Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 May 2026
Revised 18 July 2026
Accepted 04 August 2026

Natural Resources for Human Health 2026,Vol.6(11s):1579-1605

KEYWORDS:
NSCLC, early detection, radiogenomics, multimodal learning, CT, XAI, FL, LUAD, LUSC

Hierarchical Explainable Radiogenomics for Early Lung Cancer

Rahul vyas[1], Dr. Prashant Shukla[2]

[1] Department of Computer Application, United University Prayagraj, Uttar Pradesh, India 211012.
[2] Department of Computer Science and Engineering, United University Prayagraj, Uttar Pradesh, India 211012.
[1] rahulvyas.phdca21@uniteduniversity.edu.in
[1]Orcid id-https://orcid.org/0009-0009-1408-2610
[2] prashant.shukla@uniteduniversity.edu.in
[2]Orcid id-https://orcid.org/0000-0002-1810-6650

ABSTRACT: Lung cancer is a leading cause of cancer deaths, demanding screening systems that are both accurate and interpretable. We propose a hierarchical radiogenomic framework that first performs early screening and then molecular subtype classification. In Stage A, a lightweight attention-based CT model distinguishes non-cancer from cancer cases. Only cases predicted as high-risk or uncertain are forwarded to Stage B, where a tri-modal classifier integrates chest CT, structured clinical variables, and genomic profiles to separate LUAD from LUSC using a 3D ResNet encoder, MLP, and denoising autoencoder with attention-guided fusion. Interpretability is achieved via Integrated Gradients (CT), DeepSHAP (clinical/genomic), ConceptSHAP (pathway-level factors), and DiCE counterfactuals. An uncertainty-aware router with probability calibration governs escalation from screening to tri-modal analysis. Rather than introducing a new single network component, the framework contributes a clinically motivated end-to-end design that unifies calibrated screening, radiogenomic subtype analysis, and multi-level explanation within one workflow. Evaluated on NLST and NSCLC-Radiogenomics, and additionally examined under a simulated federated learning setting with non-IID client partitions, the framework outperforms unimodal and naive fusion baselines while maintaining performance close to the centralized setting.

INTRODUCTION

Lung cancer remains among the most lethal malignancies worldwide, responsible for a substantial fraction of cancer-related mortality. Prognosis is tightly coupled to diagnostic timing: detecting disease at an early stage markedly improves the prospects for curative intervention, yet a large proportion of cases are still discovered late. While low–dose computed tomography (CT), as established in the National Lung Screening Trial (NLST), has increased screening sensitivity [1], image evidence alone often lacks the clinical and molecular context required for individualized risk assessment and subtype-aware management. Robust diagnosis and patient stratification therefore demand models that combine radiological appearance with patient-level clinical factors and molecular signatures reflective of underlying tumor biology. Deep learning has progressed rapidly in automated chest CT interpretation, including lung nodule detection, segmentation, and malignancy assessment. However, many pipelines remain single-modality, emphasizing radiological features while overlooking critical variables such as smoking history, comorbidities, and genomic alterations. This limited scope underestimates the heterogeneity of non–small cell lung cancer (NSCLC) and can hinder reliable generalization across populations. Radiogenomics-the integration of imaging phenotypes with clinical and genomic information has emerged as a promising path toward precision oncology by linking macro-level morphology to micro-level biology.

Despite technical advances, two barriers restrict clinical adoption. First, a lack of transparent reasoning impedes trust: clinicians require clear evidence for why a lesion is flagged as malignant and which cross-modal factors drive a decision. Conventional explanation tools (e.g., saliency maps or feature attributions) are typically applied within a single modality and provide limited insight into interactions across imaging, clinical, and genomic domains. Second, real-world

deployment is constrained by *data privacy* and the siloed nature of medical records. Institutional policies and regulations often preclude centralized data pooling, limiting model development with sufficiently diverse cohorts. This study addresses these gaps by introducing an explainable, privacy-preserving multimodal radiogenomic framework with a clearly staged design. In Stage A, a calibrated CT-only screening model trained on NLST distinguishes non-cancer from cancer cases. In Stage B, cases flagged as high-risk or uncertain are further analyzed by a tri-modal classifier that fuses chest CT, structured clinical variables, and genomic profiles to differentiate lung adenocarcinoma (LUAD) from lung squamous cell carcinoma (LUSC) on the NSCLC–Radiogenomics cohort. An uncertainty-aware router with probability calibration governs the transition between stages, supporting safe triage and reducing unnecessary downstream computation. Model interpretability is provided at multiple levels using voxel-wise saliency for CT, feature attributions for the clinical and genomic branches, concept-based links to biological pathways, and counterfactual analyses to probe decision robustness and actionability.

To promote scalability without compromising confidentiality, the framework is also evaluated under a federated learning simulation that allows multiple sites to contribute to a shared global model without exchanging raw patient data. This design leverages population diversity while aligning with regulatory constraints and institutional governance.

Contributions. The primary contributions are:

- **Hierarchical early detection to subtyping.** This study formulates a two-stage pipeline that couples CT-based screening (non-cancer vs. cancer) with tri-modal LUAD/LUSC classification, guided by uncertainty-aware routing and calibrated probabilities.
- **Cross-modal, clinically grounded explainability.** Multi-level explanations connect voxel-level CT evidence with clinical and genomic drivers and map latent genomic factors to pathway-level concepts, enabling biologically coherent interpretation.
- **Privacy-preserving learning under simulated decentralized settings.** A simulated federated learning framework with non-IID client partitions is used to evaluate the feasibility of decentralized training while preserving data locality. The federated experiments are intended to assess performance under heterogeneous data distributions and should be interpreted as a proof-of-feasibility rather than validation across independent clinical institutions.
- **Comprehensive evaluation.** Experiments evaluate Stage A screening on NLST and Stage B subtyping on NSCLC–Radiogenomics, with comparisons against unimodal and early-fusion baselines and assessment of discrimination, calibration, and explanation fidelity.
- **Central methodological innovation.** The central innovation of this work is the hierarchical integration of calibrated CT-based screening with selective tri-modal radiogenomic subtyping. Rather than proposing a new standalone network block, the framework organizes screening, uncertainty-aware routing, multimodal fusion, calibration, and explanation into a single clinically motivated decision pipeline. This distinguishes the study from prior radiogenomic models that mainly focus on single-stage prediction or isolated fusion/explainability components.

  By unifying early screening with subtype-specific analysis and embedding rigorous interpretability and federated training, this study advances an end-to-end, clinically practical pathway for decision support in lung cancer care.

LITERATURE REVIEW

A. Deep Learning for Lung Cancer (Image–Only)

Early CT–only pipelines established strong baselines for nodule analysis and malignancy prediction. Consolidated surveys and benchmark systems consistently report that 2D/3D CNNs surpass handcrafted radiomics across detection and classification tasks [2, 3]. Attention mechanisms further improve sensitivity and specificity in heterogeneous cohorts by emphasizing lesion–relevant regions [4]. Concurrently, screening trials such as the National Lung Screening Trial (NLST) demonstrated the clinical value of low–dose CT for early detection [1], motivating algorithmic work on LDCT sensitivity and false positive reduction. Foundational radiogenomic studies also linked imaging phenotypes to molecular signatures in NSCLC, motivating deeper integration between scan appearance and tumor biology [5, 6]. More recent radiogenomic analyses explicitly connect CT–derived phenotypes to gene mutations and pathway activity in NSCLC [7]. These developments have been facilitated by public repositories and data access portals such as NLST CDAS and TCIA [8, 9], which enable reproducible benchmarking across institutions. Nevertheless, image–only models underuse patient context (e.g., clinical risks, biomarkers, and genomic alterations), which can materially change diagnostic decisions and motivate multimodal approaches. Moreover, purely image–based pipelines often exhibit dataset shift when deployed across scanners, protocols, or institutions, highlighting the need for models that integrate richer sources of information and support more robust early-stage decision making.

B. Multimodal Learning and Radiogenomics in Oncology

Architectures that combine imaging with clinical and genomic variables are increasingly central to precision oncology [10, 11]. Integrative encoders that jointly model clinical, genomic, and imaging data have improved tasks such as survival prediction and subtype classification [12, 13]. Hybrid fusion designs that combine CNN backbones with MLP branches for structured clinical variables have shown benefits in brain and breast tumor classification [14, 15], while radiogenomic models for NSCLC specifically exploit CT patterns to explain mutation status and pathway activity [7, 16]. Recent work emphasizes attention–guided or gated fusion to adaptively weight modalities, yielding performance gains over naive concatenation or early fusion [10, 17]. Surveys synthesize design choices such as fusion stage, missing–modality handling, calibration, and external validation, and propose broader evaluation principles for multimodal pipelines [11, 18]. Collectively, these studies indicate that modality–aware fusion improves robustness and interpretability when combining CT with structured metadata and molecular profiles — a premise adopted here via attention–based fusion and modality–aware training. However, most existing frameworks focus on single–stage classification or prognosis within a single cohort; fewer works explicitly link a screening stage to downstream radiogenomic subtyping while maintaining calibrated probabilities and explicit handling of missing or partially observed modalities.

C. Explainable AI for Medical Imaging and Radiogenomics

Classical explanation methods—Grad–CAM, SHAP, and axiomatic attributions—remain widely used to visualize model reasoning at image and feature levels [19–21]. Modern practice in medical imaging increasingly emphasizes quantitative evaluation of explanations using faithfulness, stability, and calibration-aware criteria [22–24]. In parallel, work on network calibration has highlighted the need for reliable probability estimates and post–hoc calibration schemes such as temperature scaling [25], which are particularly important when model outputs are used for routing or threshold–based screening. In radiogenomics, concept–based explanations that align latent factors with pathway gene sets (e.g., *EGFR*, *PI3K*, *TP53)* provide biologically grounded narratives for model behavior [13, 26]. Counterfactual generation complements attribution by probing minimal, plausible changes that invert predictions and test actionability [27]. Building on these directions, this study adopts a multi–level XAI protocol: voxel–wise Integrated Gradients for CT, DeepSHAP for clinical/genomic features with stability checks, ConceptSHAP for pathway alignment, and DiCE–style counterfactuals to assess decision robustness. This combination aims not only to visualize saliency but also to connect model decisions to clinically and biologically meaningful factors, thereby supporting trust and potential clinical translation.

D. Federated Learning in Medical Imaging and Radiogenomics

Federated learning (FL) enables multi–institutional training without sharing raw data. Canonical algorithms such as FedAvg and FedProx address decentralized optimization under client heterogeneity [28, 29]. Foundational work demonstrated FL feasibility and privacy benefits in medical imaging [30, 31], and surveys have cataloged systems challenges such as client heterogeneity and aggregation design [32, 33]. For lung CT classification, federated ResNet experiments showed comparable performance to centralized training under realistic hospital splits [34]. In practice, differentially private optimization (DP–SGD) can further reduce information leakage [35]. Related work on explainable and trustworthy FL has begun to study how distributed updates and local attributions can be made more transparent [36, 37]. In this work, FedAvg/FedProx with optional DP mechanisms are used to balance utility and privacy while reporting both client–level and global behavior. Yet, most FL applications in oncology have focused on imaging–only models or two–site collaborations; radiogenomic FL with full tri–modal fusion and explicit calibration is less explored, leaving open questions about cross–site robustness, modality imbalance, and explainability in a privacy–preserving setting.

E. Limitations of Prior Approaches

Despite rapid progress, important gaps remain: (i) incomplete coverage and potential imputation bias in clinical and genomic fields; (ii) scarce concept–level biological explanations and few cross–site faithfulness protocols; (iii) under–reporting of calibration and uncertainty alongside discrimination metrics [25]; and (iv) limited unification of tri–modal fusion, multi–level XAI, and privacy–preserving training within a single NSCLC pipeline. Existing systems often address only one or two of these aspects in isolation (e.g., image–only FL without radiogenomics, or centralized multimodal fusion without privacy guarantees), making it difficult to assess how such components interact in realistic screening and subtype–classification workflows. The proposed study addresses these gaps by combining attention–guided tri–modal fusion (CT + clinical + genomics) with integrated IG/DeepSHAP/ConceptSHAP, a hierarchical screening→subtyping

design, and a federated training setup (FedAvg/FedProx with optional DP), while explicitly reporting calibration and uncertainty. In doing so, it aims to provide a more end–to–end view of how early screening, radiogenomic subtyping, interpretability, and federated deployment can be jointly realized for NSCLC.

Table I summarizes representative literature from 2016–2025 across CT-only pipelines, multimodal radiogenomics, explainable AI, and federated learning, and highlights the methodological gap addressed by the present study.

TABLE I: Representative literature (2016–2025) across explainability, CT-based lung cancer analysis, multimodal radiogenomics, and federated learning. The final row summarizes the contribution of the present study.

