstrate that the proposed multimodal approach achieves an AUC of 0.93 and high recall for malignant nodules.** **Keywords—Lung cancer, pulmonary nodules, multimodal deep learning, 3D convolutional neural networks, semantic features, patient-level evaluation, LIDC-IDRI.**

I. Introduction

Lung cancer represents one of the most significant public health challenges worldwide and continues to be the leading cause of cancer-related mortality. According to recent epidemiological studies, survival rates depend strongly on early detection and accurate assessment of pulmonary nodules identified on chest imaging. Computed tomography (CT) has become the standard imaging modality for lung cancer screening due to its ability to identify small nodules that may indicate early-stage malignancy. The rapid growth of medical imaging data has spurred the development of automated computer-aided diagnosis systems to assist radiologists in interpreting CT scans.

In recent years, deep learning methods – particularly convolutional neural networks (CNNs) – have shown promising results in various medical imaging tasks, including lung nodule detection and classification. CNN-based approaches can automatically learn hierarchical representations from CT images and have demonstrated strong performance in distinguishing between benign and malignant nodules [1], [2].

B.-A. Banescu is with the National University of Science and Technology POLITEHNICA of Bucharest, Romania (e-mail: briana.banescu@stud.acs.upb.ro).

L. Ichim is with the National University of Science and Technology POLITEHNICA of Bucharest, Romania, and with the "Stefan S. Nicolau" Institute of Virology, Bucharest, Romania (e-mail: loretta.ichim@upb.ro).

D. Popescu is with the National University of Science and Technology POLITEHNICA of Bucharest, Romania (e-mail: dan.popescu@upb.ro).

Several recent studies have highlighted the growing impact of artificial intelligence in lung cancer diagnosis and the potential of deep learning systems to improve diagnostic accuracy and reduce radiologist workload [3].

Despite these advances, many existing methods rely exclusively on imaging data and do not incorporate additional clinical or radiological information that may be available during the diagnostic process. In clinical practice, radiologists evaluate pulmonary nodules based on several semantic characteristics such as margin definition, spiculation, texture and calcification patterns. These attributes capture important morphological properties of nodules and can provide valuable insights into malignancy risk. Previous research has shown that integrating imaging features with radiologist-defined biomarkers and other quantitative descriptors can improve predictive models for lung nodule malignancy assessment. For example, combined image-based, biomarker-based, and hybrid models on the LIDC-IDRI dataset highlight the relevance of multimodal information for malignancy prediction [4].

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI dataset) provide a valuable resource for exploring such multimodal approaches. In addition to CT scans, the dataset includes detailed annotations from multiple radiologists describing nodule characteristics through a set of semantic attributes and malignancy scores. These annotations allow the integration of expert-defined descriptors with automatically extracted image features.

In this work, we propose a multimodal deep learning framework that combines CT image features with radiologist-annotated semantic attributes for lung nodule malignancy classification. A 3D CNN is used to extract spatial features from CT patches, while a multilayer perceptron (MLP) processes six radiological descriptors related to nodule morphology and appearance. The resulting feature representations are fused to produce the final malignancy prediction.

The proposed approach is evaluated on the LIDC-IDRI dataset using a patient-level split to avoid data leakage. Experimental results demonstrate that the multimodal model achieves competitive performance, with an area under the ROC curve (AUC) of 0.93. Furthermore, an ablation study comparing CNN-only, MLP-only and the multimodal configurations highlights the complementary role of image-based and semantic features in lung nodule classification.

II. Related Work

The development of computer-aided diagnosis systems for lung cancer detection and classification has been significantly advanced by the application of artificial intelligence in medical imaging. In recent years, numerous studies have explored deep learning techniques for analyzing CT images of pulmonary nodules.

Convolutional neural networks (CNNs) have become one of the most widely used approaches for lung nodule classification due to their ability to automatically learn hierarchical features from medical images. Early deep learning studies demonstrated that CNN models can effectively extract discriminative spatial patterns from CT scans and achieve promising performance in distinguishing malignant from benign nodules. Recent works have further improved CNN-based methods by employing 3D convolutional architectures that capture volumetric information from CT patches rather than relying solely on two-dimensional slices. Several studies have investigated deep learning approaches for lung nodule analysis and diagnosis [1]-[3]. One article [1] presented a comprehensive review of deep learning methods for CT-based lung nodule analysis, while other studies emphasized the value of volumetric CNN models and related machine learning strategies for improving pulmonary nodule diagnosis [2], [3].

