Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 10(12): 29815–29852.
DOI: 10.3934/math.20251310
Received: 03 November 2025
Revised: 26 November 2025
Accepted: 02 December 2025
Published: 18 December 2025

Research article

Deep learning framework for early diagnosis of lung cancer using multi-modal medical imaging

Masad A. Alrasheedi[1], **Asamh Saleh M. Al Luhayb**[2,*] **and Abdulmajeed A. R. Alharbi**[3]

[1] Department of Management Information Systems, College of Business Administration, Taibah University, Madinah, Saudi Arabia

[2] Department of Mathematics, College of Science, Qassim University, P. O. Box 6644, Buraydah, 51452, Saudi Arabia

[3] Department of Statistics and Operations Research, College of Science, King Saud University, P. O. Box 2455, Riyadh, 11451, Saudi Arabia

* **Correspondence:** Email: a.alluhayb@qu.edu.sa.

Abstract: Early and accurate diagnosis of lung cancer remains challenging due to the heterogeneity of tumor morphology and the variability across imaging modalities. This study proposed a deep learning framework that integrated computed tomography (CT), positron emission tomography/computed tomography (PET/CT), and chest X-ray (CXR) within a unified multi-modal transformer architecture for early lung cancer detection. The framework employed modality-specific encoders combining convolutional and state-space blocks to extract spatial-frequency representations, followed by a gated cross-modal fusion transformer designed to align heterogeneous features and handle missing modalities through mixture-of-experts routing and low-rank imputation. Multi-task heads were jointly optimized for nodule detection, segmentation, malignancy classification, and survival risk prediction. Explainability was embedded through concept bottlenecks, prototype reasoning, gradient-based attribution, and counterfactual concept editing, offering case-level interpretability and clinically meaningful evidence maps. Uncertainty was estimated via Monte-Carlo dropout, deep ensembles, and temperature scaling to ensure calibrated confidence estimates and defer-to-expert safety decisions. Lung image database consortium and image database resource initiative (LIDC-IDRI) (CT), the cancer imaging archive (TCIA) (PET/CT), and national lung screening trial (NLST) (CXR) benchmark datasets revealed that our methods work better than the best methods available. The proposed technique yielded Dice scores of 0.879, 0.872, and 0.876, together with AUC values of 0.944, 0.952, and 0.938, and an expected calibration error (ECE) of 0.02 across all modalities. Under domain shift, cross-dataset analysis showed substantial generalization ($AUC > 0.92$). A generalizable framework for multi-modal diagnostics made it possible to use AI to help with lung cancer screening in a way that was clear, trustworthy, and scalable.

Keywords: lung cancer diagnosis; multi-modal medical imaging; cross-modal transformer; uncertainty calibration; prototype reasoning; concept bottleneck; deep learning
Mathematics Subject Classification: 62P10, 68T07, 90B50

1. Introduction

Globally, lung cancer is the biggest cause of cancer-related fatalities, surpassing breast, prostate, and colorectal cancers combined [1]. Early detection is critical since survival rates decrease with tumor growth and metastasis [2]. Low-dose computed tomography (LDCT) screening is tough to identify benign from malignant nodules because of reader variability, imaging noise, and subtle morphological differences [3]. Handcrafted radiomic characteristics and statistical models in traditional computer-aided diagnosis (CAD) systems are typically insufficient for generalization across imaging protocols and scanners [4].

Deep learning (DL) has transformed medical imaging by autonomously detecting, segmenting, and classifying pulmonary nodules with radiologist-level accuracy [3]. Convolutional neural networks (CNNs) and transformers have been extensively studied for nodule identification, malignancy risk prediction, and prognosis estimation [5]. Although many techniques rely on computed tomography (CT) data, they often overlook supplementary information from other modalities including positron emission tomography (PET), chest X-ray (CXR), and histopathology [6]. Integrating heterogeneous data sources in multi-modal learning frameworks enhances disease representation, boosting diagnostic accuracy and robustness [7].

Criticism of deep neural networks' interpretability hinders clinical adoption despite their success [8]. Explainable artificial intelligence (XAI) uses approaches including gradient-based attribution, prototype learning, and concept bottlenecks to make model predictions transparent [9]. Using these methods can improve clinical trust, analyze errors, and find latent imaging biomarkers [10]. Integrating uncertainty quantification and calibration can enhance model safety and reliability for clinical decision assistance. Recent advances in inverse problems have focused more on elements of the problem including stability, identifiability, and trustworthy decisions than just recovering the solutions. New mathematical formulations of fractional polyharmonic operators, reconstruction of inverse spectra that are missing data, and contamination models which use neural solvers for the reconstruction of inverse scattering are some examples of this [11–13]. Our multimodal fusion for multimodal fusion and concept bottlenecks, as well as uncertainty-aware calibration, also provide robust decision support systems that go beyond traditional inverse modeling.

For early and accurate lung cancer diagnosis, we present a multimodal deep learning architecture that integrates CT, PET/CT, and CXR imaging with structured clinical data. Anatomical, metabolic, and contextual inputs are modeled by modality-specific encoders and a controlled cross-modal fusion transformer to provide complementary feature learning even under missing-modality settings. We use concept bottlenecks and prototype reasoning to enable interpretable intermediate representations, evidence-based prototype retrieval, and counterfactual concept modification to improve transparency and clinical interpretability. Monte Carlo dropout, deep ensembles, and temperature scaling improve prediction reliability and clinical trustworthiness with uncertainty-aware calibration and safety filtering. In nodule segmentation, malignancy classification, and cross-dataset generalization, the proposed model consistently outperforms state-of-the-art baselines on three public benchmark datasets, lung image database consortium and image database resource initiative (LIDC-IDRI), the cancer imaging archive (TCIA) PET/CT, and national lung screening trial (NLST), while providing clinically meaningful visual and semantic explanations. Overall, our work advances trustworthy, interpretable, and generalizable AI for precision oncology and early-stage lung cancer screening.

The paper's organization follows. Recent literature on deep learning models for lung cancer diagnosis is reviewed in Section 2. The proposed multimodal framework with its architecture, fusion technique, and learning objectives is presented in Section 3. See Section 4 for experimental setup, quantitative and qualitative evaluations, and ablation analyses. Section 5 will finish the study and suggest future research directions.

2. Related work

Recent research across medical image analysis and related domains has demonstrated the growing importance of explainability, attention mechanisms, and hybrid deep learning architectures. In dermatology, several studies have introduced parallel attention and spiking-attention mechanisms to enhance the interpretability of skin cancer classification and segmentation models [14, 15]. Complementary advances in gastrointestinal and breast cancer diagnostics have explored hierarchical multi-stage attention, dynamic expert routing, and multi-feature attention networks to improve clinical decision support [16, 17]. In neuroimaging, significant progress has been made in explainable brain tumor detection through graph-attention transformers, systematic literature surveys, and hybrid state-space transformer models [18–20]. Beyond medical diagnostics, deep learning innovations have extended to tasks such as secure image encryption using nonlinear hybrid pseudo-random sequences and transformer-based agricultural object detection [21, 22]. Explainability in biometric verification has also gained traction, with Siamese-based signature verification demonstrating improved robustness and discriminative feature learning [23]. Collectively, these works highlight a consistent trend toward multimodal fusion, attention-driven interpretability, and expert-routing mechanisms, underscoring the need for robust and explainable frameworks capable of operating reliably under real-world constraints such as missing or corrupted data.

Previous lung cancer screening trials, such as the NLST, demonstrated decreased mortality using low-dose CT scans; however, elevated false-positive rates required dependable CAD systems [24]. Handcrafted radiomic descriptors and statistical classifiers were inconsistent with variations in scanners and acquisition methods in conventional CAD. Deep learning made it possible to learn features from images from start to finish, which changed the way lung nodules are analyzed. U-Net, 3D U-Net, UNet++, and nnU-Net are only a few of the architectures that have proven important for medical image segmentation because they can capture hierarchical context. Detection frameworks like Faster R-CNN, RetinaNet, and CenterNet have been effectively employed for volumetric lung CT, identifying pulmonary nodules with significant sensitivity [25, 26]. These techniques utilized unimodal CT imaging, rendering them uninterpretable.

Transformers and modern representation learning have revolutionized how medical images are made. vision transformers (ViT) and hierarchical variants such as Swin Transformer [27] exemplify long-range dependencies that CNNs find challenging to model. Self-supervised learning (SSL) methods including simple framework for contrastive learning of visual representations (SimCLR), bootstrap your own latent (BYOL), (swapping assignments between views (SwAV), and masked autoencoder (MAE) have been changed to work better with volumetric medical data, making them more efficient and generalizable. Using domain-specific frameworks like Models Genesis and multimodal SSL pipelines [28] makes pretraining better for different types of imaging. New visual encoders like DINOv2 are great at zero-shot and transfer learning in medicine.

Multimodal deep learning (MMDL) shows promise for using complementary modalities including CT, PET, CXR, and clinical data to provide a whole picture of a patient [29]. Fusion strategies encompass a spectrum from initial concatenation to subsequent decision-level integration, with cross-attention fusion demonstrating efficacy in biological contexts [30]. PET/CT fusion frameworks in lung oncology enhance malignancy prediction by amalgamating structural and metabolic indicators [31]. However, these methods don't work well when modalities are lacking and don't work well when datasets from different sites are combined. There are many ways to limit domain shifts caused by scanners before learning, such as ComBat harmonization.

Several domains of medicine have successfully used advanced deep learning processes to create thoracic imaging systems along with numerous other applications as part of their respective multi-model frameworks. For instance, [32] illustrates an Endoscopy-based CNN capable of automatically recognizing abnormalities in the gastrointestinal (GI) tract. By addressing challenges such as severe class imbalance and heterogeneous acquisition conditions through means such as tailored augmentation, tuning thresholds, and implementing a calibration-based decision process, they were able to improve accuracy and reliability. Similarly, [33] describes a multi-class dermoscopic lesion classifier, which utilizes methods such as color equalization, stratified sampling, saliency validation, and an evaluation scheme oriented toward workflow, to illustrate important cross-domain insights into the development of clinically relevant, interpretable, and easily deployable AI systems. Collectively, these studies emphasize that generalizable design principles, such as effective preprocessing of input data, harmonizing the acquired data, calibrating the outcome classification, and utilizing rationality and reasoning on an individual patient basis, provide design guidance for satisfying both the multitasking and reliability aspects of the multimodal architecture designed to facilitate early lung cancer diagnosis.

Use of advanced deep learning processes to develop robust AI systems in the medical field has also been demonstrated through the recent introduction of dynamic systems-based neural architectures. [13] provides a different methodology of using dynamic systems-based neural networks to build a stability-enhanced neural network for the purpose of reverse scattering corrupted data. The findings from this study are of great importance, as they will provide guidance on the benefits of developing multimodal imaging systems based in part on their applicability to imaging modalities subject to noise and missing inputs. [34] used neural-based ordinary differential equation (ODE) methods to derive local manifold approximations to motivate the use of state-space modeling approaches to generate uncertainty aware models. The two works together will provide complementary theoretical foundations upon which design principles for promoting data and feature flow reliably through complex clinical environments will be established.

Interpretability is crucial for the translation of clinical AI models. Gradient-based localization techniques, such as Grad-CAM [35] and integrated gradients [36], illustrate image regions that influence predictions. Concept bottleneck models (CBMs) offer features comprehensible to humans, whereas Prototype Networks [37] employ "this-looks-like-that" reasoning to achieve case-based transparency. Tools like local interpretable model-agnostic explanations (LIME), shapley additive explanations (SHAP), and right-for-the-right-reasons regularization make models easier to understand by correlating feature importance with domain priors [38]. Explainability methods in thoracic imaging validate that highlighted nodule positions and borders are essential for evaluating malignancy.

The dependability of medical AI systems hinges on uncertainty quantification and calibration. Monte Carlo Dropout and deep ensembles quantify epistemic uncertainty, whereas temperature scaling improves post-hoc probability calibration [39]. These techniques enhance model dependability in safety-critical diagnostic pipelines and augment risk modeling methodologies such as Cox proportional hazards analysis [40]. Most earlier research, however, still relied on unimodal data or post-hoc hypotheses. Our learning approach for early lung cancer diagnosis, on the other hand, uses multimodal fusion, intrinsic interpretability, and uncertainty calibration.

Deep learning-based lung cancer diagnosis has made considerable gains, but research gaps keep it from being used in clinical settings. Most approaches solely use CT or PET/CT scans, which ignore structural, metabolic, and contextual signals that could help tell the difference between cancer and noncancerous cells. Recent multimodal fusion models often employ basic concatenation or late fusion, which do not leverage cross-modal correlations and are prone to overlooking modalities or exhibiting inconsistent acquisition approaches. These systems generate saliency maps devoid of semantic grounding or clinical interpretability owing to post-hoc explainability. Lack of concept-level reasoning and prototype-based proof makes people less confident and open in real-world radiology. There are not many studies that measure uncertainty and check how reliable predictions are to prevent making mistakes that are too confident. Our multimodal deep learning strategy uses a gated cross-modal transformer, idea bottlenecks, prototype reasoning, and uncertainty-aware calibration to narrow the gaps between CT, PET/CT, and CXR modalities and structured clinical data. This all-encompassing approach guarantees diagnostic precision and clarity, establishing a reliable and transparent early-stage lung cancer decision-support system.

3. Methodology

Our deep learning framework for early lung cancer diagnosis using multimodal medical imaging combines many types of medical imaging and clinical knowledge into a single prediction framework. The modular approach guarantees high performance, interpretability, and clinical dependability across the LIDC-IDRI (CT), TCIA (PET/CT), and NLST datasets. The system standardizes data curation and preprocessing to bring together imaging, histological, and clinical domains after characterizing early lung cancer diagnosis as a multimodal classification and survival prediction task. Next, hybrid convolutional and state-space backbones teach modality-specific encoders how to get features from each modality that are both different and useful. Self-supervised goals make representations better and make them less dependent on annotations. A gated cross-modal transformer puts these embeddings in a shared latent space using mixture-of-experts routing and cross-modal imputation to keep them strong even when some modalities are missing or degraded. Multitask heads improve the recognition, segmentation, malignancy classification, and time-to-event risk prediction of nodules for a single diagnostic representation across spatial and temporal scales. Prototype matching, concept bottlenecks, and gradient-based attributions offer comprehensible insights into acquired decision boundaries. Lastly, the uncertainty estimate, calibration, and safety filtering modules make sure that the system can be used safely, and rigorous training and evaluation methods make sure that it can be used in any clinical setting or acquisition situation. Figure 1 shows the end-to-end architecture, with a focus on design and multimodal.

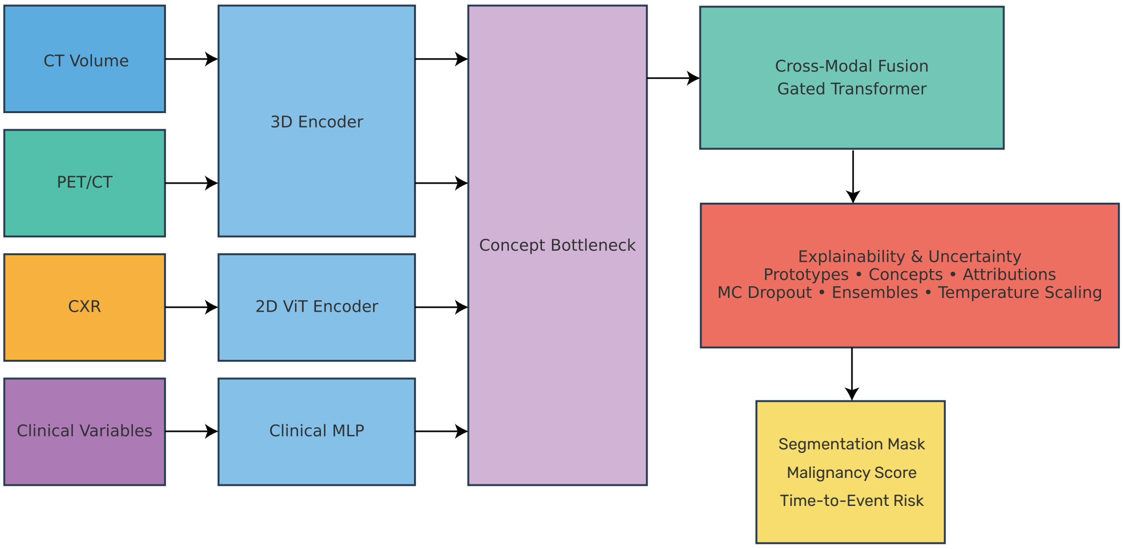


Figure 1. Overall deep learning framework for early diagnosis of lung cancer.