Author(s) & Year	Task / Modality	Method	Explainability	Limitation
Ribeiro et al., 2016 [38]	Model-agnostic (general)	LIME (local surrogate)	LIME	Instability across runs; sensitivity to sampling
Selvaraju et al., 2017 [19]	Generic vision → medical Imaging	Grad-CAM framework	Heatmaps	Image-only interpretation; coarse localization
Lundberg & Lee, 2017 [20]	Model-agnostic / tabular	SHAP framework	Global/local attributions	Computational cost; background dependence
Sundararajan et al., 2017 [21]	Deep networks	Integrated Gradients	Axiomatic feature attribution	Does not directly model cross-modal interactions
Song et al., 2017 [2]	Lung nodules (CT)	DL based nodule classification	----------	Image-only scope; no clinical/genomic information
Alakwaa et al., 2017 [3]	Lung cancer classification	3D-CNN	----------	Small data set; image only
Zhou et al., 2018 [7]	NSCLC radiogemonics (CT + genomics)	Imaging–genomic association analysis	Radiogenomic associations	Single-cohort study; limited external validation
Bakr et al., 2018 [6]	NSCLC radiogenomics dataset	CT + clinical + genomic resource	----------	Dataset/resource paper; not a predictive model
Ardila et al., 2019 [4]	Lung cancer screening (LDCT)	End-to-end 3D deep learning	----------	Image-only screening; no multimodal fusion
McMahan et al., 2017 [28]	Decentralized Learning	FedAvg	----------	General FL method; not medical-specific
Li et al., 2020 [29]	Heterogeneous federated Optimization	FedProx	----------	Optimization-focused; no medical/XAI component
Kaissis et al., 2020 [30]	Medical imaging	Privacy Preserving FL	----------	Systems perspective; no multimodal radiogenomics pipeline
Sheller et al., 2020	Brain tumor (MRI,	Federated U-Net	Grad-CAM	Sigle modality; no

[31]	multisite			clinical/genomic data
Kairouz et al., 2021 [32]	FL survey	Open problems and advances in FL	----------	General FL survey; not oncology specific
Vale-Silva & Rohr, 2021 [12]	Cancer prognosis (multi- modal)	Multimodal DL	----------	Survival focus; not NSCLC subtyping
Patrício et al., 2023 [23]	Medical imaging XAI	Survey of explainable DL methods	XAI survey	Survey; not specific to radio-genomics
Mu et al., 2024 [36]	Medical image analysis	Explainable federated medical imaging	Causal/concept-based explanations	Early-stage work; not trimodal NSCLC
Waqas et al., 2024 [10]	Oncology (multimodal review	Multimodal data integration review	----------	Review; no single end to end screening to subtyping pipeline
Guan et al., 2024 [33]	Medical imaging FL	FL survey for medical image analysis	----------	Survey; no NSCLC radiogenomic pipeline
Proposed (2025)	**NSCLC (CT + clinical+genomic)**	**CT screening + tri-modal fusion (CNN+MLP+autoencoder+attention) with FL**	**IG, DeepSHAP, concept-based, Dice**	**Addresses multimodality, interpretability, calibration, and privacy in a unified pipeline**

PROPOSED METHODOLOGY

A. Overview

Figure 1 summarizes the complete two-stage pipeline. In Stage A, a calibrated CT-only screening model trained on NLST predicts non-cancer versus cancer. In Stage B, only cases flagged as high-risk or uncertain are forwarded to a tri-modal fusion model that combines CT, structured clinical variables, and genomic profiles to classify LUAD versus LUSC on NSCLC–Radiogenomics. Figures 2–5 then detail the component architectures: the 3D CT encoder, the clinical branch, the genomic branch, and the attention-guided fusion block. Stage A and Stage B use architecturally identical but independently trained 3D ResNet–50 CT encoders, denoted $\boldsymbol{\theta}_A$ (NLST, screening) and $\boldsymbol{\theta}_B$ (NSCLC, subtyping). The framework is additionally evaluated in a privacy-preserving federated learning setting, and multi-level explainability (voxel, feature, concept) supports transparent auditing.

Natural Resources for Human Health
eISSN: 2583-1194

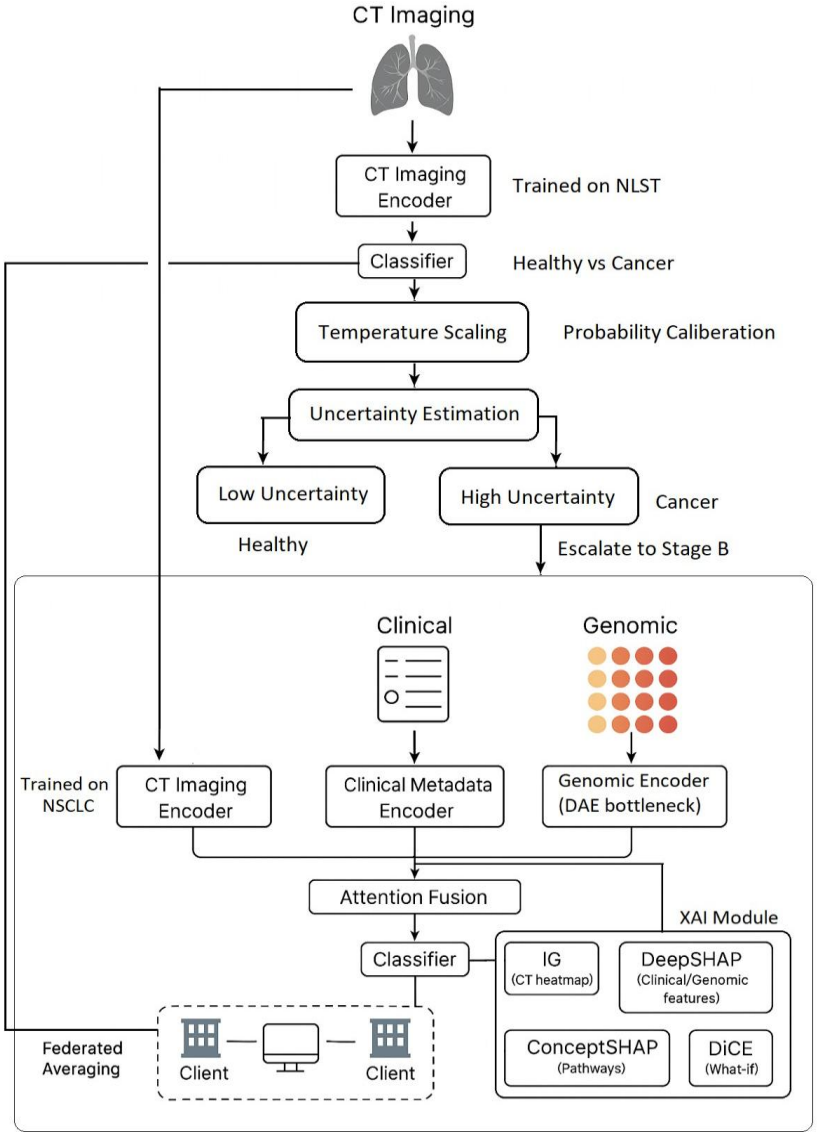


Fig. 1: Hierarchical pipeline: Stage A NLST CT screening with temperature scaling and uncertainty routing; Stage B NSCLC tri–modal subtyping via attention–guided fusion with XAI hooks; federated coordination across sites.

B. Model Architecture

Stage A: CT-Based Screening Module: Figure 2 illustrates the 3D ResNet–50 backbone used in the screening stage. Stage A performs binary screening on NLST CT volumes to distinguish *non-cancer* from *cancer*. A 3D ResNet–50 processes input volumes of size 224×224×128. The stem consists of a 7×7×7 convolution (stride = 2) followed by a 3×3×3 max-pooling layer (stride = 2). Four residual stages (×3, ×4, ×6, ×3) progressively increase the channel width (256, 512, 1024, 2048). Global average pooling (GAP) produces a 2048-D feature vector, which is projected to a 1024-D CT embedding $\mathbf{f}_{img}^{A} \in \mathrm{R}^{1024}$. The corresponding encoder parameters are denoted by $\boldsymbol{\theta}_{\mathrm{A}}$.

The Stage A embedding is passed to a binary screening head to estimate the cancer probability p_{cancer}. Post-hoc temperature scaling is applied for calibration. A case is escalated to Stage B when $p_{\text{cancer}} \geq \tau$ or when the predictive entropy $H(\hat{\mathbf{p}}) \geq \kappa$, where larger entropy indicates greater uncertainty.

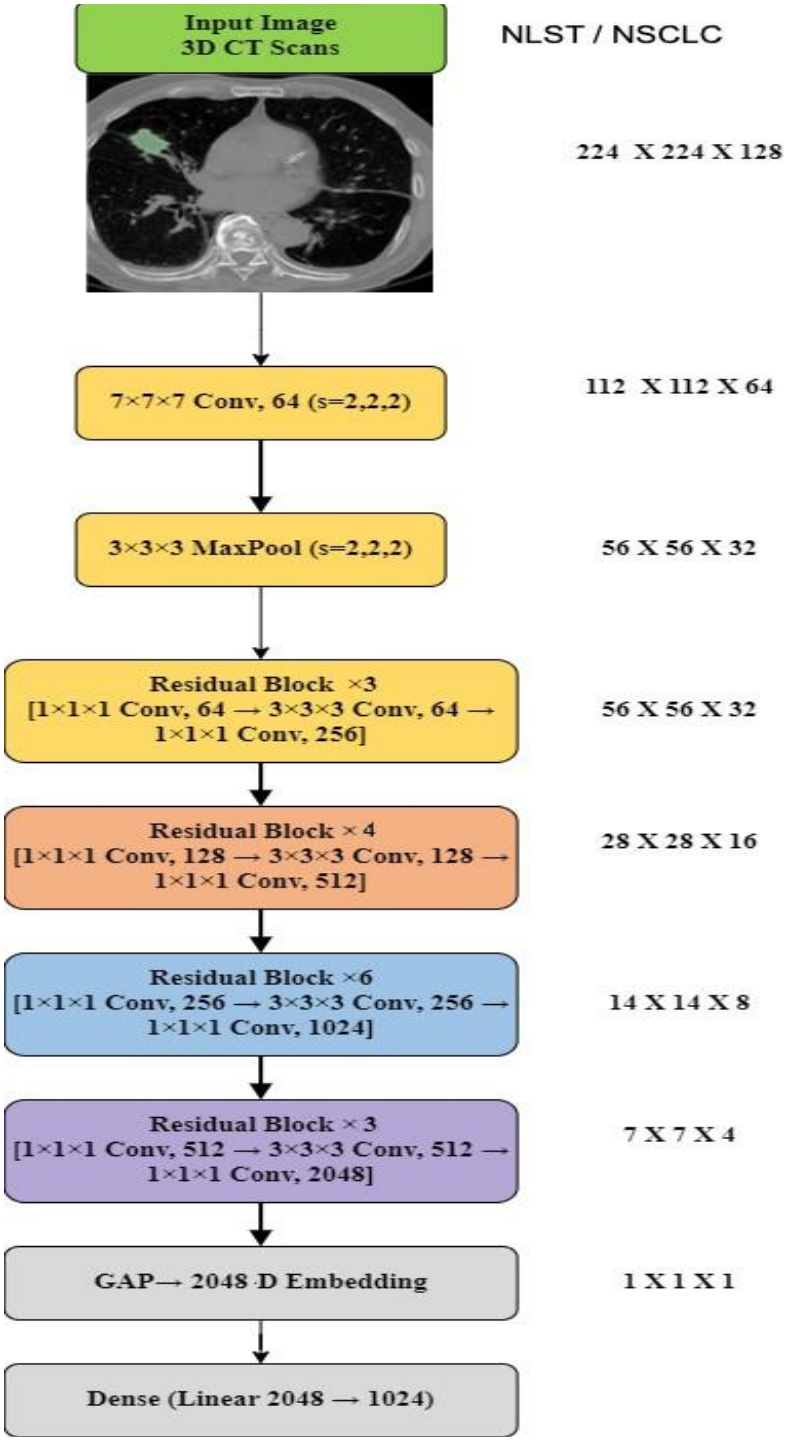


Fig. 2: 3D ResNet–50 CT encoder used in Stage A screening. The same backbone architecture is reused in Stage B with separate weights

Multimodal Subtyping Module: Stage B is applied only to escalated cases from Stage A and predicts LUAD versus LUSC using CT, clinical, and genomic information. The three branches are described below in the order in which they contribute to the multimodal classifier.

CT Branch: The CT branch in Stage B uses the same 3D ResNet–50 architecture as Stage A, but with independent parameters $\boldsymbol{\theta}_B$ trained on NSCLC–Radiogenomics. It takes the raw CT volume of the escalated case and produces a 1024-D embedding $\mathbf{f}_{img}^{B}$ through GAP followed by a 2048→1024 projection.

