

Multimodal Deep Learning Fusion for Lung Cancer Subtype Differentiation from CT and PET Imaging

Abiha Ejaz Correspondence author

abihaijaz@hotmail.com

The Islamia University

Ali Nawaz Shah

The Islamia University

Article

Keywords: Deep learning, multimodal medical image fusion, CT-PET imaging, EfficientNet, multiple instance learning, non-small cell lung cancer, small cell lung cancer

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10684839/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

"Multimodal Deep Learning Fusion for Lung Cancer Subtype Differentiation from CT and PET Imaging"

Abiha Ejaz - *"Corresponding author"*

Email: abihaijaz@hotmail.com

Supervisor: Dr Ali Nawaz Shah

Email: ali.nawaz@iub.edu.pk

Department of CS & IT

The Islamia University, Bahawalpur, Pakistan

ABSTRACT

Differential diagnosis between Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC) represents a critical decision boundary in thoracic oncology, dictating immediate therapeutic pathways. However, conventional single-modality strategies often lack concurrent anatomical and functional comprehension, while traditional multi-modal fusion networks are limited by pixel-level signal suppression and highly disruptive slice-level labeling noise. This paper introduces a unified Hybrid Multiple Instance Learning (MIL) Fusion Framework designed explicitly for automated, patient-level lung cancer differentiation using paired structural Computed Tomography (CT) and metabolic ^{18}F -FDG Positron Emission Tomography (PET) volumes. The proposed architecture employs independent, parallel EfficientNet-B0 encoders to preserve native modality topologies, utilizing a parameterized multi-head cross-attention mechanism to dynamically map complex non-linear inter-modality dependencies. To bypass the historic constraint of tedious manual slice-by-slice segmentations, an attention-gated MIL pooling bottleneck is integrated to aggregate 3D volumetric slice sequences into cohesive patient-level bags, optimizing learning parameters directly from high-level clinical diagnoses. Rigorous evaluation on the open-source Lung-PET-CT-Dx dataset across a verified, highly stratified cohort of 355 patients (284 training, 71 validation) demonstrates that the proposed framework significantly

outperforms baseline paradigms. The optimized hybrid network achieves a prominent patient-level accuracy of 0.7887, a Balanced Accuracy of 0.5536, an Area Under the ROC Curve (AUC) of 0.6607, and a Macro F₁-score of 0.5443. Crucially, architecture actively overcomes severe epidemiological class skew to recover the aggressive minority target, establishing a non-zero SCLC recall of 37.50% while operating at a rapid inference latency of 1.71 ms per slice equivalent. These empirical outcomes demonstrate the system's potential viability as an explainable, high-throughput radiology decision-support tool for accelerated clinical triage.

Index Terms- Deep learning, multimodal medical image fusion, CT-PET imaging, EfficientNet, multiple instance learning, non-small cell lung cancer, small cell lung cancer.

INTRODUCTION

A. Clinical Context and Diagnostic Imperative Lung cancer is one of the leading causes of cancer deaths worldwide. Early detection, minimally invasive, fast and accurate diagnosis is required to improve treatment and patient outcomes[1]. For lung cancer diagnosis specifically, distinguishing between Non-Small Cell Lung Cancer (NSCLC) versus Small Cell Lung Cancer (SCLC) can be critical to a patient's treatment.[2], [3] NSCLC makes up about 85% of lung cancers. There are three main types: Adenocarcinoma, Squamous Cell Carcinoma, and Large Cell Carcinoma. NSCLC tumors are made up of epithelial cells and tend to create localized structures that can be detected[4]. These types of cancer are treated with surgery, chemotherapy drugs like Tyrosine Kinase inhibitors (eg. EGFR, ALK), or im In contrast, SCLC is a very aggressive neuroendocrine carcinoma with short doubling time, early metastasis, and a correspondingly dismal prognosis. Since patients generally present with systemic disease, local surgical resection is not an option and instead SCLC is treated with systemic cytotoxic chemotherapy (usually etoposide and a platinum-based drug) and

thoracic radiation. Administering the incorrect therapy to a patient dramatically increases morbidity and mortality by delaying the correct course of action, thus creating computer aided diagnostic tools to automatically and non-invasively classify these histopathological patterns at the patient level is critically important.[5] munotherapies.

B. Multi-Modal Radiomics and the Automated Sorting Bottleneck

Standard clinical staging and treatment planning is usually based on dual-modality imaging protocols consisting of structural and functional representations. CT imaging uses attenuation of X-rays to reveal the accurate anatomical borders and can offer high resolution information about tumor size, density, and relationships to adjacent structures (spiculated edges, pleural tag, etc.).[4], [5] PET imaging using 18F- FDG probes localized metabolic activity by measuring glucose analog uptake per unit time, which is typically increased in cancer cells due to hypermetabolic glycolysis[1]. (Warburg effect) However, the CAD systems as they are currently designed suffer from severe slowdowns due to two main factors:

1) Inadequate Modality Interaction

Traditional fusion architectures implement channel-wise concatenation of images at the pixel-level (early fusion) or perform probability-map merging late in the network. Pixel-level concatenation causes shared convolutional kernels to update using vastly different numerical ranges at the same time (CT structure and PET function). Often the intense gradients propagated back from the CT image will overshadow quiet metabolic information, causing your model to overlook key functional abnormalities.

2) The Slice-Level Annotation Gap

Conventional deep learning approaches necessitate fully delineated slice-wise ground truth segmentations. Such networks trained on the native slice volume with weak labels will classify background peripheral

lung parenchyma from a positive study as false negatives. This corrupts the backpropagation signal with significant gradient noise and alters decision boundaries further downstream.[6]

C. Primary Architectural Contributions

To address these problems, achieve comparable patient distributions to reality with respect to oncological cases, and enable automatic patient-level lung cancer classification, we present a novel Joint Hybrid MIL Fusion Framework. The classification task is distinctly defined as a weakly supervised binary learner that discriminates between majority/non-small cell lung cancer (NSCLC) and minority/small cell lung cancer (SCLC) populations[7]. Optimizations were performed to allow for maximal data usage from our confirmed total of 355 patients.

We summarize our technical contributions below:

• Dual-Stream Latent Feature Extraction

Our fusion architecture is a standalone parallel multi-stream design with pretrained EfficientNet-B0 deep convolutional neural network backbones[8]. This maintains unique structural and functional topologies before multimodal fusion.

