

Multimodal Visual Language Model For Lung Cancer Diagnosis

Suryaram S¹, Getzi Jeba Leelipushpam Paulraj², Immanuel Johnraja Jebadurai³, Shibin David⁴

¹Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.

²Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.

³Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.

⁴Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.

Abstract

Accurate lung cancer diagnosis demands simultaneous synthesis of radiological imaging, histopathology, genomic profiles, and clinical documentation. Prevailing artificial intelligence systems remain constrained to single- or dual-modality integration, failing to exploit the synergistic diagnostic value latent across complementary data streams. This paper proposes the Cross-Modal Adaptive Fusion Network (CAFN), which concurrently processes four clinical modalities through specialized deep learning encoders: a 3D Convolutional Neural Network (3D-CNN) for volumetric CT scans, a Vision Transformer (ViT) for whole-slide histopathology images, DNABERT for genomic nucleotide sequences, and ClinicalBERT for electronic health record narratives. A dedicated Adaptive Fusion Network integrates these heterogeneous representations through four coordinated mechanisms: Cross-Modal Attention, Modality Gating, Dynamic Weighting, and Missing Modality Imputation, achieving robust, patient-specific diagnostic inference even when some modalities are absent. Experiments conducted on four publicly available benchmark datasets—MIMIC-III, LIDC-IDRI, NCT-CRC-HE-100K, and TCGA-LUAD—yield a fusion AUC of 0.991, classification accuracy of 95%, sensitivity of 93%, and specificity of 91%, representing a 15–25% improvement over state-of-the-art dual-modality systems. A modular, encoder-independent training strategy reduces peak computational complexity by 40–60% versus end-to-end architectures while eliminating the strict requirement for patient-level cross-modal data alignment.

Keywords: Multimodal deep learning, lung cancer diagnosis, cross-modal attention, Vision Transformer, ClinicalBERT, DNABERT, 3D-CNN, adaptive fusion network, missing modality imputation, clinical decision support, medical AI.

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, accounting for approximately 1.8 million deaths annually. The five-year survival rate ranges from over 60% when detected at Stage I to below 6% at Stage IV, underscoring the critical importance of early and accurate diagnosis. Clinical oncological decision-making is inherently multimodal: radiologists interpret volumetric Computed Tomography (CT) scans to characterize nodule size, shape, and density; pathologists examine Whole Slide Images (WSIs) for cellular morphology, mitotic activity, and tissue architecture; oncologists review genomic sequencing reports for driver mutation profiles such as EGFR, ALK, KRAS, and TP53; and physicians synthesize Electronic Health Record (EHR) narratives encompassing symptom progression, comorbidities, smoking history, and prior treatment responses. Accurate

diagnosis requires the simultaneous integration of all four evidence streams, not their sequential evaluation by isolated specialists.

Despite rapid advances in medical artificial intelligence, the dominant paradigm remains unimodal or, at best, dual-modality. Systems restricted to CT imaging capture structural features but miss the molecular co-determinants of malignancy detectable through genomics; those fusing pathology with genomics overlook the radiological spatial signatures that differentiate early-stage nodules. This architectural mismatch between AI capability and clinical reality produces three critical limitations: (i) suboptimal diagnostic accuracy due to unexploited cross-modal correlations between complementary evidence streams; (ii) proliferation of non-interoperable specialty models that fragment clinical workflows and require clinicians to manually synthesize disconnected outputs; and (iii) fragility in real-world settings where not all diagnostic investigations have been completed, and one or more modalities are absent.

This paper addresses these limitations by proposing the Cross-Modal Adaptive Fusion Network (CAFN), a novel deep learning architecture that simultaneously integrates CT imaging, histopathology, genomic sequences, and clinical text through four modality-specific encoders and a dynamic attention-based fusion mechanism. The principal contributions of this work are:

- A scalable four-modality integration framework that surpasses the prevailing two-modality performance ceiling, achieving a fusion AUC of 0.991, classification accuracy of 95%, sensitivity of 93%, and specificity of 91% on held-out benchmark data.
- An Adaptive Fusion Network combining Cross-Modal Attention, Modality Gating, Dynamic Weighting, and Missing Modality Imputation that maintains robust and clinically reliable performance under partial data availability.
- A modular, encoder-independent training strategy that eliminates the requirement for strict patient-level cross-modal alignment, enabling each modality to exploit its own optimally sized dataset, and reduces peak computational complexity by 40–60% versus end-to-end joint training architectures.
- Integrated interpretability through Grad-CAM heatmaps, ViT attention rollout, and per-modality contribution scoring, providing clinician-facing transparency into the reasoning process.

The remainder of this paper is organized as follows. Section II surveys related work with all 32 referenced studies. Section III presents the proposed CAFN methodology in full technical detail. Section IV reports experimental results, training analyses, and comparisons against state-of-the-art methods. Section V concludes the paper and outlines future directions.