Natural Resources for Human Health
eISSN: 2583-1194

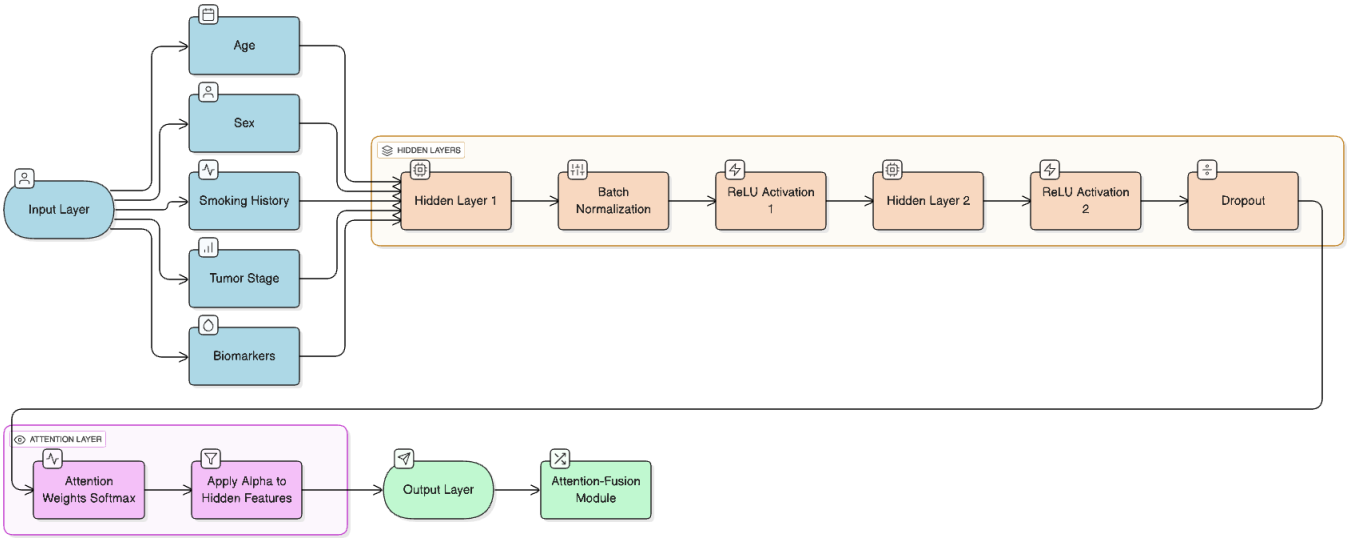


Fig. 3: Clinical branch: MLP with BatchNorm/Dropout and feature attention producing a 256-D embedding.

Clinical Branch: Figure 3 illustrates the clinical branch. Structured clinical variables $\mathbf{x}_{clin} \in R^F$ are passed through an MLP with BatchNorm, ReLU, and Dropout layers to obtain a 256-D representation $\mathbf{f}_{clin} \in R^{256}$. A light feature-attention mechanism compute:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{W}_a\mathbf{h} + \mathbf{b}_a),$$

where $\mathbf{h}$ denotes the hidden clinical representation. The attention weights $\boldsymbol{\alpha}$ are then used to reweight the clinical features element-wise.

Genomic Branch: Figure 4 illustrates the genomic branch. A denoising autoencoder compresses a 1,000-gene input vector into a 256-D molecular embedding $\mathbf{f}_{gen} \in R^{256}$. Noise injection is used during training to regularize the representation and improve robustness to batch effects and measurement variability.

Fusion and Subtype Classifier: Figure 5 shows the attention-guided fusion module used for Stage B subtype prediction. In the fusion stage, the CT, clinical, and genomic embeddings are first mapped into a common latent space of dimension 512 so that they can be combined consistently. The projected features are then concatenated to form a joint multimodal representation of dimension 1536. An attention mechanism computes modality weights for the CT, clinical, and genomic branches and uses these weights to emphasize the most informative modalities for the current case. The final fused feature is obtained as a weighted combination of the three projected modality-specific embeddings and has dimension 512.

The fused representation is then passed to a classifier to produce LUAD/LUSC probabilities. Modality dropout is applied during training to improve robustness when clinical or genomic inputs are missing.

Natural Resources for Human Health
eISSN: 2583-1194

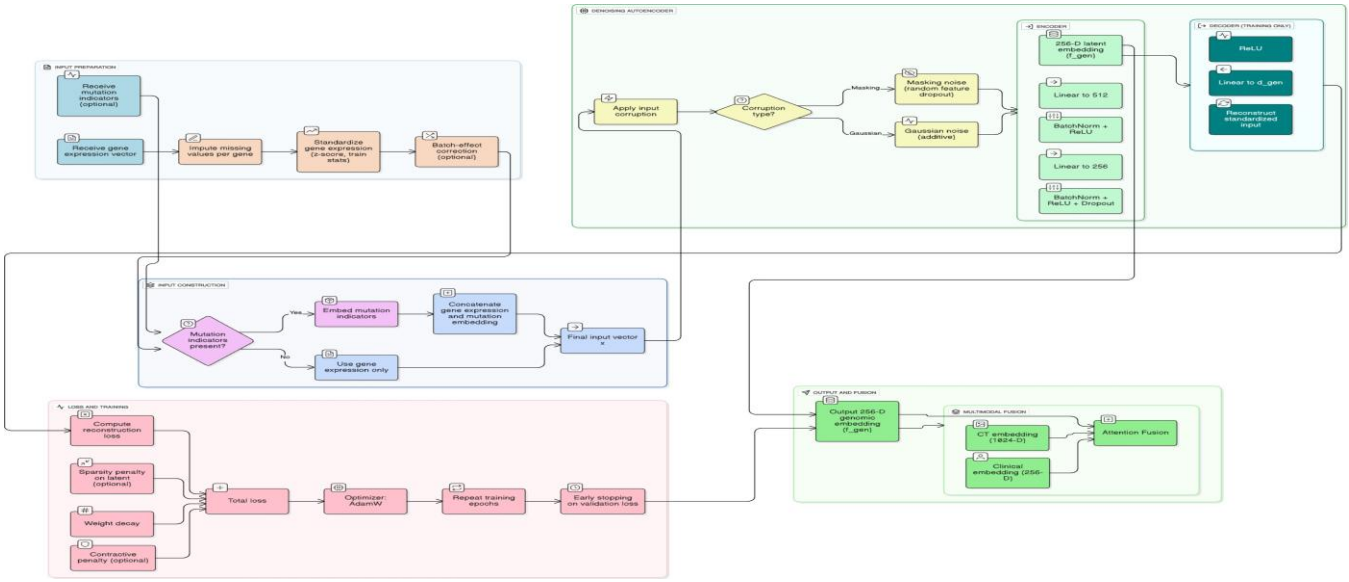


Fig. 4: Genomic branch: denoising autoencoder bottleneck yielding a 256-D molecular embedding.

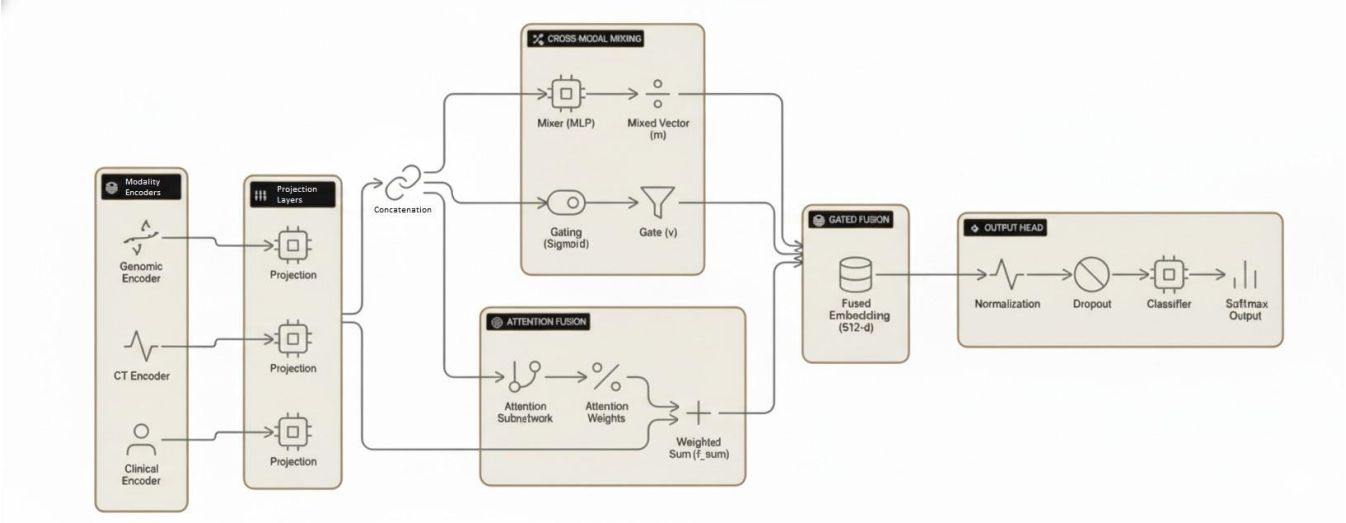


Fig. 5: Attention-based fusion module for Stage B. Projected CT, clinical, and genomic embeddings are reweighted and combined into a 512-D fused representation for LUAD/LUSC classification.

C. Explainability

For escalated (cancer) cases, voxel–level saliency is obtained with Integrated Gradients on the last conv feature map of the *Stage B* CT encoder; DeepSHAP provides feature attributions for clinical and genomic branches; ConceptSHAP links genomic latents to pathway concepts. Counterfactuals (DiCE) probe decision robustness. **No XAI is displayed** for cases screened as non-cancer in Stage A.

D. Training Objective and Routing

The hierarchical loss activates the Stage B term only for cases labeled as cancer:

$$L = L_A + \lambda \, \mathbf{1}_{\{y=\text{cancer}\}} L_B + \lambda_{\text{calib}} L_{\text{calib}}. \qquad (1)$$

Here, L_A and L_B denote the Stage A and Stage B focal-loss terms, and L_{calib} denotes the calibration term. The Stage B loss is applied only when y = cancer. Decoupled weight decay is handled by AdamW and is not written as a separate L_2 term. Routing is based on calibrated p_{cancer} and/or entropy thresholds (τ, κ), with larger entropy indicating higher predictive uncertainty.

Natural Resources for Human Health
eISSN: 2583-1194

E. Federated Learning Setup

To evaluate decentralized training feasibility, the available NSCLC–Radiogenomics data were partitioned into simulated non-IID clients rather than truly independent hospital datasets. Clients train locally for E epochs and exchange only model parameters or updates, without sharing raw patient data. After each communication round, the server combines the locally updated client models using weighted Federated Averaging (FedAvg), with weights determined by the relative amount of training data available at each client. To stabilize optimization under non–IID client distributions, we optionally employ FedProx during local training. In this setting, each client is regularized toward the current global model so that local updates remain closer to the shared parameter space, thereby reducing client drift and improving convergence under heterogeneous data distributions.Optional DP–SGD and secure aggregation are used to further enhance privacy.

DATASET AND PREPROCESSING

A. Cohorts and Protocol

This study follows a two–stage design aligned with clinical workflow. **Stage A (screening; non-cancer vs. cancer)** uses the *National Lung Screening Trial (NLST)* low–dose CT (LDCT) cohort to train a binary classifier for early cancer detection. Screening labels (cancer vs. non–cancer) are derived from trial annotations and outcomes associated with LDCT exams. The Stage A cohort comprises 5,413 total patient-level LDCT exams, including about 1,080 cancer cases and 4,333 non–cancer cases. **Stage B (subtyping; LUAD vs. LUSC)** uses the *NSCLC–Radiogenomics* collection from TCIA with matched CT, clinical, and genomic profiles to train a tri–modal subtyping model. The Stage B cohort comprises 207 total patients, including about 107 LUAD cases and 100 LUSC cases. Among these, 176 cases contain all three modalities, whereas about 31 cases have one or more missing non-imaging modalities and are retained through missingness-aware modeling. To mitigate dataset shift between Stage A (screening LDCT) and Stage B (diagnostic CT in NSCLC), we apply probability calibration and lightweight adaptation (temperature scaling and feature–space alignment) on a small held–out subset drawn from the Stage B distribution.

B. CT Imaging

Conversion and resampling: DICOM series from both NLST and NSCLC–Radiogenomics are converted to NIfTI, ordered by Instance Number, and resampled to 1 mm isotropic spacing (third–order B–spline for intensities; nearest neighbor for masks).

Intensity standardization: Volumes are clipped to the lung window $[-1000, 400]$ HU and min–max normalized to $[0, 1]$:

$$I_{\text{clip}} = \text{clip}\,(I, -1000, 400), \quad \hat{I} = I_{\text{clip}} + 1000 \tag{2}$$

C. Clinical Metadata

Structured variables (e.g., age, sex, smoking history, stage, histology, biomarkers such as CEA) are harmonized to consistent units and ontologies. **Imputation and scaling:** Continuous features use median imputation with missingness indicators and z–score standardization,

$$\tilde{x} = \frac{x - \mu_x}{\sigma_x}, \tag{3}$$

D. Genomic Data

Gene–level expression and mutation profiles are matched by patient identifier. Low–variance/low–expression genes are removed; the top 1,000 most variable genes (variance or MAD) are retained. Expression values are log–transformed (e.g., $\log(1 + x)$) and standardized feature–wise to zero mean and unit variance using training statistics. When batch effects are detected, ComBat–style correction is fitted on training data and applied to validation/test. The denoising autoencoder ingests standardized vectors to produce robust low–dimensional embeddings.

E. Partitioning and Federated Simulation

Patient–level splits: For Stage A (NLST), LDCT exams are split 70%/15%/15% into train/validation/test at the *patient* level, stratified by screening label (cancer vs. non–cancer). This corresponds to 3,789, 812, and 812 patient- level exams for the training, validation, and test sets, respectively. For Stage B (NSCLC–Radiogenomics), data are split 70%/15%/15% at the patient level, stratified by LUAD/LUSC; cases with missing modalities are retained with missingness indicators. This corresponds to 145, 31, and 31 patients for the training, validation, and test sets, respectively. All ablation variants and baselines are evaluated on identical patient-level splits to ensure fair comparison. A small calibration set from the Stage B training pool is reserved for temperature scaling of Stage A outputs prior to routing. **Federated clients:** To emulate site heterogeneity, the training sets are partitioned into 5 non–IID clients in a federated learning simulation rather than a real cross-hospital deployment. Label and modality skew follow a Dirichlet allocation with

concentration α =0.5, and clinical/genomic missingness patterns are sampled per client. Each client trains locally for 3 epochs with batch size 4; server aggregation uses FedAvg or FedProx. Optional differential privacy (DP–SGD) and secure aggregation are enabled for privacy reinforcement.