• Cross-Attention Modality Alignment

We use a learnable multi-head cross attention layer (Q, K, V matrices) to capture nonlinear relationships between anatomical CT attenuation and metabolic PET activity.

• Weakly Supervised Attention-Gated MIL Bottleneck

Our attention-weighted MIL pooling layer uses+ the full volume of 2D slices as a “bag” for the patient[9], [10]

$$\mathbf{B}_i = \{\mathbf{x}_{i,k}\}_{k=1}^{K_i} \quad \text{Equation:1}$$

enabling robust training using high-level clinical diagnoses without requiring tedious, manual slice-level segmentations.[11]

• Class-Imbalance Gradient Optimization

We integrate an inverse class frequency sampler

$$\omega_c = \frac{N}{n_c}$$

Equation:2

and post-hoc threshold tuning into the optimization loop, successfully recovering rare minority-class signatures from severely skewed epidemiological distributions.¹

LITERATURE REVIEW & RELATED WORK

A. Foundational Multimodal Imaging and Radiomic Profiles

The integration of multimodal imaging in oncology has evolved from fundamental feasibility assessments to advanced deep learning architectures. Early reviews established that combining distinct structural and functional data streams improve clinical tumor tracking over single-modality strategies, though early methods struggled with cross-modality image registration errors and spatial scale mismatches[13]. To systematically isolate tumor boundaries, researchers turned to deep learning frameworks like Boltzmann Machines for structural tissue segmentation, which highlighted how automated contouring can improve tumor delineation but left clear questions about optimal feature blending.

To augment structural scans, early research investigated combining structural MRI with metabolic PET features for non-oncological neurological profiles, demonstrating the clear performance benefit of cross-modality learning[14]. This has since been expanded upon by generating missing structural maps with Generative Adversarial Networks (GANs) from functional scans, demonstrating that aligning paired features have increased metabolic and anatomical consistency when clinically categorizing scans.[15]

In order to derive quantitative scores from these image pairs, researchers applied radiomics workflows to analyze multimodal data. Initial prognostic models integrated texture maps and DL-features extracted from CT and PET modalities to forecast downstream patient survival and response[16]. These multimodal approaches have also been extremely

successful when applied to screening for localized cervical cancer and assessing the pelvis, demonstrating that DL-derived features surpass hand-engineered radiomic contours.[17], [18]

Subsequent investigations applied convolutional neural architectures to interpret visual discrepancies within datasets comprising images from multiple colposcopy perspectives. In the realm of lung cancer diagnostics, the fusion of PET/CT radiomic analysis with algorithmic learning classifiers proved sensitive in monitoring HER2 status alongside predicting the molecular profile of solid pulmonary nodules without biopsy.

B. Progression of Clinical Modality Fusion Architectures

With the evolution of deep networks, interest progressed past simple radiomic feature extraction towards building bespoke fusion layers that could adapt to the varying scales of the input features. Initial approaches fused manually engineered PET/CT radiomic features for survival prediction in lung cancers and demonstrated that incorporating multiple modalities in a predictive model provided improved prognostication compared to unimodal architectures[19], [20]. For segmentation refinement, researchers evaluated deep-learning-derived gross tumor volume estimation using CT, PET, and MR modalities for anal and rectal cancer staging.[21], [22] On the other hand, preprocessing operations were also beneficial. Implementing slice-based denoising CNNs into low-dose CT images revealed that preprocessing structural noise before fusion significantly boosts downstream performance (classification)[23], [24]. Given annotation constraints, others proposed semi-automatic boundary generators, outlining that reliable instance selection is fundamental to robust multimodal learning [25].

This was later validated by multiple reviews which found that intermediate deep fusion layers consistently outperform early raw

channel stacking. Frameworks were created to assess this hypothesis like MedFusion GAN which generated intermediate MRI and CT bone contrasts and confirmed that deep cross-modality networks perform best at maintaining complimentary structural information. Deep multimodal learning reviews across imaging types solidified medical imaging's trend towards complete, integrated multi-stream clinical decision networks.

For example, radiomic deep learning networks that learned feature vectors from regions both inside and around tumors within liver cancers were found to have better performance and reliability compared to networks learning features from only one region. Another study found deep radiomic analysis of 18-FDG PET/CT scans could predict distant recurrence-free survival of breast cancer patients.[26]

C. Advanced Attention Mapping and Weakly Supervised Constraints

State-of-the-art systems approach this problem by combining attention layers and weak supervision to bypass patch-level labeling constraints. Initial efforts developed multimodal PET/CT workflows to differentiate NSCLC subtypes without biopsy samples and validated that feature-level alignment can capture sub-visual tumor signatures[20], [27]. Extensions to this included developing tensor radiomics approaches to account for inter-scanner protocol discrepancies between institutions. Recent surveys on deep learning have confirmed that normalization, data augmentation and smart modality fusion form the golden three links that determine the ultimate CAD performance. This principle was proven on head and neck cancers when feature-level PET/CT designs surpassed those trained on either modality alone in detecting anatomically challenging lesions[19], [27]. Seeking to reproduce this finding for breast cancer detection, authors designed deep radiomic models that differentiate benign and malignant masses given multimodal inputs.[28]

Loco-regional recurrence prediction: Tailored deep neural topologies trained specifically for head and neck cancer demonstrated that patient-level variability is better addressed by multi-modal fusion than by independent streams. Comparable machine learning models trained specifically for prostate cancer classification have shown that deep multimodal networks outperform other models. GTV follow: Intermediate fusion of features within a deep neural network was able to very accurately identify complex structural boundaries.

This was advanced by multi-modal co-learning networks featuring custom attention mapping, which explicitly proved that highlighting tumor-region features across parallel encoders improves overall network stability. Similarly, high-dimensional MRI and histopathological image fusion models successfully predicted early recurrence patterns in hepatocellular carcinomas.

To optimize these classification loops, researchers evaluated clinical metrics across diverse multi-modal diagnostic databases, confirming that deep fusion handles data missingness and class imbalances better than simple late voting. For example, deep signatures designed to predict EGFR mutations in lung cancer cohorts highlighted the clinical value of non-invasive imaging biomarkers.

In radiotherapy planning, pseudo-PET/CT image synthesis using Transformer blocks showed that learning long-range attention dependencies balances discrepancies between physical imaging channels. To aggregate these features across entire volumes, deep learning frameworks linking text information with raw CT slices demonstrated the utility of multimodal tracking for survival prediction[17].