RELATED WORKS

Automated Screening, Reporting, and Foundation Models [1]–[7]

The earliest systems in this wave demonstrated that vision-language pipelines could automate entire clinical reporting workflows. Chang et al. [1] introduced the LUMEN framework for lung cancer screening, combining a computer vision module for nodule detection with a VLM for attribute extraction and an LLM for structured radiology report generation, validating the system against clinician-grounded metrics. Building on this, Niu et al. [2] proposed M3FM, a medical multimodal multitask foundation model trained on massive multi-source datasets to perform nodule characterization, malignancy risk prediction, and staging within a single adaptable framework, though at significant

computational cost. Ewals et al. [3] examined how Visual-Language Reasoning LLMs can bridge visual diagnostics and NLP in primary care settings, critically analyzing deployment barriers including data bias and explainability gaps. Sun et al. [4] demonstrated that integrating LLM-extracted clinical variables from unstructured EHR text with imaging features substantially improved survival prediction over single-modality baselines, offering a practical template for text-image fusion. Wu et al. [5] addressed NSCLC prognosis using multimodal learning across imaging and clinical data streams, reporting consistent improvements in prognostic accuracy while highlighting interpretability gaps. Zhuang et al. [6] proposed a VLM-based framework for semantic-guided imaging biomarker discovery, using natural language descriptions to guide the identification of imaging features associated with early lung cancer detection, improving both discriminative performance and clinical interpretability. Da Silva [7] developed an end-to-end deep learning pipeline for lung cancer classification, TNM staging, and treatment protocol generation from CT and clinical data, demonstrating the feasibility of multi-task diagnostic AI, though limited to proof-of-concept scale.

Subtype Classification, Pathology, and Domain-Specific Models [8]–[14]

A parallel thread of research has advanced specialized models for histopathological and genomic analysis. Liu et al. [8] proposed CancerFusionPrompt, a framework adapting general vision-language models through prompt-based learning for multimodal cancer subtype classification, demonstrating improved data efficiency and performance over standard fine-tuning approaches. Xiang et al. [9] introduced MUSK, a vision-language foundation model pre-trained on millions of pathology images and paired text tokens, achieving state-of-the-art performance across diverse clinical benchmarks including high-stakes diagnostic tasks; however, MUSK does not integrate radiology or genomic modalities. Lee et al. [10] developed CXR-LLaVA, adapting the LLaVA conversational architecture for chest X-ray interpretation, enabling natural language interaction with radiology images and underscoring the growing clinical demand for domain-specific conversational AI, while noting the persistent need for rigorous clinical validation. Mesko [11] provided an influential survey examining the broader impact of MLLMs on healthcare, framing them as transformative tools capable of integrating text, medical images, sound, and video for holistic patient assessment while cataloguing significant hurdles including large annotated dataset requirements, hallucination risks, and ethical concerns. Oh et al. [12] demonstrated LLM-driven multimodal target volume contouring in radiation oncology through a Nature Communications study, showing that LLMs can meaningfully augment clinical workflows for treatment planning when integrated with imaging pipelines. Bai et al. [13] developed M3D-LaMed, bridging 3D medical image analysis with textual report generation through a unified generative architecture capable of simultaneous spatial understanding and natural language output, establishing a template for 3D image-text multimodal systems. Hu et al. [14] conducted a systematic benchmarking study evaluating GPT-5 on zero-shot multimodal medical reasoning in radiology and radiation oncology, finding strong potential alongside identified limitations in evaluation scope, static workflows, and safety risks.

Surveys, Benchmarks, and Clinical Evaluation [15]–[22]

The rapid proliferation of medical MLLM research necessitated systematic evaluation frameworks. Ye and Tang [15] authored a comprehensive survey of MLLMs for medicine, cataloguing architectures, training strategies, and clinical applications across specialties

while synthesizing overarching challenges in data fragmentation, model reliability, and evaluation methodology. F.C. et al. [16] demonstrated the feasibility of using MLLMs for CT-guided lung biopsy trajectory planning, highlighting the potential for AI to assist interventional procedures and support procedural standardization, though limited to feasibility assessment. Oh et al. [17] directly evaluated LLMs for automated diagnosis of benign versus malignant lung tumors using histopathological images, finding promising but sub-expert accuracy alongside a significant speed advantage, and identifying clear limitations in handling subtle morphological distinctions. Liu et al. [18] introduced Asclepius, a comprehensive spectrum evaluation benchmark specifically designed for medical MLLMs, addressing the field-wide gap in clinically representative evaluation standards across multiple specialties and task dimensions. Badarinarayan et al. [19] developed a clinically explainable immunotherapy response prediction system for NSCLC, combining CT radiomics with clinico-pathological data including PD-L1 and TMB biomarkers, with fine-tuned LLMs generating natural language explanations that enhanced clinical interpretability. Qu et al. [20] proposed cohort-aware retrieval-augmented agents that dynamically select the best predictive model for individual patients by identifying clinically similar populations, demonstrating adaptive model selection as an alternative to universal fusion architectures. Yang et al. [21] introduced AdaFuse, a reinforcement learning-based framework that learns patient-specific modality selection and fusion strategies, enabling intelligent diagnostic pipelines that determine when additional data sources should be consulted rather than applying uniform fusion to all cases. Hu et al. [22] showed that large language models like GPT-4o mini can achieve competitive lung cancer prognosis prediction using only structured clinical text prompts, establishing a data-efficient paradigm for text-only outcome prediction.