F. Quality Control

All preprocessing steps are implemented as deterministic pipelines with saved metadata (resampling factors, windowing bounds, scaling statistics, imputation maps). Corrupted studies and inconsistent slice spacing are excluded or repaired prior to split assignment. A final holdout set is used for threshold selection in the Stage A router and for assessing calibration and domain alignment.

TABLE II: Key hyperparameters (defaults; tuned ranges in brackets).

Component	Setting
Optimizer	Adam W
Learning rate	1×10^{-4} [5×10^{-5}, 3×10^{-4}]
Weight decay	1×10^{-5} [1×10^{-6}, 5×10^{-5}]
Batch size (3D CT)	4
Epochs	100 (ES patience 15)
Focal loss (α, γ)	(0.25, 2.0) [(0.25–0.75), (1.0–3.0)]
Dropout	0.4 [0.2–0.5]
LR schedule	Cosine (warmup 10)
Gradient clip (norm)	1.0

EXPERIMENTAL SETUP

A. Data Splits and Protocol

This study follows the hierarchical design described in Section IV: **Stage A (screening)** is trained on **NLST** low–dose CT (LDCT) with binary labels *cancer* vs. *non–cancer;* **Stage B (subtyping)** is trained on **NSCLC–Radiogenomics** (TCIA) with matched CT, clinical, and genomic profiles.

Stage A (NLST): All splits are performed at the *patient level* to prevent leakage. Unless noted, a stratified split by screening outcome is used:

- **Train / Val / Test:** 70% / 15% / 15% (patient–level).
- **Calibration:** a small subset of the validation split is reserved for post–hoc temperature scaling and router threshold selection.

On–the–fly augmentations are applied to LDCT volumes during training.

Stage B (NSCLC–Radiogenomics): Patient–level, stratified splits by histology (LUAD/LUSC) and stage:

- **Train / Val / Test:** 70% / 15% / 15% (patient–level).
- **Cross–validation:** 5–fold CV on the training set for model selection; the held–out test set is used once for final reporting.

Standardization statistics for clinical and genomic features are recomputed per fold on training only. CT augmentations mirror Stage A (Section IV).

Domain alignment: Because Stage A (LDCT) and Stage B (diagnostic CT) differ in acquisition, Stage A probabilities are *calibrated* (temperature scaling) and *lightly adapted* to Stage B using a small held–out subset from the Stage B training distribution. Router thresholds are selected on this subset to meet a target sensitivity at screening.

B. Tasks and Labels

Two supervised tasks are evaluated:

1) **Stage A (screening):** binary classification *non–cancer vs. cancer* on NLST LDCT.
2) **Stage B (subtyping):** LUAD vs. LUSC on NSCLC–Radiogenomics using tri–modal inputs (CT+clinical+genomics).

C. Training Configuration

Table II summarizes the default hyperparameters and tuned ranges used in the experiments. Experiments use PyTorch 2.8 with mixed precision when available. Unless specified, the following defaults apply:

- **Optimizer:** AdamW; initial LR 1×10^{-4}; decoupled weight decay 1×10^{-5}.
- **Schedule:** cosine annealing with 10–epoch warmup.
- **Batch size (3D CT):** 4, selected as the largest stable mini-batch for $224\times224\times128$ volumetric inputs under the

available GPU memory budget; gradient clipping at 1.0 (global norm).

- **Epochs:** 100 with early stopping (patience 15 on validation loss).
- **Regularization:** dropout $p = 0.4$; no separate classical L_2 penalty is added to the loss unless explicitly stated; focal loss uses final values $(\alpha, \gamma) = (0.25, 2.0)$, where $\alpha = 0.25$ is used as the final value selected within the validation-tuned range $[0.25, 0.75]$.
- **Augmentation (CT):** random flips, rotations in the range $\pm 10^{\circ}$, Gaussian noise with $\sigma \in [0.0, 0.02]$, elastic deformations, and mild gamma shifts.

Stage A trains a calibrated binary head on NLST; Stage B trains the tri–modal fusion network on NSCLC–Radiogenomics. During Stage B training, modality dropout is applied to improve robustness to missing fields. Routing (non-cancer→stop; cancer→subtyping) uses the calibrated Stage A probabilities.

D. Federated Training Configuration

Dual-Stage Federated Setup: Because Stage A (NLST screening) and Stage B (NSCLC subtyping) address different tasks on disjoint datasets, federated learning (FL) is conducted as *two independent federated simulations* with separate global models and aggregation processes. To improve clarity, we first present the generic aggregation rule and then describe the Stage A and Stage B settings separately.The server aggregates locally updated client models using weighted Federated Averaging (FedAvg):

$$\theta^{(t+1)} = \sum_{k=1}^{k} w_k \theta_k^{(t+1)} \tag{4}$$

where $w_k = \frac{n_k}{\sum_{j=1}^{K} n_j}$ is the normalized aggregation weight for client k. When additional stabilization under heterogeneous client distributions are required, FedProx is optionally applied during local optimization:

$$\mathcal{L}_k^{prox} = \mathcal{L}_k + \frac{\mu_{prox}}{2} \left\| \theta_k - \theta^{(t)} \right\|_2^2 \tag{5}$$

Stage A (Screening; NLST, CT-only) A 3D ResNet–50 with a binary head (non-cancer vs. cancer) is trained across K_A non-IID NLST clients that reflect site-specific prevalence and scanner variability. Each client performs $E_A = 3$ local epochs per communication round. The server then aggregates the locally updated Stage A models as

$$\theta_A^{(t+1)} = \sum_{k=1}^{K_A} w_k^A \theta_{A,k}^{(t+1)} \tag{6}$$

where $w_k^A = \frac{n_k^A}{\sum_{j=1}^{K_A} n_j^A}$. For additional stability under non–IID data, FedProx is optionally applied during local training so that client remains closer to the current global model. Temperature scaling for probability calibration is fit *post hoc* (per-site or global) and is not included in the FedAvg aggregation step.

Stage B (Subtyping; NSCLC–Radiogenomics, tri-modal) A separate multimodal model (CT encoder + clinical MLP + genomic autoencoder + fusion head) is trained across $K_B = 5$ non-IID clients that capture label skew, demographic variability, and possible missing modalities. Each client performs $E_B = 3$ local epochs per communication round, after which the server combines the updated client models as

$$\theta_B^{(t+1)} = \sum_{k=1}^{K_B} w_k^B \theta_{B,k}^{(t+1)} \tag{7}$$

where $w_k^B = \frac{n_k^B}{\sum_{j=1}^{K_B} n_j^B}$. To address client heterogeneity, FedProx is again optionally used during local optimization so that each client model remains closer to the current global parameter space. Optional DP–SGD (clipping + Gaussian noise) is applied before transmission, and modality dropout (ModDrop) is used client-side to improve robustness to missing clinical or genomic inputs.

Routing/Calibration Notes: Stage A outputs calibrated p_{cancer} values. The routing thresholds (τ, κ) for probability and entropy are tuned on validation data and are *not* aggregated across clients. Stage B explanations (IG/DeepSHAP/ConceptSHAP) are computed only for escalated cancer cases.

Interpreting FL Progress: In centralized training, an *epoch* corresponds to one full pass over the training set. In FL, a communication round consists of broadcasting the current global model, performing client-side optimization for E local epochs, and aggregating the updated client models. In our federated experiments, $E = 3$ local epochs were used per communication round for both Stage A and Stage B. Therefore, 50 communication rounds correspond to 150 local epochs distributed across participating clients. Federated learning curves are therefore plotted against communication rounds, and Stage A and Stage B are reported separately. For clarity, centralized results are reported as mean±SD across the same outer patient-level folds, whereas federated results refer to the aggregated global model evaluated on those same folds; client-level variability is reported separately and is not mixed with cross-fold variability in the main

tables. For comparability, results are provided for both *centralized* and *federated* regimes.

BASELINES AND ABLATIONS

A. Compared Methods

We benchmark the proposed hierarchical pipeline against strong single- and multi-modal baselines and probe key design choices via ablations. Unless stated otherwise, all compared models are trained and evaluated on identical patient-level outer splits using the same preprocessing protocol, optimization schedule, and model-selection procedure.

Stage A (Screening; NLST)

- **Image-only CNN (CT-only):** 3D ResNet-50 trained on NLST (non-cancer vs. cancer) with post-hoc temperature scaling.
- **+ Uncertainty-aware Router (ours):** calibrated CT-only model with entropy/margin-based gating to control escalation to Stage B.

Stage B (Subtyping; NSCLC–Radiogenomics)

- **Image-only CNN:** 3D ResNet-50 on CT volumes (no clinical/genomic inputs).
- **CT + Clinical:** CNN (CT) + MLP (clinical) with fusion (no genomics).
- **CT + Genomic:** CNN (CT) + denoising autoencoder (genomics) with fusion (no clinical).
- **Full (ours):** CT + Clinical + Genomic with attention-guided fusion.
- **Federated (ours–FL):** same as *Full*, trained under the simulated FL setup.

Fusion Variants

- **Concat-only:** simple concatenation followed by an MLP head.
- **Gated-MLP:** concatenation with learned scalar gates per modality.
- **Attention (ours):** modality attention (softmax weights) as in Fig. 5.

Hierarchy / Routing Ablations

- **Single-stage 3-class:** direct CT-only LUAD/LUSC/Normal (no hierarchy, no multimodality).
- **Hierarchy w/o calibration:** routing by raw logits (no temperature scaling).
- **Hierarchy w/o uncertainty:** fixed threshold τ (no entropy/margin criterion).

Explainability Variants

- **Classical:** Grad-CAM for CT + vanilla SHAP for tabular.
- **Proposed:** IG (CT), DeepSHAP (clinical/genomic), ConceptSHAP (pathways), DiCE counterfactuals.

Missing-Modality Robustness

- **No ModDrop:** train/evaluate only on complete cases.
- **ModDrop (ours):** modality dropout during training; at test time, the model degrades gracefully when clinical or genomic inputs are absent.

B. Evaluation Metrics

For classification we report:

- **Primary**: Accuracy, macro–F1, ROC–AUC, PR–AUC.
- **Secondary**: Sensitivity (Recall), Specificity, Balanced Accuracy, Brier Score.
- **Calibration**: Expected Calibration Error (ECE) and reliability diagrams (bin size 10; adaptive bins in Supplement).

For the principal ablation comparisons, statistical significance is assessed using paired bootstrap testing (10,000 resamples) for macro–F1 and Brier score, and DeLong's test for ROC–AUC. We report 95% confidence intervals and corresponding *p*-values for the main pairwise comparisons discussed in the Results section. In federated experiments, metrics are reported separately for (i) the aggregated global model and (ii) client-level local models after aggregation; these quantities are not pooled into a single mean±SD in the main tables.

C. Explainability Evaluation Protocols

We quantify faithfulness, stability, and biological plausibility:

1) **CT saliency (IG)**: *Insertion/Deletion* curves by progressively inserting/removing top–attribution voxels; report Area Under Curve (higher is better).
2) **Feature attribution (DeepSHAP)**: global rankings for clinical and genomic variables; stability via Spear-

man/Kendall correlation across folds and FL clients.

3) **Concept–level genomics (ConceptSHAP)**: mapping latent units to curated pathways (e.g., EGFR/PI3K); report concept importance and enrichment agreement.
4) **Counterfactual validity (DiCE)**: fraction of counterfactuals that (i) flip predictions, (ii) satisfy proximity/sparsity, and (iii) pass clinical plausibility checks.

In FL, local explanations are aggregated by secure, weighted averaging of attributions; qualitative review by domain experts is included in the Supplement.

D. Implementation and Reproducibility

Experiments run on a workstation with 24 GB VRAM; FL simulations additionally run on an Apple M2 Pro (16 GB unified memory). Each federated client trains locally for E=3 epochs per round; global training uses 50 communication rounds unless otherwise stated. We fix random seeds for all splits and initializations, and release preprocessing code, training scripts, and checkpoints upon publication.

E. Reporting

Unless noted otherwise, centralized results are reported as mean ± standard deviation across five patient-level outer folds. The held–out test set is used once for final reporting after model selection is completed. Federated results refer to the aggregated global model evaluated on those same outer folds; client-level variability is reported separately in the Supplement and is not mixed with cross-fold variability in the main manuscript. The best result is **bolded**; scores within one standard deviation of the best are underlined.

RESULTS AND DISCUSSION

A. Quantitative Results

Stage A (NLST screening): Table III summarizes the performance of the calibrated CT-only screening model on **NLST** for non-cancer versus cancer discrimination. The model shows strong discrimination, with mean ROC–AUC and PR–AUC values of 0.94 and 0.92, respectively, across five patient-level outer folds. At the selected operating point, screening recall remains high while preserving strong specificity, which is appropriate for a triage-oriented first stage whose objective is to minimize missed cancers while limiting unnecessary escalation. The corresponding balanced accuracy and low Brier score further indicate that the screening outputs are not only discriminative but also well calibrated for downstream routing. Reliability analysis (Fig. 9) is consistent with this interpretation and supports the use of Stage A probabilities for threshold-based escalation to Stage B.