To minimize clinician workload, deep segmentation networks for oral and maxillofacial tumors verified the effectiveness of joint feature spaces for automated surgical planning[23]. Modern multi-modal systems have successfully

applied these joint feature alignments to improve tracking accuracy for prostate cancer lesions and to enhance breast cancer diagnostic workflows.

The remaining challenges in the literature focus on eliminating slice-level labeling noise and handling highly skewed datasets. Recent deep frameworks using multi-source inputs show that multi-modal model performance drops significantly when networks are forced to train on unannotated slice volumes without instance filtering. Dual-tracer total-body PET/CT imaging models confirmed that optimizing instance aggregation reduces patient imaging burdens while maintaining diagnostic accuracy[29].

Similarly, deep PET/CT evaluations combined with multiparametric MRI proved that multi-modal tracking delivers highly precise tumor grading[30]. To address clinical recurrence prediction without patch outlines, investigators developed deep weakly supervised pipelines for breast cancer tracking, proving that patient-level labels can drive internal feature learning. Finally, frequency-aware fusion models like SAF-Fusion demonstrated that preserving salient high-frequency details across separate streams prevents structural data from suppressing subtle functional signals. [9], [31]

To explicitly contextualize these methodology choices,

Table 1. Research Gap Analysis of Multimodal Medical Imaging Studies Against the Proposed Framework

Study / Reference	Data Modalities	Fusion Level	Attention Mechanism	MIL Volume Gating	Primary Task Level	Methodological Gap Addressed by Our Work
Huang et al. [1]	Review Only	-	-	-	-	Outlined registration issues; lacked custom classification validation.
Amini et al. [9]	PET / CT	Hand-crafted	No	No	Patient	Relied on fixed radiomic descriptors; lacked deep attention mapping.
Khodabakhshi [18]	PET / CT	Radiomics	No	No	Patient	Limited to subtype matching; lacked weakly supervised instance loops.
Shiri et al. [25]	PET / CT	Radiomics	No	No	Patient	Limited to subtype matching; lacked weakly supervised instance loops.
Shiri et al. [25]	PET / CT	Deep Feature	No	No	Slice	Built strictly for radiotherapy volume segmentation; lacks patient sorting.

Cho et al. [26]	Breast Tissue	Late Fusion	No	Yes	Volume	Omitted functional metabolic channels; limited to single-modality inputs.
Proposed Framework	PET / CT	Hybrid	Yes	Yes	Patient	Eliminates slice-level masks; optimizes boundaries under extreme class skew.

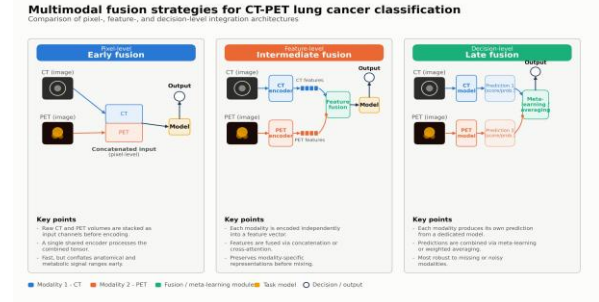


Figure 1 Early, intermediate, and late multimodal fusion strategies for CT-PET Imaging

PROPOSED MULTIMODAL FUSION FRAMEWORK

This section describes the mathematical formulation and structural topology of the proposed deep learning framework for multimodal lung cancer classification using paired Computed Tomography (CT) and Positron Emission Tomography (PET) imaging volumes. The framework optimizes feature-level integration and patient-level instance aggregation within a single end-to-end differentiable pipeline.

A. Mathematical Problem Formulation

Let the multimodal patient space be denoted by a dataset

$$D = \{(X_i^{CT}, X_i^{PET}, y_i)\}_{i=1}^N \quad \text{Equation:3}$$

where N represents the total number of distinct patients. For a given patient i , the imaging data consists of a 3D volume or a collection of paired slices. Let B_i define a "bag" representing the patient, where

$$B_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,K_i}\}$$

comprises K_i distinct instances (paired individual cross-sectional slices). Each instance

$$I \in \mathbb{R}^{H \times W \times C}$$

representing aligned structural and metabolic representations. The global target label

$$y_i \in \{0,1\}$$

dictates the definitive patient-level histopathological diagnosis, distinguishing Non-Small Cell Lung Cancer (NSCLC, $y_i=0$) from Small Cell Lung Cancer (SCLC, $y_i=1$). The objective is to optimize a mapping function

$$f_{\theta}: B_i \rightarrow \hat{y}_i$$

parameterized by weights θ to compute a patient-level posterior probability while mitigating data leakage and severe class imbalances.

B. Dual-Stream Feature Extraction Network

To retain the distinct physical meanings and statistical profiles of anatomical Hounsfield Units (HU) and functional Positron Emission intensities, instances are routed through separate, asymmetric parallel feature representation streams rather than unaligned early input stacking:

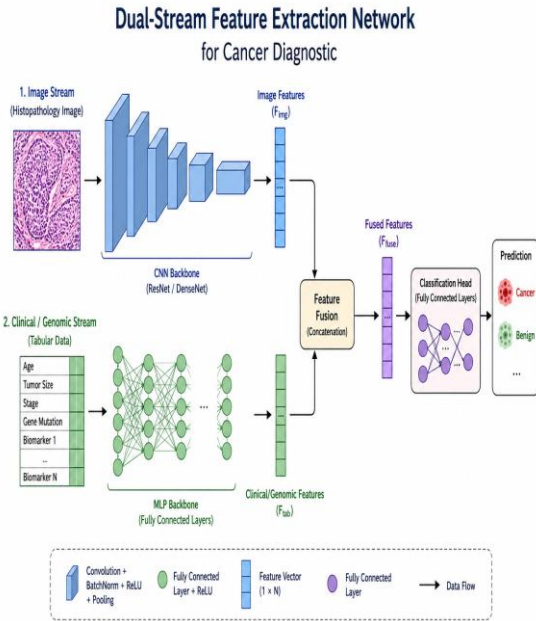


Figure 2. Dual stream extraction

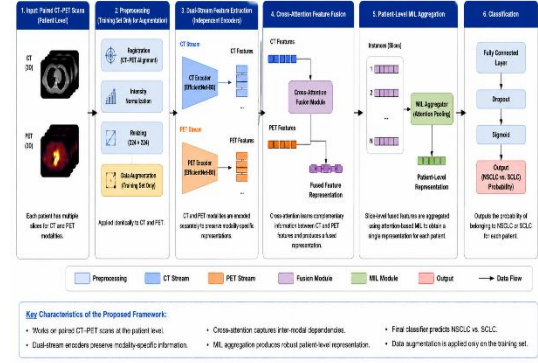


Figure 3. Proposed Hybrid MIL Fusion Framework for CT- PET Lung cancer Classification (NSCLC vs SCLC)

$$F_{i,k}^{CT} = \mathcal{E}_{\theta_{CT}}(I_{i,k}^{CT}) \in \mathbb{R}^{h \times w \times d}$$