Despite these advances, image-only models may struggle to capture certain diagnostic cues typically used by radiologists when evaluating pulmonary nodules. This limitation has motivated research into integrating additional sources of information into predictive models.

In clinical practice, radiologists assess pulmonary nodules based on several morphological and semantic characteristics, including margin sharpness, spiculation, internal texture, and calcification patterns. These descriptors reflect expert knowledge and provide valuable indicators of malignancy risk.

Radiomics approaches aim to extract quantitative features from medical images that describe shape, texture, and intensity patterns. Numerous studies have shown that radiomic features can improve lung cancer prediction and characterization models. Recent work has also shown that radiomics-based classifiers can achieve competitive performance for pulmonary nodule malignancy assessment on LIDC-IDRI-derived data. Liu et al. developed an interpretable radiomics-based model and reported strong discriminative performance for benign-versus-malignant nodule classification [5]. One of the foundational radiomics frameworks is introduced [6], showing that quantitative imaging features can provide clinically meaningful information for cancer characterization. In addition to automatically extracted radiomic features, some datasets provide semantic annotations generated by radiologists.

Recent studies have shown that multimodal learning strategies can improve lung nodules and lung cancer diagnosis by integrating complementary information sources, such as CT imaging, nodule-related descriptors, and clinical variables [7], [8].

Recent reviews have further emphasized the importance of multimodal artificial intelligence in medical imaging, highlighting its potential to improve robustness and diagnostic reliability [9]. In parallel, explainable artificial intelligence methods have gained increasing attention to improve the interpretability and transparency of lung cancer prediction systems [10]. Furthermore, multimodal AI frameworks integrating imaging and structured non-imaging information have demonstrated promising results for lung cancer diagnosis [11].

Motivated by these findings, the proposed paper investigates a multimodal approach that integrates deep image features extracted from CT patches with radiologist-defined semantic attributes available in the LIDC-IDRI dataset. By combining visual and expert-defined descriptors, the proposed framework aims to leverage complementary information sources for improved lung nodule malignancy classification.

III. Methodology

A. Dataset Used

The experiments in this study were conducted using the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [12]. This publicly available dataset contains thoracic computed tomography (CT) scans collected from multiple medical institutions and annotated by experienced radiologists. The dataset has been widely used in research on lung nodule detection and malignancy classification. Examples of images used in our study are presented in Fig. 1.

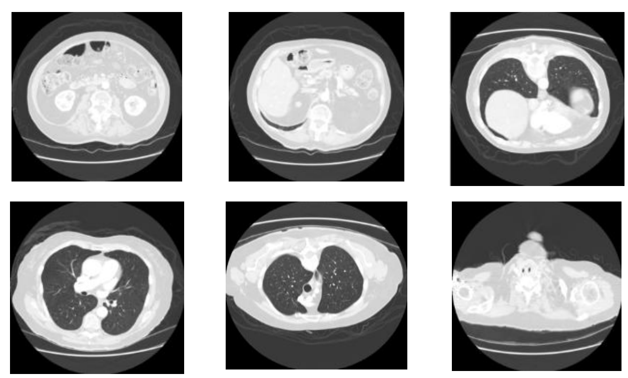

Figure 1. Examples of CT slices from LIDC-IDRI) dataset [12].

The LIDC-IDRI dataset includes CT scans from more than 1,000 patients, together with detailed annotations provided independently by four radiologists. Each radiologist identifies pulmonary nodules and assigns a set of semantic attributes describing nodule appearance. These annotations include several morphological descriptors such as subtlety, calcification, sphericity, margin definition, spiculation, and texture. In addition, each nodule is assigned a malignancy score ranging from 1 to 5, where higher values indicate a higher likelihood of malignancy.

The availability of both CT imaging data and expert-defined semantic descriptors makes LIDC-IDRI particularly suitable for studying multimodal approaches to lung nodule classification.

B. CT Volume Preprocessing

To ensure consistency across the dataset, all CT volumes were first preprocessed prior to feature extraction. The DICOM series corresponding to each patient was loaded and converted into three-dimensional volumes. Since CT scans in the dataset were acquired using different imaging protocols, the volumes were resampled to a uniform spatial resolution of $1 \times 1 \times 1$ mm³ using linear interpolation.