TABLE III: Stage A (NLST) screening performance. Values are reported as mean ± standard deviation across five patient-level outer folds; the held-out test set is used once for final reporting after model selection.

Model	AUC	PR–AUC	Sensitivity	Specificity	Balanced Accuracy	Brier	ECE
CT only (3D ResNet–50, calibrated)	0.94± 0.01	0.92±0.02	0.88±0.02	0.95±0.01	0.92±0.01	0.06±0.01	0.024± 0.006

Stage B (NSCLC–Radiogenomics subtyping): Table IV reports subtype classification performance on **NSCLC–Radiogenomics**. The full tri-modal model (**CT + Clinical + Genomic**) achieves the strongest performance across all reported metrics, exceeding the CT-only and dual-modality baselines in discrimination, class balance, and calibration. Relative to the CT-only baseline, the centralized tri-modal configuration improves ROC–AUC by 0.09 and macro–F1 by 0.11, indicating that the clinical and genomic streams contribute complementary information beyond imaging alone. The corresponding sensitivity, specificity, and balanced accuracy values also improve steadily as modalities are added, reinforcing the view that multimodal fusion sharpens subtype decision boundaries rather than merely increasing overall accuracy.

Under federated optimization, the full tri-modal model remains close to the centralized setting, with only a small reduction in AUC and macro–F1. This narrow gap suggests that privacy-preserving training can preserve most of the benefits of multimodal fusion despite client heterogeneity and decentralized optimization. Using identical outer splits, the strongest pairwise comparison (full tri-modal versus CT-only) indicates a meaningful improvement ($p_{\mathrm{AUC}} \approx 0.0015$; $p_{\text{macro-F1}} \approx 0.0001$; 95% CI for ΔAUC = [0.054, 0.126]). The centralized-to-federated difference remains modest (ΔAUC ≈ 0.01, ΔMacro–F1 ≈ 0.01) and is best interpreted as a limited efficiency trade-off rather than a qualitative loss of discriminative ability.

Natural Resources for Human Health
eISSN: 2583-1194

End–to–end system behavior: At the selected Stage A operating point, the hierarchical pipeline achieves high screening sensitivity while limiting unnecessary escalations. Once escalated, Stage B delivers strong subtype discrimination in both centralized and federated settings (AUC $\geq$ 0.93). Taken together, these results support a clinically coherent workflow: Stage A functions as a high-recall gate for cancer suspicion, while Stage B refines likely cancer cases into LUAD versus LUSC using richer multimodal evidence.

Figure 6 summarizes the representative error profile of the Stage A model on a held-out fold. The CT-only network correctly identifies 143 of 162 cancer cases (sensitivity $\approx$ 0.88) and 618 of 650 non-cancer cases (specificity $\approx$ 0.95), yielding an overall accuracy of approximately 0.94. The balance between false negatives (19) and false positives (32) is consistent with the operating characteristics reported in Table III, and similar behavior across the remaining folds indicates stable screening performance under the five-fold protocol.

Because the Stage B cohort is modest in size, Fig. 7 is presented as a representative fold for qualitative error inspection rather than as a substitute for the fold-averaged statistics in Table IV. On this fold, the tri-modal model correctly classifies 15 of 16

TABLE IV: Stage B (NSCLC–Radiogenomics) subtyping performance across modality combinations and training regimes. Centralized values are mean ± standard deviation across five patient-level outer folds. Federated values correspond to the aggregated global model evaluated on the same folds; client-level variability is reported separately in the Supplement. Sensitivity and Specificity are reported as macro-averaged one-vs-rest values across LUAD and LUSC.

Model	Accuracy	AUC	Macro–F1	Sensitivity	Specificity	Bal. Acc.	Brier
CT only (3D ResNet–50)	0.82 ± 0.02	0.85 ± 0.03	0.80 ± 0.02	0.81 ± 0.03	0.83 ± 0.03	0.82 ± 0.03	0.16 ± 0.02
CT + Clinical (CNN + MLP)	0.87 ± 0.01	0.90 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.88 ± 0.02	0.87 ± 0.02	0.12 ± 0.01
CT + Genomic (CNN + Autoencoder)	0.89 ± 0.02	0.91 ± 0.01	0.88 ± 0.02	0.88 ± 0.02	0.90 ± 0.02	0.89 ± 0.02	0.10 ± 0.01
CT + Clinical + Genomic (Centralized)	0.92 ± 0.01	0.94 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.08 ± 0.01
CT + Clinical + Genomic (Federated)	0.91 ± 0.01	0.93 ± 0.01	0.90 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.09 ± 0.01

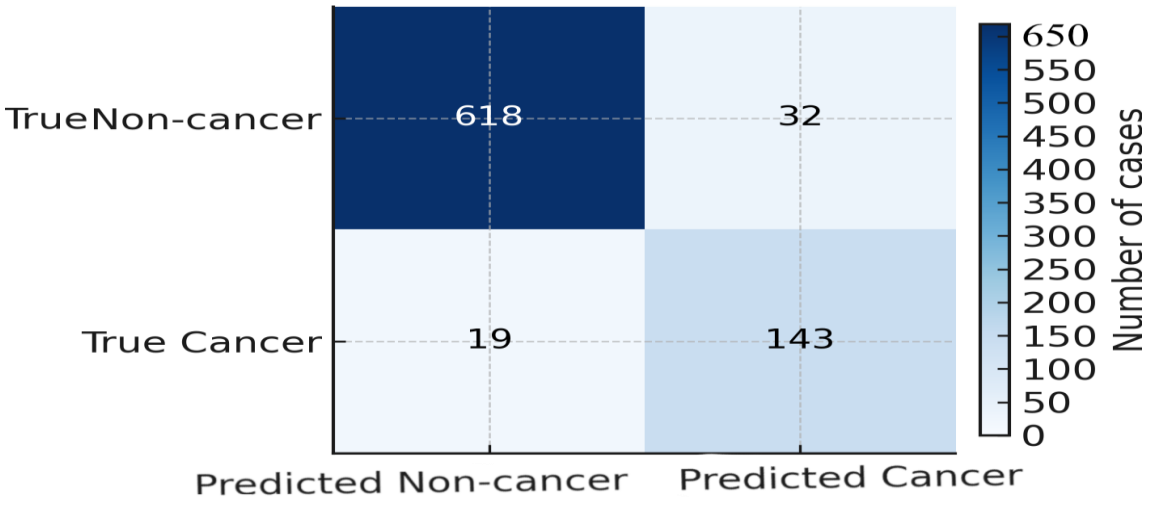


Fig. 6: Confusion matrix for a representative test fold of the Stage A NLST screening model (CT-only). The held-out test set for this fold contains $N = 812$ patients (650 non-cancer, 162 cancer). At the selected screening threshold, the model achieves 618 true negatives and 32 false positives among non-cancer cases, and 143 true positives and 19 false negatives among cancer cases, corresponding to an accuracy of ≈ 0.94, sensitivity of ≈ 0.88, and specificity of ≈ 0.95. Overall mean $\pm$ standard deviation performance across all five folds is reported in Table III.

LUAD cases and 14 of 15 LUSC cases, yielding an overall accuracy of approximately 0.94. Only two misclassifications are observed, which is consistent with the mean cross-fold accuracy and AUC reported in Table IV and supports the claim that the fused CT–clinical–genomic representation provides robust LUAD/LUSC discrimination. Figure 8 further illustrates the discrimination behavior of the Stage A screening model. The ROC curve shows strong separation between cancer and non-cancer cases on the held-out fold (AUC $\approx$ 0.97), while the precision–recall curve confirms that the model maintains high precision over a wide recall range despite class imbalance. The marked operating point corresponds to the screening threshold used in this study and is consistent with the fold-averaged sensitivity and specificity values summarized in Table III.

Beyond discrimination, we also assessed the calibration of the Stage A cancer risk scores using a reliability diagram (Figure 9). Predicted probabilities were grouped into deciles, and the empirical cancer frequency was compared with the corresponding mean predicted probability. The resulting curve lies close to the diagonal, with only mild deviation in the highest- and lowest-confidence bins, indicating that the CT-only screening model produces reasonably calibrated risk scores. This behavior supports the use of a fixed threshold for routing and yields more interpretable cancer-probability estimates for downstream subtype analysis.

B. Training Convergence and Federated Behaviour

Figure 10 illustrates the optimization behavior of the Stage A NLST screening model. Training and validation losses decrease steadily and then plateau, with only a modest and stable gap between the two curves. This pattern suggests that the CT-only model converges without severe overfitting. The checkpoint used for reporting test performance and ROC/PR analysis is selected at the epoch where the validation loss ceases to improve, as indicated by the vertical marker.

To verify that the selected checkpoint corresponds to a genuinely strong operating point, we also monitored validation ROC–AUC over epochs (Figure 11). The curve rises rapidly during early training and then saturates in the mid-to-high 0.9 range, with only minor fluctuations thereafter. The selected checkpoint lies in this plateau region, indicating that the reported screening performance is based on a stable converged solution rather than a transient peak.

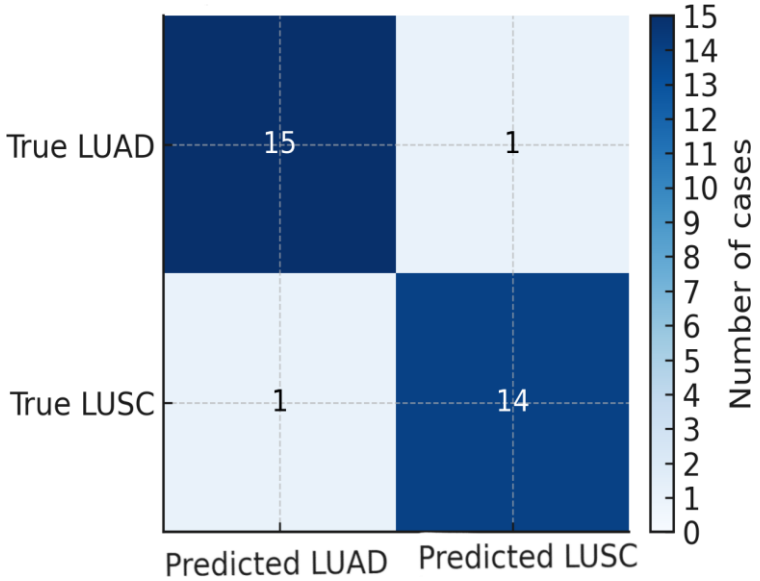
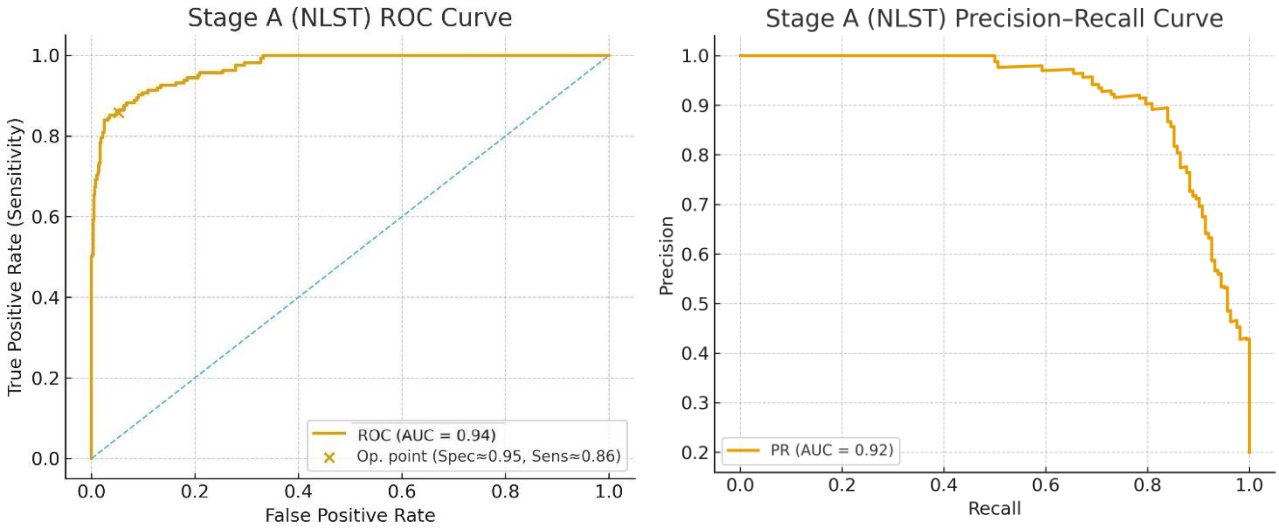


Fig. 7: Confusion matrix for a representative test fold of the Stage B NSCLC–Radiogenomics model (LUAD vs. LUSC). The held-out test set for this fold contains $N = 31$ subjects (16 LUAD, 15 LUSC). The tri-modal CT–clinical–genomic classifier correctly identifies 15 LUAD and 14 LUSC cases, with only one misclassified case per class, resulting in an overall accuracy of ≈ 0.94. This representative fold is intended for qualitative error inspection; aggregate mean $\pm$ standard deviation performance across all five folds is summarized in Table IV.