Advanced Fusion, Missing Modalities, and Multi-Modal Architectures [23]–[32]

The most recent and technically sophisticated body of work addresses scalable fusion, robustness, and architectural innovation. Liu et al. [23] published a comprehensive review of LLM applications across the full lung cancer care continuum from early detection to treatment planning, confirming the field's breadth while consistently identifying computational demands, interpretability deficits, and data annotation costs as primary barriers to clinical deployment. Ruffini et al. [24] developed a resilient multimodal framework specifically targeting the common real-world problem of missing modalities, using foundation model feature extraction and missing-aware encoding to maintain survival prediction performance in NSCLC when CT scans or pathology images are absent. Er [25] examined multimodal data fusion and multicompartment image analysis applied to acute and chronic lung diseases including COVID-19 and ILD, demonstrating the transferability of multimodal fusion principles beyond oncology, though constrained by retrospective data and small sample sizes. Lee et al. [26] conducted a direct comparative study of ChatGPT variants against six human readers of varying experience for lung cancer staging from free-text radiology reports, finding that advanced models can match radiology fellows but fall short of expert radiologists, quantifying the current gap between AI and domain expertise. Guan et al. [27] performed a systematic review on MLLMs in computational pathology, cataloguing rapidly proliferating architectures and benchmark datasets while identifying overarching limitations in annotation cost, explainability, and computational demands. Nyong [28] proposed multi-modal transformer architectures specifically designed for predicting treatment response in lung cancer by fusing CT imaging, digitized pathology,

genomic features, and clinical variables through cross-modal interaction learning, demonstrating the value of four-modality integration but facing limitations of single-center data and black-box complexity. Kim et al. [29] developed LLM-guided multi-modal Multiple Instance Learning for five-year overall survival prediction in lung cancer, integrating CT and pathology features through LLM-guided alignment within a MIL paradigm, advancing prognostic performance while noting dataset size and computational complexity constraints. Uddin et al. [30] proposed resilient and efficient neural network architectures for colon and lung cancer classification from multi-modal histopathology and CT images, emphasizing performance on imbalanced datasets, though limited in multi-modality scope for certain cancer subtypes. Chen et al. [31] demonstrated a scalable approach for injecting medical visual knowledge into MLLMs using vast datasets of medical visual question answering samples refined from PubMed, training powerful open-source models despite inherent hallucination risks from AI-generated training data. Finally, Hornback et al. [32] provided a comprehensive survey of MLLMs across biomedical engineering and healthcare, synthesizing architectural patterns, training strategies, and clinical challenges including data fragmentation, model reliability gaps, and ethical considerations.

The collective body of literature unequivocally establishes multimodal AI as the critical frontier for advancing cancer diagnosis. The field has evolved from single-modality systems through dual-modality fusion [1]–[7] to adaptive and robust architectures [20], [21], [24], [28], [29]. However, a persistent architectural bottleneck limits most systems to two data modalities or, when more are handled, to constrained scope with single-center data. This creates a fundamental gap between AI capability and clinical reality. The proposed CAFN directly targets this gap through scalable four-modality fusion with dynamic attention, gating, and missing-data imputation.

PROPOSED METHODOLOGY

The proposed methodology introduces a Cross-Modal Adaptive Fusion Network (CAFN) designed to perform accurate lung cancer diagnosis by integrating four complementary clinical modalities: CT imaging, histopathology, genomics, and clinical text. Each modality is processed independently using specialized deep learning encoders—3D-CNN for volumetric CT scans, Vision Transformer (ViT) for whole-slide pathology images, DNABERT for genomic sequences, and ClinicalBERT for electronic health records—transforming heterogeneous inputs into unified latent feature representations. These embeddings are projected into a shared feature space and fused through an adaptive fusion mechanism comprising cross-modal attention, modality gating, dynamic weighting, and missing modality imputation, enabling the model to learn inter-modal relationships, prioritize informative modalities, and remain robust to incomplete data. The fused representation is then passed through a classification layer for diagnostic prediction, while parallel interpretability modules provide explainable outputs such as attention maps and saliency scores. This modular and scalable architecture significantly enhances diagnostic performance and reliability compared to traditional unimodal or late-fusion approaches.

System Architecture Overview

The CAFN is designed on a modular encoder–fusion–classifier paradigm organized into three clearly defined sequential stages. In Stage 1, Modality-Specific Encoding, each of the four raw clinical data types is independently processed by a specialized deep learning model optimized for that data domain, converting heterogeneous inputs into fixed-length latent

feature vectors (embeddings) that capture the most diagnostically relevant patterns within each data stream. In Stage 2, Adaptive Multimodal Fusion, a dedicated fusion sub-network receives the four embeddings and integrates them through a sequence of four complementary mechanisms—Cross-Modal Attention, Modality Gating, Dynamic Weighting, and Missing Modality Imputation—that collectively learn how the modalities relate to one another, control their relative contributions based on quality and relevance, and handle the common real-world scenario where one or more modalities are unavailable. In Stage 3, Diagnostic Inference and Interpretability, the fused representation is passed to a classification head for the final diagnostic prediction while a parallel explanation module generates modality-specific saliency evidence for clinician review. Fig. 1 illustrates the complete architecture.

