Clinical Prior Guided Cross-Modal Hierarchical Fusion for Histological Subtyping of Lung Cancer in CT Scans

Chenchen Fan[1,†], Ahmed Elazab[2,†], Songqi Zhang[1], Yuxuan Wang[1]
Qinghua Liang[1], Danna Li[1], Yongquan Zhang[1,*], Ying Xiang[1]
Bo Liu[3], and Changmiao Wang[4,*]

[1] Zhejiang University of Finance and Economics
[2] Shenzhen University
[3] Northwest China Research Institute of Electronic Equipment
[4] Shenzhen Research Institute of Big Data
zyq@zufe.edu.cn, cmwangalbert@gmail.com

Abstract. Accurate lung cancer localization and classification in computed tomography (CT) images are vital for effective treatment. However, existing approaches still face challenges such as redundant information in CT images, ineffective integration of clinical prior knowledge, and difficulty in distinguishing subtle histological differences across lung cancer subtypes. To address these, we propose Cross-Modal Detection Auxiliary Classification (CM-DAC), a framework enhancing classification accuracy. It employs a YOLO-based slice detection module to extract lesion areas, which are processed using the Multimodal Contrastive Learning Pretrain (MCLP) module, minimizing redundancy. Specifically, MCLP aligns 3D patches with clinical records via a cross-modal hierarchical fusion module, integrating image and clinical features through attention mechanisms and residual connections. Additionally, we employ multi-scale fusion strategies to further enhance histological distinction by capturing features at different resolutions. Simultaneously, a text path expands category labels into semantic vectors using a medical ontology-driven text augmentation approach. These vectors are encoded and aligned with fusion feature vectors. Experimental results on both private and public datasets confirm that the proposed CM-DAC outperforms competitive methods, achieving superior classification performance. The code is available at https://github.com/fancccc/CM-DAC.

Keywords: Lung Cancer Subtyping · Hierarchical Attention Fusion · Medical ontology.

1 Introduction

Lung cancer remains the leading cause of cancer-related deaths worldwide, accounting for over 18% of such fatalities in 2023 [18]. Early and accurate diagnosis

[1] † These authors contributed equally to this work.
[2] * Corresponding author: zyq@zufe.edu.cn, cmwangalbert@gmail.com

of pulmonary nodules, which are critical precursors to lung cancer, is essential for improving patient survival rates [14,2]. Although Computed Tomography (CT) scans are widely used imaging modalities for nodule evaluation due to their ability to capture detailed spatial characteristics of small lesions, their clinical usefulness is limited by significant challenges. Firstly, a single CT scan produces hundreds of cross-sectional slices, many of which contain redundant information, such as normal tissue interference and artifacts. Secondly, pulmonary nodules vary greatly in morphology, density, and margin characteristics. These issues contribute to diagnostic errors by radiologists, with rates of misdiagnosis and missed diagnoses reported between 10% and 26% [20]. These limitations highlight the urgent need for the development of automated, high-precision algorithms for diagnosing pulmonary nodules.

The rapid advancement of deep learning has significantly advanced pulmonary nodule analysis in CT imaging [12]. Early approaches relying on manually engineered features with conventional neural networks [9] faced limitations in generalizability due to their dependence on subjective feature selection. The emergence of Convolutional Neural Networks marked a paradigm shift, with Faster R-CNN [16] demonstrating the feasibility of automated 2D nodule detection. However, its inability to model 3D contextual relationships across CT slices remained a critical constraint. Subsequent 3D architectures like 3D ResNet [6] addressed this spatial modeling challenge [13], while transfer learning strategies partially mitigated data scarcity issues [8]. Nevertheless, as noted in recent studies [5], these single-modal approaches still struggle with classifying subtle lesions and fail to leverage complementary clinical data modalities such as genomic profiles or pathology reports. Recent multimodal frameworks attempt to bridge this gap. SAMA [1] pioneers cross-modal fusion through self-attention mechanisms between CT and RNA Sequencing (RNA-seq) data, while Contrastive Language–Image Pretraining (CLIP) [15]-inspired models like CLIP-Lung [10] and CMMF [4] explore vision-text alignment. However, as evidenced by TMSS [17], current methods face inherent limitations static feature interaction that overlooks lesion evolution patterns. MultiSurv [19] further highlights the persistent semantic disconnect between low-level imaging features and high-order clinical concepts. In summary, current methods face three main challenges, including redundant information in CT images, ineffective integration of clinical prior knowledge, and difficulty in distinguishing subtle histological differences across lung cancer subtypes.