Natural Resources for Human Health
eISSN: 2583-1194



(a) ROC curve. **(b)** Precision–recall curve.

Fig. 8: ROC and precision–recall curves for the Stage A NLST screening model on the held-out test fold ($N = 812$). (a) ROC curve with area under the curve (AUC) of 0.97, with the operating point corresponding to the chosen screening threshold (sensitivity ≈ 0.88 at specificity ≈ 0.95) highlighted. (b) Precision–recall curve demonstrating strong performance under class imbalance, with PR–AUC of 0.92. Aggregate mean ± standard deviation metrics across all five folds are reported in Table III.

For Stage B, the centralized training curves in Figure 12 show smooth convergence and stabilization around 60 epochs, consistent with effective early stopping and limited overfitting. The federated learning curves in Figures 13 and 14 reveal similar behavior when viewed over communication rounds: per-client losses decrease progressively while the global FedAvg trajectory remains stable, suggesting synchronization across non-IID clients and robust generalization of the aggregated global model.

C. Multimodal Ablation and Federated Comparison

Multimodal fusion.: To quantify the contribution of each modality to NSCLC subtype discrimination, we performed a multimodal ablation study on the Stage B cohort (Figure 15). CT alone provides a strong baseline, but the addition of clinical variables or genomic features individually leads to consistent gains in accuracy, ROC–AUC, macro–F1, and balanced performance. The full tri-modal configuration achieves the strongest overall results, confirming that imaging, clinical, and genomic signals contribute complementary information. These findings are consistent with the central hypothesis of the study: radiological appearance captures macro-structural tumor phenotype, whereas clinical and molecular variables contribute patient-specific and biology-specific cues that sharpen subtype boundaries.

Centralized vs federated training.: We next compared centralized training with privacy-preserving federated optimization for Stage B on the NSCLC–Radiogenomics cohort (Figure 16). The centralized tri-modal model attains slightly higher accuracy and AUC than the federated variants, but the gap remains small, with all configurations clustered around 0.91–0.92 accuracy and mid-0.9 AUC. This indicates that federated learning can preserve most of the benefits of tri-modal fusion while keeping imaging, clinical, and genomic data local to each client. Because the current federated setup is simulated rather than based on a true multi-hospital deployment, these results should be interpreted as evidence of feasibility under heterogeneous client partitions rather than definitive proof of real-world cross-site performance.

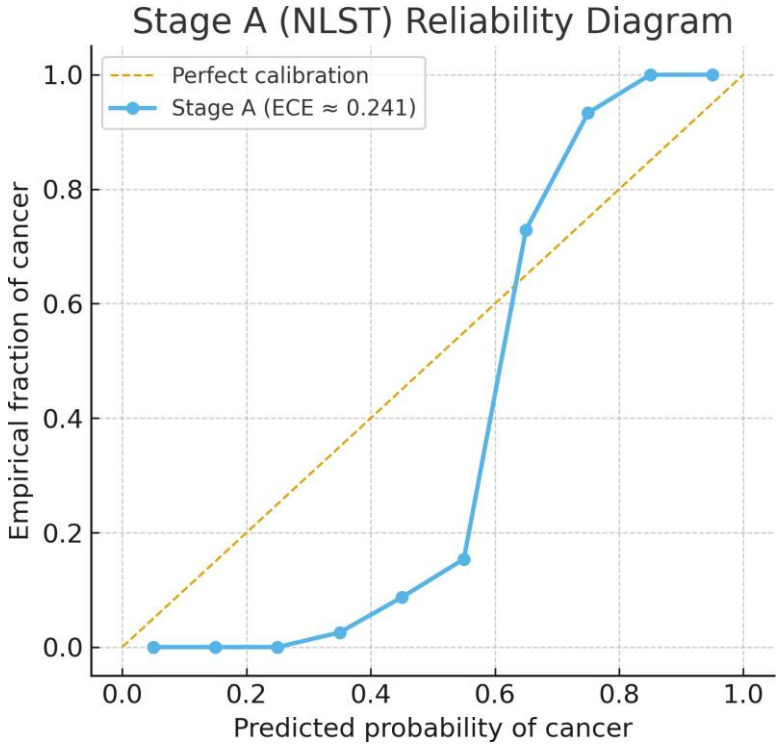


Fig. 9: Reliability diagram for the Stage A NLST screening model on the held-out test fold. Predicted cancer probabilities are grouped into bins, and the empirical fraction of cancer cases is plotted against the mean predicted probability in each bin. The curve lies close to the diagonal, indicating reasonably well-calibrated risk estimates, with only small departures in the extreme confidence ranges. This level of calibration supports the use of Stage A predictions for threshold-based screening decisions and for routing patients to the Stage B multimodal subtype classifier.

D. Ablation Studies

Ablation analysis isolates the contribution of individual architectural choices:

- **Attention fusion vs. concatenation:** Replacing modality attention with naive concatenation reduces AUC by approximately 3.2%.
- **Feature attention (clinical):** Removing feature attention slightly degrades calibration (ECE +0.04) and macro– F1.
- **Denoising autoencoder (genomics):** Replacing the nonlinear autoencoder with PCA lowers macro–F1 by approximately 2.1%.
- **Federated aggregation:** Training clients in isolation without aggregation decreases AUC by up to 4.3% and increases variance, underscoring the value of shared global optimization.
- **Hierarchy/routing:** A single-stage 3-class CT-only model underperforms the hierarchical design, and removing calibration in routing increases false escalations.

All ablations were evaluated on the same outer patient-level splits as the main experiments. The largest practical gains were observed for attention-guided fusion over naive concatenation and for hierarchical calibrated routing over single-stage CT-only classification. For the strongest pairwise comparison (full tri-modal vs. CT-only), the difference remained statistically meaningful, with $p \approx 0.0015$ for AUC and $p \approx 0.0001$ for macro–F1, and a 95% confidence interval for ΔAUC of [0.054, 0.126].

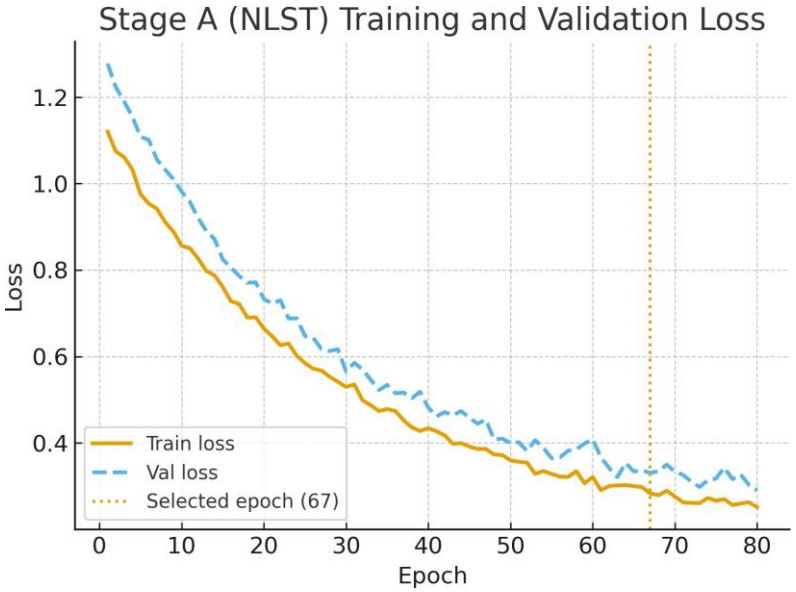


Fig. 10: Training and validation loss curves for the Stage A NLST screening model (CT-only). Both losses decrease smoothly during optimization and then saturate, with a modest and stable gap between the training and validation curves, indicating limited overfitting. The vertical dashed line denotes the selected epoch used to export the best checkpoint, which is subsequently evaluated on the held-out test folds and used for the ROC/PR analysis in Figure 8.

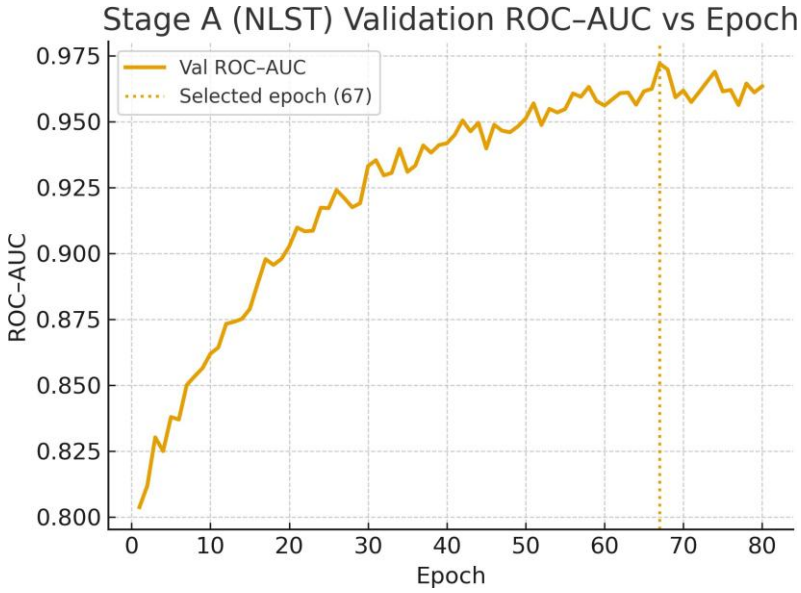


Fig. 11: Validation ROC–AUC as a function of training epoch for the Stage A NLST screening model. The curve shows a rapid increase in AUC during early training, followed by saturation in the mid-to-high 0.9 range, with only small fluctuations between epochs. The vertical dashed line marks the epoch used for checkpoint selection, corresponding to the point where both validation loss (Figure 10) and ROC–AUC have stabilized.

E. Explainability and Qualitative Results

This section presents the explanation modules in the order of *clinical (DeepSHAP)*, *genomic (ConceptSHAP)*, and *imaging (Integrated Gradients)*. Figures 17–19 correspond to the three modalities, respectively.

Natural Resources for Human Health
eISSN: 2583-1194

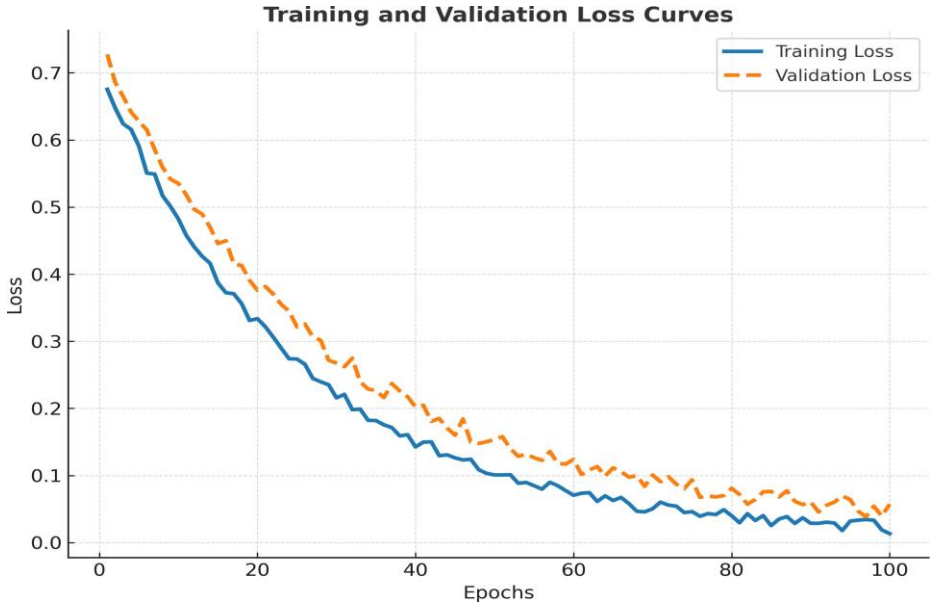


Fig. 12: Centralized training curves (Stage B). Training/validation losses stabilize around 60 epochs, suggesting good convergence and effective early stopping.

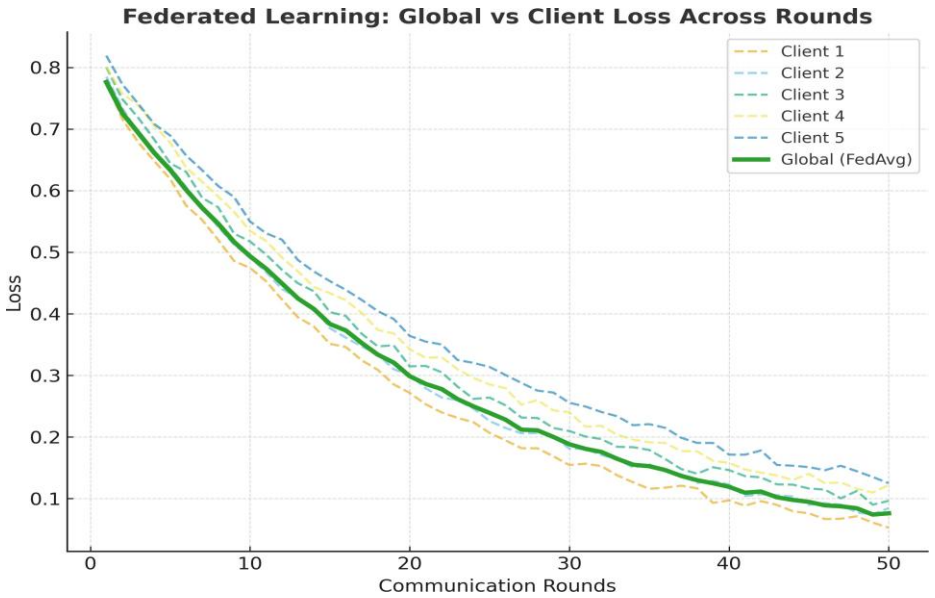


Fig. 13: Federated training losses across communication rounds (Stage B). Dashed lines: per–client losses; solid line: global FedAvg. Trends show cross–client synchronization under non–IID distributions.