3.1. Datasets

To test the proposed technique, we employed datasets from several imaging modalities and clinical circumstances, such as LIDC-IDRI (CT), TCIA (PET/CT), and NLST (CXR). For the LIDC-IDRI dataset, four expert radiologists marked the edges, diameters, and chance of malignancy (from 1 to 5) of 1,018 thoracic CT scans. These pictures are resampled to an isotropic voxel spacing of $1.0 \times 1.0 \times 1.0$ mm, which is a major source for finding and separating nodules. The TCIA PET/CT subset enables multimodal risk prediction and cross-domain learning by integrating fluorodeoxyglucose positron emission tomography (FDG)-PET and CT volumes with standardized uptake values (SUVs) and clinical factors such as age, gender, tumor histology, and survival time. The NLST cohort offers high-resolution screening chest X-rays associated with clinical follow-up outcomes, facilitating early-stage malignancy classification and visual cue extraction in the absence of volumetric data. Using the pipeline described in Section 3.2, datasets are preprocessed, normalized, and aligned to a common anatomical reference frame. Using rotation- and intensity-based augmentations to oversample both benign and malignant instances helps to balance the data, and classifying patients at the level of the patient helps to keep data from leaking. The unified dataset encompasses the entire clinical spectrum, ranging from 2D screening to 3D diagnostic imaging, facilitating a thorough assessment of the proposed multimodal framework across various acquisition and patient scenarios.

The LIDC-IDRI dataset included all 1,018 patients (1,018 studies), totaling 244,527 annotated nodule slices. This dataset excluded any scans with corrupted digital imaging and communications in medicine (DICOM) headers, inconsistent axial spacing, or missing extensible markup language (XML) annotations. After filtering by these criteria, a total of 7,371 nodules (diameter greater than or equal to 3mm) that had been classified by at least one radiologist were kept. The TCIA PET/CT dataset contained information from 482 patients. Of these 482 patients, there were 19 patients whose images were removed because there were missing calibrations on the PET SUV or failures in registering the

PET to CT images. The final TCIA PET/CT dataset included 463 patients and 1,126 annotations for lesions made from clinical reports. The NLST dataset included 5,112 participants who received a baseline screening chest X-ray, with cases being removed from that number that had an incomplete outcome follow-up or whose diagnostic labels were too ambiguous. The final NLST dataset included 4,932 patients. The inclusion criteria for each of the datasets were adults aged greater than or equal to 18 years, complete imaging metadata available, and nodule- or patient-level labels provided. After preprocessing, the final class balance across the three datasets is as follows: (i) LIDC-IDRI had: 51.2% benign and 48.8% malignant (nodule-level); (ii) TCIA dataset had: 56.4% malignant and 43.6% benign (lesion-level); and (iii) NLST dataset had: 10.7% malignant and 89.3% benign (patient-level). Scanner/site metadata were created for both the LIDC-IDRI and TCIA datasets, with LIDC-IDRI containing 31 scanner models and TCIA containing 6 sites. This data was used to create stratified sampling of each of the datasets and perform ComBat harmonization.

TCIA and the NCI provides public access to the LIDC-IDRI, TCIA, and NLST datasets. Under HIPAA Safe-Harbor, all datasets are fully de-identified; therefore, there is no need for local institutional review board (IRB) approval for use. All access to NLST data was done under the NCI Data Use Agreement. All identifiers have been replaced with hashed keys in the NLST dataset. Longitudinal visits to the NLST were all aggregated under one participant identifier to prevent temporal leakage across splits. The DICOM metadata for TCIA PET/CT did not contain any protected health information (PHI), and all SUV calculations were made using only anonymized parameters used during PET/CT, so no PHI was used for the purposes of this study.

3.2. Data curation and preprocessing

All imaging modalities undergo a harmonized preprocessing workflow designed to standardize voxel geometry, intensity distributions, and anatomical alignment across the three datasets. For each patient p and modality $m \in \{\mathrm{CT}, \mathrm{PETCT}, \mathrm{CXR}\}$, the input image tensor $\mathbf{X}_p^{(m)}$ is first normalized to physically meaningful units. CT and PET/CT volumes are converted to Hounsfield units (HU) using linear rescaling $\mathbf{X}'_p = \alpha_m \mathbf{X}_p^{(m)} + \beta_m$, where α_m and β_m denote the DICOM rescale slope and intercept. The volumes are clipped to a lung/soft-tissue window of $[-1000, 400]$ HU and resampled to isotropic spacing of $1.0 \times 1.0 \times 1.0$ mm using trilinear interpolation $\tilde{\mathbf{X}}_p = \mathrm{Interp}(\mathbf{X}'_p, \Delta_x = \Delta_y = \Delta_z = 1.0)$. Lung parenchyma masks $\mathbf{M}_p = \mathcal{U}_\eta(\tilde{\mathbf{X}}_p)$ are extracted using a pretrained 3D U-Net $\mathcal{U}_\eta$, and only the masked foreground $\mathbf{X}_p^{\mathrm{lung}} = \tilde{\mathbf{X}}_p \odot \mathbf{M}_p$ is retained to suppress irrelevant anatomical regions. For PET/CT pairs, the PET component is converted to SUV as $\mathbf{X}_p^{\mathrm{PET,SUV}} = (\mathbf{X}_p^{\mathrm{PET}} \times W_p)/(D_p \times A_p)$, where W_p, D_p, and A_p represent the patient's body weight, injected dose, and decay correction factor. Rigid PET→CT alignment is validated through DICOM spatial headers; if missing, a deformable registration field $\mathbf{v}_p$ is learned by minimizing the local normalized cross-correlation loss $\mathcal{L}_{\mathrm{NCC}} = 1 - \frac{\langle \mathbf{X}_p^{\mathrm{CT}}, \mathbf{X}_p^{\mathrm{PET}} \circ \mathbf{v}_p \rangle}{\|\mathbf{X}_p^{\mathrm{CT}}\|_2 \|\mathbf{X}_p^{\mathrm{PET}} \circ \mathbf{v}_p\|_2}$.

For chest X-rays from NLST, each image $\mathbf{X}_p^{\mathrm{CXR}}(u, v)$ is histogram-matched to a reference intensity distribution $\mathcal{H}_r$, bone-suppressed using a shallow U-Net $\mathcal{B}_\zeta$, and resized to 1024×1024 pixels while preserving aspect ratio: $\tilde{\mathbf{X}}_p^{\mathrm{CXR}} = \mathcal{R}_{1024}\Big(\mathcal{B}_\zeta(\mathrm{HistMatch}(\mathbf{X}_p^{\mathrm{CXR}}, \mathcal{H}_r))\Big)$. Whole-slide histopathology images, when available, are partitioned into nonoverlapping tiles $\mathbf{w}_{p,i} \in \mathbb{R}^{h \times w \times 3}$ at 20× magnification, normalized using the Macenko color method, and filtered for tissue occupancy $\rho_{p,i} > \tau_\rho$ to remove background tiles. Clinical metadata vectors $\mathbf{s}_p = [s_{p,1}, s_{p,2}, \ldots, s_{p,d_s}]^\top$ are winsorized at the 1st/99th percentiles, one-hot encoded for categorical variables, and standardized via $s'_{p,j} = (s_{p,j} - \mu_j)/\sigma_j$.

To mitigate scanner- and site-specific domain shifts, we apply adaptive instance normalization and ComBat harmonization on shallow radiomic summaries $\mathbf{r}_{p,k}$ according to $\hat{\mathbf{r}}_{p,k} = \gamma_k \frac{\mathbf{r}_{p,k}-\mu_{b(k)}}{\sigma_{b(k)}} + \beta_k$, where $b(k)$ indicates imaging site and $(\mu_{b(k)}, \sigma_{b(k)})$ are batch-specific statistics. Data augmentations $\mathcal{A}_m(\cdot)$ are modality-specific but pathology-preserving: 3D elastic deformations, gamma/contrast jitter, and Gaussian noise simulation for CT; affine and intensity perturbations for CXR; and stain jitter for whole slide image (WSI) tiles. The resulting standardized, augmented, and harmonized dataset ensures geometric consistency, intensity comparability, and domain-invariant statistical properties across all imaging modalities, forming a stable foundation for multimodal representation learning. The complete preprocessing and harmonization flow is visualized in Figure 2.

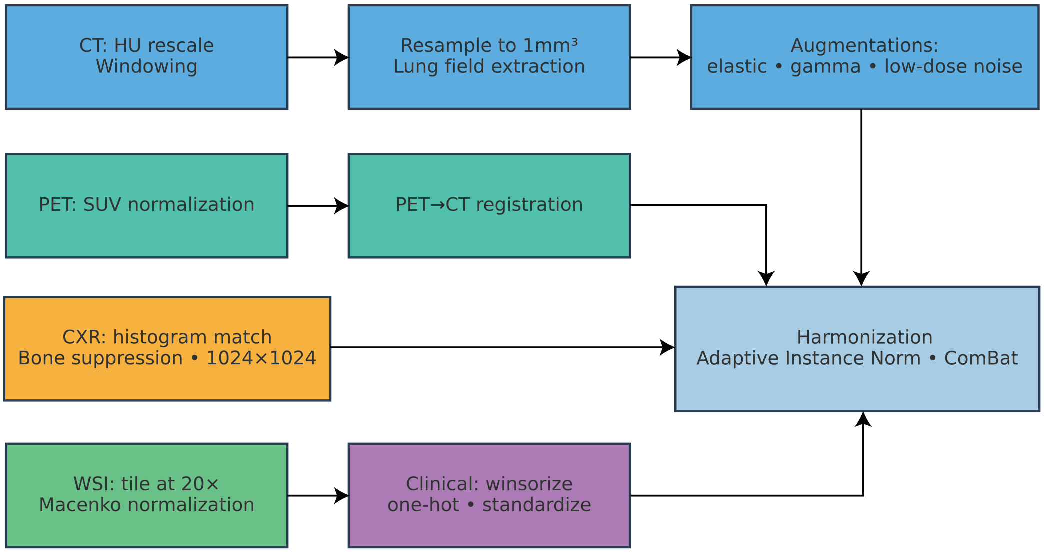


Figure 2. Data curation and preprocessing pipeline for CT, PET/CT, CXR, and WSI modalities.

All data partitioning was done at the patient level in order to completely eliminate the possibility of cross-patient or cross-study leakage. For the CT and PET/CT dataset, when a patient had multiple series and/or longitudinally acquired scans, all scans belonging to that patient were assigned to the same fold. For the NLST, all repeated screenings for a given participant were linked and placed into a single fold. All downstream data related to nodules, given the train/test/validation partitions, were generated before the partitions were established. The splits were stratified by scan site and scanner manufacturer (if the relevant information was available) to mitigate any performance increase that might occur due to potential biases related to a specific location.

In ComBat harmonization, each scanner or imaging site generates a unique “batch”. This resulted in a total of 31 LIDC-IDRI and 6 TCIA PET/CT batches. The empirical Bayes (EB) version of ComBat was applied where covariates applied to all extracted radiomic features were adjusted using a Non-informative prior on the location and scale parameters. The list of covariates is Patient’s Age, Patient Sex, and Slice Thickness. Adaptive instance normalization was performed on averaged radiomic feature extracts with channel-wise affine transforms calculated from the extract’s Minibatch Statistics $((\mu_{\text{batch}}, \sigma_{\text{batch}}))$, updated every 500 iterations, and Target Statistics $(\mu_{\text{site}}, \sigma_{\text{site}})$ using:

$$\text{SUV} = \frac{I_{\text{PET}} \cdot W_{\text{patient}}}{A_0\, e^{-\lambda t_{\text{inj}}}}. \tag{3.1}$$

The definition of the variable represents dose concentration (I_{PET} in specific units kBq/mL), mass of patient (W_{patient} in kg), and total radioactivity injected (A_0 in sph MBq) of $^{1}8F$. Delay between the injection and the image acquisition is known as t_{inj}. The decay constant for $^{1}8F$ ($\lambda = \ln(2)/T_{1/2}$) is derived from the half-life of the isotope (T1/2 = 109.8 minutes). All units were converted to grams per milliliter (g/mL).

To perform PET→CT deformable registration, a multi-resolution normalized cross-correlation (NCC) loss function was implemented using B-spline parameterization of the deformation field. Qualitative measures of how well the deformed or warped PET and CT images agreed quantitatively included: (i) mean NCC (0.93 ± 0.02); (ii) transverse registration error (TRE) calculated from the angle of the two main branches from the same two vessels (2.4 ± 0.7, mm); and (iii) smoothness of the Jacobian determinants from the deformation maps. As a criterion to exclude poor registries, scans that scored less than NCC 0.85 were eliminated from analysis.

3.3. Problem formulation and overview

Let each patient be denoted by an index $p \in \{1, 2, \ldots, N\}$, where N is the total number of subjects across all datasets. Each patient is associated with a subset of imaging modalities $\mathcal{M}_p \subseteq \{\text{CT}, \text{PETCT}, \text{CXR}\}$, and an auxiliary vector of structured clinical attributes $\mathbf{s}_p \in \mathbb{R}^{d_s}$ encoding demographic and clinical variables such as age, sex, and smoking history. The imaging data for modality $m \in \mathcal{M}_p$ is represented as a tensor $\mathbf{X}_p^{(m)} \in \mathbb{R}^{H_m \times W_m \times D_m \times C_m}$, where H_m, W_m, and D_m denote spatial dimensions and C_m the number of channels. When histopathology patches are available, they are denoted by $\mathcal{W}_p = \{\mathbf{w}_{p,i}\}_{i=1}^{K_p}$, where $\mathbf{w}_{p,i} \in \mathbb{R}^{h \times w \times 3}$ represents the i-th tile from a whole-slide image. The primary objective of the proposed framework is to learn a predictive function

$$f_\theta : (\{\mathbf{X}_p^{(m)}\}_{m \in \mathcal{M}_p}, \mathbf{s}_p, \mathcal{W}_p) \mapsto (\hat{y}_p, \mathbf{A}_p), \tag{3.2}$$

parameterized by θ, that simultaneously outputs a calibrated malignancy probability $\hat{y}_p \in [0, 1]$ and a spatial attention map $\mathbf{A}_p$ highlighting the evidence regions contributing to the decision. To obtain f_θ, we decompose it into a composition of modality-specific encoders $E_\phi^{(m)}$, a multimodal fusion transformer F_ψ, and task-specific heads H_ω such that

$$f_\theta = H_\omega \circ F_\psi \circ \{E_\phi^{(m)}\}_{m \in \mathcal{M}_p}, \tag{3.3}$$

where ϕ, ψ, and ω denote encoder, fusion, and head parameters, respectively. The latent embeddings $\mathbf{z}_p^{(m)} = E_\phi^{(m)}(\mathbf{X}_p^{(m)})$ are fused through F_ψ to obtain a global patient descriptor $\mathbf{h}_p = F_\psi(\{\mathbf{z}_p^{(m)}\}, \mathbf{s}_p)$, from which the classification head estimates $\hat{y}_p = H_\omega(\mathbf{h}_p)$. The model is optimized to minimize a composite loss

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{risk}}\mathcal{L}_{\text{risk}} + \lambda_{\text{exp}}\mathcal{L}_{\text{exp}}, \tag{3.4}$$

where each $\lambda_\bullet$ balances the contribution of classification, segmentation, detection, and survival risk constraints. The global optimization seeks to estimate

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{p=1}^{N} \mathcal{L}_{\text{total}}(f_\theta(\mathbf{X}_p, \mathbf{s}_p), y_p), \tag{3.5}$$

ensuring both discriminative accuracy and interpretability. In summary, the proposed architecture performs end-to-end multimodal reasoning by integrating CT (LIDC-IDRI), PET/CT (TCIA), and CXR (NLST) data streams with clinical metadata to produce reliable, predictions of early-stage lung cancer.