After resampling, the CT volumes were converted into Hounsfield units (HU), and intensity values were normalized to reduce variations across scans. This preprocessing step ensures that the neural network receives standardized input data during training.

The LIDC-IDRI XML annotations provide radiologist-defined nodule contours at the slice level rather than explicit volumetric masks. For each annotated nodule, the contour points were extracted from the XML files, mapped to the corresponding CT slices, and converted into filled 2D binary regions. These slice-wise contour-derived regions were then aggregated along the axial direction to reconstruct a 3D binary nodule representation, which served as the spatial reference for patch extraction and as supervision for segmentation [12].

C. Nodule Localization and Patch Extraction

After reconstructing the contour-derived 3D nodule representation, pulmonary nodules were localized using the coordinates provided in the LIDC-IDRI XML annotations. For each annotated nodule, the center position was computed from the reconstructed contour-derived nodule extent across slices. Using these coordinates, a 3D patch of size $64 \times 64 \times 64$ voxels centered on the nodule was extracted from the preprocessed CT volume.

If the extracted region extended beyond the boundaries of the volume, zero-padding was applied to maintain a consistent patch size. The extracted patches were stored as NumPy arrays and used as input for the deep learning model.

D. Radiological Semantic Attributes

In addition to the image patches, the semantic attributes annotated by radiologists were used as auxiliary input features. From the available annotations, six attributes describing nodule characteristics were selected: subtlety, calcification, sphericity, margin, spiculation, and texture. These attributes capture important morphological and structural properties of pulmonary nodules and are frequently used by radiologists when assessing malignancy risk.

Since multiple radiologists evaluated each nodule, the attribute values were aggregated by computing the median score across radiologists. This approach reduces inter-observer variability and provides a robust representation of the expert assessment.

E. Dataset Splitting

To ensure a fair evaluation of the proposed model, the dataset was split at the patient level rather than at the nodule level. This prevents nodules from the same patient from appearing in both training and testing sets, which could otherwise lead to data leakage and overly optimistic performance estimates.

The dataset was divided into training, validation, and testing subsets using an approximate 80%–10%–10% split of patients. The training set was used for model optimization, the validation set for hyperparameter tuning and threshold selection, and the test set for final performance evaluation.

F. Proposed Multimodal Architecture

To improve lung nodule malignancy classification, we propose a multimodal deep learning architecture (Fig. 2) that integrates image-based features extracted from CT scans with semantic attributes annotated by radiologists. The proposed framework consists of two main branches: a 3D CNN for extracting visual features from CT patches and an MLP for processing radiological semantic descriptors. The outputs of the two branches are combined through feature fusion and used to perform the final malignancy classification.

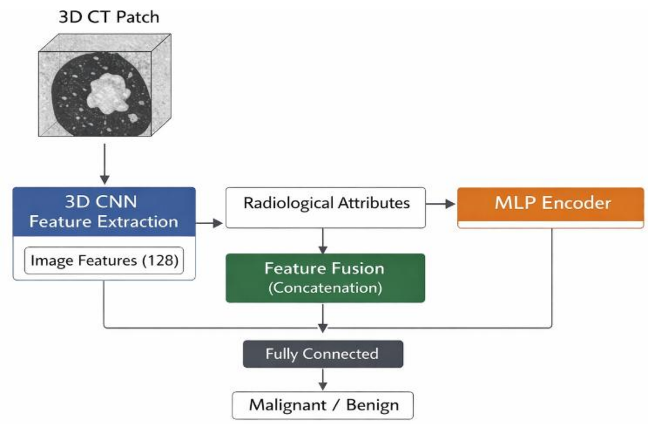

Figure 2. Proposed multimodal architecture.

G. CNN Branch for CT Image Feature Extraction

The first branch of the model processes volumetric CT patches centered on pulmonary nodules. Each input patch has a spatial resolution of $64 \times 64 \times 64$ voxels and represents a localized three-dimensional region of the lung containing the nodule.

A 3D CNN is used to extract hierarchical spatial features from the CT patch. The CNN architecture consists of multiple convolutional blocks, each composed of a 3D convolution layer followed by batch normalization, a ReLU activation function, and max pooling for spatial down-sampling. These layers allow the network to learn discriminative representations that capture the shape, texture, and structural properties of lung nodules.