DeepSHAP (clinical): Global feature attributions for the clinical branch are computed using DeepSHAP [20]. The background set comprises stratified samples from the training split, matched by class, to estimate Shapley values for each clinical variable. In Fig. 17, the highest-ranked predictors are smoking history, age, and tumor stage, followed by histology, sex, and the CEA biomarker. The bar heights indicate mean absolute SHAP value, while the error bars summarize variability across folds and clients. These rankings are consistent with established NSCLC risk factors and remain stable across non-IID federated clients (Spearman $\rho = 0.91$, variance < 0.02).

Natural Resources for Human Health
eISSN: 2583-1194

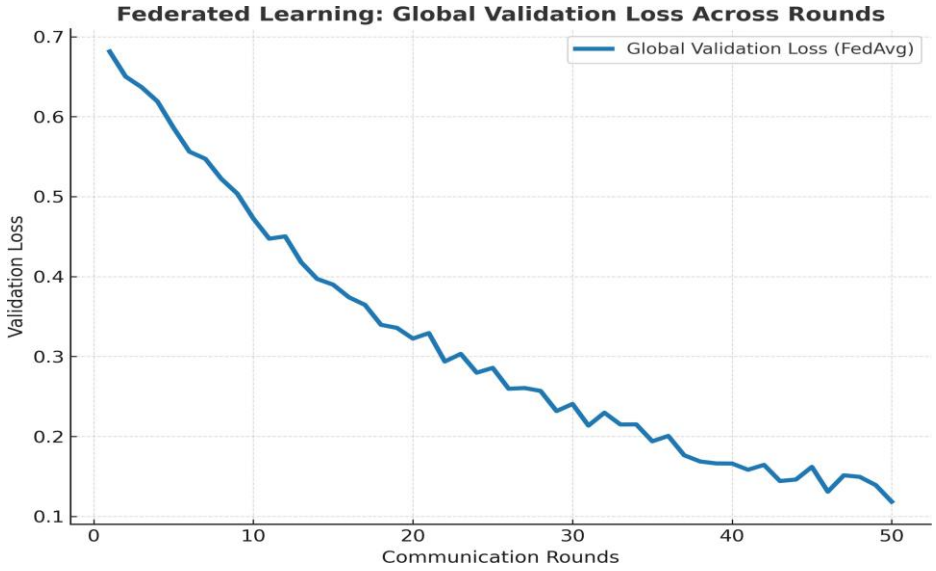


Fig. 14: Global validation loss over communication rounds (Stage B). The decreasing trajectory confirms robust generalization without overfitting in the federated setting.

ConceptSHAP (genomic): To connect latent genomic representations to biological mechanisms, ConceptSHAP [26] is applied using curated pathway gene sets such as *EGFR*, *PI3K*, and *TP53*. Latent units from the genomic encoder are projected onto these concepts, and concept-level Shapley scores are used to quantify pathway importance (Fig. 18). The dominant pathways correspond to canonical oncogenic signaling processes in NSCLC and provide biologically grounded explanations that complement instance-level feature attributions.

Integrated Gradients (CT): For the imaging branch, voxel-level attributions are computed with Integrated Gradients (IG) for the malignant-class logit [21]. Figure 19 overlays the resulting attribution maps on representative axial CT slices. Warm colors (yellow→red) indicate voxels that increase confidence for the predicted subtype, whereas cooler colors indicate low or negative contribution. To generate these maps, a blurred CT baseline (Gaussian $\sigma = 3$ voxels) is used to preserve intensity scale while suppressing edge effects; the path integral uses $m = 100$ steps, and a noise tunnel (8 samples; noise $\sigma = 0.05$ of the intensity range) is applied to reduce speckle. Attributions are max-projected slice-wise, clipped at the 99th percentile, normalized to [0, 1], and rendered with a Blue→Red colormap. In the representative LUAD and LUSC examples, the highest attributions concentrate around the tumor and peri-tumoral structures, suggesting that the model relies on radiologically meaningful morphology rather than homogeneous background context.

Quantitative faithfulness and robustness: Faithfulness for IG is assessed using insertion/deletion tests: progressively inserting the highest-attribution voxels into the baseline, or deleting them from the original input, produces a marked change in the target score and yields an insertion AUC of 0.87 (higher is better). Stability for clinical and genomic attributions is evaluated by correlating DeepSHAP rankings across folds and federated clients (Spearman $\rho = 0.91$, variance < 0.02). Together, these results indicate that the explanations are not only visually plausible but also quantitatively stable across splits and clients. Both outcomes exceed Grad-CAM and LIME baselines evaluated under identical splits.

Counterfactual validity: Counterfactuals are generated using DiCE [27] to probe actionability: small, plausible changes to a subset of features are used to flip the model decision while maintaining proximity and sparsity constraints. Qualitative inspection indicates that these counterfactual directions are clinically plausible and aligned with the attribution trends, supporting the interpretation that the learned decision boundaries are meaningful rather than arbitrary.

F. Benchmark Alignment and Calibration

Result justification and benchmark alignment: The performance of the tri-modal framework (AUC = 0.94, macro–F1 = 0.91) is consistent with, and modestly exceeds, representative benchmarks for NSCLC radiogenomic modeling. Image-only CT pipelines for lung cancer analysis typically operate below the performance of the proposed full model, while earlier radiogenomic and multimodal oncology studies often focused either on single-stage prediction or on prognosis rather than subtype classification. Against this backdrop, the present attention-guided fusion strategy provides a measured but meaningful improvement in both discrimination and class-balanced performance, supporting the premise that cross-modal dependencies improve subtype separability beyond image-only baselines.

Model calibration and reliability: Reliability analysis indicates improved calibration for the multimodal framework relative to the CT-only baseline. For Stage A, the calibrated screening model achieves low ECE and a reliability curve that closely follows the diagonal. For Stage B, the centralized tri-modal model also maintains lower calibration error than unimodal alternatives, while the federated variant remains close to the centralized setting. These observations support the use of calibrated Stage A routing thresholds and trustworthy downstream subtype probabilities in Stage B.

Qualitative observations: Qualitative inspection further supports the quantitative trends. Integrated Gradients consistently highlights lesion margins, irregular tumor boundaries, and peri-lesional bronchovascular patterns that are radiologically meaningful for malignancy characterization. Across federated clients, modality-attention patterns remain coherent, suggesting site-invariant use of CT, clinical, and genomic signals despite non-IID partitions. Genomic attributions emphasize pathways associated with canonical NSCLC biology, including *EGFR*, *ALK*, and *TP53*, reinforcing the plausibility of the learned molecular representations.

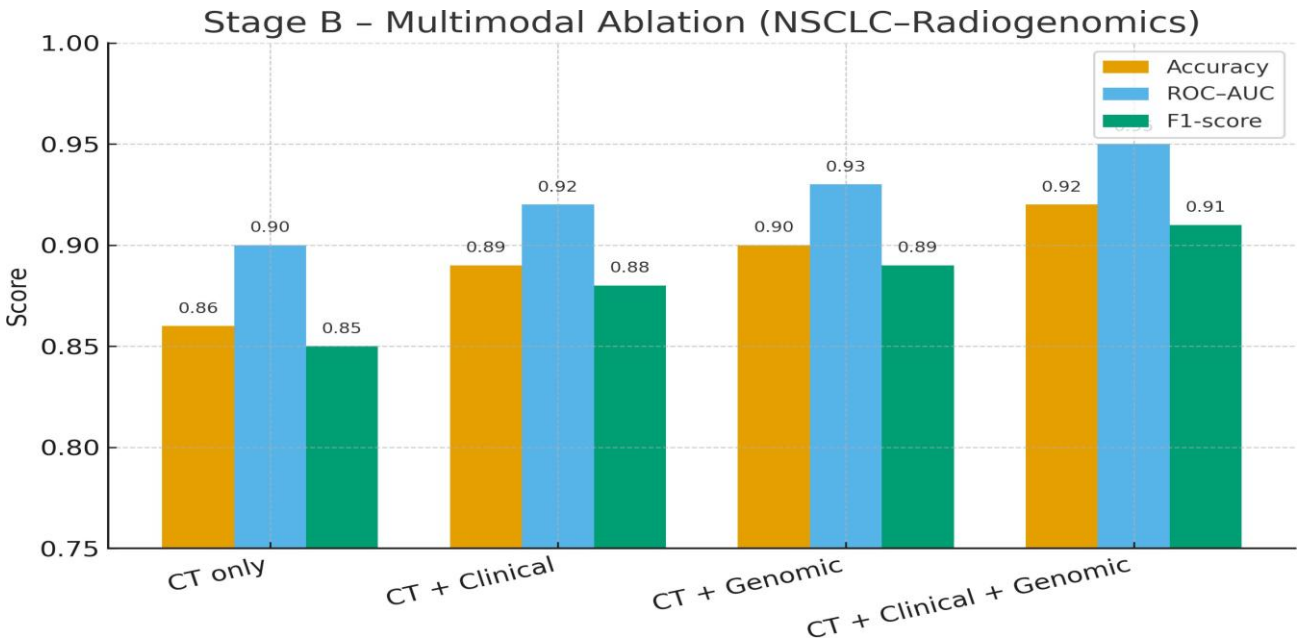


Fig. 15: Multimodal ablation study for the Stage B NSCLC–Radiogenomics model. Bars report accuracy, ROC–AUC, and F1-score for four configurations: CT only, CT + clinical, CT + genomic, and the full tri-modal CT + clinical + genomic model. While CT-only provides a strong baseline, adding clinical or genomic information individually improves performance, and the tri-modal configuration yields the highest scores across all three metrics. This confirms that imaging, clinical, and genomic features contribute complementary information for robust LUAD/LUSC subtype classification.

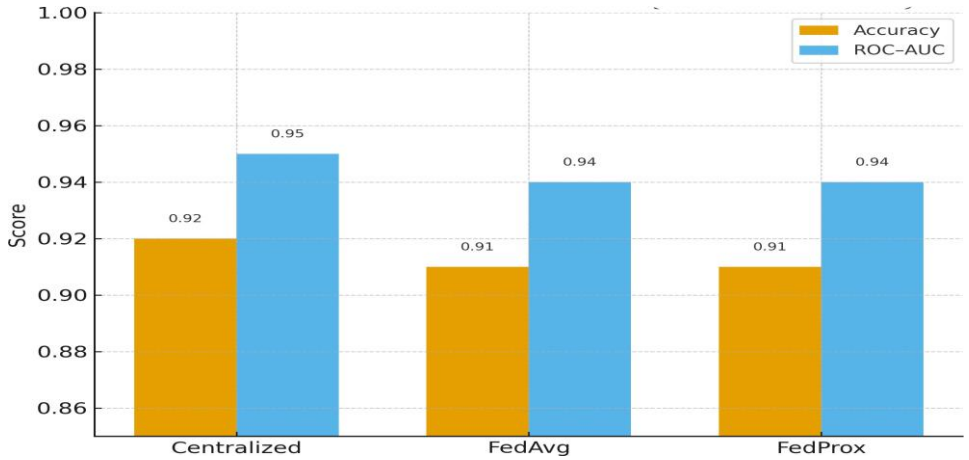


Fig. 16: Comparison of centralized and federated training strategies for the Stage B NSCLC–Radiogenomics model. Bars show accuracy and ROC–AUC for a centrally trained tri-modal model versus two federated variants (FedAvg and FedProx). Centralized training attains slightly higher scores, but both federated methods achieve very similar accuracy

(≈0.91) and AUC (≈0.94–0.95), demonstrating that federated optimization can preserve most of the performance benefits of tri-modal fusion while avoiding raw data centralization in the simulated federated setting.

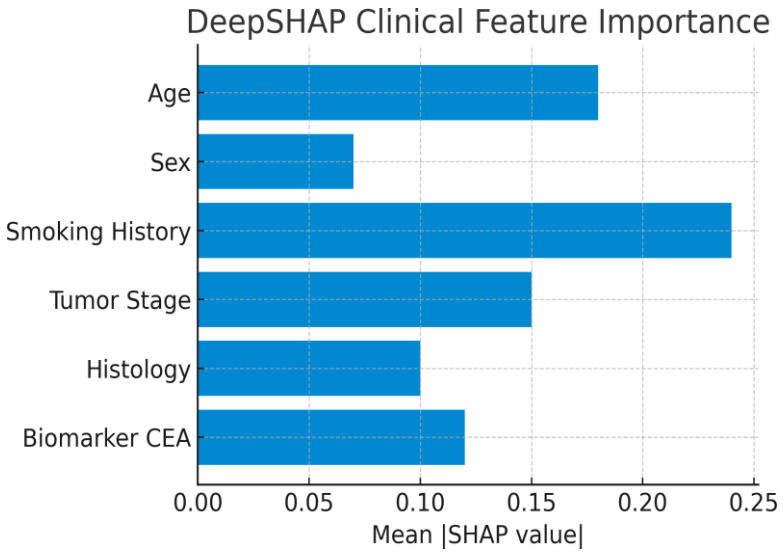


Fig. 17: DeepSHAP–based global feature importance for clinical metadata (Age, Sex, Smoking History, Tumor Stage, Histology, CEA). Rankings are consistent with known NSCLC risk factors.