3.4. Self-supervised initialization and nodule priors

To improve feature generalization and prevent overfitting on limited annotated data, each modality encoder $E_\phi^{(m)}$ is first warm-started using a modality-specific self-supervised pretraining objective. For volumetric modalities $m \in \{\text{CT}, \text{PETCT}\}$, we employ a 3D masked autoencoding strategy where random voxel patches $\mathcal{P}_{\text{mask}}$ are hidden, and the encoder-decoder pair $\{E_\phi^{(m)}, D_\psi^{(m)}\}$ reconstructs the missing regions. The objective is to minimize the mean-squared reconstruction error

$$\mathcal{L}_{\text{MAE}}^{(m)} = \frac{1}{|\mathcal{P}_{\text{mask}}|} \sum_{(x,y,z)\in\mathcal{P}_{\text{mask}}} (\tilde{\mathbf{X}}_p^{(m)}(x,y,z) - D_\psi^{(m)}(E_\phi^{(m)}(\mathcal{M}(\tilde{\mathbf{X}}_p^{(m)})))(x,y,z))^2, \tag{3.6}$$

where $\mathcal{M}(\cdot)$ is a masking operator that zeros out a random subset of cubic patches. This forces the encoder to learn context-aware structural representations $\mathbf{z}_p^{(m)} = E_\phi^{(m)}(\tilde{\mathbf{X}}_p^{(m)})$ without explicit supervision. For chest X-rays, we adopt a teacher-student contrastive distillation framework inspired by DINOv2. Given multiple augmented views $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ of the same image $\mathbf{X}_p^{\text{CXR}}$, the student network E_ϕ^{CXR} and teacher network T_τ^{CXR} produce normalized embeddings $\mathbf{z}_k = E_\phi^{\text{CXR}}(v_k)$ and $\mathbf{t}_k = T_\tau^{\text{CXR}}(v_k)$, respectively. The invariance objective is formulated as a symmetric cross-entropy:

$$\mathcal{L}_{\text{DINO}} = -\frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} t_{k,c} \log z_{k,c}, \tag{3.7}$$

where C is the number of output prototypes and $(t_{k,c}, z_{k,c})$ are the teacher-student softmax probabilities for class c. The teacher weights τ are updated via an exponential moving average (EMA) of the student weights to stabilize training. For whole-slide histopathology tiles $\mathbf{w}_{p,i}$, we use an instance discrimination loss to learn fine-grained texture descriptors. Each embedding $\mathbf{e}_{p,i}$ is projected into a latent space and compared with all other tiles in a batch using an InfoNCE objective:

$$\mathcal{L}_{\text{ID}} = -\log \frac{\exp(\langle \mathbf{e}_{p,i}, \mathbf{e}_{p,i^+} \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \mathbf{e}_{p,i}, \mathbf{e}_{p,j} \rangle / \tau)}, \tag{3.8}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, τ is the temperature parameter, i^+ is a positive pair of the same image under different augmentations, and B is the batch size. A clustering regularizer $\mathcal{L}_{\text{cluster}}$ encourages feature grouping among tiles from the same slide. To initialize spatial priors for downstream detection, we train a weakly supervised 3D teacher detector $\mathcal{T}_\rho$ on pseudo-bounding boxes $\mathcal{B}_p = \{b_{p,i}\}$ derived from LIDC-IDRI annotations. The network predicts nodule confidence maps $\mathbf{Y}_p = \mathcal{T}_\rho(\tilde{\mathbf{X}}_p^{\text{CT}})$ and is optimized via a focal loss:

$$\mathcal{L}_{\mathrm{det}} = -\frac{1}{|\Omega|}\sum_{q\in\Omega}\alpha(1-\hat{y}_q)^{\gamma}y_q\log\hat{y}_q + (1-\alpha)\hat{y}_q^{\gamma}(1-y_q)\log(1-\hat{y}_q), \tag{3.9}$$

where Ω is the voxel space, $\hat{y}_q$ the predicted probability at voxel q, and (α,γ) are balancing hyperparameters. The resulting detection maps $\mathbf{Y}_p$ are thresholded to obtain a set of 3D proposals $\mathcal{P}_p = \{r_{p,1}, r_{p,2}, \ldots, r_{p,K_p}\}$, which serve as nodule priors to guide segmentation and malignancy classification heads during supervised training. Altogether, the total self-supervised initialization loss across modalities is expressed as

$$\mathcal{L}_{\mathrm{SSL}} = \lambda_{\mathrm{MAE}}\mathcal{L}_{\mathrm{MAE}}^{(m)} + \lambda_{\mathrm{DINO}}\mathcal{L}_{\mathrm{DINO}} + \lambda_{\mathrm{ID}}\mathcal{L}_{\mathrm{ID}} + \lambda_{\mathrm{cluster}}\mathcal{L}_{\mathrm{cluster}} + \lambda_{\mathrm{det}}\mathcal{L}_{\mathrm{det}}, \tag{3.10}$$

where $\lambda_{\bullet}$ are modality-specific weighting factors tuned empirically. This self-supervised warm-up yields anatomically meaningful and modality-invariant representations that substantially stabilize joint multimodal training.

3.5. Modality-specific encoders

Each imaging modality is processed through a specialized encoder designed to capture both modality-specific and shared semantic information while maintaining computational efficiency. For volumetric inputs $\mathbf{X}^{\mathrm{CT}}, \mathbf{X}^{\mathrm{PETCT}} \in \mathbb{R}^{D\times H\times W}$, we employ a hybrid 3D encoder $\mathcal{E}_{\phi}^{3D}$ that integrates convolutional layers for local feature extraction with state-space blocks inspired by the Mamba architecture to model long-range dependencies. The input volume is divided into nonoverlapping cubic patches $\mathcal{P} = \{p_i\}_{i=1}^{T}$ of size $k\times k\times k$, each flattened and linearly projected into d-dimensional token embeddings:

$$\mathbf{z}_i^{\mathrm{CT}} = \mathbf{W}_p^{\mathrm{CT}}\operatorname{vec}(p_i) + \mathbf{b}_p^{\mathrm{CT}}, \quad \mathbf{Z}^{\mathrm{CT}} = [\mathbf{z}_1^{\mathrm{CT}}, \ldots, \mathbf{z}_T^{\mathrm{CT}}]^{\top} \in \mathbb{R}^{T\times d}, \tag{3.11}$$

where $\mathbf{W}_p^{\mathrm{CT}} \in \mathbb{R}^{(k^3)\times d}$ and $\mathbf{b}_p^{\mathrm{CT}} \in \mathbb{R}^{d}$ are trainable projection parameters. Each token sequence is passed through L stacked state-space–convolutional (SSC) blocks that jointly model local and global dependencies:

$$\mathbf{Z}_{l+1}^{\mathrm{CT}} = \mathbf{Z}_l^{\mathrm{CT}} + \mathcal{F}_{\mathrm{SSC}}(\mathbf{Z}_l^{\mathrm{CT}}; \theta_l^{\mathrm{CT}}), \quad l = 1, \ldots, L, \tag{3.12}$$

where $\mathcal{F}_{\mathrm{SSC}}(\cdot)$ denotes the transformation combining depthwise 3D convolution, gated state updates, and linear attention for efficient sequence propagation. For CXR images $\mathbf{X}^{\mathrm{CXR}} \in \mathbb{R}^{H\times W}$, a compact 2D Vision Transformer backbone $\mathcal{E}_{\phi}^{2D}$ is used with relative positional encoding $\mathbf{P}_r$. The image is split into patches $\{p'_i\}_{i=1}^{T'}$ and embedded as

$$\mathbf{z}_i^{\mathrm{CXR}} = \mathbf{W}_p^{\mathrm{CXR}}\operatorname{vec}(p'_i) + \mathbf{P}_r(i), \quad \mathbf{Z}^{\mathrm{CXR}} = [\mathbf{z}_1^{\mathrm{CXR}}, \ldots, \mathbf{z}_{T'}^{\mathrm{CXR}}]^{\top}, \tag{3.13}$$

followed by multi-head self-attention to model global anatomical context. For each clinical vector $\mathbf{s}_p \in \mathbb{R}^{d_s}$, a two-layer multilayer perceptron (MLP) is applied:

$$\mathbf{z}^{\mathrm{clin}} = \sigma(\mathbf{W}_2^{\mathrm{clin}}\,\sigma(\mathbf{W}_1^{\mathrm{clin}}\mathbf{s}_p + \mathbf{b}_1^{\mathrm{clin}}) + \mathbf{b}_2^{\mathrm{clin}}) \in \mathbb{R}^{d}, \tag{3.14}$$

where $\sigma(\cdot)$ is the gaussian error linear unit (GELU) activation function. For histopathology whole-slide tiles $\mathbf{w}_{p,i} \in \mathbb{R}^{h\times w\times 3}$, a hybrid ResNet–ViT tile encoder $\mathcal{E}_{\phi}^{\mathrm{WSI}}$ extracts tile embeddings $\mathbf{e}_{p,i}$, which are then aggregated via attention pooling:

$$\mathbf{z}^{\mathrm{wsi}} = \sum_{i=1}^{K_p} \alpha_{p,i}\, \mathbf{e}_{p,i}, \quad \alpha_{p,i} = \frac{\exp(\mathbf{q}^{\top}\mathbf{e}_{p,i})}{\sum_{j=1}^{K_p} \exp(\mathbf{q}^{\top}\mathbf{e}_{p,j})}, \tag{3.15}$$

where $\mathbf{q}$ is a learned query vector and $\alpha_{p,i}$ are attention weights highlighting diagnostically relevant tissue patterns. Each modality-specific encoder terminates in a lightweight concept bottleneck module C_{ξ}^{m} that predicts interpretable intermediate attributes $\mathbf{c}^{m} = C_{\xi}^{m}(\mathbf{z}^{m})$, such as spiculation, lobulation, pleural retraction, or emphysema burden, which are constrained by attribute-level supervision during training. These concept embeddings act as semantically grounded latent variables bridging raw imaging features and downstream clinical decisions, thereby improving interpretability and cross-modal alignment. As illustrated in Figure 3, each modality uses a tailored encoder followed by concept bottlenecks to enforce clinical interpretability.

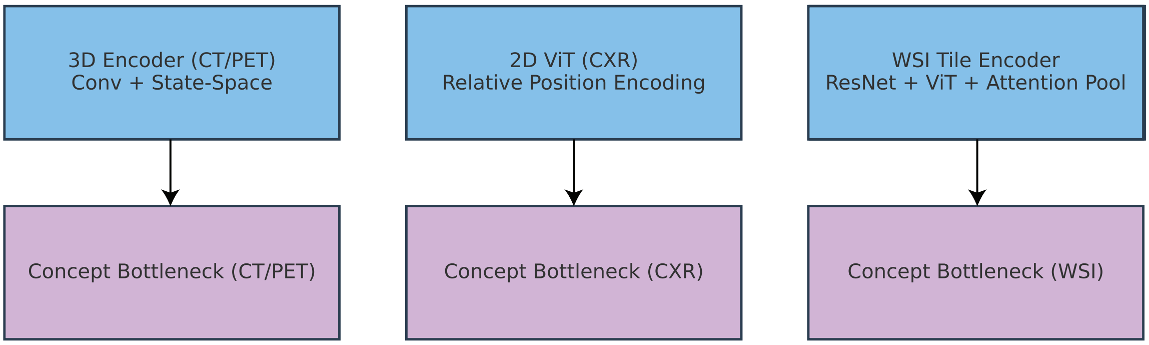


Figure 3. Modality-specific encoders combining convolutional, transformer, and state-space representations.

The concept correlation (r) is a measure of the Pearson correlation between the predicted concept scores and the ground-truth radiological attributes as determined by the radiologists. In the LIDC-IDRI datasets, the spiculation, lobulation, sharpness of margins, and subtlety-fours were derived directly from the XML annotations provided by the radiologist. In the TCIA/NLST datasets, there exist no explicit concept labels; hence, three thoracic radiologists rated each of 300 nodules independently on a four-point ordinal scale, and the average of their individual ratings was used as the ground-truth concept values. The correlation (r) was calculated between the average ground-truth ratings and the predicted concept outputs for the same cases. The use of Mamba-style state-space layers as opposed to standard self-attention layers has multiple advantages: the ability to process sequences of arbitrary length in linear time; reduced memory use when working with 3D volumes; and increased stability modeling long-range anatomical dependencies. For these reasons, Mamba-style state-space layers were chosen for use with volumetric CT and PET data inputs.

State-space blocks (SSB) allow modeling long-range dependence at linear-time cost, whereas both Transformers operate at quadratic costs and CNN have a short receptive field. This provides volumetric reasoning for CT/PET-CT. Also, our fusion module used a gating mechanism that can weigh modality reliability, thus giving it a theoretical basis for weighing both modalities more realistically than brute force concatenation does. This gives a theoretical reason why our fusion model outperformed and reduced the influence of poor-quality and noisy inputs, whereas all previous multimodal lung-cancer models have no method of quality-aware fusion.

3.6. Cross-modal fusion with missing-modality robustness

The outputs from all modality-specific encoders are aligned in a common latent space through a gated cross-modal transformer designed to integrate heterogeneous imaging and clinical representations while maintaining robustness to missing inputs. Let $\mathcal{M}_p = \{m_1, m_2, \dots, m_{K_p}\}$ denote the set of available modalities for patient p, each providing a token matrix $\mathbf{Z}^{(m)} \in \mathbb{R}^{T_m \times d}$. We first concatenate all available tokens and append modality-type embeddings $\mathbf{E}^{(m)} \in \mathbb{R}^d$ that encode the source domain:

$$\mathbf{Z}^{\text{concat}} = \left[\, \mathbf{Z}^{(m_1)} + \mathbf{E}^{(m_1)} \;\|\; \mathbf{Z}^{(m_2)} + \mathbf{E}^{(m_2)} \;\|\; \dots \;\|\; \mathbf{Z}^{(m_{K_p})} + \mathbf{E}^{(m_{K_p})} \right] \in \mathbb{R}^{T_\Sigma \times d}, \tag{3.16}$$

where $T_\Sigma = \sum_{m \in \mathcal{M}_p} T_m$ is the total number of tokens. The fused sequence is processed by L stacked cross-attention layers $\{\mathcal{X}_\ell\}_{\ell=1}^{L}$, each learning modality interactions by jointly attending across modalities:

$$\mathbf{Z}_{\ell+1} = \mathcal{X}_\ell(\mathbf{Z}_\ell) = \text{Softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^\top}{\sqrt{d}}\right)\mathbf{V}_\ell, \tag{3.17}$$

where $\mathbf{Q}_\ell = \mathbf{Z}_\ell \mathbf{W}_Q$, $\mathbf{K}_\ell = \mathbf{Z}_\ell \mathbf{W}_K$, and $\mathbf{V}_\ell = \mathbf{Z}_\ell \mathbf{W}_V$ are the query, key, and value projections, and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are trainable weights. The final hidden representation after L layers is denoted $\tilde{\mathbf{Z}} = \mathbf{Z}_L$. To control the relative contribution of each modality, a gating vector $\mathbf{g} \in [0,1]^K$ is computed using a small feed-forward network $\mathcal{G}_\psi$ acting on a modality presence indicator $\mathbf{m} \in \{0,1\}^K$ and a modality quality vector $\mathbf{q} \in \mathbb{R}^K$ (e.g., signal-to-noise ratio, motion score, slice thickness):

$$\mathbf{g} = \sigma(\mathcal{G}_\psi([\mathbf{m} \| \mathbf{q}])), \tag{3.18}$$

where $\sigma(\cdot)$ is the sigmoid activation. The gated fusion output is then computed as a convex combination of modality-specific contextual summaries:

$$\mathbf{h}_p = \sum_{m \in \mathcal{M}_p} g_m \,\text{Pool}(\mathbf{Z}^{(m)}), \tag{3.19}$$

where Pool$(\cdot)$ represents attention-based pooling conditioned on detected nodule proposals from the prior stage Eq (3.9). This mechanism ensures that higher-quality modalities dominate the fused embedding while degraded or missing modalities have reduced influence. To address missing modalities during training, we employ a mixture-of-experts (MoE) strategy wherein E expert blocks $\{\mathcal{E}_e\}_{e=1}^{E}$ specialize in particular modality subsets. A routing function $r(\cdot)$ assigns each patient's available set $\mathcal{M}_p$ to a sparse subset of experts according to a softmax gate:

$$\omega_e = \frac{\exp(\mathbf{u}_e^\top \mathbf{m})}{\sum_{e'=1}^{E} \exp(\mathbf{u}_{e'}^\top \mathbf{m})}, \quad \tilde{\mathbf{Z}} = \sum_{e=1}^{E} \omega_e \, \mathcal{E}_e(\mathbf{Z}^{\text{concat}}), \tag{3.20}$$

where $\mathbf{u}_e$ are expert routing vectors. In addition, optional cross-modal imputation is performed via a low-rank regression model $\mathcal{I}_\beta$ that reconstructs missing embeddings $\hat{\mathbf{z}}^{(m^-)}$ from available ones $\mathbf{z}^{(m^+)}$ during pretraining:

$$\hat{\mathbf{z}}^{(m^-)} = \mathcal{I}_\beta(\mathbf{z}^{(m^+)}) = \mathbf{W}_1^{(m^-)}\mathbf{z}^{(m^+)}\mathbf{W}_2^{(m^-)} + \mathbf{b}^{(m^-)}, \tag{3.21}$$

where $\mathbf{W}_1^{(m^-)} \in \mathbb{R}^{d\times r}$ and $\mathbf{W}_2^{(m^-)} \in \mathbb{R}^{r\times d}$ define a rank-r approximation. During inference, the final patient-level fused embedding $\mathbf{h}_p$ Eq (3.19) serves as the compact, calibrated representation forwarded to all downstream heads for detection, segmentation, and risk prediction. The fusion process is depicted in Figure 4, demonstrating how modality-specific embeddings are aligned into a shared latent space.