After several convolutional layers, global feature aggregation is performed using adaptive average pooling, which reduces the spatial feature maps to a compact feature vector. This vector represents the learned visual embedding of the nodule and serves as the image-based representation used in the multimodal fusion stage.

H. MLP Branch for Radiological Attribute Processing

In addition to the image-based features extracted by the CNN branch, the proposed framework incorporates semantic descriptors annotated by radiologists in the LIDC-IDRI dataset. These attributes describe morphological properties of pulmonary nodules that are commonly used in clinical assessment.

Six semantic attributes are used as input features: subtlety, calcification, sphericity, margin, spiculation, and texture. Each attribute is assigned a discrete score in the range [1, 5], where the values correspond to radiologists' assessment of specific morphological characteristics. To reduce inter-observer variability, the final attribute value for each nodule is obtained by computing the median score across radiologists. The semantic attributes are represented as a feature vector (1)

$$x_{attr} = [a_1, a_2, a_3, a_4, a_5, a_6] \in \mathbb{R}^6 \quad (1)$$

where

$$a_i \in [1,5]$$

represents the score of the i-th radiological descriptor.

The feature vector is processed using a multilayer perceptron (MLP) composed of fully connected layers with nonlinear activation functions. The transformation performed by a fully connected layer can be expressed as (2)

$$h = If(Wx + b) \quad (2)$$

where:

- x represents the input vector
- W denotes the weight matrix
- b represents the bias term
- If is a nonlinear activation function (ReLU in this study)

The MLP progressively transforms the semantic attribute vector into a compact latent representation (3)

$$f_{attr} = MLP(x_{attr}) \quad (3)$$

where f_{attr} denotes the learned semantic feature representation.

I. Multimodal Feature Fusion

The multimodal representation is obtained by combining the feature vectors produced by the CNN and MLP branches.

Let f_{img} be the feature vector extracted from the CT patch by the CNN and f_{attr} is the feature vector generated by the MLP from radiological attributes. The fused feature representation is obtained by concatenation (4)

$$f_{fusion} = [f_{img} \parallel f_{attr}] \quad (4)$$

where || denotes the concatenation operation.

The fused vector is then processed by a fully connected classifier that produces the final malignancy prediction.

J. Malignancy Label Definition

The LIDC-IDRI dataset [11] provides a malignancy rating assigned by radiologists on a scale from 1 to 5, where:

- 1 – highly unlikely to be malignant
- 5 – highly suspicious for malignancy

In this work, the malignancy score is not used as an input feature in the model. Instead, it is used only to define the ground-truth label for training and evaluation.

The binary label y is defined as (5)

$$y = \begin{cases} 1, & \text{if malignancy score} > 3 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $y = 1$ indicates a malignant nodule and $y = 0$ indicates a benign nodule.

This formulation ensures that the model learns to predict malignancy based solely on CT image features and radiological descriptors, without directly using the malignancy score during inference.

K. Loss Function

The network parameters are optimized using the binary cross-entropy loss with logits, defined as (6)

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (6)$$

where:

- y_i is the ground truth label
- p_i is the predicted probability of malignancy
- N is the number of samples.

To address class imbalance in the dataset, a positive class weighting factor is applied during training.

IV. Experimental Setup

The proposed multimodal model was implemented using the PyTorch deep learning framework. The network was trained using mini-batch stochastic optimization with the Adam optimizer. The initial learning rate was set to 1×10^{-4}, and the model was trained for 20 epochs.

A batch size of 8 samples was used during training due to the high memory requirements of processing volumetric CT patches. The model parameters were initialized randomly and optimized using backpropagation.

To reduce overfitting and improve generalization, dropout regularization with a probability of 0.3 was applied in the fully connected classification layers.

The model was trained using the binary cross-entropy loss with logits, which is commonly used for binary classification tasks. Since the dataset exhibits class imbalance between benign and malignant nodules, a class weighting strategy was employed during training.

A positive class weight was introduced in the loss function to penalize misclassification of malignant nodules more strongly. This strategy encourages the model to reduce false negative predictions, which is particularly important in medical diagnosis.

To improve model robustness and increase data diversity during training, a simple data augmentation strategy was applied to the CT patches. Specifically, random spatial flipping of the patches along the axial plane was used with a probability of 0.5.

This augmentation technique helps the model learn invariant representations of nodules and reduces the risk of overfitting to specific spatial orientations.