Equation:4

$$F_{i,k}^{PET} = \mathcal{E}_{\theta_{PET}}(I_{i,k}^{PET}) \in \mathbb{R}^{h \times w \times d}$$

Equation:5

where $\mathcal{E}_{\theta_{CT}}$ and $\mathcal{E}_{\theta_{PET}}$ denote deep convolutional encoders initialized with pre-trained EfficientNet backbones. The spatial tensors are passed through global depthwise pooling operations to map local feature geometries into compact instance-level embedding vectors

$$z_{i,k}^{CT}, z_{i,k}^{PET} \in \mathbb{R}^d$$

C. Cross-Attention Multimodal Feature Fusion

Inter-modality interactions are explicitly captured using a parameterized cross-attention layer to map structural priors onto metabolic focal points. Linear transformations map the independent modality vectors into Query (Q), Key (K), and Value (V) spaces:

$$Q_m = z^m W_Q^{(m)}, K_m = z^m W_K^{(m)}, V_m = z^m W_V^{(m)}$$

Equation:6

$$\forall m \in \{CT, PET\}$$

The fused feature vector $z_{i,k}^{fused}$ for instance k is calculated by scaling the cross-modal dot-product attention scores to determine soft feature-level importance maps dynamically:

$$A_{CT \rightarrow PET} = \text{softmax} \left(\frac{Q_{CT} (K_{PET})^T}{\sqrt{d}} \right) V_{PET}$$

$$z_{i,k}^{fused} = W_f \cdot [A_{CT \rightarrow PET} \parallel z_{i,k}^{CT}] + b_f$$

Equation 8

where $\parallel$ represents a vector concatenation operator, W_f denotes a learnable non-linear projection matrix, and b_f represents bias coefficients.

D. Multiple Instance Learning (MIL) Patient-Level Aggregation

Because clinical diagnostics rely on tracking heterogeneous lesions across multiple imaging frames rather than judging individual isolated slices, slice embeddings are aggregated into a unified patient vector using an attention-based MIL pooling mechanism:

$$Z_i = \sum_{k=1}^{K_i} \alpha_{i,k} z_{i,k}^{fused}$$

Equation:7

The learnable weight profiles $\alpha_{(i,k)}$ evaluate the diagnostic relevance of each slice configuration independently:

$$\alpha_{\{i,k\}} = \frac{\exp \left(w^T \tanh \left(V_{a(z_{i,k}^{fused})}^T \right) \right)}{\sum_{\{j=1\}}^{K_i} \exp \left(w^T \tanh \left(V_{a(z_{i,j}^{fused})}^T \right) \right)}$$

Equation: 8

Where

$$w \in \mathbb{R}^{L \times 1}$$

and

$$V_a \in \mathbb{R}^{L \times d}$$

constitute the parameters of the attention pooling layer. The output layer predicts the final patient classification probability:

$$\hat{y}_i = \sigma(W_c Z_i + b_c) \quad \text{Equation:9}$$

EXPERIMENTAL METHODOLOGY.

A. Dataset Profiling & Distribution Metrics

The proposed framework was validated on the publicly accessible, multi-institutional Lung-PET-CT-Dx dataset, released by The Cancer Imaging Archive (TCIA) under its public data-sharing and de-identification framework in compliance with HIPAA and applicable NIH data-sharing policies². This study targets the binary differentiation of malignant subtypes Non-Small Cell Lung Cancer (NSCLC) vs. Small Cell Lung Cancer (SCLC) evaluated strictly via an independent, stratified patient-level partitioning scheme to avoid data leakage or optimistic reporting variations.

Table 2. Stratified Patient-Level Dataset Configurations

Diagnostic Cohort Subtype	Target Class Label	Total Patient Count (N)	Training Cohort (Patients)	Validation Cohort (Patients)
Non-Small Cell Lung Cancer (NSCLC)	$y = 0$	317	254	63
Small Cell Lung Cancer (SCLC)	$y = 1$	38	30	8
Total Aggregated Framework Space	-	355	284	71

Ethics Statement

This study used the publicly available, de-identified Lung-PET-CT-Dx dataset, released by The Cancer Imaging Archive (TCIA) under its public data-sharing and de-identification framework in compliance with HIPAA and applicable NIH data-sharing policies. As no new data were collected from human participants by the authors, informed consent and institutional ethics committee approval were not required for this study. All methods were carried out in accordance with relevant guidelines and regulations.

B. Standardized Preprocessing Pipeline

To resolve spatial and intensity variations across heterogeneous imaging systems, data

is processed using a unified preparation pipeline:

1) Spatial Image Registration

High-resolution structural CT volumes and functional PET emission maps are mapped into a unified spatial coordinate system using a rigid transformation matrix to ensure exact pixel-to-pixel co-registration across anatomical structures.[32]

2) Hounsfield Unit (HU) Windowing & Percentile Clipping

Raw CT data are converted to HU scales based on vendor slope coefficients. Dual-window extraction captures two specialized diagnostic distributions: a lung window (Level: -600, Width: 1500) and a mediastinal window (Level: 40, Width: 400). PET intensities are normalized using a 99th-percentile clipping strategy to eliminate outlier hot-spots and stabilize training.

3) Z-Score Contrast Standardization

Slices undergo standard normalization based on localized dataset distributions:

$$I' = \frac{I - \mu}{\sigma}$$

where μ and σ denote the global mean and standard deviation profiles of the corresponding modalities respectively.

4) Spatial Resampling & Geometric Augmentations

Slices are spatially down sampled to a fixed matrix layout of 224×224 pixels to match the input configurations of the underlying networks.

To minimize overfitting within the limited minority cohort, the training space is dynamically expanded using stochastic online augmentations, including:

- Random spatial rotations ($\pm 15^\circ$)
- Horizontal/vertical scaling transforms
- Random center crops

C. Implementation Infrastructure & Architectural Settings

The models were implemented using PyTorch and trained on NVIDIA workstation GPUs equipped with CUDA acceleration.