To tackle these challenges, this paper introduces an end-to-end CM-DAC framework that seamlessly integrates detection with classification models. The detection stage employs the YOLOv11 [7] model to extract 2D slices from the original CT images, facilitating the training of the detection model. From the identified lesion locations, fixed-size 3D patches are cropped and fed into a 3D ResNet network for feature extraction. During feature fusion, a novel attention mechanism is utilized to merge multi-scale image features with prior clinical information, resulting in semantically enriched cross-modal features. For text labels, medical terminology is used for text augmentation, and a text encoder is

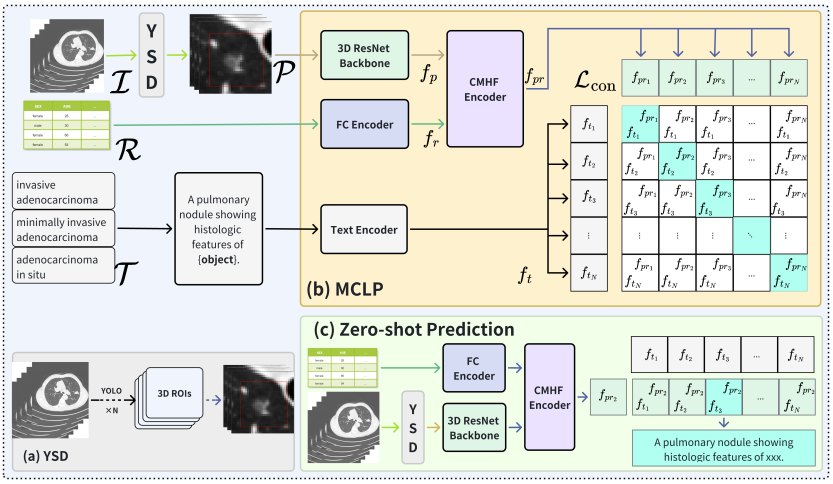

Fig. 1: Overview of the CM-DAC framework: (a) YSD module for CT image detection and ROI extraction; (b) MCLP module for multimodal-text contrastive learning; (c) Zero-shot prediction for end-to-end classification results.

applied for encoding these labels. Contrastive learning then aligns the labels with their corresponding feature vectors. In the inference phase, the network processes the raw CT image and prior clinical information. Initially, the detection module extracts the lesion area, followed by Region Of Interest (ROI) cropping and category classification using the Multimodal Contrastive Learning Pretraining (MCLP) module, calculate the cosine similarity with pretrained label weight and find the maximum as the result. Our main contributions include: (1) Detecting target regions and removing irrelevant areas to reduce interference from redundant information in CT images; (2) Aligning clinical priors with lesion image features to guide the model in prioritizing relevant image characteristics; and (3) Implementing multi-scale feature fusion to enhance lesion recognition across different scales.

2 Proposed Method

Our method, as illustrated in Fig. 1, processes three input modalities, $\{\mathcal{I}, \mathcal{R}, \mathcal{T}\}$ representing 3D CT images, clinical records, and textual labels, respectively. The YOLO-based Slice Detection (YSD) module analyzes $\mathcal{I}$ to generate lesion regions ($\mathcal{P}$) from $\mathcal{I}$, which are encoded as f_p using a 3D ResNet in MCLP. Meanwhile, $\mathcal{R}$ is encoded by a fully connected network as f_r. The Cross-Modal Hierarchical Fusion (CMHF) module integrates the features f_r and f_p. Simultaneously, $\mathcal{T}$ is expanded into radiological descriptions and encoded via a biomedical text encoder, enabling contrastive learning between the multimodal-text pairs. Dur-

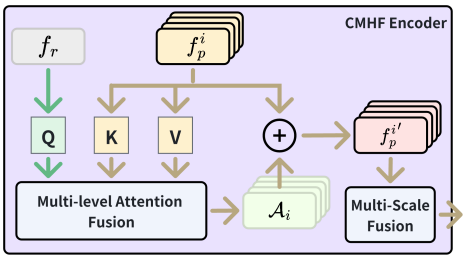

Fig. 2: The architecture of the CMHF encoder, which is used to integrate clinical records and images through feature fusion.

ing the zero-shot inference phase, the system matches the detected lesions with textual prototypes in the embedding space, producing category predictions.