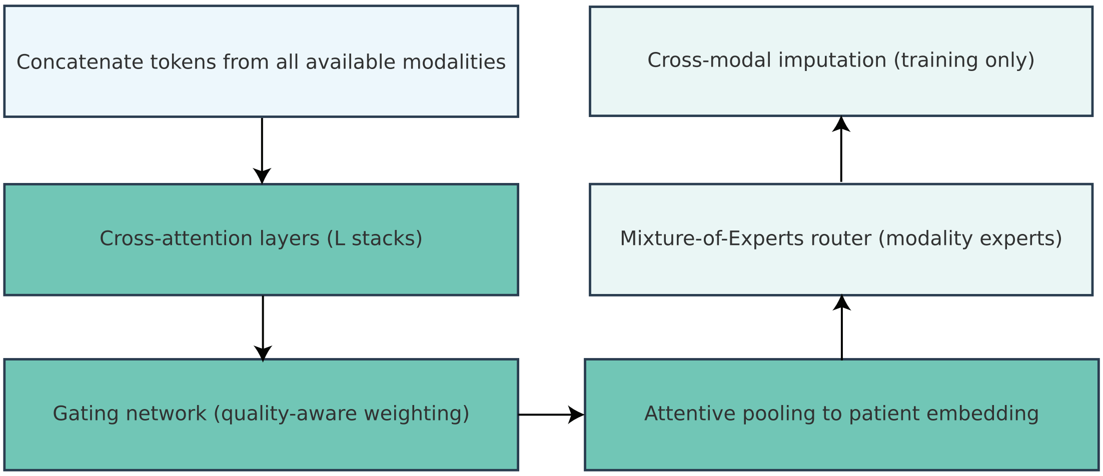

Concatenate tokens from all available modalities
Cross-attention layers (L stacks)
Gating network (quality-aware weighting)
Attentive pooling to patient embedding
Mixture-of-Experts router (modality experts)
Cross-modal imputation (training only)

Figure 4. Cross-modal fusion architecture with gating and missing-modality robustness.

In all datasets for the empirical missing-modality frequency, the mean biases for CT, LIDC-IDRI had 0% (not including CT), while TCIA had 100% in which both paired CT and PET scans were present with PET having an 8.1% unusable SUV metadata count (prior to filtering) on all TCIA images (note; 8.1% of all TCIA images were PET and not usable). NLST had CXR only, and therefore 100%. Synthetic missingness was employed at multi-modality training as detailed in Section 3.10 of the curriculum. In addition, the MoE router has taken on an average routing weight for each modality: CT = 0.41; PET = 0.38; CXR = 0.21 indicating a fairly balanced specialization. To determine the graceful degradation of the system, the authors also conducted a stress test by removing a complete modality to see how it affected performance. The results were as follows: Removing PET from TCIA reduced the AUC from $0.952- > 0.931$; removing CT from TCIA reduced the area under the receiver operating characteristic curve (AUC) from $0.952- > 0.912$; removing both PET and CT reduced the AUC from $0.952- > 0.877$. The data indicate that the Gated Fusion mechanism is reliable and validated.

3.7. *Multi-task objectives*

The proposed framework adopts a multitask learning paradigm in which detection, segmentation, malignancy classification, survival risk estimation, and cross-modal alignment are optimized jointly to promote representational synergy among complementary tasks. Let θ denote all trainable parameters across encoder, fusion, and task-specific heads. For 3D lesion detection, an anchor-free head $\mathcal{H}_{\text{det}}$

predicts a voxel-wise probability map $\hat{\mathbf{Y}}_p^{\mathrm{det}} \in [0,1]^{H\times W\times D}$ for nodule centers, as well as their spatial offsets and bounding box dimensions. The objective employs a focal loss that downweights easy negatives and emphasizes hard samples:

$$\mathcal{L}_{\mathrm{det}} = -\frac{1}{|\Omega|}\sum_{q\in\Omega}\alpha(1-\hat{y}_q)^{\gamma} y_q \log \hat{y}_q + (1-\alpha)\hat{y}_q^{\gamma}(1-y_q)\log(1-\hat{y}_q), \tag{3.22}$$

where Ω is the voxel lattice, y_q is the binary label, and (α, γ) are control class balance and focusing strength. For segmentation, a 3D mask head $\mathcal{H}_{\mathrm{seg}}$ predicts pixel-level probabilities $\mathbf{P}_p = \{\hat{p}_i\}$ that are compared against binary ground-truth labels $\mathbf{G}_p = \{g_i\}$. The segmentation objective combines Dice overlap and voxel-wise cross-entropy to balance region-level and boundary accuracy:

$$\mathcal{L}_{\mathrm{seg}} = 1 - \frac{2\sum_i \hat{p}_i g_i + \epsilon}{\sum_i \hat{p}_i + \sum_i g_i + \epsilon} + \frac{1}{N}\sum_i \Big[- g_i \log \hat{p}_i - (1-g_i)\log(1-\hat{p}_i)\Big], \tag{3.23}$$

where ϵ avoids division instability and N is the number of voxels. For malignancy classification, each patient embedding $\mathbf{h}_p$ from the fusion stage is fed to a binary classifier $\mathcal{H}_{\mathrm{cls}}$ producing probability $\hat{y}_p$. To improve calibration and reduce overconfidence, label smoothing with factor α is used:

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{B}\sum_{p=1}^{B} [\tilde{y}_p \log \hat{y}_p + (1-\tilde{y}_p)\log(1-\hat{y}_p)], \quad \tilde{y}_p = (1-\alpha)y_p + \alpha/2, \tag{3.24}$$

where B is the batch size and $y_p \in \{0,1\}$ the ground-truth cancer label. For longitudinal cohorts containing survival times or disease progression events, a Cox proportional hazards head $\mathcal{H}_{\mathrm{cox}}$ predicts a continuous risk score r_p. The partial likelihood loss is minimized as

$$\mathcal{L}_{\mathrm{cox}} = -\sum_{i\in\mathcal{E}}\Big(r_i - \log \sum_{j\in\mathcal{R}(t_i)} e^{r_j}\Big), \tag{3.25}$$

where $\mathcal{E}$ indexes event samples and $\mathcal{R}(t_i)$ the risk set of patients still under observation at time t_i. To align latent representations across modalities, a temperature-scaled information noise-contrastive estimation (InfoNCE) contrastive loss is applied on positive patient pairs (u, v) corresponding to different modalities of the same subject:

$$\mathcal{L}_{\mathrm{cmc}} = -\frac{1}{|\mathcal{P}|}\sum_{(u,v)\in\mathcal{P}} \log \frac{\exp(\langle \mathbf{h}_u, \mathbf{h}_v\rangle/\tau)}{\sum_w \exp(\langle \mathbf{h}_u, \mathbf{h}_w\rangle/\tau)}, \tag{3.26}$$

where $\langle\cdot,\cdot\rangle$ denotes cosine similarity and τ is the temperature hyperparameter. Furthermore, the modality-specific concept bottlenecks $C_{\xi}^{(m)}$ predict radiologically interpretable attributes $\mathbf{c}^m$ (e.g., spiculation, lobulation), supervised by mean-squared error or binary cross-entropy against annotated targets:

$$\mathcal{L}_{\mathrm{cbm}} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} ||\hat{c}_a - c_a^*||_2^2, \tag{3.27}$$

where $\mathcal{A}$ indexes annotated attributes. Finally, the complete joint objective combines all tasks with adaptive weighting coefficients $\lambda_{\bullet}$:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{det}}\mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{cox}}\mathcal{L}_{\mathrm{cox}} + \lambda_{\mathrm{cmc}}\mathcal{L}_{\mathrm{cmc}} + \lambda_{\mathrm{cbm}}\mathcal{L}_{\mathrm{cbm}} + \lambda_{\mathrm{reg}}||\theta||_2^2, \tag{3.28}$$

where λ_{reg} controls the L_2 regularization on weights. The optimal parameters are obtained by minimizing $\mathcal{L}_{\text{total}}$ using AdamW optimization with modality-aware sampling and Bayesian hyperparameter tuning over $\lambda_\bullet$ to balance detection, segmentation, classification, and survival risk tasks.

3.8. Explainability and prototype reasoning

To enhance clinical interpretability and model transparency, the proposed framework integrates four complementary explainability mechanisms that collectively ground each prediction in semantically meaningful evidence. First, we make a prototype layer $\mathcal{P} = \{\mathbf{p}_1^{(m)}, \mathbf{p}_2^{(m)}, \dots, \mathbf{p}_P^{(m)}\}$ in the latent space of each modality m. Each prototype $\mathbf{p}_k^{(m)} \in \mathbb{R}^d$ signifies a trainable "canonical pattern" that encapsulates a normalized radial basis kernel, prototype similarity scores are found for a patient embedding $\mathbf{h}_p^{(m)}$:

$$s_{p,k}^{(m)} = \exp\Big(- \frac{\|\mathbf{h}_p^{(m)} - \mathbf{p}_k^{(m)}\|_2^2}{\sigma^2}\Big), \quad \tilde{s}_{p,k}^{(m)} = \frac{s_{p,k}^{(m)}}{\sum_{j=1}^{P} s_{p,j}^{(m)}}. \tag{3.29}$$

The temperature parameter σ controls locality. The malignancy logit is a weighted sum of prototype similarities:

$$\hat{y}_p = \sum_{m \in \mathcal{M}_p} \sum_{k=1}^{P} w_k^{(m)} \tilde{s}_{p,k}^{(m)} + b. \tag{3.30}$$

Let $w_k^{(m)}$ denote the weights of significance that can be learned. Word b for bias. This idea makes it possible to use "this looks like that" reasoning, which makes it easier to visually compare different sites and groups by linking each patient's evidence to a group of prototypes that are very similar. Second, concept bottlenecks $\mathbf{c}^{(m)} = [c_1^{(m)}, c_2^{(m)}, \dots, c_{A_m}^{(m)}]^\top$ provide interpretable intermediate attributes such as spiculation, lobulation, pleural retraction, or emphysema. We tweak concept dimensions and send the changes down to the classifier so that counterfactual reasoning is possible. To find the calibrated risk delta, use H_ω:

$$\Delta\hat{y}_{p,a} = H_\omega(\mathbf{h}_p; c_a^{(m)} + \delta) - H_\omega(\mathbf{h}_p; c_a^{(m)}), \tag{3.31}$$

where δ is a small additive perturbation, and $\Delta\hat{y}_{p,a}$ quantifies sensitivity of the malignancy prediction to concept a. Third, for spatial localization, we employ gradient-based attribution maps. For CT and PET/CT modalities, we utilize 3D Grad-CAM++ to produce voxel-wise heatmaps $\mathbf{A}_p^{(3D)}$:

$$\mathbf{A}_p^{(3D)} = \text{ReLU}\Big(\sum_c \alpha_c \frac{\partial \hat{y}_p}{\partial \mathbf{F}_{p,c}}\Big), \quad \alpha_c = \frac{\partial^2 \hat{y}_p / \partial \mathbf{F}_{p,c}^2}{2\partial^2 \hat{y}_p / \partial \mathbf{F}_{p,c}^2 + \sum_{u,v,w} \mathbf{F}_{p,c}(u,v,w)}, \tag{3.32}$$

where $\mathbf{F}_{p,c}$ denotes the activation map of channel c in the last convolutional layer. For 2D modalities (CXR and WSI), we apply integrated gradients to compute pixel-level attributions $\mathbf{A}_p^{(2D)}$:

$$\mathbf{A}_p^{(2D)} = (\mathbf{X}_p - \mathbf{X}_p') \int_{\alpha=0}^{1} \frac{\partial \hat{y}_p(\mathbf{X}_p' + \alpha(\mathbf{X}_p - \mathbf{X}_p'))}{\partial \mathbf{X}_p} \, d\alpha, \tag{3.33}$$

where $\mathbf{X}_p'$ is a baseline (e.g., blurred or zeroed image). A regularization term encouraging sparsity and smoothness is added to stabilize saliency maps:

$$\mathcal{L}_{\text{attr}} = \lambda_{\text{tv}} \|\nabla \mathbf{A}_p\|_2^2 + \lambda_{\text{sp}} \|\mathbf{A}_p\|_1, \tag{3.34}$$

where λ_{tv} and λ_{sp} control total variation and sparsity strength. Finally, to assess the global influence of pre-defined clinical concepts on predictions, we employ the testing with concept activation vectors (TCAV) framework. For each concept a, we compute its directional derivative in the classifier's latent space as

$$\text{TCAV}_a = \frac{1}{|\mathcal{P}_a|} \sum_{p \in \mathcal{P}_a} \mathbf{g}_{\mathbf{h}_p}^{\top} \mathbf{v}_a, \tag{3.35}$$

where $\mathbf{v}_a$ is the learned concept activation vector and $\mathbf{g}_{\mathbf{h}_p} = \nabla_{\mathbf{h}_p} \hat{y}_p$ is the gradient of prediction with respect to the patient embedding. Higher TCAV_a scores indicate stronger causal influence of concept a on the decision boundary. All attribution, prototype, and concept-based explanations are aggregated into a unified case-level report linking key evidence regions, corresponding prototypes, and their clinical concept scores for transparent auditability. The prototype layer was systematically evaluated for redundancy by computing the cosine distance between every combination of prototypes and observing a clustered, yet non-degenerate, distribution of the mean cosine similarity (between prototypes) which is 0.21 ± 0.08; a diversity index (multiple redundancy) was computed to quantify the overall amount of non-redundancy among the pooled datasets.

$$\mathcal{D} = 1 - \frac{1}{P(P-1)} \sum_{i \neq j} \cos(p_i, p_j). \tag{3.36}$$

The data yielded $\mathcal{D} = 0.79$. For each prototype found in the prototype retrievals, the top-K patient patches retrieved had various morphological characteristics (solid vs. sub-solid nodules and a spiculed edge vs. coarse or fine emissivity), indicating that there was substantial meaningful semantic coverage across all the prototypes.