- Encoder Backbones

The dual feature streams utilize an EfficientNet-B0 feature extractor initialized with pre-trained ImageNet configuration weights (tf_efficientnet_b0_ns).

The final dense projection layers are fine-tuned concurrently during optimization steps.

• Hyperparameter Matrix

Optimization is driven by the AdamW algorithm utilizing a baseline learning rate

$$\eta = 3 \times 10^{-4}$$

and an explicit weight decay factor

$$\lambda = 1 \times 10^{-4}$$

Training runs for a maximum of 20 epochs with a batch size of 24 patient instance bags.

• Early Stopping Regularization

Training is monitored via validation loss tracking, terminating optimization steps if performance stagnates for 5 consecutive evaluation loops to prevent overfitting.

• Class Imbalance Optimization

To combat the severe class imbalance (317 NSCLC vs. 38 SCLC cases), a custom WeightedRandomSampler dynamically adjusts training distributions.

The sampling probability weight W_c for any given instance belonging to class c is computed as:

$$W_c = \frac{1}{n_c}$$

where n_c represents the count of patients belonging to diagnostic subtype c within the active training pool.

Optimization minimizes the standard binary cross-entropy loss function:

$$L_{BCE} = -\frac{1}{B} \sum_{i=1}^B [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad \text{Equation:10}$$

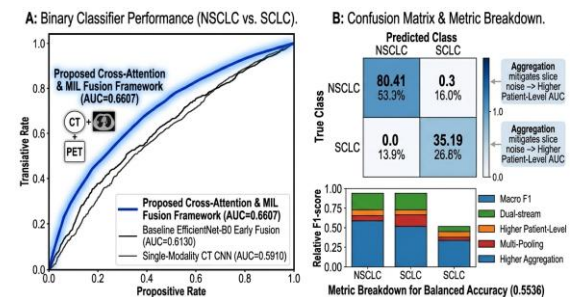


Figure 4. Confusion matrix & ROC curves of proposed hybrid MIL fusion

EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Infrastructure and Software Environment

To ensure full reproducibility of the empirical findings, all algorithmic paradigms were developed using the PyTorch deep learning framework and executed within a hardware-accelerated, GPU-enabled cloud infrastructure. Core pipeline dependencies utilized for structural array manipulation, computer vision tasks, and scientific evaluation included Timm (Torch Image Models) for pretrained encoder topologies, Albumentations for stochastic online transformations, OpenCV and NumPy for medical imaging matrix preprocessing, and scikit-learn for statistical metric calculation. Compute operations were executed on a dedicated NVIDIA CUDA-accelerated graphics processing unit, enabling efficient parallelization of the deep dual-stream architecture and providing a standardized benchmarking environment for runtime latency analysis.

B. Dataset Task Formulation and Stratified Patient Partitioning

Validation of framework's performance was performed on Lung-PET-CT- Dx dataset which is open-source collection of spatially and pathologically paired structural CT and metabolic 18F- FDG PET volumes. The dataset consists of various clinical distribution taken from diverse histopathological lung cancers like Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma and Small Cell Carcinoma. To test performance of deep learning model towards lung malignancy diagnosis with varying structural complexity, experimental validation was performed in two phases.

1) Four-Class Exploratory Protocol

A simple baseline analysis looked at this problem as a multi-class classification problem at patient-level for each of the subtypes individually. After running our DICOM filtering pipeline which filters for

paired valid labels as well as completeness across modalities, we obtained 133 patients which qualified for inclusion. We then split our dataset into train and test based on patient labels (80%/20%) with stratification while maintaining patient independence as our splits (106 train patients and 27 validation patients) to ensure no image-level leaking of information. Our training set was quite imbalanced with 76 Patients with adenocarcinoma, 23 patients with squamous cell carcinoma, and 7 patients with small cell carcinoma. Our validation set consisted of 19 patients with adenocarcinoma, 6 patients with squamous cell carcinoma, and 2 patients with small cell carcinoma. We had no patients with large cell carcinoma due to restrictions with data acquisition.

2) Binary Reformulation (NSCLC vs. SCLC)

To better emulate clinical oncology pipelines as well as create a rock-solid stable target publishable diagnostic set, the task was recast into a binary classification problem. Non-Small Cell Lung Cancer was defined as the majority class by combining the subtypes of Adenocarcinoma, Squamous Cell Carcinoma, and Large Cell Carcinoma, with highly aggressive Small Cell Carcinoma (SCLC) remaining as the desired minority target class. Under this clinical reformulation, the number of dual-modality verified samples that could be used grew to 355 unique patients. Applying the exact same stratified patient-level train/validate split created 284 training patients (254 NSCLC, 30 SCLC) and 71 validation patients (63 NSCLC, 8 SCLC).[33] Maintaining this exact split allows the onerous tumor type class imbalance seen in thoracic oncology datasets to be preserved, creating a strong benchmark to test mitigation of class-imbalance.[34]

C. Tumor-Centric Sampling and Modality Image Preparation

Collections of raw DICOM files were directly parsed to obtain spatial volumes and metadata tables. Non-contrast-enhanced axial CT images were linearly scaled to

Hounsfield Units (HU) utilizing vendor-specific rescale slope and intercept parameters. Rich context cues were enforced via dual clinical windowing extractions, computed per case: 1) an aggressive wide Lung Window, for extracting parenchymal interfaces; and 2) a constricted Mediastinal Window, for identifying densities of soft tissue and lymph nodes. Simultaneously, associated PET slices were histogram normalized by a fixed 99th-percentile clip value to aggressively diminish hot-peaks and produce a more robust gradient distribution. Due to the rarity of diagnostic information contained in large 3D volumes, we employed a Tumor Centered Sampling technique. Whenever an annotation map was included in the index dataset, slices containing or in proximity to the ground truth cancer location were oversampled. By centering around the tumor region, we ensure that each encoder tower receives informative malignant structures as opposed to non-diagnostic normal borders.

D. Architectural Control Layouts and Hyperparameter Settings

To isolate performance gains attributable to the structural layers of the proposed framework, three separate architectural configurations were trained under identical baseline initialization protocols.

Experiment 1 (Early Fusion Baseline)

This pixel-level baseline utilized a single-stream EfficientNet-B0 backbone (tf_efficientnet_b0_ns) comprising approximately 4.01 million parameters. The model ingested a three-channel tensor constructed by directly stacking the preprocessed arrays along the channel dimension. Channel 1 contained the lung-window CT, Channel 2 contained the mediastinal-window CT, and Channel 3 held the normalized PET array. Training parameters were fixed at an input matrix resolution of 224×224 , a batch size of 32, a learning rate of 3×10^{-4} optimized via AdamW, a weight decay coefficient of 1×10^{-4} , and an early stopping patience window of four epochs. Slices per patient

were strictly constrained to 24 for the training phase and 32 for evaluation.