The performance of the proposed model was evaluated using several standard metrics for binary classification: accuracy, precision, recall, F1-score, Area Under the ROC Curve (AUC). The "positive" criterion was considered malignant, and for benign, the "negative" criterion. Considering the evaluations in the categories TP (true positive), TN (true negative), FP (false positive), and FN (false negative), the performance indicators in Table I were calculated.

TABLE I. PERFORMANCE INDICATORS

Metrics	Formula
Accuracy (*Acc*)	$Acc = \frac{TP+TN}{TP+FP+TN+FN}$
Precision (*Pre*)	$Pre = \frac{TP}{TP+FP}$
Recall (*Rec*)	$Rec = \frac{TP}{TP+FN}$
F1 Score (*F1*)	$F1 = \frac{2TP}{2TP+FP+FN}$

Among these metrics, recall for malignant nodules is particularly important in medical applications, as false negative predictions may lead to missed diagnoses. Therefore, special attention was given to analysing the model's ability to correctly identify malignant nodules.

To evaluate the contribution of each modality, an ablation study was conducted by training and evaluating three different model configurations: a CNN-only model using CT image patches, an MLP-only model using radiological semantic attributes, and a Multimodal CNN + MLP model. This analysis allows us to assess the individual and combined contributions of image-based and semantic features to the overall classification performance.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed multimodal architecture was evaluated on the LIDC-IDRI dataset using the patient-level data split described in Section III. The model achieved strong performance in distinguishing between benign and malignant lung nodules.

The overall classification performance was assessed using several standard evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). The multimodal model achieved an AUC of approximately 0.87, demonstrating its ability to effectively discriminate between benign and malignant nodules.

In medical diagnosis tasks, correctly identifying malignant nodules is particularly important, as false negative predictions may lead to delayed treatment. The proposed model achieved a recall of approximately 0.83 for malignant nodules, indicating that most malignant cases were successfully detected.

The receiver operating characteristic (ROC) curve for the proposed model is shown in Fig. 3, illustrating the trade-off between true positive and false positive rates across different classification thresholds. To further analyse the model's performance, a confusion matrix was computed on the test set. The confusion matrix provides insight into the number of correctly and incorrectly classified samples. The results indicate that the proposed model successfully identifies most malignant nodules, while maintaining a relatively low number of false positive predictions. Although some benign nodules are incorrectly classified as malignant, this behaviour is often acceptable in medical screening systems where reducing false negatives is considered more critical than minimizing false positives.

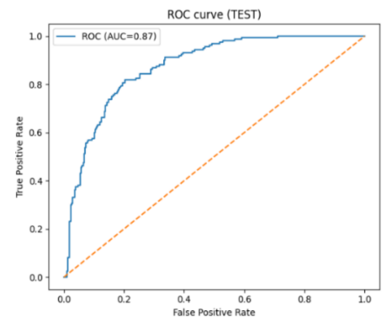

Figure 3. Receiver operating characteristic (ROC) curve of the proposed multimodal CNN+MLP model evaluated on the test set.

The confusion matrix visualization is presented in Fig. 4 and provides a detailed overview of classification outcomes.

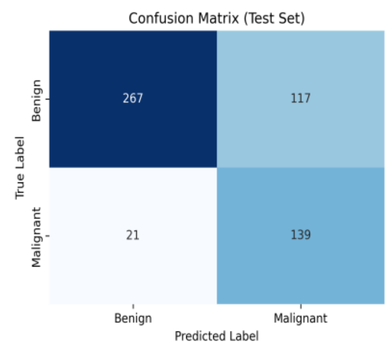

Figure 4. Confusion matrix showing the classification performance of the proposed multimodal model on the test dataset.

To better understand the contribution of each modality, an ablation study was performed by evaluating three different model configurations:

- a CNN-only model using CT image patches.
- an MLP-only model using radiological semantic attributes.
- the proposed multimodal CNN+MLP architecture.

Table II summarizes the performance obtained by each model.

TABLE II. ABLATION STUDY COMPARISON

Model	AUC	Recall (Malignant)	Precision (Malignant)
CNN – only	0.85	0.79	0.76
MLP – only	0.89	0.88	0.81
CNN + MLP	**0.93**	**0.90**	**0.84**

The CNN-only model achieved the lowest performance, with an AUC of 0.39, indicating that image features extracted solely from small CT patches were insufficient to reliably distinguish between benign and malignant nodules in this experimental setting. This result suggests that visual features alone may not fully capture the complex morphological characteristics associated with malignancy.