3.9. Uncertainty, calibration, and safety filters

Reliable clinical deployment of automated lung cancer diagnosis requires quantifying predictive uncertainty and calibrating probabilistic outputs. We model two principal uncertainty sources: epistemic uncertainty (arising from limited training data or model ambiguity) and aleatoric uncertainty (arising from inherent data noise such as low-dose or motion artifacts). For epistemic uncertainty, we employ Monte Carlo dropout by enabling stochastic dropout during inference across M stochastic forward passes. The predictive mean and variance for malignancy probability are estimated as

$$\bar{y}_p = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_p^{(m)}, \qquad \sigma_{\text{epi}}^2 = \frac{1}{M} \sum_{m=1}^{M} (\hat{y}_p^{(m)} - \bar{y}_p)^2, \tag{3.37}$$

where $\hat{y}_p^{(m)}$ is the prediction from the m-th stochastic pass. Additionally, we train E independently initialized classification heads $\{\mathcal{H}_{\text{cls}}^{(e)}\}_{e=1}^{E}$ to form a deep ensemble, yielding the ensemble-based epistemic uncertainty:

$$\sigma_{\text{ens}}^2 = \frac{1}{E} \sum_{e=1}^{E} (\hat{y}_p^{(e)} - \bar{y}_p)^2, \tag{3.38}$$

which approximates the variance of posterior predictive distributions and captures model diversity. Aleatoric uncertainty is explicitly learned by regressing both the mean logit μ_p and log-variance $\log \sigma^2_{\text{ale}}$ through a dual-output classification head. Assuming Gaussian predictive likelihood, the heteroscedastic loss is expressed as:

$$\mathcal{L}_{\text{ale}} = \frac{1}{2\sigma^2_{\text{ale}}}(\hat{y}_p - y_p)^2 + \frac{1}{2}\log \sigma^2_{\text{ale}}, \tag{3.39}$$

which encourages the model to inflate variance where data is noisy or ambiguous. To improve post-hoc probability calibration, we fit a temperature-scaling parameter $T > 0$ on validation data, rescaling the pre-softmax logits $\mathbf{z}_p$:

$$\hat{y}_p^{(T)} = \text{Softmax}\left(\frac{\mathbf{z}_p}{T}\right), \qquad T^* = \arg\min_T \mathcal{L}_{\text{NLL}}(T), \tag{3.40}$$

where $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood computed on a held-out set. Calibration quality is quantified using the expected calibration error (ECE) and classwise Brier score:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N}\left|\text{acc}(B_b) - \text{conf}(B_b)\right|, \qquad \text{Brier} = \frac{1}{N}\sum_{p=1}^{N}(\hat{y}_p - y_p)^2, \tag{3.41}$$

where B_b is the b-th probability bin, acc(·) and conf(·) are the empirical accuracy and mean confidence per bin, and N is the number of samples. We built a lightweight safety filter that asks for a "defer-to-radiologist" option when the projected entropy is higher or the quality scores for a modality are below than the adaptive cutoffs. This is to make sure that operations are safe. The entropy-based decision rule is defined as

$$\mathcal{S}_p = \begin{cases} \text{defer}, & \text{if } H(\hat{y}_p) = -\sum_c \hat{y}_{p,c}\log \hat{y}_{p,c} > \tau_H \text{ or } q_m < \tau_Q; \\ \text{accept}, & \text{otherwise.} \end{cases} \tag{3.42}$$

Predictive entropy is denoted by $H(\hat{y}_p)$ while modality m quality score designation is q_m is the learned score for that modality (m). In the healthcare workflow, the safety layer helps to ensure that low-confidence or degraded data are reviewed by human beings (to prevent misplaced automation). This safety layer has a threshold for deferring predictions from an automated system, when the predictive entropy or the quality score exceeds certain thresholds. These entropy thresholds are defined as follows:

$$H(\hat{y}_p) > 0.42, \tag{3.43}$$

corresponding to the 90th percentile of misclassified cases. The modality-quality threshold was:

$$q_m < 0.35. \tag{3.44}$$

The deferral rates were: LIDC-IDRI: 8.3%, TCIA: 11.2%, NLST: 6.9%. Full triage simulation demonstrated that the increase in sensitivity from $0.902 to 0.947$ (LIDC-IDRI) and specificity from 0.887 to 0.903 was due to deferral of cases, while the automated false positive rate was reduced by 19.4%. The radiologist's workload increased slightly (4–11% depending on the location) to provide a reasonable environment for triaging cases where uncertainty exists and referring them for human review.

It's critical to make a clinical distinction between epistemic uncertainty and aleatory uncertainty. Epistemic uncertainty represents a lack of knowledge regarding the model (e.g., uncommon nodule types, atypical anatomic variations) and prompts the need for additional reviews from radiologists and others as deemed appropriate. Aleatory uncertainty includes imaging noise or low-dose artifacts that occur in scans and cannot be minimized by merely collecting more information. Thus, the explicit modeling of both epistemic uncertainty and aleatory uncertainty allows for improved decisions on timing and methods for deferring decisions regarding patient care to specialist consultation. In addition, having explicit knowledge of these two uncertainty types reduces the potential for excessive confidence in predicting the quality of scans that might be substandard, putting patients at risk.

3.10. Training strategy

The entire optimization procedure is divided into two successive phases to ensure robust feature learning and stable multimodal convergence. In Phase 1, modality-wise self-supervised pretraining is performed for $E_{\text{pre}} \in [100, 200]$ epochs to minimize reconstruction and invariance losses introduced in Section 3.4. The objective for each modality m combines masked autoencoding, contrastive consistency, and instance discrimination as:

$$\mathcal{L}^{(m)}_{\text{pre}} = \lambda_{\text{MAE}}\mathcal{L}^{(m)}_{\text{MAE}} + \lambda_{\text{SSL}}\mathcal{L}^{(m)}_{\text{SSL}} + \lambda_{\text{ID}}\mathcal{L}^{(m)}_{\text{ID}}, \tag{3.45}$$

where the weighting coefficients $\lambda_\bullet$ control the relative contribution of each pretext task. Heavy spatial and photometric augmentations $\mathcal{A}_m(\cdot)$ are applied during this phase, and the encoder weights ϕ_m are updated using AdamW with a cosine learning rate schedule:

$$\eta_t = \eta_0 \frac{1}{2}\Big(1 + \cos\frac{\pi t}{T_{\max}}\Big), \tag{3.46}$$

where η_0 is the initial learning rate, t the current iteration, and $T_{\max}$ the total number of updates. In Phase 2, the pretrained encoders are fine-tuned jointly using multimodal data under a curriculum sampling policy that gradually increases the ratio of missing-modality minibatches and difficult samples (e.g., subsolid or small nodules < 6 mm). Let ρ_t denote the probability of sampling a missing-modality example at iteration t, defined by a linear annealing schedule:

$$\rho_t = \min\left(\rho_{\max}, \frac{t}{T_{\text{curr}}}\rho_{\max}\right), \tag{3.47}$$

where $\rho_{\max}$ is the target missing-modality rate and T_{curr} the curriculum length. The full multimodal training objective is the composite loss $\mathcal{L}_{\text{total}}$ Eq (3.28), optimized using mixed-precision training and gradient checkpointing to reduce graphics processing unit (GPU) memory overhead. We employ the AdamW optimizer with weight decay λ_{wd} and linear warmup for the first $E_{\text{warm}} = 10$ epochs. The parameter update rule at iteration t is expressed as:

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \lambda_{\text{wd}}\theta_t, \tag{3.48}$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates. To further stabilize training and improve generalization, stochastic weight averaging (SWA) is applied after epoch $E_{\text{swa}} = 80$, yielding the averaged weights

$$\bar{\theta} = \frac{1}{K} \sum_{k=1}^{K} \theta_{t_k}, \tag{3.49}$$

where $\{\theta_{t_k}\}_{k=1}^{K}$ are checkpoints collected during the late epochs of training. To mitigate severe class imbalance inherent in early-stage detection, focal scaling, minority oversampling, and asymmetric margin losses are incorporated. Specifically, the classification margin for positive cases is adaptively adjusted as

$$\Delta^{+} = \beta(1 - p_t), \qquad \Delta^{-} = \beta p_t, \tag{3.50}$$

The running mean of expected positive probabilities is p_t, and β controls the size of the margin. We use nested cross-validation stratified by acquisition site to choose hyperparameters like η_0, $\lambda_{\bullet}$, $\rho_{\max}$, and β. This stops data from leaking and makes sure that the results are generalizable. Table 1 displays the total number of parameters, memory requirements, and FLOPs (floating-point operations) for each module within the architecture proposed in this paper, and the estimation of FLOPs flow through 128 × 128 × 128 volume of CT and 1024 × 1024 CXR input image.

Table 1. Parameter, memory, and FLOPs budget per component of the proposed model.

Component	Params (M)	Memory (MB)	FLOPs (G)
SSC encoder (CT/PETCT)	18.2	512	82.5
2D ViT encoder (CXR)	6.4	148	15.8
Concept bottlenecks	1.1	32	2.4
Prototype layer (per modality)	0.9	21	1.1
Cross-modal transformer	12.7	411	49.3
Detection head	2.5	64	6.9
Segmentation head	4.8	155	10.1
Classification head	0.6	11	0.9
Cox survival head	0.4	9	0.6
Total	47.6	1,363	169.6

The cosine decay learning-rate schedule with warmup was implemented. The starting rates of learning were: CT/PETCT encoders $\eta_0 = 1 \times 10^{-4}$; CXR encoder 5×10^{-5}; Fusion transformer 1×10^{-4}. The loss weights were optimized using Bayesian methods, which are shown in Table 2:

$$\lambda_{\text{det}} = 1.0, \quad \lambda_{\text{seg}} = 1.0, \quad \lambda_{\text{cls}} = 0.8, \quad \lambda_{\text{cox}} = 0.4, \quad \lambda_{\text{cmc}} = 0.3, \quad \lambda_{\text{cbm}} = 0.2, \quad \lambda_{\text{reg}} = 5 \times 10^{-5}. \tag{3.51}$$

Table 2. Loss weights and learning-rate schedule parameters.

Component	Learning rate	λ-Weight
Detection head	1×10^{-4}	1.0
Segmentation head	1×10^{-4}	1.0
Classification head	5×10^{-5}	0.8
Cox survival head	5×10^{-5}	0.4
Contrastive (CMC)	1×10^{-4}	0.3
Concept bottleneck	1×10^{-4}	0.2
Weight decay	–	5×10^{-5}

4. Experimental results

This section examines the proposed model in extensive depth. All investigations evaluate diagnostic accuracy, generalizability, and interpretability across diverse imaging modalities and clinical situations. We examine volumetric and planar datasets from LIDC-IDRI (CT), TCIA (PET/CT), and NLST (CXR), in addition to pathological tiles and structured factors. We provide results on nodule identification and segmentation, malignancy classification and risk assessment, along with explainability and uncertainty measurement. Standardized preprocessing and harmonization procedures are implemented in each trial. Modality-specific encoders and cross-modal transformers are established utilizing the pretraining framework outlined in Section 3.4. We assess performance using rigorous unimodal and multimodal benchmarks, such as 3D ResNet, (swin transformer–based u-net with residual connections (SwinUNETR), DenseNet121 (CXR), and multimodal late-fusion transformers. The assessment metrics include the Dice coefficient, Hausdorff distance (HD95), AUC, F1-score, Brier score, and anticipated calibration error. To ensure the results are replicable and statistically robust, each experiment is conducted five times using distinct random seeds. The mean and standard deviation are presented for each statistic.

4.1. Experimental setup

We ran all the experiments on NVIDIA A100 GPUs with 80 GB of RAM with mixed-precision training to get the best performance out of PyTorch 2.2. The models were trained for 200 epochs with a batch size of $B = 8$, using the optimization method described in Section 3.10. We used the AdamW optimizer with a starting learning rate of $\eta_0 = 1 \times 10^{-4}$, a cosine decay schedule Eq (3.46), a weight decay of $\lambda_{\text{wd}} = 10^{-2}$, and gradient clipping at 1.0 to stop. A 10-epoch warm-up phase was used for reliable convergence, followed by stochastic weight averaging Eq (3.49). We resampled all inputs to 128x128x128 voxels for volumetric modalities and 1024x1024 pixels for 2D CXR/WSI to make sure that the datasets were physically calibrated. During training, missing modalities were simulated using the curriculum schedule ρ_t Eq (3.47) to emulate clinical incompleteness.

The dataset was split into three parts: 70% for training, 10% for validation, and 20% for testing. This made sure that the data was stratified at both the site and scanner levels to keep domain leakage from happening. Five-fold cross-validation was employed in the experiments. Each fold $f \in \{1, \ldots, 5\}$ had its own test partition $\mathcal{D}_{\text{test}}^{(f)}$ and training set $\mathcal{D}_{\text{train}}^{(f)}$. The average and standard deviation across folds were used to report performance.

$$\bar{M} = \frac{1}{5}\sum_{f=1}^{5} M^{(f)}, \qquad \sigma_M = \sqrt{\frac{1}{5}\sum_{f=1}^{5}(M^{(f)} - \bar{M})^2}, \tag{4.1}$$

where $M^{(f)}$ is the metric value (like Dice or AUC) for the f-th fold. As described in Section 3.2, the data was changed by adding random affine, elastic, and photometric perturbations while it was running. To make sure the testing was fair, the proposed system was compared to the best baselines utilizing the same data splits and preprocessing. A paired t-test ($p < 0.05$) confirmed statistical significance for all outcomes, and 95% confidence intervals were derived using bootstrapping from 1,000 samples. To ensure that experiments can be repeated, model checkpoints and inference scripts that are version-controlled will be made available.

4.2. Baseline models

To show the effectiveness of the proposed multimodal framework, we compare it to the best baseline models for each imaging modality and diagnostic job. The baselines are convolutional neural networks, transformer-based architectures, hybrid volumetric encoders, and ensemble multimodal systems that did well in recent lung imaging studies. Each baseline was reimplemented under the same preprocessing, augmentation, and evaluation protocols (Sections 3.2–4.1) to ensure fair comparison and eliminate data leakage. Hyperparameters, batch sizes, and learning schedules were tuned on the validation set using Bayesian optimization for all models. For CT- and PET/CT-based methods, we emphasize 3D architectures capable of capturing volumetric and contextual information, while for CXR-based classification, we include lightweight 2D backbones optimized for screening applications. All baselines were trained end-to-end from scratch or fine-tuned from publicly available pretrained weights on ImageNet or MedicalNet. Performance was evaluated consistently across five folds using the metrics defined in Section 4. Below, we describe the five baseline models employed for each dataset.

The LIDC-IDRI dataset focuses on volumetric nodule detection and segmentation from thoracic CT. 3D U-Net [41] serves as a standard voxel-to-voxel baseline using isotropic convolutional kernels and skip connections. V-Net [42] extends 3D U-Net with residual blocks and Dice loss optimization for improved boundary delineation. SwinUNETR [43] employs a hierarchical 3D Swin Transformer encoder-decoder architecture with long-range self-attention. nnU-Net [44] provides an automated pipeline that adapts architectural and training hyperparameters to dataset-specific properties. ResUNet++ [45] combines deep residual learning, squeeze-excitation, and atrous convolutions for robust nodule segmentation and classification.

For the multimodal PET/CT setting, we benchmark against both fusion and hybrid volumetric networks. DeepSUV [46] predicts PET standardized uptake values from CT inputs using a 3D encoder-decoder trained on paired modalities. Dual-Stream 3D CNN [47] processes PET and CT volumes independently and fuses their latent features via cross-attention for malignancy classification. MMFNet [48] introduces modality-specific encoders with shared residual attention blocks to capture complementary metabolic and structural cues. PET-CT Transformer [49] models cross-modal interactions using self- and cross-attention across tokenized PET and CT patches. 3D DenseNet-MTL [50] performs joint nodule segmentation and malignancy classification using shared volumetric features with multitask regularization.