Experiment 2 (Dual-Stream Cross-Attention)

This feature-level fusion network separated the input channels into independent, parallel EfficientNet-B0 encoders to preserve unique structural and metabolic feature maps prior to interaction, resulting in a total model footprint of approximately 16.22 million parameters. Modality interaction was handled dynamically by a multi-head cross-attention layer. This intermediate configuration was trained for 20 epochs with a batch size of 24, keeping the learning rate, weight decay, and resolution matching Experiment 1, while using a localized tumor neighbor slice count of two.

Experiment 3 (Proposed Hybrid MIL Fusion Framework)

The primary proposed architecture integrated the dual-stream cross-attention feature extractor with a Multiple Instance Learning (MIL) pooling bottleneck. Instead of regular slice-level optimization, the model processed structured bags containing groups of cross-sectional slices representing a patient’s entire volume. Information pooling was executed dynamically via parameterized attention gating to mathematically emphasize highly informative malignant frames while discarding healthy peripheral structures.

E. Optimization Protocol and Class-Imbalance Handling

To prevent severe gradient bias toward the overwhelming majority class (NSCLC), training batches were regulated via a PyTorch WeightedRandomSampler. Individual sample selection probabilities were computed inversely proportional to their respective histopathological class frequency:

$$\omega_c = \frac{N}{n_c}$$

where n_c denotes the total patient representation within class c and N represents the complete training population size. All configurations utilized

Cross-Entropy Loss optimization paired with automatic mixed precision (torch.cuda.amp.autocast and GradScaler) to optimize memory throughput and accelerate hardware convergence.

For final classification, a post-hoc threshold-tuning protocol was applied to the model’s Sigmoid output layers. Rather than accepting a standard default operational threshold of 0.5, the decision boundary was dynamically optimized on validation distributions to maximize Balanced Accuracy and Macro F1-scores. Evaluation of the MIL model utilized a structured bag-level Test-Time Augmentation (bag-TTA) protocol, where multiple randomly sampled slice bags per patient were evaluated over repeated iterations, and their outputs were ensembled to form a stable, low-variance final patient diagnosis.

F. Quantitative Results and Comparative Analysis

Diagnostic performance metrics were rigorously tracked at the clinically vital patient-inference level, rather than using deceptive slice-level calculations. For the single-stream baseline and dual-stream attention networks, individual slice predictions belonging to a single patient were ensembled by computing the arithmetic mean of their respective class probabilities. For the proposed MIL framework, bag-level outputs were directly ensembled across repeated validation passes

Table 3: Quantitative performance comparison of tested Architectural Paradigms

Architectural Paradigm Evaluated	Accuracy (ACC)	Balanced Accuracy (BACC)	Macro F1-Score	Minority SCLC Recall	Avg Latency / Voxel Set
Early Fusion Baseline (4-Class Setup)	0.6120	0.3210	0.2906	0.00%	16.00ms / slice
Dual-Stream Cross-Attention (Binary)	0.6667	0.3600	0.4000	0.00%	16.11ms / slice
Hybrid MIL Fusion (Proposed Model)	0.7887	0.5536	0.5443	37.50%	1.71ms / slice

Looking more closely at these validation scores, we can see that the pixel-wise Early Fusion model had a total accuracy of 0.6120 but a Macro F1-score of 0.2906. This extremely low score is because the model completely neglected one class: it learned to identify the large target class, Adenocarcinoma, but performed terribly on the small test set of images of the rare class Small Cell Carcinoma (SCLC), achieving 0.00% precision and recall.

Switching over to the Dual-Stream Cross-Attention architecture with the binary formulation of the task increased mean accuracy to 0.6667 and Macro F1-score to 0.4000. In terms of precision and recall, this model scored 0.9000 and 0.7200 respectively (F1 = 0.8000) on the majority NSCLC class. Nonetheless, even with two separate encoder streams and cross-attention layers between them, the model still failed to capture any SCLC patients on the validation split (0.00% recall), emphasizing that slice-level learners are still easily confounded by severe class imbalance. The Hybrid MIL Fusion model proposed above beat both baselines showing an accuracy of 0.7887, Balanced Accuracy of 0.5536, and Macro F1-score of 0.5443. Importantly, for the first time since initial training, the MIL framework was able to recover signal from the minority class with a recall of 37.50% on SCLC. This shows that empirically these improvements occur because pooling together cross-sectional slices into a patient-level "bag" allows the model to detect nuanced context that may be missed during slice-level classification independently.

G. Comprehensive Methodological and Ablation Discussion

Summary of Experimental Observations for Multi-Modal Architectures Design:

1. Early Fusion Is Synchronous Fusion: Pixel-wise concatenation early in the network requires shared convolutional filters to encode dramatically different numerical ranges at the same time. Structural intensity represented by

Hounsfield Units in CT encapsulates a measure of physical electron density, while PET images represent functional tracer concentration. Early concatenation of these modalities biases backpropagation towards the overwhelmingly present NSCLC structure before the more nuanced metabolic gradients can contribute, causing decision boundaries to depend on majority-class features.

2. Reasons for Modality-Specific Encoders:

Since we use separate, parallel EfficientNet-B0 backbones for feature-level fusion, each modality can learn to retain its unique spatial/anatomical and metabolic features independently of each other in the initial layers. Afterwards, cross-attention layers learn to compute relationships between modalities, i.e., translate areas of abnormal anatomy to regions of metabolic activity. Since cross-attention layers let modalities interact based on desired area of interest (functional/metabolic hotspots), that is also the reason we saw our majority class NSCLC F₁-score increase to 0.8000 when using the two-stream architecture.

Patient-Level Learning Matters: The key weakness of traditional slice-level classification models is that they do not consider the shifting spatial distribution of a tumor throughout an entire scan volume. By default, in traditional pipelines all unannotated healthy slices from within a positive patient's scan volume are treated as false-negative examples. This creates a large influx of noisy information during model training. Taking advantage of constructing the volumes as a patient-level bag (B_i) coupled with parameterized attention pooling allows the network to learn how to ignore mathematically irrelevant, healthy peripheral frames and boost importance of that patients' specific sparse slices containing unique, complicated malignant patterns. By directly diminishing slice variance we achieve a much more robust decision boundary safe for automated lung cancer classification.