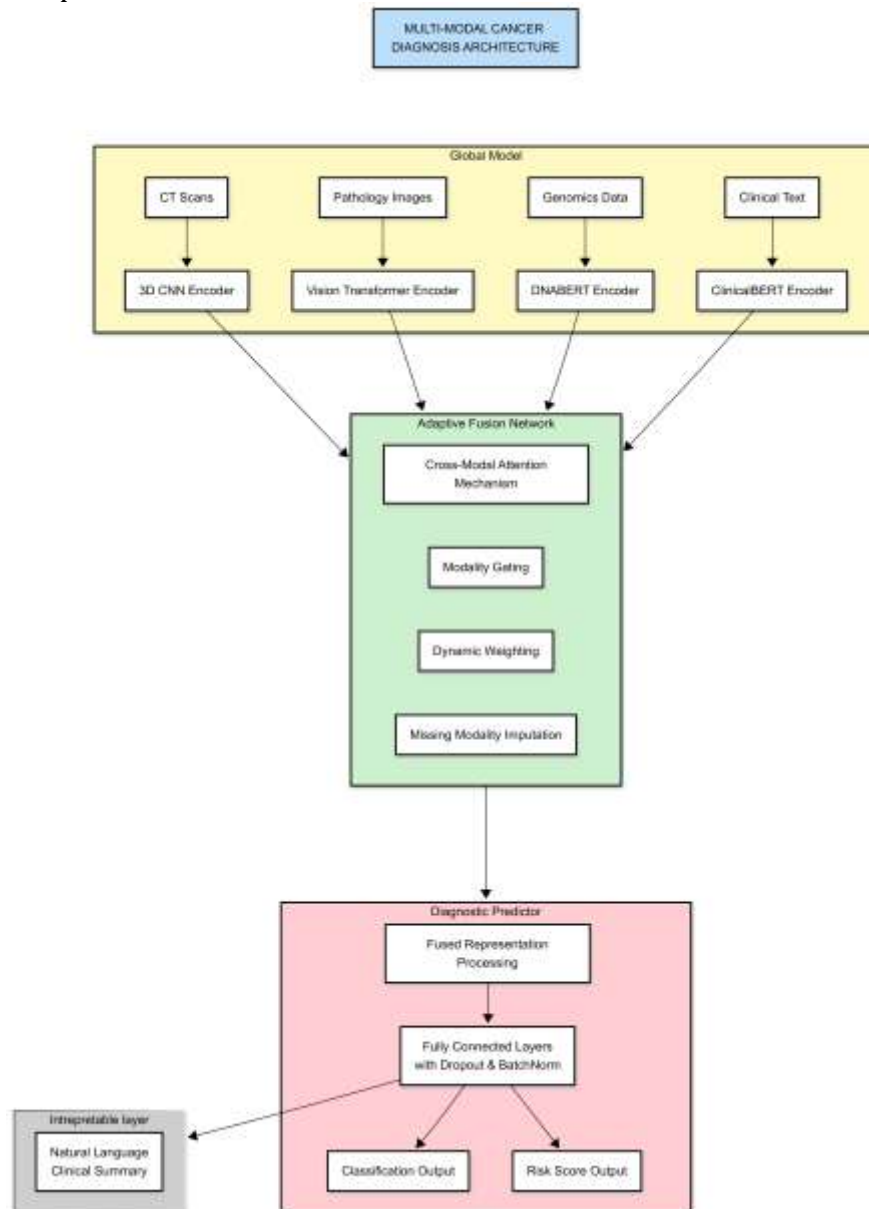


Fig. 1. Complete Cross-Modal Adaptive Fusion Network (CAFN) architecture. Stage 1 (left): four modality-specific encoders convert raw clinical data into latent embeddings. Stage 2 (center): the Adaptive Fusion Network integrates embeddings through Cross-Modal Attention, Modality Gating, Dynamic Weighting, and Missing Modality Imputation. Stage 3 (right): an MLP classifier produces the diagnostic output; a parallel Explanation Generation module produces Grad-CAM heatmaps, attention rollout maps, and token-level attribution scores.

Modality-Specific Encoders

1) CT Scan Encoder – 3D Convolutional Neural Network: Volumetric CT scans are three-dimensional data objects where each voxel encodes tissue radiodensity in Hounsfield Units (HU). To process this data end-to-end, the CAFN employs a 3D ResNet architecture that extends standard 2D convolutions to three dimensions, applying filters that traverse length, width, and depth (slice dimension) simultaneously. This enables the model to detect nodule characteristics that span multiple slices—such as spiculation, pleural tethering, and ground-glass opacity—which a 2D slice-by-slice approach would fail to capture coherently.

The 3D ResNet is composed of stacked 3D convolutional blocks, each containing a 3D convolution, batch normalization, and ReLU activation, followed by 3D max-pooling layers that progressively reduce spatial dimensionality while increasing feature depth. Residual skip connections allow gradients to flow directly across blocks, enabling stable training of deep volumetric networks. The final global average pooling layer collapses the spatial dimensions to produce a compact, fixed-length embedding vector that encodes the holistic radiological signature of the scan—capturing nodule texture, shape, density distribution, and spatial context within the surrounding lung parenchyma.