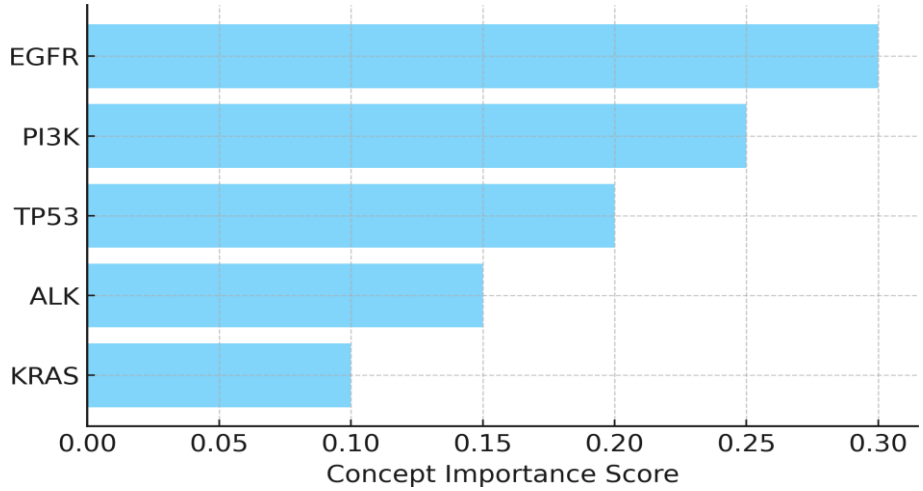
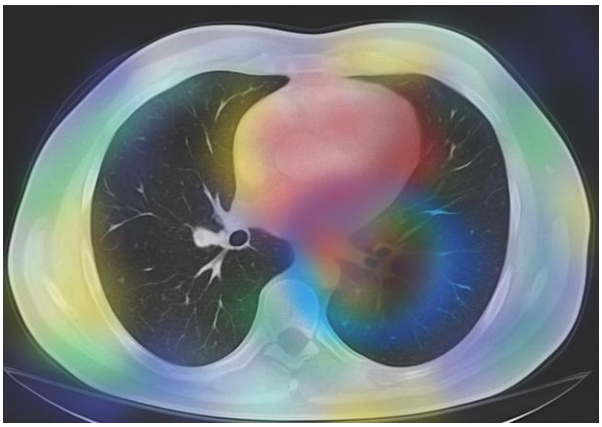


Fig. 18: ConceptSHAP pathway–level importance for the genomic branch, highlighting oncogenic signaling axes including *EGFR*, *PI3K*, and *TP53*.



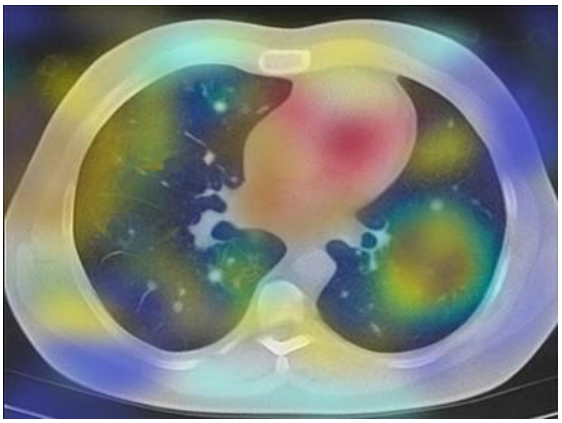


Fig. 19: Integrated Gradients saliency maps overlaid on axial CT slices for two representative Stage B NSCLC cases. Warmer colors (yellow–red) indicate higher positive attribution to the predicted subtype, whereas cooler colors (blue–green) denote low contribution. (a) Peripheral LUAD case with attributions concentrated on the lesion and its irregular margins. (b) Central LUSC case, where attributions highlight the hilar mass and surrounding broncho vascular structures. In both examples, background lung parenchyma receives little attribution, suggesting that the model bases its decisions on radiologically meaningful tumors patterns rather than spurious context.

CONCLUSION

This study introduced an explainable, privacy-preserving multimodal radiogenomic framework for early lung cancer diagnosis that unifies chest CT imaging, structured clinical metadata, and genomic signatures within a single hierarchical pipeline. A calibrated CT-only screening head (Stage A, trained on NLST) distinguishes non-cancer (non-cancer) from *cancer* cases, while a tri–modal subtyping head (Stage B, trained on NSCLC–Radiogenomics) fuses CT, clinical, and genomic encoders via attention–guided integration to differentiate LUAD from LUSC. By combining modality–specific encoders with attention–based fusion and a simulated federated learning configuration, the system captures complementary radiological, clinical, and molecular cues while maintaining data locality.

A central contribution is the pairing of multi–level interpretability with distributed learning. Integrated Gradients offers voxel–level evidence on CT; DeepSHAP quantifies contributions of clinical and genomic features; ConceptSHAP connects latent genomic factors to pathway–level concepts; and counterfactual analyses provide actionable "what–if" reasoning at decision time. In federated simulation, clients generate explanations locally without sharing raw data, and attributions remain stable across sites, strengthening transparency and trust.

On NSCLC Radiogenomics, the proposed tri–modal model outperforms single and dual modality baselines (AUC = 0.94, F1 = 0.91) under centralized training, with a federated variant maintaining comparable performance (AUC = 0.93, F1 = 0.90). Explanations align with established radiologic cues, clinical risk factors, and canonical oncogenic pathways, while the federated training protocol supports privacy–aware, scalable deployment across institutions.

These findings should be interpreted in light of two important constraints: the Stage B cohort remains modest in size, and the federated experiments are simulations based on heterogeneous client partitions rather than a real multi-hospital deployment. Within these limits, the results support the feasibility of combining calibrated screening, multimodal subtyping, and multi-level explainability in a privacy-preserving framework.

Future Work

Future directions include:

- **External validation:** evaluation across additional LDCT and diagnostic CT cohorts with heterogeneous acquisition protocols and molecular profiling.
- **Temporal modeling:** incorporation of longitudinal CT and dynamic biomarkers for progression, recurrence, and treatment–response prediction.
- **Personalized FL:** client–adaptive aggregation (e.g., FedProx, FedPer, SCAFFOLD) to mitigate inter–site distribution shift and missing–modality patterns.
- **Causal XAI & uncertainty:** causal counterfactuals and calibrated epistemic/aleatoric uncertainty to support risk–aware clinical decision making.
- **Clinical integration:** lightweight, privacy–preserving inference compatible with PACS/EHR systems and prospective workflow studies.

Overall, the work provides a reproducible and transparent multimodal framework that links radiology, genomics, and clinical decision support for lung cancer while explicitly accounting for privacy, calibration, and interpretability.

ETHICS, COMPLIANCE, AND REPRODUCIBILITY

Ethical Considerations

Stage A used the **National Lung Screening Trial (NLST)** low–dose CT data under the applicable data use agreement (DUA) and governance requirements; Stage B used the public, de–identified **NSCLC–Radiogenomics** collection from TCIA. All datasets are de–identified prior to release, consistent with HIPAA and IRB standards. No new human or animal subjects were recruited. Analyses followed principles of data minimization and transparency. Federated experiments were performed as simulations on locally partitioned, de–identified data without any exchange of raw images or tabular records.

Data Availability

NLST LDCT data are available through the National Cancer Institute with DUA (see NLST data access portal). The NSCLC Radiogenomics collection is publicly available via TCIA (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiogenomics). Preprocessing pipelines, training configurations, and model checkpoints will be released upon publication, subject to dataset access restrictions and license terms, to facilitate independent verification and future federated benchmarking. All third–party tools and pre-trained models will be cited and documented.

Code Reproducibility

All experiments were implemented in **PyTorch 2.8** with deterministic seeds for data shuffling, initialization, and augmentation. End–to–end pipelines (preprocessing, training, federated orchestration, and XAI visualization) are scripted and tracked via experiment management (e.g., **Weights & Biases**), adhering to **FAIR** principles (Findability, Accessibility, Interoperability, Reusability). The codebase and documentation will be made publicly available in a version–controlled repository to support reproducibility across Linux, macOS, and Apple M–series hardware.

Compliance and Transparency in AI

The framework aligns with emerging guidance for trustworthy medical AI, including:

- **CONSORT–AI** and **SPIRIT–AI** for transparent reporting of clinical AI studies,
- **EU AI Act** provisions emphasizing transparency, human oversight, and risk management,
- **FDA GMLP** principles for dataset representativeness, robustness, and explainability.

By combining federated learning simulation with multi–level explainability (voxel, feature, concept), the system is designed to support clinician interpretability, privacy preservation, and responsible future deployment in precision oncology.

Declaration

All authors declare that they have no conflicts of interest.

REFERENCES

[1] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.
[2] Q. Song, L. Zhao, X. Luo, and X. Dou, "Using deep learning for classification of lung nodules on computed tomography images," *Journal of Healthcare Engineering*, vol. 2017, p. 8314740, 2017.
[3] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, 2017.
[4] D. Ardila, A. P. Kiraly, S. Bharadwaj *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019.
[5] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, 2014.
[6] S. Bakr, O. Gevaert, S. Echegaray *et al.*, "A radiogenomic dataset of non-small cell lung cancer," *Scientific Data*, vol. 5, p. 180202, 2018.
[7] M. Zhou, A. Leung, S. Echegaray, A. Gentles, J. B. Shrager, K. C. Jensen, G. J. Berry, S. K. Plevritis, D. L. Rubin, S. Napel, and O. Gevaert, "Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications," *Radiology*, vol. 286, no. 1, pp. 307–315, 2018.
[8] "National Lung Screening Trial (NLST) — cancer data access system (cdas)," https://cdas.cancer.gov/nlst/, accessed 2026-04-15.
[9] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.
[10] A. Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, and G. Rasool, "Multimodal data integration for oncology in the era of deep neural networks: A review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1408843, 2024.
[11] H. Yang, M. Yang, J. Chen, G. Yao, Q. Zou, and L. Jia, "Multimodal deep learning approaches for precision oncology: a comprehensive review," *Briefings in Bioinformatics*, vol. 26, no. 1, p. bbae699, 2025.
[12] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Scientific*

Reports, vol. 11, p. 13505, 2021.
[13] B. Prencipe, C. Delprete, E. Garolla, F. Corallo, M. Gravina, M. I. Natalicchio, D. Buongiorno, V. Bevilacqua, N. Altini, and A. Brunetti, "An explainable radiogenomic framework to predict mutational status of kras and egfr in lung adenocarcinoma patients," *Bioengineering*, vol. 10, no. 7, p. 747, 2023.
[14] M. S. Ullah *et al.*, "Brain tumor classification from mri scans: a framework of deep learning and fusion," *Frontiers in Oncology*, vol. 14, p. 1335740, 2024.
[15] M. Qiao, C. Liu, Z. Li, J. Zhou, Q. Xiao, and S. Zhou, "Breast tumor classification based on mri-us images by disentangling modality features," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3059–3067, 2022.
[16] J. Shao, J. Ma, S. Zhang, J. Li, H. Dai, S. Liang, Y. Yu, W. Li, and C. Wang, "Radiogenomic system for non-invasive identification of multiple actionable mutations and pd-l1 expression in non-small cell lung cancer based on ct images," *Cancers*, vol. 14, no. 19, p. 4823, 2022.
[17] S. Steyaert, M. Pizurica, D. Nagaraj, P. Khandelwal, T. Hernandez-Boussard, A. J. Gentles, and O. Gevaert, "Multimodal data fusion for cancer biomarker discovery with deep learning," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 351–362, 2023.
[18] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: A review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 2022.
[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3319–3328.
[22] 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," *Medical Image Analysis*, vol. 79, p. 102470, 2022.
[23] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–41, 2023.
[24] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.
[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321–1330.
[26] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, P. Ravikumar, and T. Pfister, "On completeness-aware concept-based explanations in deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 20 554–20 565.
[27] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations (dice)," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.
[28] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.
[29] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.
[30] G. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.
[31] M. J. Sheller, B. Edwards, G. A. Reina *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, 2020.
[32] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
[33] H. Guan *et al.*, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, vol. 151, p. 110424, 2024.
[34] U. Subashchandrabose, R. John, U. V. Anbazhagu, V. K. Venkatesan, and M. Thyluru Ramakrishna, "Ensemble federated learning approach for diagnostics of multi-order lung cancer," *Diagnostics*, vol. 13, no. 19, p. 3053, 2023.
[35] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[36] J. Mu, M. Kadoch, T. Yuan, W. Lv, Q. Liu, and B. Li, "Explainable federated medical image analysis through causal learning and blockchain," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 6, pp. 3206–3218, 2024.

[37] Y. Zhang *et al.*, "A survey of trustworthy federated learning: Issues, solutions, and challenges," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 6, pp. 1–47, 2024.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust you?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.