In contrast, the MLP-only model, which uses radiological semantic attributes such as subtlety, calcification, sphericity, margin, spiculation, and texture, achieved substantially better performance, with an AUC of 0.86 and a malignant recall of 0.83. These results highlight the strong predictive value of expert-defined radiological descriptors that capture clinically relevant morphological patterns.

The proposed multimodal CNN+MLP model achieved the best overall performance, reaching an AUC of 0.93, a malignant recall of 0.90, and a precision of 0.84. These results indicate that combining image-based representations with radiological attributes provides complementary information that improves the model's ability to detect malignant nodules.

Table III presents a comparison between the proposed method and several recent approaches for lung nodule malignancy classification on the LIDC-IDRI dataset.

TABLE III. COMPARISON WITH RELATED WORK

Study	Method	AUC	Notes
[4]	3D CNN	0.80	Images only
[5]	Radiomics + RF	0.88	Handcrafted features
Proposed Method	3D CNN + Radiologist Attributes	0.93	Multimodal fusion

VI. CONCLUSION

In this study, we proposed a multimodal deep learning framework for lung nodule malignancy classification that integrates volumetric CT image features with radiologist-annotated semantic attributes from the LIDC-IDRI dataset. The proposed architecture combines a 3D CNN for extracting spatial representations from CT patches with an MLP that processes radiological descriptors describing nodule morphology. Experimental results demonstrate that the multimodal model achieves competitive performance, reaching an AUC of approximately 0.93 and a high recall for malignant nodules. These findings suggest that incorporating expert-defined semantic descriptors can significantly enhance computer-aided diagnosis systems for lung cancer.

Future work will explore the integration of automatic nodule segmentation and more advanced multimodal fusion strategies to further improve classification performance.

REFERENCES

[1] Y. Gu, J. Chi, J. Liu, L. Yang, H. Zhang, X. Li, X. Lu, and J. Yang, "A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning," *Computers in Biology and Medicine*, vol. 137, Art. no. 104806, 2021.

[2] H. Jin, Z. Zhang, X. Zhang, and X. Zhang, "Machine learning techniques for pulmonary nodule diagnosis: A review," *Biomedical Signal Processing and Control*, vol. 80, Art. no. 104351, 2023.

[3] G. Kang, K. Liu, B. Hou, and N. Zhang, "3D multi-view convolutional neural networks for lung nodule classification," *PLOS ONE*, vol. 12, no. 11, 2017.

[4] K. Mehta, J. Jain, A. Sengupta, and T. Chakraborti, "Lung nodule classification using biomarkers, volumetric radiomics, and 3D CNNs on LIDC-IDRI," *Journal of Digital Imaging*, vol. 34, pp. 181–193, 2021.

[5] J. Liu, J. Chen, C. Reyes-Aldasoro, and E. E. Bron, "Lung nodule classification using radiomics model trained on synthetic low-dose CT images," *Computer Methods and Programs in Biomedicine*, vol. 257, Art. no. 108095, Dec. 2024.

[6] H. J. Aerts, E. Velazquez, R. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014.

[7] T. A. Shaik, X. Tao, C. Dann, Y. Li, and L. Galligan, "A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom," *Information Fusion*, vol. 102, 2024.

[8] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Computers in Biology and Medicine*, vol. 177, Art. no. 108635, 2024.

[9] J. Duan, Y. Xiong, Y. Li, and W. Ding, "Deep learning based multimodal biomedical data fusion: An overview and comparative review," *Information Fusion*, vol. 112, Art. no. 102536, 2024.

[10] C. Wang, J. Liu, et al., "Towards reliable and explainable AI model for pulmonary nodule diagnosis," *Biomedical Signal Processing and Control*, vol. 88, Part B, Art. no. 105646, 2024.

[11] E. Oncu and F. Ciftci, "Multimodal AI framework for lung cancer diagnosis: Integrating CNN and ANN models for imaging and clinical data analysis," *Comput Biol Med*, vol. 193, Art no. 110488. Jul. 2025.

[12] S. G. Armato 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*, vol. 38, no. 2, pp. 915–931, Jan. 2011.