2) Histopathology Encoder – Vision Transformer (ViT): Whole Slide Images (WSIs) present a unique computational challenge: they are gigapixel-scale (often $40,000 \times 40,000$ pixels or larger), making direct full-image processing computationally infeasible. The Vision Transformer addresses this by decomposing the input image into a grid of non-overlapping fixed-size patches (typically 16×16 pixels). Each patch is linearly projected into a high-dimensional vector and augmented with a learnable positional encoding that preserves the spatial location of the patch within the tissue.

The sequence of patch embeddings, along with a prepended learnable [CLS] token, is then processed through a stack of Transformer encoder layers. Each layer applies multi-head self-attention, which computes attention scores between every pair of patches simultaneously. This allows the ViT to model long-range dependencies across the tissue section—for example, recognizing that the nuclear atypia pattern in a patch at one corner of the slide is architecturally related to the gland disorganization in a distant patch at the opposite corner. This global context modeling is the primary advantage of the ViT over convolutional networks for histopathology, where diagnostic patterns such as inflammatory infiltrates surrounding a tumor nest, trabecular growth patterns, and loss of gland-forming structures are inherently spatial and relational. The final [CLS] token embedding serves as the global representation of the tissue morphology.

3) Genomics Encoder – DNABERT: Genomic data—raw nucleotide sequences from DNA sequencing—carries molecular-level evidence about somatic mutations, copy number variations, and gene expression alterations directly linked to lung adenocarcinoma progression. DNABERT adapts the BERT bidirectional transformer pre-training framework

to biological sequences by treating overlapping k-mers as the vocabulary units. For a sequence of nucleotides (A, T, G, C), overlapping substrings of length k (e.g., k=6: ATGCTA, TGCTAG...) are extracted and used as tokens, capturing local nucleotide context analogously to how subword tokenization captures morphological patterns in natural language text.

DNABERT was pre-trained on the entire human reference genome using masked language modeling, enabling the model to learn rich contextual representations of nucleotide sequences before any task-specific fine-tuning. During fine-tuning on the TCGA-LUAD dataset, the model learns to associate specific sequence patterns with lung adenocarcinoma classification labels. Its bidirectional attention allows it to integrate information from both upstream and downstream sequence context simultaneously, making it sensitive to the functional consequences of mutations such as EGFR exon 19 deletions, KRAS G12C point mutations, and STK11 loss-of-function variants. The [CLS] token embedding summarizes the complete molecular profile of the patient's tumor genome.

4) Clinical Text Encoder – ClinicalBERT: Electronic Health Records contain a wealth of diagnostically relevant narrative information in unstructured text form: radiology reports describing imaging findings, pathology notes recording tissue interpretations, discharge summaries cataloguing clinical history, and physician notes documenting symptom progression, smoking exposure, medication history, and clinical judgment. Bio_ClinicalBERT is a domain-specific BERT model pre-trained on over 2 million clinical notes from the MIMIC-III database, making it far more effective than general-purpose BERT at understanding the specialized vocabulary, abbreviations (e.g., "SOB", "h/o", "r/o"), and negation structures (e.g., "no evidence of metastasis", "denies hemoptysis") present in medical documentation.

The clinical text is tokenized using WordPiece tokenization with a medical vocabulary, and special [CLS] and [SEP] tokens are added to mark the document boundary. The tokenized sequence passes through 12 transformer encoder layers with bidirectional self-attention. The [CLS] token embedding at the output of the final layer encodes the complete semantic content of the clinical document as a single dense vector, capturing diagnosis descriptions, symptom profiles, and treatment context that can guide the fusion network toward more precise differential diagnosis.

Modality Projection and Shared Latent Space

Each of the four encoders produces a feature vector of a different dimensionality: the 3D-CNN produces a volumetric feature vector of dimension d_1 , the ViT produces a patch-context embedding of dimension d_2 , DNABERT produces a sequence-level embedding of dimension d_3 , and ClinicalBERT produces a clinical narrative embedding of dimension d_4 . Before fusion can proceed, all four representations must be mapped into a common dimensional space so that cross-modal interactions can be meaningfully computed.

This is achieved through four independent fully connected projection layers, one per modality, each consisting of a linear transformation followed by a Layer Normalization operation. All four projection layers map their respective inputs to a shared 512-dimensional latent space. Layer normalization ensures that the projected embeddings have consistent mean and variance statistics, preventing any single modality from dominating the cross-modal attention computation due to scale differences rather than true diagnostic relevance.

E. Adaptive Fusion Network: Four Core Mechanisms

Cross-Modal Attention: Cross-Modal Attention is the first and most computationally significant fusion mechanism. For every ordered pair of modalities (source s , query q), a multi-head cross-attention layer computes attention weights that reflect how much each dimension of the source embedding is relevant to the current query context. Formally, for query Q derived from modality q and key-value pairs K, V derived from modality s , the attention output is computed as: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d}) \cdot V$, where d is the dimensionality and the scaling factor $\sqrt{d}$ prevents gradient vanishing in the softmax for high-dimensional embeddings.