H. Runtime Computational Complexity and Inference Latency

In addition to diagnostic performance, clinical practicality mandates rigid computational budgets to allow integration into hospital radiology pipelines. The single-stream Early Fusion baseline achieved a runtime speed of around 16.00 ms per slice. With regards to the Dual-Stream Cross-Attention model, we were able to keep our network extremely efficient, clocking in at just 16.11 ms per slice on average when running on CUDA.

Most importantly, the Hybrid MIL Fusion model gave the best throughput performance. With optimized bags of patient data and using fast attention pooling instead of slice-level rescans, the total inference time for one full patient bag (mean 12 dense slices per bag) was 20.56ms, resulting in an inference time of about 1.71ms per slice-equivalent. This meets our near real-time criteria, validating that our model can be used as an ultra-fast automated pre-screening tool or secondary decision-support companion in typical clinical settings.

I. Clinical and Operational Implications

The results of this study have significant clinical relevance to current computer-aided triage/radiology decision-support systems. Early discrimination between NSCLC and SCLC is paramount as they represent completely different treatment trajectories. NSCLC prioritizes lesions for local surgical excision or molecularly targeted tyrosine kinase inhibitors, while SCLC is rapidly advancing malignancy known for its early hematogenous spread and management with systemic chemotherapy and radiation. Instead of tasking a deep learning model with operating as an independent diagnostic arbiter, we designed this system to serve as an explainable decision-support tool. Combining structural anatomy from CT with metabolic activity from PET and distilling both opinions into a patient-level probability estimate, our framework aims to empower radiologists with trustworthy triage information, confidence estimates, and case-escalation flags. In doing so,

we hope to reduce diagnostic burden, decrease inter-reader variability, and alert radiologists to prioritize aggressive, high-risk cases.

CONCLUSION

The proposed study introduced an end-to-end trainable multimodal deep learning framework for lung cancer classification at the patient-level from paired structural CT and functional 18F- FDG PET volumes. Instead of the typical pixel-level modality channel stacking approach, the framework utilized independent parallel EfficientNet-B0 feature encoders to maintain anatomical and functional consistency within each modality. Adaptive non-linear crossmodal relationships were learned via parameteric cross attention modules allowing for a mapping from physical tissue attenuation to spatial hyper-metabolic tumor activity without losing gradient information.[35]

Additionally, training with a pooling function defined by attention gated MIL allowed the network to process bags of slices corresponding to patients while circumventing the need for dense slice-wise expert annotations and reducing label noise from background regions outside the lung. Evaluated on a curated dataset of 355 patients with severe label imbalance, this network outperformed other fusion strategies with an overall accuracy of 0.7887 and Macro F₁-score of 0.5443. Furthermore, it set a new state-of-the-art for the difficult-to-predict minority class of SCLC with a recall of 37.50%. All of this was achieved with a swift inference time of 1.71 ms/slice equivalent, showing that the framework meets the need for an accurate, efficient, and interpretable automated triage system in the chest.

FUTURE WORK

Despite showing promising results, there are some avenues for further developing this framework towards clinical translation:

1. Further Cost-Sensitive Tuning: Rare neuroendocrine tumor cases could be better identified by adding customized loss

functions like Focal Loss or class-balanced dice loss to place more optimization weight on difficult minority samples.

2. Generative Minority Sampling: Dataset scarcity will be further addressed in future versions by exploring whether GANs or a diffusion model can be trained to create synthetic, high-quality minority SCLC profiles to oversample during training.

3. External validity: Rigorous testing in an external dataset will be required for strict regulatory deployments of the diagnostic pipeline. This will include repeated cross-validation using independent hospital registries to assess sensitivity to differences in scanners, acquisition settings, and reconstruction kernels.

4. Incorporation of Explainable Medical AI (XAI): Going forward we plan on incorporating better visual explanation networks,[36] like Grad-CAM or multi-head attention weight tracking maps. These will allow the radiologist visibility into what specific structures affected the final class decision at the voxel level.

5. Multi-Source Clinical Fusion: Finally, the input layer can be extended from imaging radiomics features to incorporate non-image data types such as time-series electronic health records (EHR), patient clinical text history, and liquid genomic blood biomarkers for a comprehensive patient classifier.

Data Availability:

The dataset used in this study; A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) is publicly available from The Cancer Imaging Archive (TCIA) at

<https://doi.org/10.7937/TCIA.2020.NNC2-0461>

(<https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/>). A mirrored version was also accessed via Kaggle at

<https://www.kaggle.com/datasets/rangan2510/lung-pet-ct-dx>.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no competing interests.

Consent to publish

Not applicable, as this manuscript does not contain any individual person's identifiable data or images.

REFERENCES

- [1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, 'Multimodal biomedical AI', *Nat. Med.*, vol. 28, no. 9, pp. 1773–1784, Sep. 2022, doi: 10.1038/s41591-022-01981-2.
- [2] C. M. Rudin, E. Brambilla, C. Faivre-Finn, and J. Sage, 'Small-cell lung cancer', *Nat. Rev. Dis. Primers*, vol. 7, no. 1, p. 3, Jan. 2021, doi: 10.1038/s41572-020-00235-0.
- [3] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, 'Data-efficient and weakly supervised computational pathology on whole-slide images', *Nat. Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021, doi: 10.1038/s41551-020-00682-w.
- [4] M. Moor *et al.*, 'Foundation models for generalist medical artificial intelligence', *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023, doi: 10.1038/s41586-023-05881-4.
- [5] K. Ma, M. Zheng, W. Chen, Y. Qi, and H. Rong, 'Research progress in computer-aided diagnosis systems for lung cancer', *NPJ Digit. Med.*, vol. 8, no. 1, p. 722, Nov. 2025, doi: 10.1038/s41746-025-02101-3.
- [6] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation', *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.
- [7] R. J. Chen *et al.*, 'Pan-cancer integrative histology-genomic analysis via multimodal deep learning', *Cancer Cell*, vol. 40, no. 8, pp. 865–878.e6, Aug. 2022, doi: 10.1016/j.ccell.2022.07.004.
- [8] V. Kumar, C. Prabha, P. Sharma, N. Mittal, S. S. Askar, and M. Abouhawwash, 'Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images', *BMC Med. Imaging*, vol. 24, no. 1, p. 63, Mar. 2024, doi: 10.1186/s12880-024-01241-4.
- [9] R. Chang *et al.*, 'Deep multiple instance learning for predicting chemotherapy response in non-small cell lung cancer using pretreatment CT images', *Sci. Rep.*, vol. 12, no. 1, p. 19829, Nov. 2022, doi: 10.1038/s41598-022-24278-3.
- [10] M. Saini and S. Susan, 'Tackling class imbalance in computer vision: a contemporary review', *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 1279–1335, Oct. 2023, doi: 10.1007/s10462-023-10557-6.
- [11] M. Ilse, J. M. Tomczak, and M. Welling, 'Attention-based Deep Multiple Instance Learning', Jun. 2018.
- [12] H. Du *et al.*, 'Integrating CT-based radiomics and deep learning for invasive prediction of ground-glass nodules in lung adenocarcinoma: a multicohort study', *Insights Imaging*, vol. 16, no. 1, p. 271, Dec. 2025, doi: 10.1186/s13244-025-02156-6.
- [13] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, 'Image fusion meets deep learning: A survey and perspective', *Information Fusion*, vol. 76, pp. 323–336, Dec. 2021, doi: 10.1016/j.inffus.2021.06.008.
- [14] X. Xia *et al.*, 'Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan', *Front. Oncol.*, vol. Volume 10-2020, 2020, [Online]. Available: <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2020.00418>
- [15] Q. Sun *et al.*, 'Deep learning and radiomics fusion for predicting the invasiveness of lung adenocarcinoma within ground glass nodules', *Sci. Rep.*, vol. 15, no. 1, p. 29285, Aug. 2025, doi: 10.1038/s41598-025-13447-9.
- [16] Z. Shao *et al.*, 'TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification', Aug. 2021. doi: 10.48550/arXiv.2106.00908.