2.1 YOLO-base Slices Detection Module

The training process for the model's detection branch, as shown in Fig. 1(a), begins with an input $\mathcal{I} \in \mathbb{R}^{W \times H \times D}$. The detection module then proceeds through several stages to generate ROI patches $\mathcal{P} \in \mathbb{R}^{P \times P \times P}$. First, lesion-aware slice selection is performed, extracting 2D axial slices $\mathcal{S} = \{s_i\}_{i=1}^{N}$ from $\mathcal{I}$ that contain lesions annotated by radiologists. M_{det} is trained on the extracted slices $\mathcal{S}$. For each predicted 2D bounding box $\mathbf{b}_k$ in slice s_j, the 3D lesion centroid (x_c, y_c, z_c) is calculated. This involves axial continuity validation, requiring an Intersection over Union of at least 0.7 across three consecutive slices. Finally, the ROI volume extraction step crops cubic subvolumes $\mathcal{P}$ centered at (x_c, y_c, z_c) with an edge length of 32mm, including a 15 mm margin around the lesion. The 32mm crop size aligns with clinical standards (>30mm = mass) and dataset stats (mean: 23.69mm; 75th %ile: 29mm).

2.2 Mutilmodal Contrastive Learning Pretrain Module

As illustrated in Fig. 1(b), our MCLP module utilizes specialized encoders for three distinct input modalities. Initially, the 3D ResNet backbone encodes $\mathcal{P}$, producing multi-scale features: $f_p = \{f_p^i\}_{i=1}^{4}$, where each f_p^i is represented as $\mathbb{R}^{B \times 256 \times W_i \times H_i \times D_i}$. These features are extracted from the last four layers of the ResNet. Meanwhile, $\mathcal{R}$ is processed through a fully-connected network to yield compact representations: $f_r \in \mathbb{R}^{B \times 256}$. For $\mathcal{T}$, clinical template expansion translates labels like "invasive adenocarcinoma" into more detailed descriptions, such as "A pulmonary nodule showing histologic features of invasive adenocarcinoma." These narratives are encoded by a CLIP text encoder to generate $f_t \in \mathbb{R}^{B \times 256}$.

Next, we integrate the clinical records and CT features using the CMHF encoder, as shown in Eq. 1, resulting in the fused feature representation:

$$f_{pr} = \Psi(f_p, f_r) \in \mathbb{R}^{B\times 256}, \tag{1}$$

where $\Psi(\cdot)$ signifies our CMHF Encoder, as depicted in Fig. 2. This module processes f_p and f_r through hierarchical fusion mechanisms for seamless integration. For the tabular clinical features f_r, we use parallel projection heads to generate scale-specific embeddings:

$$f_r^i = \sigma(\mathbf{W}_i^{(2)} \cdot \sigma(\mathbf{W}_i^{(1)} f_r + \mathbf{b}_i^{(1)}) + \mathbf{b}_i^{(2)}), \tag{2}$$

where σ denotes the ReLU activation function, and $\{\mathbf{W}_i^{(k)}, \mathbf{b}_i^{(k)}\}$ are learnable parameters for each scale i.

For each set of features $\{f_p^i, f_r\}$, we introduce a Multi-level Attention Fusion mechanism. A significant innovation in our module is the scale-adaptive attention mechanism, which enhances the integration of information across different scales at each hierarchical level i, as outlined in Algorithm 1.

Algorithm 1 Multi-level Attention Fusion

Input: f_p^i, f_r, α_i, dimensions D_i, H_i, W_i, batch size B
Compute $N_i = D_i \times H_i \times W_i$, $\quad K_i \leftarrow f_p^i$, $\quad V_i \leftarrow f_p^i$
Step 1: Spatial-to-Sequence Transformation
Reshape CT features: $f_p^i \leftarrow \text{reshape}(f_p^i, (N_i, B, 256))$
Step 2: Dynamic Clinical Expansion
Project clinical embeddings: $Q_i \leftarrow$Eq. 2(f_r)
Step 3: Context-Aware Attention
Compute cross-modal interaction:

$$\mathcal{A}_i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

Step 4: Residual Fusion
Combine attention maps with original CT features:

$$f_p^{i'} = f_p^i + \alpha_i \cdot \text{Reshape}(\mathcal{A}_i)$$

Return: $f_p^{i'}$

In the Multi-scale Aggregation module, the fused features $\{f_p^{i'}\}$ undergo adaptive pooling and are then combined with the enhanced clinical features. This procedure is expressed by the equation:

$$f_{pr} = \mathbf{W}_p \cdot \text{Concat}\left(\{\text{Pool}(\{f_p^{i'}\})\}, f_r\right), \tag{3}$$

resulting in the final joint representation $f_{pr} \in \mathbb{R}^{B\times 256}$, which is utilized for subsequent tasks. This architecture provides several benefits: it ensures precise align-

ment between anatomical structures and clinical indicators, preserves modality-specific patterns through residual learning, and enables adaptive weighting of cross-modal evidence across spatial hierarchies.

Finally, the framework improves modality alignment using contrastive learning. The loss function is defined as:

$$\mathcal{L}_{\text{con}} = -\sum_{i=1}^{B} \log \frac{\exp(\langle f_{pr_i}, f_{t_i} \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle f_{pr_i}, f_{t_j} \rangle / \tau)}, \tag{4}$$

where τ represent the temperature hyperparameter (we set $\tau = 0.07$, a standard choice in contrastive learning), and $\langle \cdot, \cdot \rangle$ represents cosine similarity.

2.3 Zero-shot Prediction

The inference process is illustrated in Fig. 1(c). The inputs include raw CT images and clinical information. During the detection phase, each slice of the 3D image is processed, and results from all layers are combined. The position of the target region along the vertical axis is determined by calculating the IoU. Subsequently, a lesion block is cropped based on this position. The cropped lesion block is then fed into the MCLP module for classification. Within this module, the similarity between the fused features and all category vectors is calculated. The category vector with the highest similarity is selected as the predicted result. In conclusion, the model outputs the type of lesion. We evaluate model performance using Accuracy (Acc), Precision (Pre), Recall (Rec), F1-Score (F1), and AUC. For multi-class classification, all metrics are computed using macro averaging.

3 Experiments

3.1 Dataset and Implementation

Private Dataset. The dataset was collected from our collaborating hospital, and consists of 1,614 cases of lung adenocarcinoma from 1,430 anonymous patients. Each case includes CT scan images, clinical data, and bounding boxes indicating tumor locations. The dataset is categorized into three groups: invasive adenocarcinoma, microinvasive adenocarcinoma, and adenocarcinoma in situ, comprising 53.5%, 24.1%, and 22.4% of the cases, respectively. The clinical data were obtained through various pathological tests and were carefully evaluated by experienced medical professionals. This data includes demographic details such as patient age and gender, along with lesion-specific morphological characteristics, including tumor margin, density, shape, and location.

The open-source LPCD dataset, as referenced in [11], includes the same modalities as our dataset. It comprises 342 lung cancer cases, categorized into adenocarcinoma, small cell carcinoma, large cell carcinoma, and squamous cell carcinoma. Due to the limited number of squamous cell cases, the large cell and squamous cell categories are combined, creating a three-class dataset. The class

Table 1: Comparison of the proposed method with baseline methods and state-of-the-art approaches on the LPCD and the private datasets (%). **Bold** text highlights the best indicator, while underlined text represents the second-best. ↑ indicates an increase, and ↓ indicates a decrease.

Data	Method	Acc	Pre	Rec	F1	AUC
LPCD [11]	ResNet18 [6]	71.01	23.67	33.33	27.68	50.00
	ViT [3]	71.01	23.67	33.33	27.68	64.87
	nnMamba [5]	72.06	24.02	33.33	27.92	46.05
	CLIP-Lung [10]	72.46	42.04	49.08	44.95	76.73
	TMSS [17]	79.71	76.67	61.59	**63.33**	79.62
	MultiSurv [19]	78.26	49.89	**65.31**	56.46	78.87
	CM-DAC (Ours)	**80.88**↑1.17	**88.57**↑11.9	64.00↓1.31	63.06↓0.27	**84.91**↑5.29
Private	ResNet18 [6]	60.29	29.00	31.97	30.40	75.04
	ViT [3]	69.35	61.88	61.81	61.84	80.97
	nnMamba [5]	65.42	47.55	58.30	48.85	85.08
	CLIP-Lung [10]	71.52	65.05	65.26	64.64	85.60
	TMSS [17]	78.95	74.79	76.58	75.45	88.96
	MultiSurv [19]	74.92	70.28	68.46	68.72	88.45
	CM-DAC (Ours)	**80.06**↑1.11	**76.05**↑1.26	**77.62**↑1.04	**76.71**↑1.26	**90.26**↑1.30