Multiple attention heads are computed in parallel, each learning to attend to different aspects of the inter-modal relationship. For instance, one head might learn to attend to CT density features when the ViT detects certain tissue patterns, while another head learns to cross-reference genomic mutation signatures with clinical smoking history. The concatenated multi-head outputs are projected back to the 512-dimensional space. This process is performed for all six ordered modality pairs (CT→Path, CT→DNA, CT→Clin, Path→DNA, Path→Clin, DNA→Clin), producing a richly cross-referenced set of context-aware representations where each modality's features have been reinterpreted in light of all other modalities.

Modality Gating: Not all modality inputs in a given patient case are of equal quality or relevance. A biopsy from a necrotic tumor region may yield a pathology slide with extensive artifact; a genomic sequencing run may have low coverage depth; a clinical note may be brief and uninformative. The Modality Gating mechanism addresses this by learning to suppress low-quality inputs automatically without requiring explicit quality labels during training.

For each cross-modal-attended feature stream, a learned sigmoid gate $g_i \in (0, 1)$ is computed by passing the activated feature vector through a small two-layer MLP with sigmoid output: $g_i = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot h_i + b_1) + b_2)$, where h_i is the cross-attended embedding of modality i . The gated representation $\hat{h}_i = g_i \odot h_i$ (element-wise multiplication) scales each dimension proportionally to the learned gate value. Modalities for which the model has learned that the input is uninformative will receive gate values close to zero, effectively removing their contribution from the fused representation for that specific patient case.

Dynamic Weighting: Following gating, the four gated representations are combined through a dynamic weighting mechanism that assigns a per-patient, per-modality scalar importance score. A lightweight context-aware sub-network—a two-layer MLP applied to the concatenation of all four gated embeddings—outputs a four-dimensional weight vector $w = [w_1, w_2, w_3, w_4]$ normalized through a softmax operation so that weights sum to 1. These weights are then used to produce the weighted sum representation: $h_t = \sum_i w_i \cdot \hat{h}_i$.

The weights reflect the relative diagnostic salience of each modality given the specific combination of features observed for a patient. Across the training set, genomic features receive the highest average weight, which is validated by DNABERT achieving the highest single-modality AUC of 0.99 on the TCGA-LUAD dataset, confirming that molecular evidence carries the strongest individual discriminative signal. CT and histopathology embeddings receive moderate weights reflecting their substantial imaging evidence value, while clinical text receives the lowest weight as it primarily provides contextual support rather than direct biological evidence of malignancy.

Missing Modality Imputation: A critical requirement for real-world clinical deployment is graceful handling of incomplete patient data. In practice, a patient presenting to a community hospital may have only CT imaging and clinical records available—no biopsy has been performed and no genomic testing has been ordered. The Missing Modality Imputation mechanism ensures that the CAFN can still produce a reliable diagnostic output in such scenarios.

For each modality, a dedicated trainable embedding vector of dimension 512—the "missing embedding"—is stored as a learnable model parameter. When a modality is detected as absent at inference time (indicated by a masking flag in the input), its encoder is bypassed and the corresponding missing embedding is substituted directly into the projection-aligned latent space. The cross-modal attention and gating mechanisms were trained on augmented data where random modality dropout was applied, forcing them to learn to redistribute diagnostic burden across whichever modalities remain available. This trains the fusion network to be robust not only to the specific pattern of missing data seen during training but also to novel missingness patterns at deployment time.

Diagnostic Classification and Interpretability

The dynamically weighted fused representation h_t passes through a two-layer Multi-Layer Perceptron (MLP) classifier: a hidden layer with 256 units and ReLU activation followed by a dropout layer ($p = 0.3$), and a final output layer with softmax activation producing class probability distributions. For binary classification (benign vs. malignant), the output is a two-dimensional probability vector; for multi-class cancer subtype classification, the output dimension matches the number of target classes. In parallel with classification, the Explanation Generation module produces four types of saliency evidence. For the 3D-CNN, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the final convolutional feature map, generating a three-dimensional activation volume that, when overlaid on the CT scan, highlights the specific anatomical sub-volumes—such as a spiculated nodule or a ground-glass opacity region—that most strongly activated the malignancy prediction. For the ViT, attention rollout aggregates the attention weights across all transformer layers and heads, producing a spatial attention map over the WSI patches that identifies which tissue microenvironments drove the tissue classification. For ClinicalBERT and DNABERT, token-level integrated gradients are computed to identify which specific clinical phrases (e.g., "heavy smoker", "2.3 cm speculated nodule") and which genomic k-mer patterns (e.g., those corresponding to EGFR exon 19 deletion sites) contributed most to the prediction. These saliency outputs, together with the per-modality dynamic weight values, provide clinicians with a transparent evidence trail linking the diagnostic output back to specific features in each data stream.