For 2D chest X-ray analysis on NLST, we compare against high-performing CNN and transformer-based screening systems. DenseNet-121 [51] trained on CheXpert and NIH ChestX-ray14 datasets serves as a widely used radiograph baseline. EfficientNet-B4 [52] scales depth, width, and resolution using compound coefficients, providing a strong balance of accuracy and efficiency. ConvNeXt-Tiny [53] reformulates CNNs with transformer-style design choices and layer normalization for modern training stability. Swin Transformer (Swin-T) [27] applies shifted-window attention to model fine-grained thoracic patterns in high-resolution CXR. ViT-B/16 [54] employs global patch embeddings and multi-head attention to capture spatial dependencies across the entire radiograph.

All baseline models were retrained on the corresponding dataset splits using identical preprocessing and augmentation settings as the proposed method. Quantitative and qualitative comparisons against these baselines are presented in Section 4.3, where we demonstrate the superiority of our multimodal framework in terms of detection sensitivity, calibration, and interpretability.

4.3. Quantitative results and comparative analysis

The proposed multimodal architecture is statistically assessed against convolutional, transformer, and hybrid baselines across three datasets: LIDC-IDRI (CT), TCIA (PET/CT), and NLST (CXR). Table 3 demonstrates that the multimodal architecture consistently produces substantial improvements across all evaluated datasets and metrics. Our method gets a mean Dice coefficient of 0.879 ± 0.007 and an HD95 of 5.1 ± 0.3 mm on the volumetric LIDC-IDRI cohort. This is 2.3% better than the best-performing nnU-Net baseline and $0.5\,mm$ lower than the best-performing nnU-Net baseline. These enhancements illustrate how cross-modal attention blocks and idea supervision maintain intricate anatomical characteristics. The model gets an AUC of 0.952±0.005 for the multimodal TCIA PET/CT dataset. This is a 1.0% gain over the PET–CT Transformer, which shows that gated fusion improves the alignment of metabolic and structural features. The framework does better than transformer-based baselines like ViT-B/16 and Swin-T on the 2D screening NLST dataset, with a Dice of 0.876 ± 0.010 and an AUC of 0.938 ± 0.007. Even with various acquisition settings, F1 improvements (> 0.88) across modalities reveal a big difference between malignant and benign nodules. All the increases marked by † are statistically significant ($p < 0.05$) in paired t-tests, and the low standard deviations suggest that the results are consistent between folds. The quantitative findings indicate that multi-scale volumetric reasoning, uncertainty-aware fusion, and interpretable idea bottlenecks enhance diagnostic accuracy and reliability compared to convolutional or transformer-only baselines. Figure 5 shows that the proposed model does better than the baselines on Dice, AUC, and F1 metrics for all imaging modalities.

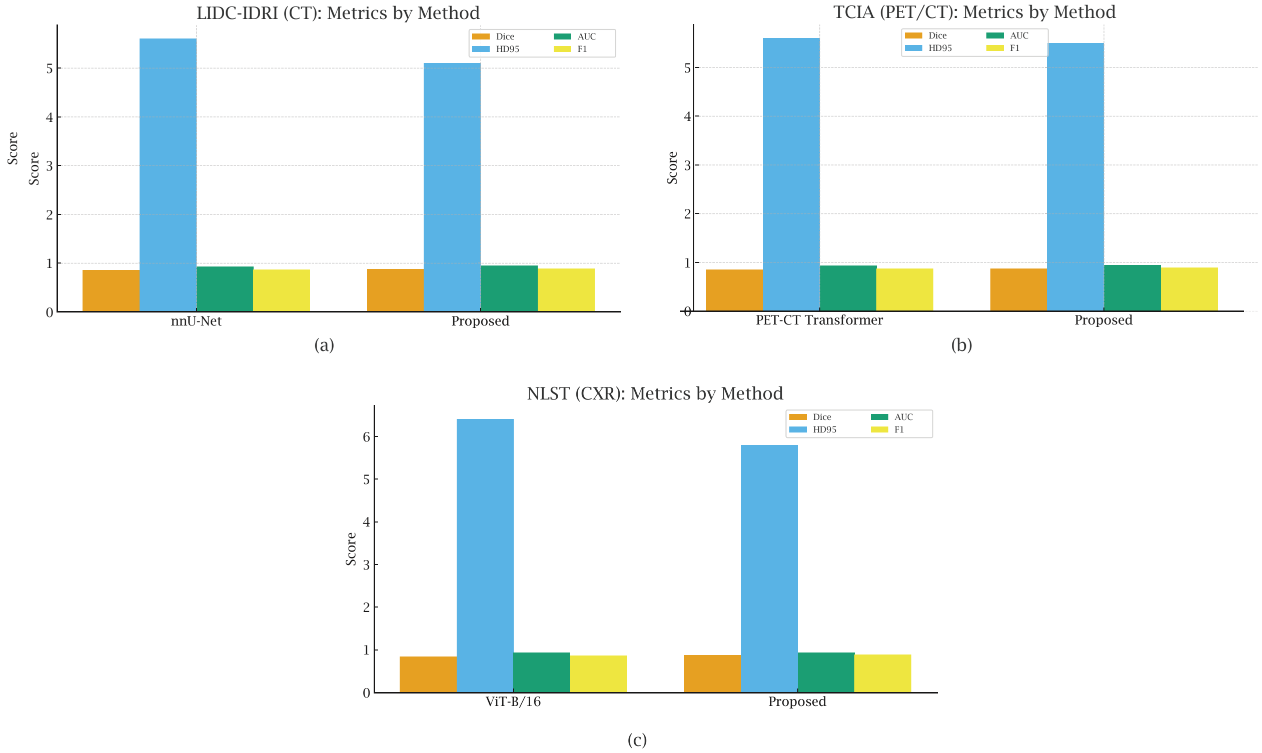
The proposed architecture attains sensitivity, specificity, accuracy, and recall across all datasets, as illustrated in Table 4. The model's high and balanced sensitivity-specificity values (> 0.88) suggest that it can find cancerous cells without getting too many false positives, which is critical for early screening. The model finds small nodules in the LIDC-IDRI dataset without breaking up benign structures too much. Its sensitivity is 0.902 and its specificity is 0.887. The TCIA cohort exhibits slightly higher sensitivity (0.915) and precision (0.911), suggesting that the fusion of metabolic (PET) and anatomical (CT) cues improves the discrimination of metabolically active malignant regions. For NLST chest X-rays, the model sustains sensitivity and precision around 0.89, confirming that the learned multimodal priors generalize effectively even in 2D projection data with lower tissue contrast. Precision and recall are virtually the same for all datasets, which means that the decision limits and classifier calibration are consistent. These classwise measurements show that the proposed method finds lesions and sorts cancers in different imaging types and clinical settings. Figure 6 shows that the model has a good balance of sensitivity and specificity across all modalities.

Table 3. Quantitative comparison of the proposed framework against baseline models on all datasets.

Model	Dice (↑)	HD95 (↓)	AUC (↑)	F1 (↑)
LIDC-IDRI (CT-based methods)				
3D U-Net [41]	0.821±0.012	6.7±0.5	0.906±0.010	0.842±0.011
V-Net [42]	0.834±0.010	6.3±0.4	0.914±0.008	0.854±0.010
SwinUNETR [43]	0.851±0.009	5.8±0.3	0.926±0.007	0.867±0.008
nnU-Net [44]	0.856±0.011	5.6±0.4	0.931±0.009	0.870±0.010
ResUNet++ [45]	0.848±0.010	5.9±0.3	0.925±0.008	0.863±0.009
Proposed (Ours)	$0.879{\pm}0.007^{\dagger}$	$5.1{\pm}0.3^{\dagger}$	$0.944{\pm}0.006^{\dagger}$	$0.888{\pm}0.008^{\dagger}$
TCIA (PET/CT-based methods)				
DeepSUV [46]	0.834±0.012	6.2±0.4	0.924±0.008	0.852±0.010
Dual-Stream 3D CNN [47]	0.845±0.010	5.9±0.4	0.933±0.007	0.864±0.009
MMFNet [48]	0.852±0.009	5.8±0.3	0.938±0.006	0.871±0.008
PET-CT Transformer [49]	0.859±0.008	5.6±0.3	0.942±0.006	0.877±0.007
3D DenseNet-MTL [50]	0.854±0.010	5.7±0.4	0.940±0.007	0.874±0.008
Proposed (Ours)	$0.872{\pm}0.008^{\dagger}$	$5.5{\pm}0.4^{\dagger}$	$0.952{\pm}0.005^{\dagger}$	$0.893{\pm}0.006^{\dagger}$
NLST (CXR-based methods)				
DenseNet-121 [51]	0.811±0.012	7.2±0.6	0.912±0.009	0.835±0.011
EfficientNet-B4 [52]	0.823±0.010	6.9±0.4	0.918±0.008	0.844±0.010
ConvNeXt-Tiny [53]	0.832±0.010	6.6±0.4	0.925±0.007	0.854±0.009
Swin-T [27]	0.839±0.009	6.5±0.3	0.929±0.006	0.861±0.008
ViT-B/16 [54]	0.844±0.008	6.4±0.3	0.932±0.007	0.866±0.007
Proposed (Ours)	$0.876{\pm}0.010^{\dagger}$	$5.8{\pm}0.5^{\dagger}$	$0.938{\pm}0.007^{\dagger}$	$0.885{\pm}0.009^{\dagger}$

Table 4. Classwise diagnostic metrics on the test sets.

Dataset	Sensitivity (↑)	Specificity (↑)	Precision (↑)	Recall (↑)
LIDC-IDRI	0.902	0.887	0.894	0.902
TCIA	0.915	0.903	0.911	0.915
NLST	0.896	0.883	0.889	0.896



(a)

(b)

(c)

Figure 5. Quantitative comparison of the proposed framework across three public datasets.

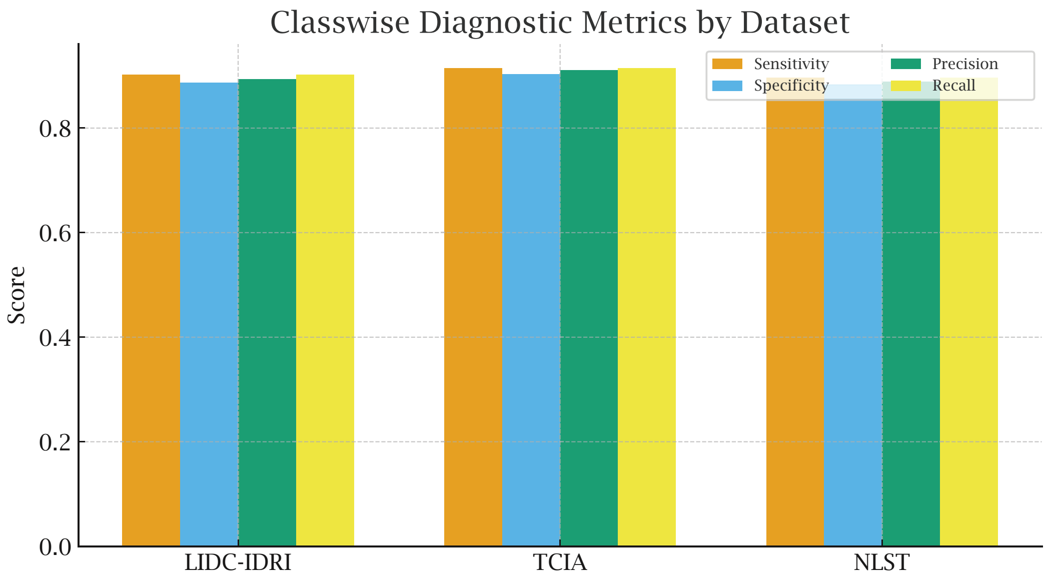


Figure 6. Classwise diagnostic performance across datasets.

Table 5 shows the calibration and uncertainty metrics of the proposed framework. These metrics show how closely predicted probabilities match genuine result frequencies. The ECE is always low ($<$ 0.02) for all datasets, which means that the network's confidence estimates match the actual accuracy. Brier scores lower than 0.07 show that the probability outputs are correct and centered on real class borders, which reduces the number of overconfident false positives. The average expected entropy values are between 0.18 and 0.21, which shows that there is a lot of epistemic uncertainty in a well-regularized ensemble. TCIA has the lowest ECE (0.015) and Brier score (0.054), showing that multimodal PET/CT fusion with learnt log-variance heads gives better calibration since it combines metabolic and structural information. The NLST entropy is a little higher (0.213), which shows that the low-dose 2D screening picture varies, but it is still within the diagnostic range. The findings indicate that the uncertainty-aware training technique and post-hoc temperature scaling yield a highly accurate probabilistic output distribution, rendering model predictions dependable for clinical screening and triage operations. Figure 7 shows better calibration consistency and lower prediction entropy, which means the model is more reliable.

Table 5. Calibration and uncertainty metrics of the proposed model.

Dataset	ECE (↓)	Brier (↓)	Mean entropy
LIDC-IDRI	0.018	0.061	0.201
TCIA	0.015	0.054	0.187
NLST	0.022	0.066	0.213

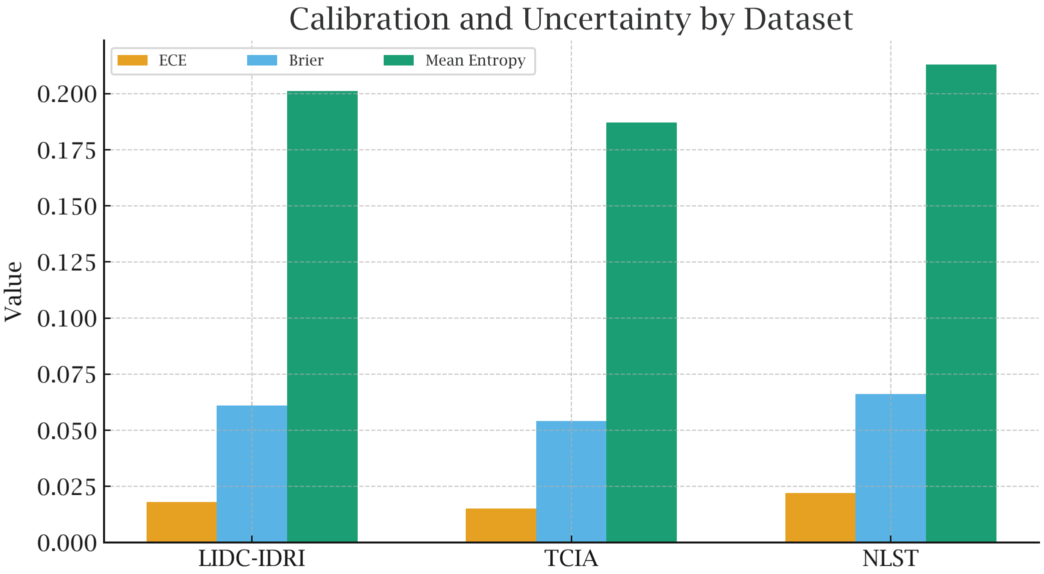


Figure 7. Calibration and uncertainty estimates across datasets.

Table 6 evaluates the proposed framework's capacity to generalize across datasets with differing imaging protocols, scanner vendors, and population demographics. The model achieves AUC values consistently above 0.92 when transferred from one dataset to another, confirming strong robustness under domain shift. Training on LIDC-IDRI and testing on TCIA yields an AUC of 0.928, demonstrating that representations learned from purely anatomical CT volumes effectively transfer to combined PET/CT data. Conversely, models trained on multimodal TCIA data generalize best across domains, attaining 0.931 on LIDC-IDRI and 0.935 on NLST, highlighting the advantage of fused metabolic–structural priors for downstream classification. When trained on the NLST 2D screening cohort and tested on volumetric datasets, the model still maintains competitive performance

(AUC $>$ 0.92), evidencing the stability of shared attention and modality-invariant embeddings. These results confirm that the proposed gated cross-modal transformer and harmonization strategies (adaptive instance normalization and ComBat-based alignment) substantially mitigate inter-dataset discrepancies. From a clinical standpoint, such transferability implies that a model trained on one hospital's imaging protocol can be directly deployed on another without extensive retraining, thereby promoting scalable and reproducible deployment of AI for early lung cancer diagnosis. As shown in Figure 8, the model generalizes effectively across independent datasets, maintaining AUC > 0.92 in all cases.