distributions are 70.7% adenocarcinoma, 17.3% small cell carcinoma, and 12.0% for the merged large and squamous cell carcinoma category.

Implementation Details. The CT data are first resampled to 1mm isotropic resolution, then preprocessed by applying a window level of -600 and a window width of 1500, followed by standardization. Clinical records are handled using one-hot encoding for categorical variables and min-max normalization for numerical variables to ensure consistency across features. The experiments are conducted in an environment using Python 3.10.12, PyTorch 2.3.1+cu121, and NVIDIA L20 GPUs. The Adam optimizer is employed with a learning rate of 0.00025 and a weight decay of 0.0001. Additionally, the learning rate is adjusted using the ReduceLROnPlateau scheduler, and training is performed over 200 epochs. To improve the model's generalizability and ensure comprehensive utilization of the data, independent 5-fold cross-validation was performed on each dataset. Basic augmentations, such as flipping and rotation, were applied to address class imbalance.

3.2 Experiment Results

Detection Module Analysis. Our YOLOv11-based lesion localization module achieved optimal single-class detection performance with the lightweight YOLOv11n variant (mAP50=0.701). However, its effectiveness diminished in multi-class scenarios (mAP50=0.467), revealing inherent limitations in differentiating subtle inter-class variations, a critical challenge addressed by our subsequent cross-modal classification framework.

Table 2: Ablation Results on the private dataset (%). CMHF is replaced by Addition, and CL is replaced by Classifier.

Method	Acc	Pre	Rec	F1	AUC
w/o $\mathcal{R}$	65.42	61.10	62.09	60.99	81.57
w/o CMHF CL	75.39	70.28	70.98	69.96	87.72
w/o CMHF	76.95	72.07	73.73	72.70	88.78
w/o CL	78.19	74.31	72.73	73.37	90.63
CM-DAC (Ours)	**80.06**	**76.05**	**77.62**	**76.71**	**90.26**

Comparison with Other Methods. This study validates the effectiveness of CM-DAC using both the public LPCD dataset and our private dataset. As shown in Table 1, the multimodal approach (CLIP-Lung [10], TMSS [17], MultiSurv [19]) significantly outperforms single-modal baseline models (ResNet [6], ViT [3], nnMamba [5]) on both datasets, with improvements in Acc, Pre, Rec, F1, and AUC. On the LPCD dataset, the multimodal approach improves accuracy by 8.82% and AUC by 20.04%. On our private dataset, accuracy and AUC improve by 10.71% and 5.18%, respectively, highlighting the importance of integrating clinical text and image data for effective modeling.

CM-DAC, utilizing cross-modal hierarchical alignment mechanism, achieves superior performance, surpassing the second-best method by 1.17% in accuracy, 11.9% in precision, and 5.29% in AUC on the LPCD dataset, with an accuracy of 80.88%, precision of 88.57%, and AUC of 84.91%. This model also demonstrates a significant advantage in reducing false positives, as indicated by its precision, which is 3 times the standard (26.64%) deviation higher than other methods. For the private dataset, CM-DAC leads in all metrics. It achieves an accuracy of 80.06% and an AUC of 90.26%, with improvements of 1.11% and 1.30%, respectively, over the second-best TMSS model. All metrics show gains of over 1%, demonstrating the model's robust cross-institution generalization capability. These results suggest that the feature fusion strategy guided by pathological semantics effectively coordinates the complementary information from image representations and clinical data, thereby enhancing diagnostic accuracy and ensuring model robustness.