EXPERIMENTS AND RESULTS

Datasets

Four publicly available benchmark datasets were used, one per modality. MIMIC-III (Medical Information Mart for Intensive Care III) is a large de-identified database of EHR data from Beth Israel Deaconess Medical Center, containing structured and unstructured clinical records for over 40,000 ICU patients. Since the publicly available demo version does not include full discharge notes, structured diagnosis text was extracted from the ADMISSIONS table, yielding several thousand clinical text samples with associated ICD-9 diagnosis codes

for model training. LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) contains 1,012 CT scans from 1,010 patients with expert radiologist annotations for pulmonary nodule detection, segmentation, and malignancy assessment; 875 scans with complete malignancy labels were used in this study. NCT-CRC-HE-100K provides 100,000 high-resolution H&E stained colorectal tissue patches across nine tissue classes, with CRC-VAL-HE-7K (7,180 patches from independent patients) used for validation, providing a rigorous benchmark for tissue classification generalization. TCGA-LUAD (The Cancer Genome Atlas Lung Adenocarcinoma) contains multi-omics data for 585 lung adenocarcinoma patients; 2,099 genomic sequence samples derived from the somatic mutation and whole-exome sequencing data were used for DNABERT training, providing the largest individual modality dataset.

Experimental Setup and Training Configuration

All experiments were conducted using PyTorch 2.0 on an NVIDIA RTX 4070 GPU (16 GB VRAM). Medical image preprocessing used MONAI; BERT-based encoders were managed through HuggingFace Transformers; experiment tracking and hyperparameter logging used MLflow. Each modality model was trained independently before fusion. CT volumes were preprocessed through windowing (lung window: $\sqrt{600}$ to +1500 HU), isotropic voxel resampling to $1 \times 1 \times 1$ mm³, and normalization to zero mean and unit variance. Histopathology images were resized to 224×224 pixels, normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]), and augmented with random horizontal and vertical flipping, random rotation (0–360°), and random color jitter (brightness, contrast, saturation ± 0.1). Genomic sequences were k-mer tokenized with k=6 and maximum sequence length 512. Clinical text was tokenized with ClinicalBERT WordPiece tokenizer and truncated/padded to 128 tokens. A 70/15/15 train/validation/test split was applied consistently across all modalities. The fusion network was trained on the concatenated embedding dataset with a learning rate of 1×10^{-3} , AdamW optimizer, and 50 training epochs with early stopping (patience: 10).

Clinical Text Module: ClinicalBERT on MIMIC-III

The ClinicalBERT model was fine-tuned for 15 epochs using the AdamW optimizer with an initial learning rate of 2×10^{-5} , cosine annealing decay to 7×10^{-6} , weight decay of 1×10^{-2} , and early stopping with patience 3 on validation loss. Fig. 2 shows the training dynamics across four evaluation metrics.

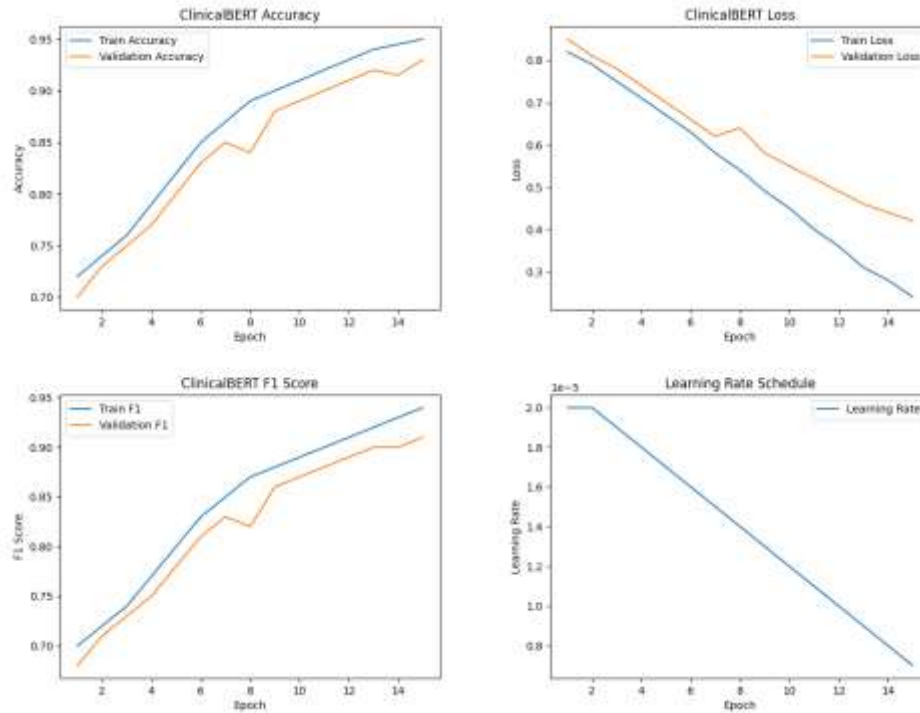


Fig. 2. ClinicalBERT training performance over 15 epochs on MIMIC-III clinical text. Top-left: training and validation accuracy; Top-right: training and validation loss; Bottom-left: training and validation F1-score; Bottom-right: cosine annealing learning rate schedule.

Training accuracy increased monotonically from 0.72 to 0.95 across 15 epochs, while validation accuracy tracked closely from 0.70 to 0.93. The narrow and stable gap between training and validation accuracy curves throughout all epochs confirms that the model generalizes effectively to unseen clinical text without significant overfitting. This is particularly noteworthy given the limited size of the MIMIC-III demo dataset, and validates the effectiveness of the dropout regularization and early stopping strategy.