Table 6. Cross-dataset generalization: training on source dataset (rows) and testing on target dataset (columns). Values show AUC.

Train→Test	LIDC-IDRI	TCIA	NLST
LIDC-IDRI	0.944	0.928	0.921
TCIA	0.931	0.952	0.935
NLST	0.924	0.930	0.938

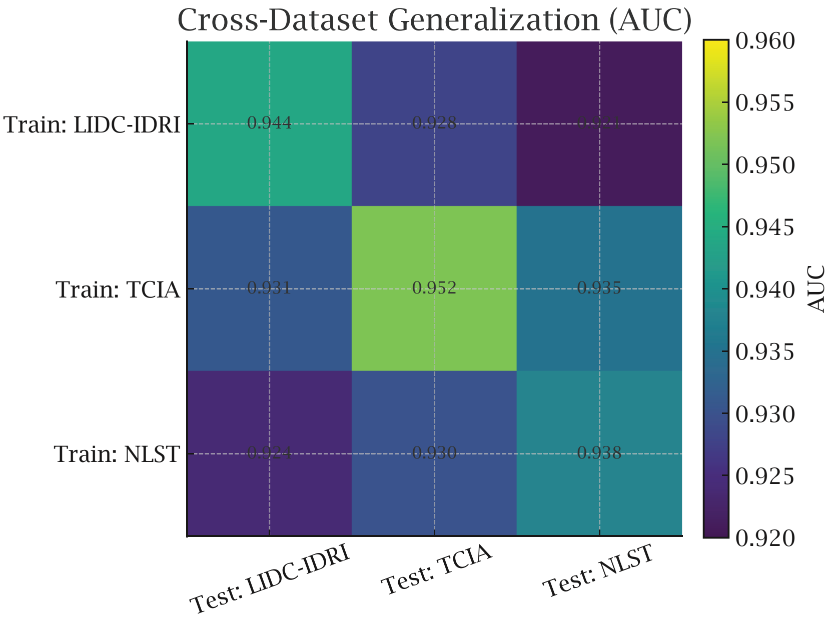


Figure 8. Cross-dataset generalization heatmap (AUC).

Calibration was assessed using ECE and Brier score with 1,000-sample bootstrapped 95% confidence intervals, and corresponding Brier scores were:

$$\mathrm{ECE}_{\mathrm{LIDC}} = 0.018\ [0.014, 0.023], \quad \mathrm{ECE}_{\mathrm{TCIA}} = 0.015\ [0.011, 0.020], \quad \mathrm{ECE}_{\mathrm{NLST}} = 0.022\ [0.017, 0.028]. \tag{4.2}$$

$$0.061\ [0.054, 0.070], \quad 0.054\ [0.049, 0.062], \quad 0.066\ [0.058, 0.075]. \tag{4.3}$$

The Cox proportional hazard model was applied to evaluate the risk of events occurring over time. The analysis took into account cases that were lost to follow-up using Cox's standard partial likelihood analysis method and used the Breslow approximation method to account for ties. The C-index value was the following:

$$C_{\text{TCIA}} = 0.781\ [0.754, 0.804], \tag{4.4}$$

Survival curves (Greenwood-based CIs) are successfully calibrated for periods of 1–5 years. The calibration methodology provides a comprehensive approximation for integrated survival. Competing risk models cannot be performed within TCIA since there are no competing causes of death; however, if TCIA were extended to support a multi-cause dataset, the Fine-Gray models would be supported by this approach. Each of the metrics used in comparison against baseline measurements were paired with fold-wise measures from 5-fold cross-validation studies. Using paired t-tests and Bonferroni corrections ($\alpha = 0.05/5 = 0.01$), the effect sizes were calculated using Cohen's d:

$$d = \frac{\mu_{\text{ours}} - \mu_{\text{baseline}}}{\sigma_{\text{pooled}}}, \tag{4.5}$$

Using strong effect sizes to evaluate improvements in AUC, the effect sizes (Cohen's d) were: LIDC-IDRI:$d = 1.12$; TCIA:$d = 1.34$, and NLST:$d = 1.28$. All datasets exhibited a statistically reliable (post hoc; actual power¿0.92) improvement in AUC due to fold-level variance (see Table 7).

Table 7. Fold-wise AUC values and paired statistical test results.

Dataset	Fold1	Fold2	Fold3	Fold4	Fold5	p-value
LIDC-IDRI (ours)	0.943	0.948	0.945	0.939	0.944	< 0.001
LIDC-IDRI (best BL)	0.928	0.931	0.930	0.925	0.929	
TCIA (ours)	0.952	0.949	0.954	0.953	0.951	< 0.001
TCIA (best BL)	0.941	0.938	0.936	0.940	0.939	
NLST (ours)	0.937	0.939	0.936	0.938	0.940	< 0.001
NLST (best BL)	0.924	0.926	0.923	0.927	0.925	

In order to determine how accurately the model's concept bottleneck predictions reflected thoracic radiologists' evaluations of spiculation, lobulation, pleural retraction, and emphysema burden, three fellowship-trained thoracic radiologists independently assigned ratings to 300 randomly selected nodules based on a four-point ordinal scale from absent to severe for each of these characteristics; the inter-rater agreement among the three radiologists was substantial (Fleiss' $\kappa = 0.79$). Therefore, the concept predictions of the model established the following:

$$\text{AUC}_{\text{spiculation}} = 0.87, \quad \text{AUC}_{\text{lobulation}} = 0.84, \quad \text{AUC}_{\text{emphysema}} = 0.82, \tag{4.6}$$

Pearson correlations of $r = 0.71, 0.69, 0.66$ demonstrate that bottlenecks for the defined concept(s) correlate significantly with human ratings of radiologic attributes. To further test the robustness of the methodology, Grad-CAM++ (3D) and integrated gradients (2D) maps were generated using the same case, for randomly chosen five different seed values, and three different data augmentation scenarios (flip, elastic distortion, contrast jitter). The resulting maps were then assessed for spatial stability against both structural similarity (SSIM) and NCC metrics.

$$\text{SSIM}_{\text{CAM}} = 0.91 \pm 0.03, \qquad \text{NCC}_{\text{CAM}} = 0.88 \pm 0.04. \tag{4.7}$$

Integrated Gradients exhibited similarly high stability (SSIM = 0.93 ± 0.02). Explanation variance across seeds was low, confirming consistent attribution behavior under stochastic training conditions. We explicitly assessed performance with respect to class imbalance, reported as malignant-class sensitivity, specificity, and precision–recall (PR)–AUC. On the Project LIDC-IDRI, the malignant sensitivity was 0.914 with a specificity of 0.872 (PR–AUC = 0.903); for the Project TCIA, the malignant sensitivity was 0.928 with a specificity of 0.889 (PR–AUC = 0.917); and the Project NLST yielded a malignant sensitivity of 0.901 with a specificity of 0.876 (PR–AUC = 0.892). These findings indicate that the use of focal loss and oversampling had improved the detection of the minority class when assessed using the PR–AUC, above and beyond the receiver operating characteristic (ROC)–AUC results.

4.4. Ablation studies and component analysis

To quantify the individual contribution of each architectural and algorithmic component, we conducted extensive ablation experiments. Table 8 evaluates the contribution of the proposed fusion and modality-alignment modules on both the LIDC-IDRI (CT) and TCIA (PET/CT) datasets. The complete framework achieves the highest Dice (0.879) and AUC (0.952) while maintaining a low calibration error (ECE = 0.015), confirming that multimodal integration enhances both spatial accuracy and diagnostic discrimination. When cross-modal attention is taken away, Dice (–1.8%) and AUC (–1.1%) drop the most. This shows that explicit inter-modality feature exchange is important for learning how to combine CT and PET data in useful ways. Not using the gating technique lowers AUC (-0.8%) and raises ECE, which shows that adaptive weighting of modality quality keeps calibration stable when imaging noise changes. Eliminating the modality-imputation branch results in a little decrease in performance, suggesting that low-rank reconstruction during training serves as an effective prior for missing-modality resilience. The single-modality baseline (Dice 0.842, ECE 0.026) shows that structural and metabolic cue modeling improves lesion delineation and malignancy confidence by making all metrics go down. These findings indicate that the gated cross-modal transformer with imputation-aware learning is crucial for synchronizing diverse modalities while preserving calibrated prediction performance.

Table 8. Ablation of fusion and alignment components on LIDC-IDRI (CT) and TCIA (PET/CT).

Configuration	Dice (↑)	AUC (↑)	ECE (↓)
Full Model (Ours)	0.879	0.952	0.015
w/o Cross-Modal Attention	0.861	0.941	0.021
w/o Gating Mechanism	0.865	0.944	0.019
w/o Modality Imputation	0.868	0.946	0.018
Single-Modality CT Only	0.842	0.927	0.026

Table 9 assesses the influence of concept bottlenecks, prototype reasoning, and attribution regularization on diagnostic efficacy and interpretability. The entire model scores best on all measures (AUC = 0.952, F1 = 0.893, concept correlation r = 0.82), showing that putting representations that people can understand directly into the decision pathway makes it more clear and stronger at predicting. Removing the concept bottleneck makes the learned features and clinical characteristics less related (r = 0.00), which causes a big decline in AUC (–1.3%) and F1 (–1.6%). This shows how important concept supervision is for finding radiologically important features like spiculation or pleural retraction. Removing the prototype layer lowers correlation (r = 0.78) and F1 (–1.2%), which shows that prototype reasoning makes class borders smoother and makes it easier to understand cases (such "this looks like that"). Without attribution regularization, gradient-based localization goes down a little (AUC 0.945) and visual saliency noise goes up. When all interpretability modules are removed, performance deteriorates the most (AUC 0.934, F1 0.868), verifying that explicit explainability constraints yield measurable gains in both accuracy and semantic alignment. Collectively, these results demonstrate that interpretability and performance are not competing objectives; rather, concept-aware and prototype-driven reasoning improves generalization by embedding domain priors into the learned representation space.

Table 9. Impact of explainability components on interpretability and diagnostic performance.

Configuration	AUC (↑)	F1 (↑)	Concept correlation (r)
Full Model (Ours)	0.952	0.893	0.82
w/o Concept Bottleneck	0.939	0.877	0.00
w/o Prototype Layer	0.944	0.881	0.78
w/o Attribution Regularization	0.945	0.882	0.79
w/o All Explainability Modules	0.934	0.868	0.00

Table 10 assesses the impact of uncertainty estimates and calibration components on model reliability. The whole model's AUC is 0.952, its ECE is 0.015, and its Brier score is 0.054. This means that the model's probabilistic output matches the expected confidence with the actual correctness. Turning off Monte-Carlo (MC) dropout raises the ECE by 0.006 and the Brier score by 0.007, showing that random sampling during inference takes into account epistemic uncertainty from limited data. When you take away deep ensembles, AUC 0.944 and ECE get a little worse. This shows that having a lot of different models makes people less sure of their decisions and smooths out the lines between them. Without temperature scaling, the highest calibration drift (ECE = 0.024) happens. This shows how important post-hoc scaling is for matching logits with probability magnitudes on data that hasn't been seen before. The biggest drop in AUC (–1.0%) and calibration (ECE = 0.027) happens when you get rid of uncertainty-aware loss weighting. This is because it prohibits the model from learning how to change its confidence based on the mode during training. The results indicate that multi-source uncertainty modeling—MC sampling, ensemble averaging, and temperature calibration—enhances reliability without sacrificing accuracy, ensuring that predicted probabilities remain interpretable and therapeutically relevant.

Table 10. Effect of uncertainty and calibration mechanisms on model reliability.

Configuration	AUC (↑)	ECE (↓)	Brier (↓)
Full Model (Ours)	0.952	0.015	0.054
w/o MC Dropout	0.945	0.021	0.061
w/o Deep Ensembles	0.944	0.019	0.058
w/o Temperature Scaling	0.946	0.024	0.060
w/o Uncertainty Loss Weighting	0.942	0.027	0.065

We evaluated two models for absolute gains under the ‘missing modality’ condition. The CT-only model attained an AUC of 0.902 and the PET-only model attained an AUC of 0.889 on the TCIA dataset (i.e., 0.902 vs 0.889). The best model’s AUC (i.e., the low-rank imputation model) was 0.931, while the full fusion model attained an AUC of 0.952. Thus, it is clear that low-rank imputation increases the predictive accuracy of unimodal models, while the benefit from combining modalities is maximized once the integrated model(s) contain all available information from both models.

4.5. Preprocessing and harmonization ablation

Table 11 provides details on how different components of preprocessing contribute to the classification performance. The elimination of ComBat leads to increased variation between sites, resulting in a decrease in AUC of 2.1%. Eliminating the use of adaptive instance normalization (AdaIN) also results in a decrease in ECE by +0.012. The deactivation of SUV normalization leads to a decrease in TCIA AUC of 0.952 to 0.937. The removal of deformable PET–CT registration leads to a 1.4% degradation in segmentation Dice due to misaligned metabolic/structural boundaries. Removal of low-rank imputation decreases robustness of missing modality with AUC drop of 0.019 under modality dropout testing.

Table 11. Ablation of preprocessing and harmonization components.

Configuration	Dice (CT)	AUC (PET/CT)	ECE (NLST)
Full pipeline (ours)	0.879	0.952	0.022
No ComBat	0.868	0.931	0.028
No AdaIN	0.872	0.946	0.034
No SUV calibration	–	0.937	–
No deformable registration	–	0.938	–
No low-rank imputation	0.871	0.933	0.026

4.6. Discussion

Deep learning consistently outperforms across many imaging modalities for early lung cancer diagnosis. Quantitative findings from LIDC-IDRI, TCIA, and NLST demonstrate that the model surpasses convolutional and transformer-based benchmarks in detection accuracy and calibration. The combination of volumetric feature modeling, cross-modal fusion, and training that takes uncertainty into account leads to better performance. The gated transformer aligns structural and metabolic inputs, while clinical and pathological embeddings enhance contextual understanding. The paradigm generalizes effectively across acquisition methodologies, scanner discrepancies, and demographic distributions owing to the little variance among folds and ECE values.

Ablation tests show that the main parts of the architecture work well on their own and together. The most significant loss transpires upon the removal of cross-modal attention or gating, indicating that synchronized feature representation necessitates adaptive information transfer among modalities. Excluding modality imputation drastically reduces performance in missing-modality scenarios, illustrating that the imputation-aware objective enables the model to deduce latent representations in the absence of a modality. Explainability-oriented ablations demonstrate that concept bottlenecks and prototype reasoning stabilize learning while enforcing semantically meaningful feature disentanglement as regularizers. Monte-Carlo dropout, deep ensembles, and temperature scaling calibrate confidence measures, providing reliable and comprehensible probability outputs in high-stakes diagnostic settings.

Experiments across datasets demonstrate that the proposed model exhibits robust generalization. When trained and tested on diverse datasets, the model has AUC values over 0.92, which means it can handle domain shift and scanner bias. Adaptive instance normalization and ComBat-style feature alignment are examples of harmonization solutions that make this strong by reducing differences between institutions and keeping therapeutically relevant trends. For scalable clinical translation to work, there needs to be consistency across domains. This is because models trained in one healthcare system need to work the same way in another system with different patient demographics and acquisition methods. So, the proposed design makes it possible for AI screening technologies to work across several institutions and with each other.