Ablation Study. To verify the effectiveness of each module, we conducted ablation experiments, and the results are presented in Table 2. Firstly, removing the clinical information resulted in a significant decrease in the model's accuracy to 65.42%, underscoring the importance of clinical data in our model. We then performed ablation on the CMHF and Contrastive Learning (CL) modules. Excluding the CMHF module led to a 1.25% decrease in accuracy, while the removal of the CL module resulted in a 2.49% drop. More critically, when both modules were excluded together, the accuracy fell by 4.05%. This indicates that both the CMHF and CL modules not only contribute significantly to the model's performance individually but also have a strong synergistic effect when combined, leading to a substantial overall improvement. Finally, the complete

network achieved optimal results across all metrics, validating the effectiveness of each module in enhancing the model's performance.

4 Conclusion

This paper presents the end-to-end CM-DAC network for lung cancer diagnosis. The network leverages the YSD module to address the issue of redundant information in CT images. The CMHF module adopts an attention mechanism and multi-scale feature fusion to align clinical priors with CT images and spatial features at different scales, ensuring the effective utilization of both modal information and their combined features. This approach helps capture subtle differences between categories. Extensive experiments and ablation studies demonstrate superior performance in identifying subtle histological differences across subtypes. In the future, we intend to incorporate large language models to process raw clinical data, further enhancing the model's usability and robustness.

Acknowledgments. This work was supported by the Guangxi Key R&D Project (No. AB24010167), the Project (No. 20232ABC03A25), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617, 2022A1515110570), Shenzhen Longgang District Science and Technology Innovation Special Fund (No. LGKCYLWS2023018), Futian Healthcare Research Project (No. FTWS002), Medical Scientific Research Foundation of the Guangdong Province of China(A2023158,C2023106), and Shenzhen Medical Research Fund (No. C2401036).

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

References

1. Ai, Y., Liu, J., Li, Y., Wang, F., Du, X., Jain, R.K., Lin, L., Chen, Y.W.: Sama: A self-and-mutual attention network for accurate recurrence prediction of non-small cell lung cancer using genetic and ct data. IEEE Journal of Biomedical and Health Informatics (2024)
2. Blandin Knight, S., Crosbie, P.A., Balata, H., Chudziak, J., Hussell, T., Dive, C.: Progress and prospects of early detection in lung cancer. Open Biology **7**(9), 170070 (2017)
3. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020)
4. Fan, C., Liu, L., Wang, Y., Li, D., Liang, Q., Elazab, A., Liu, Z., Hu, J., Tian, Y., Zhang, Y., et al.: Deep neural network for lung adenocarcinoma subtype from multimodal fusion of imaging and clinical data. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
5. Gong, H., Kang, L., Wang, Y., Wan, X., Li, H.: nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. arXiv preprint arXiv:2402.03526 (2024)
6. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555 (2018)

7. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO (Jan 2023), `https://github.com/ultralytics/ultralytics`
8. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)
9. Kuruvilla, J., Gunavathi, K.: Lung cancer classification using neural networks for ct images. Computer Methods and Programs in Biomedicine **113**(1), 202–209 (2014)
10. Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 403–412. Springer (2023)
11. Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., Wang, D.: A large-scale ct and pet/ct dataset for lung cancer diagnosis [dataset]. The Cancer Imaging Archive **10** (2020)
12. Nasrullah, N., Sang, J., Alam, M.S., Mateen, M., Cai, B., Hu, H.: Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors **19**(17), 3722 (2019)
13. Ning, J., Zhao, H., Lan, L., Sun, P., Feng, Y.: A computer-aided detection system for the detection of lung nodules based on 3d-resnet. Applied Sciences **9**(24), 5544 (2019)
14. Ost, D.E., Gould, M.K.: Decision making in patients with pulmonary nodules. American Journal of Respiratory and Critical Care Medicine **185**(4), 363–372 (2012)
15. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
17. Saeed, N., Sobirov, I., Al Majzoub, R., Yaqub, M.: Tmss: an end-to-end transformer-based multimodal network for segmentation and survival prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 319–329. Springer (2022)
18. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: A Cancer Journal for Clinicians **73**(1), 17–48 (2023)
19. Vale-Silva, L.A., Rohr, K.: Long-term cancer survival prediction using multimodal deep learning. Scientific Reports **11**(1), 13505 (2021)
20. Zahari, R., Cox, J., Obara, B.: Mitigating diagnostic errors in lung cancer classification: A multi-eyes principle to uncertainty quantification. IEEE Journal of Biomedical and Health Informatics **28**(11), 6828–6839 (2024)