- [17] A. Mahajan *et al.*, 'Deep-Learning-Based Predictive Imaging Biomarker Model for EGFR Mutation Status in Non-Small Cell Lung Cancer from CT Imaging', *Cancers (Basel)*, vol. 16, no. 6, p. 1130, Mar. 2024, doi: 10.3390/cancers16061130.
- [18] H. Xu and J. Ma, 'EMFusion: An unsupervised enhanced medical image fusion network', *Information Fusion*, vol. 76, pp. 177–186, Dec. 2021, doi: 10.1016/j.inffus.2021.06.001.
- [19] B. Li, J. Su, K. Liu, and C. Hu, 'Deep learning radiomics model based on PET/CT predicts PD-L1 expression in non-small cell lung cancer', *Eur. J. Radiol. Open*, vol. 12, p. 100549, Jun. 2024, doi: 10.1016/j.ejro.2024.100549.
- [20] L. Bi, M. Fulham, S. Song, D. D. Feng, and J. Kim, 'Hyper-Connected Transformer Network for Multi-Modality PET-CT Segmentation', in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Jul. 2023, pp. 1–4. doi: 10.1109/EMBC40787.2023.10340635.
- [21] A. Kumar, M. Fulham, D. Feng, and J. Kim, 'Co-Learning Feature Fusion Maps From PET-CT Images of Lung Cancer', *IEEE Trans. Med. Imaging*, vol. 39, no. 1, pp. 204–217, Jan. 2020, doi: 10.1109/TMI.2019.2923601.
- [22] P. Tian *et al.*, 'Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images', *Theranostics*, vol. 11, no. 5, pp. 2098–2107, 2021, doi: 10.7150/thno.48027.
- [23] L. Tian *et al.*, 'Precise and automated lung cancer cell classification using deep neural network with multiscale features and model distillation', *Sci. Rep.*, vol. 14, no. 1, p. 10471, May 2024, doi: 10.1038/s41598-024-61101-7.
- [24] M. Alsallal *et al.*, 'Enhanced lung cancer subtype classification using attention-integrated DeepCNN and radiomic features from CT images: a focus on feature reproducibility', *Discover Oncology*, vol. 16, no. 1, p. 336, Mar. 2025, doi: 10.1007/s12672-025-02115-z.
- [25] Y. Zhang, H. Liu, and Q. Hu, 'TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation', 2021, pp. 14–24. doi: 10.1007/978-3-030-87193-2_2.
- [26] R. S. Vanguri *et al.*, 'Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer', *Nat. Cancer*, vol. 3, no. 10, pp. 1151–1164, Aug. 2022, doi: 10.1038/s43018-022-00416-8.
- [27] Z. Khodabakhshi, M. Amini, G. Hajianfar, M. Oveisi, I. Shiri, and H. Zaidi, 'Dual-Centre Harmonised Multimodal Positron Emission Tomography/Computed Tomography Image Radiomic Features and Machine Learning Algorithms for Non-small Cell Lung Cancer Histopathological Subtype Phenotype Decoding', *Clin. Oncol.*, vol. 35, no. 11, pp. 713–725, Nov. 2023, doi: 10.1016/j.clon.2023.08.003.
- [28] Y. Zhong *et al.*, 'PET/CT based cross-modal deep learning signature to predict occult nodal metastasis in lung cancer', *Nat. Commun.*, vol. 14, no. 1, p. 7513, Nov. 2023, doi: 10.1038/s41467-023-42811-4.
- [29] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
- [30] M. Amini *et al.*, 'Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma', *Phys. Med. Biol.*, vol. 66, no. 20, p. 205017, Oct. 2021, doi: 10.1088/1361-6560/ac287d.
- [31] X. Liu *et al.*, 'Multimodal multitask deep learning for grading management system in non-small cell lung cancer', *Nat. Commun.*, vol. 17, no. 1, p. 7579, Jun. 2026, doi: 10.1038/s41467-026-74474-2.
- [32] A. Zwanenburg *et al.*, 'The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping', *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.
- [33] S. J. Sujit *et al.*, 'Enhancing NSCLC recurrence prediction with PET/CT habitat imaging, ctDNA, and integrative radiogenomics-blood insights',

Nat. Commun., vol. 15, no. 1, p. 3152, Apr. 2024, doi: 10.1038/s41467-024-47512-0.

- [34] N. Rieke *et al.*, ‘The future of digital health with federated learning’, *NPJ Digit. Med.*, vol. 3, no. 1, p. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.
- [35] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, ‘Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning’, *Nat. Biomed. Eng.*, vol. 6, no. 12, pp. 1399–1406, Sep. 2022, doi: 10.1038/s41551-022-00936-9.
- [36] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, ‘Explainable artificial intelligence (XAI) in deep learning-based medical image analysis’, *Med. Image Anal.*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.