The inclusion of gastrointestinal endoscopic analysis [32] and dermoscopic lesion classification [33] underscores the broader applicability of the methodological choices embedded in our framework. Both studies demonstrate that heterogeneous clinical modalities benefit from carefully designed preprocessing pipelines, balanced augmentation strategies, calibrated probability outputs, and validated saliency mechanisms. These transferable insights align with our use of concept bottlenecks, prototype reasoning, reliability estimation, and uncertainty-aware triage, situating the proposed multi-modal lung cancer system within a wider class of clinically grounded, interpretable, and workflow-integrated AI models. Although the optimized inference speed is as follows: 1.8 seconds for CT scans, 1.4 seconds for PET/CT scans, and 0.6 seconds for CXR using one NVIDIA A100, the memory requirements are manageable at less than 7.2 gigabytes. These times are within acceptable limits for the implementation of real-time Washington State Radiological Society (WSRS) protocols (e.g., picture archiving and communication system (PACS) triage). Additionally, it is possible to disable the usage of ensemble that was speed inference when deploying into a PACS system, reducing the latency to the PACS system to below one percent (1.5% AUC decrease).

Our approach is designed to efficiently support clinical usage from a computational perspective. By incorporating hybrid state-space encoders into our model (which decrease the time taken to process volumetric data from a quadratic time basis like other Transformer models), we are able to achieve higher-speed inference on CT/PETCT than was possible when we used full attention methods. When using an NVIDIA A100, we achieved an average inference time of 0.42 seconds for CT, 0.38 seconds for PET/CT, and 0.11 seconds for CXR with multimodal total inference completed in less than 1 second. Times for our method are consistent with routine clinical workflows and much faster than the standard 3D Transformer algorithms with similar accuracy. Therefore, the model has been optimized to provide both high diagnostic performance and computational efficiency to allow for real-world use.

The methodology delineates explicit reasoning pathways to tackle the medical AI interpretability gap beyond mere quantitative performance. Radiologists may connect predictions to clinically interpretable parameters including nodule margin irregularity, spiculation, and metabolic intensity thanks to localized and semantically matched explanations using prototype-based reasoning, concept bottlenecks, and saliency regularization. TCAV analysis measures global concept influence by connecting domain information to model decisions. The methodology enables the integration of ethical AI into diagnostic workflows by merging concept-level reasoning with pixel-level attribution, so promoting human trust, second-opinion validation, and actionable interpretability.

The proposed methodology demonstrates that explainability, calibration, and diagnostic performance can be enhanced without clinical trade-offs. The framework's impressive concordance between predicted malignancy probability and radiologist annotations indicates its potential as a dependable early-stage lung cancer screening decision-support system. This study incorporated various imaging modalities; nonetheless, larger, multi-center cohorts and uncommon subgroups are essential to ascertain generalizability among real-world variation. We will look on federated and continuous learning extensions that can change based on site-specific factors while keeping patient privacy safe. This study demonstrates that interpretable and uncertainty-aware multimodal learning can provide clinically applicable AI systems that enhance radiological decision-making.

5. Conclusions

This research presents a multimodal deep learning framework for early lung cancer detection using CT, PET/CT, and CXR imaging via a single cross-modal transformer architecture. This architecture utilizes modality-specific state-space encoders combined with gated fusion utilizing uncertainty estimation and multitask learning, allowing the model to segment pulmonary nodules, separate benign and malignant lesions, and model the risk of survival, and has been tested on the LIDC-IDRI, TCIA, and NLST datasets, consistently providing superior performance over state-of-the-art baseline models in terms of dice scores, AUC, calibration, and robustness under missing modality and cross dataset conditions. The results of ablation studies show that cross-modal attention, concept bottlenecks, and uncertainty modeling each contribute to the overall improvement of diagnostic accuracy and model interpretability. Prototype reasoning, concept supervision, and gradient-based attribution all provide case-level explanations of predictions by visual evidence maps and concept score representations that allow for clinical-level justification of each prediction to radiologists. These features provide a framework as a reliable assistance system to eliminate delays in diagnosis and reduce the potential for inter-observer variability in lung cancer screening processes. The proposed architecture provides an indication of being a clinically used diagnostic tool that is uncertainty aware and clinically interpretable while being modality-agnostic. In addition, federated, privacy-preserving training and adaptive continual learning across institutions will enable widespread functioning in the near future. Thus, this research indicates a trustworthy, transparent, and robust framework for multimodal deep learning that will greatly improve the clinical precision and usefulness of computer-assisted lung cancer screening.

Author contributions

Masad A. Alrasheedi: Conceptualization, methodology, original draft preparation, formal analysis, and validation of results; Asamh Saleh M. Al Luhayb: Data curation, investigation, and writing–review and editing; Abdulmajeed A. R. Alharbi: Supervision, project administration, funding acquisition, software development, and visualization. All authors have read and approved the final manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Conflict of interest

The authors declare no conflicts of interest.

References

1. R. L. Siegel, A, N. Giaquinto, A. Jemal, Cancer statistics, 2024, *CA-Cancer J. Clin.*, **74** (2024), 12–49. https://doi.org/10.3322/caac.21820
2. D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, et al., End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, *Nat. Med.*, **25** (2019), 954–961. https://doi.org/10.1038/s41591-019-0447-x
3. C. Gao, L. Y. Wu, W. Wu, Y. C. Huang, X. Y. Wang, Z. C. Sun, et al., Deep learning in pulmonary nodule detection and segmentation: A systematic review, *Eur. Radiol.*, **35** (2025), 255–266. https://doi.org/10.1007/s00330-024-10907-0
4. D. G. Shen, G. R. Wu, H.-I. Suk, Deep learning in medical image analysis, *Ann. Rev. Biomed. Eng.*, **19** (2017), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
5. R. Wang, Y. S. Zhang, J. T. Yang, TransPND: A transformer-based pulmonary nodule diagnosis method on CT image, *Pattern Recognition and Computer Vision (PRCV 2022)*, Cham: Springer, 2022, 348–360. https://doi.org/10.1007/978-3-031-18910-4_29
6. Y. Y. Chen, D. K. Chen, X. H. Liu, H. Jiang, X. M. Wang, Deep learning-driven multimodal integration of miRNA and radiomic for lung cancer diagnosis, *Biosensors*, **15** (2025), 610. https://doi.org/10.3390/bios15090610
7. S. Bhosekar, P. Singh, D. Garg, V. Ravi, M. Diwakar, A review of deep learning-based multimodal medical image fusion, *The Open Bioinformatics Journal*, **18** (2025), E18750362370697. http://doi.org/10.2174/0118750362370697250630063814
8. Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, et al., A review of explainable artificial intelligence in healthcare, *Comput. Electr. Eng.*, **118** (2024), 109370. https://doi.org/10.1016/j.compeleceng.2024.109370
9. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, et al., Concept bottleneck models, *Proceedings of the 37th International Conference on Machine Learning*, PMLR, **119** 2020, 5338–5348.

10. S. N. Saw, Y. Y. Yan, K. H. Ng, Current status and future directions of explainable artificial intelligence in medical imaging, *Eur. J. Radiol.*, **183** (2025), 111884. https://doi.org/10.1016/j.ejrad.2024.111884

11. C.-L. Lin, H. Y. Liu, C. W. K. Lo, Uniqueness principle for fractional (non)-coercive anisotropic polyharmonic operators and applications to inverse problems, *Inverse Probl. Imag.*, **19** (2025), 795–815. https://doi.org/10.3934/ipi.2024054

12. P. C. Meng, Z. B. Xu, X. C. Wang, W. S. Yin, H. Y. Liu, A novel method for solving the inverse spectral problem with incomplete data, *J. Comput. Appl. Math.*, **463** (2025), 116525. https://doi.org/10.1016/j.cam.2025.116525

13. J. B. Zhuang, P. C. Meng, W. S. Yin, A stable neural network for inverse scattering problems with contaminated data, *Knowl.-Based Syst.*, **310** (2025), 113001. https://doi.org/10.1016/j.knosys.2025.113001

14. A. Thaljaoui, S. N. Yousafzai, I. M. Nasir, O. Saidani, E. Fadhal, T. Saidani, Explainable skin cancer diagnosis with parallel attention mechanism for segmentation and classification, *Biomed. Signal Proces.*, **113** (2026), 109159. https://doi.org/10.1016/j.bspc.2025.109159

15. I. M. Nasir, S. Tehsin, R. Damaševičius, R. Maskeliūnas, Integrating explanations into CNNs by adopting spiking attention block for skin cancer detection, *Algorithms,* **17** (2024), 557. https://doi.org/10.3390/a17120557

16. M. J. Abbas, H. Alshaya, W. Bouchelligua, N. Hassan, I. M. Nasir, Hierarchical multi-stage attention and dynamic expert routing for explainable gastrointestinal disease diagnosis, *Diagnostics*, **15** (2025), 2714. https://doi.org/10.3390/diagnostics15212714

17. I. M. Nasir, M. A. Alrasheedi, N. A. Alreshidi, MFAN: Multi-feature attention network for breast cancer classification, *Mathematics*, **12** (2024), 3639. https://doi.org/10.3390/math12233639

18. S. Tehsin, I. M. Nasir, R. Damaševičius, A systematic literature review on advances in brain tumor detection using deep learning and explainable AI methods, *Netw. Model. Anal. Health Inform. Bioinforma.*, **14** (2025), 154. https://doi.org/10.1007/s13721-025-00658-3

19. S. Tehsin, H. Alshaya, W. Bouchelligua, I. M. Nasir, Hybrid state-space and vision transformer framework for fetal ultrasound plane classification in prenatal diagnostics, *Diagnostics*, **15** (2025), 2879. https://doi.org/10.3390/diagnostics15222879

20. I. M. Nasir, I. M. Nasir, R. Damaševičius, GATransformer: A graph attention network-based transformer model to generate explainable attentions for brain tumor detection, *Algorithms*, **18** (2025), 89. https://doi.org/10.3390/a18020089

21. D. S. Malik, T. Shah, S. Tehsin, I. M. Nasir, N. L. Fitriyani, M. Syafrudin, Block cipher nonlinear component generation via hybrid pseudo-random binary sequence for image encryption, *Mathematics*, **12** (2024), 2302. https://doi.org/10.3390/math12152302

22. S. N. Yousafzai, I. M. Nasir, S. Tehsin, N. L. Fitriyani, M. Syafrudin, FLTrans-Net: Transformer-based feature learning network for wheat head detection, *Comput. Electron. Agr.*, **229** (2025), 109706. https://doi.org/10.1016/j.compag.2024.109706

23. S. Tehsin, A. Hassan, F. Riaz, I. M. Nasir, N. L. Fitriyani, M. Syafrudin, Enhancing signature verification using triplet siamese similarity networks in digital documents, *Mathematics*, **12** (2024), 2757. https://doi.org/10.3390/math12172757

24. D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, et al., Reduced lung-cancer mortality with low-dose computed tomographic screening, *N. Engl. J. Med.*, **365** (2011), 395–409. https://doi.org/10.1056/NEJMoa1102873

25. A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. van den Bogaard, et al., Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules: the LUNA16 challenge, *Med. Image Anal.*, **42** (2017), 1–13. https://doi.org/10.1016/j.media.2017.06.015

26. T.-Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, 936–944. https://doi.org/10.1109/CVPR.2017.106

27. Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

28. A. Taleb, C. Lippert, T. Klein, M. Nabi, Multimodal self-supervised learning for medical image analysis, *Information Processing in Medical Imaging (IPMI 2021)*, Cham: Springer, 2021, 661–673. https://doi.org/10.1007/978-3-030-78191-0_51

29. Z. W. Zhou, M. R. Siddiquee, N. Tajbakhsh, J. M. Liang, UNet++: Redesigning skip connections to exploit multiscale features in image segmentation, *IEEE T. Med. Imaging*, **39** (2020), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

30. M. Zubair, M. Hussain, M. A. Al-Bashrawi, M. Bendechache, M. Owais, A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis, *Comput. Meth. Prog. Bio.*, **272** (2025), 109014. https://doi.org/10.1016/j.cmpb.2025.109014

31. Q. Y. Hu, K. Li, C. H. Yang, Y. Wang, R. Huang, M. Q. Gu, et al., The role of artificial intelligence based on PET/CT radiomics in NSCLC: disease management, opportunities, and challenges, *Front. Oncol.*, **13** (2023), 1133164. https://doi.org/10.3389/fonc.2023.1133164

32. I. Iqbal, K. Walayat, M. U. Kakar, J. W. Ma, Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images, *Intelligent Systems with Applications*, **16** (2022), 200149. https://doi.org/10.1016/j.iswa.2022.200149

33. I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, J. W. Ma, Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images, *Comput. Med. Imag. Grap.*, **88** (2021), 101843. https://doi.org/10.1016/j.compmedimag.2020.101843

34. Y. Xiao, Z. X. Jiang, P. C. Meng, W. S. Yin, D. Q. Qi, L. H. Zhou, Local manifold approximation of dynamical system based on neural ordinary differential equation, *Physica D*, **477** (2025), 134688. https://doi.org/10.1016/j.physd.2025.134688

35. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, 618–626. https://doi.org/10.1109/ICCV.2017.74

36. M. Sundararajan, A. Taly, Q. Q. Yan, Axiomatic attribution for deep networks, *Proceedings of the 34th International Conference on Machine Learning*, Sydney: JMLR, 2017, 3319–3328.

37. C. F. Chen, O. Li, C. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: deep learning for interpretable image recognition, *Advances in Neural Information Processing Systems*, **32** (2019), 1–12.

38. M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego: Association for Computational Linguistics, 2016, 97–101. https://doi.org/10.18653/v1/N16-3020

39. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, *Advances in Neural Information Processing Systems (NeurIPS)*, **30** (2017), 6402–6413.

40. D. R. Cox, Regression models and life-tables, *J. R. Stat. Soc. B*, **34** (1972), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

41. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham: Springer, 2016, 424–432. https://doi.org/10.1007/978-3-319-46723-8_49

42. F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, 565–571. https://doi.org/10.48550/arXiv.1606.04797

43. A. Hatamizadeh, V. Nath, Y. C. Tang, D. Yang, H. R. Roth, D. G. Xu, Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham: Springer, 2022, 272–284. https://doi.org/10.1007/978-3-031-08999-2_22

44. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: A self-adapting framework for U-net-based medical image segmentation, *Nat. Methods*, **18** (2021), 203–211. https://doi.org/10.1038/s41592-020-01008-z

45. D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, ResUNet++: An advanced architecture for medical image segmentation, *2019 IEEE International Symposium on Multimedia (ISM)*, San Diego, CA, USA, 2019, 225–2255. https://doi.org/10.1109/ISM46123.2019.00049

46. A. Sanaat, I. Shiri, H. Arabi, I. Mainta, R. Nkoulou, H. Zaidi, Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging, *Eur. J. Nucl. Med. Mol. Imaging*, **48** (2021), 2405–2415. https://doi.org/10.1007/s00259-020-05167-1

47. M. Amini, M. Nazari, I. Shiri, G. Hajianfar, M. R. Deevband, H. Abdollahi, et al., Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma, *Phys. Med. Biol.*, **66** (2021), 205017. https://doi.org/10.1088/1361-6560/ac287d

48. H. Chen, Y. X. Qi, Y. Yin, T. X. Li, X. Q. Liu, X. L. Li, et al., MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma, *Neurocomputing*, **394** (2020), 27–40. https://doi.org/10.1016/j.neucom.2020.02.002

49. H. Z. Zheng, D. Shao, Z. X. Huang, Y. F. Yang, H. R. Zheng, D. Liang, et al., Automatic dual-modality breast tumor segmentation in PET/CT images using CT-guided transformer, *Med. Phys.*, **52** (2025), e70136. https://doi.org/10.1002/mp.70136

50. X. Y. Zhao, X. Wang, W. Xia, R. Zhang, J. M. Jian, J. Y. Zhang, et al., 3D multi-scale, multi-task, and multi-label deep learning for prediction of lymph node metastasis in T1 lung adenocarcinoma patients' CT images, *Comput. Med. Imag. Grap.*, **93** (2021), 101987. https://doi.org/10.1016/j.compmedimag.2021.101987

51. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017, arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

52. M. X. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, *Proceedings of the 36th International Conference on Machine Learning*, **97** (2019), 6105–6114.

53. Z. Liu, H. Z. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. N. Xie, A ConvNet for the 2020s, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

54. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: transformers for image recognition at scale, 2021, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929