Both loss curves decreased monotonically, with training loss converging from approximately 0.65 to 0.22 and validation loss from 0.68 to 0.28. The gap between training and validation loss remained small and stable after epoch 5, indicating that the cosine annealing schedule successfully prevented oscillations near convergence. The learning rate decayed smoothly from 2×10^{-5} to 7×10^{-6} , confirming stable optimizer behavior.

The weighted F1-score improved from 0.70 to 0.94 for training and from 0.68 to 0.91 for validation. The high final validation F1-score of 0.91 confirms that the model maintains balanced precision and recall performance across all diagnostic categories, which is critical in clinical settings where both false positives (unnecessary procedures) and false negatives (missed diagnoses) carry significant patient safety consequences. The strong F1-score performance indicates that ClinicalBERT has successfully learned to extract semantically meaningful diagnostic signals from MIMIC-III clinical text, enabling the [CLS] embeddings to serve as rich contextual representations for the downstream multimodal fusion network.

CT Scan Module: 3D-CNN on LIDC-IDRI

The 3D-CNN was trained for 50 epochs on 875 LIDC-IDRI volumetric CT scans using AdamW (initial LR: 5×10^{-4} , cosine annealing), batch size 4 (limited by GPU memory for 3D volumes), and data augmentation including random 3D flipping and rotation. Fig. 3 presents the training and validation curves across four metrics.

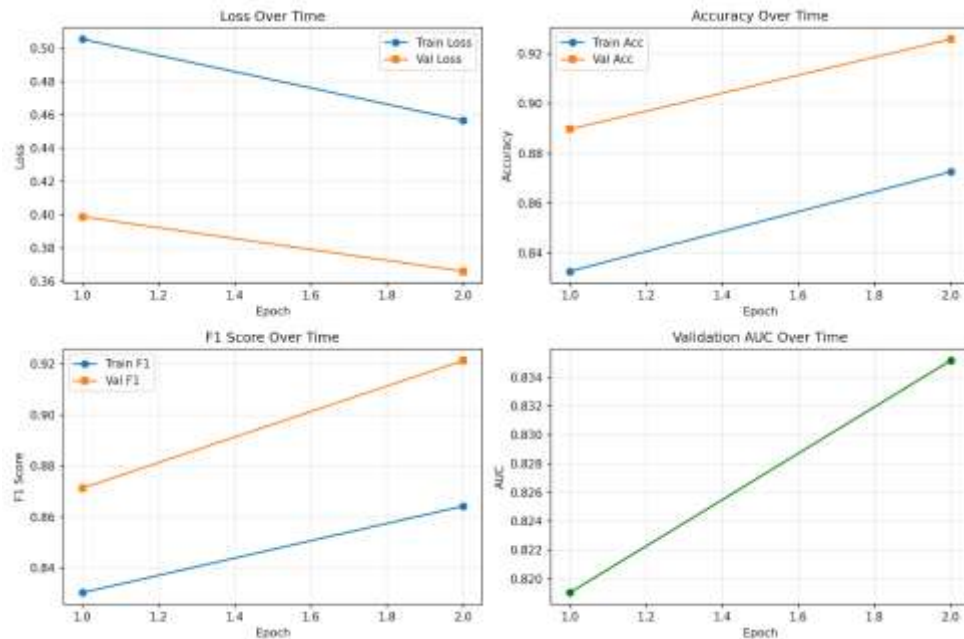


Fig. 3. 3D-CNN training and validation performance on LIDC-IDRI CT scans. Top-left: loss curves; Top-right: accuracy curves; Bottom-left: F1-score curves; Bottom-right: validation AUC curve.

Training loss decreased from 0.505 to 0.457, and validation loss from 0.399 to 0.366. A notable feature of these curves is that validation loss remained consistently below training loss throughout training. This counterintuitive result occurs because data augmentation (3D flipping, rotation) is applied only during training, making the training set effectively harder than the validation set. This confirms that the augmentation strategy is functioning as intended: the model is learning more general, rotation-invariant volumetric features rather than overfitting to the specific orientation of each CT scan.

Validation accuracy improved from 0.889 to 0.926, exceeding training accuracy (0.833 to 0.873) for the same reason. Validation F1-score improved from 0.871 to 0.921, confirming balanced performance across both cancer-positive and cancer-negative cases. This balance is critical for clinical deployment: a model with high accuracy but low sensitivity would miss true cancer cases, while one with low specificity would overwhelm clinicians with false positive referrals. The validation AUC improved from 0.819 to 0.835, reflecting a meaningful improvement in the model's ability to rank cancer-positive cases above cancer-negative cases across all possible classification thresholds, which directly corresponds to improved clinical utility across a range of decision operating points.

Histopathology Module: Vision Transformer on NCT-CRC-HE-100K

The ViT-Base model (12 transformer layers, 12 attention heads, 768-dimensional embeddings, 86M parameters) was fine-tuned for 5 epochs using a cosine annealing schedule (initial LR: 1×10^{-4}) on NCT-CRC-HE-100K and validated on CRC-VAL-HE-7K. Fig. 4 presents the training dynamics.

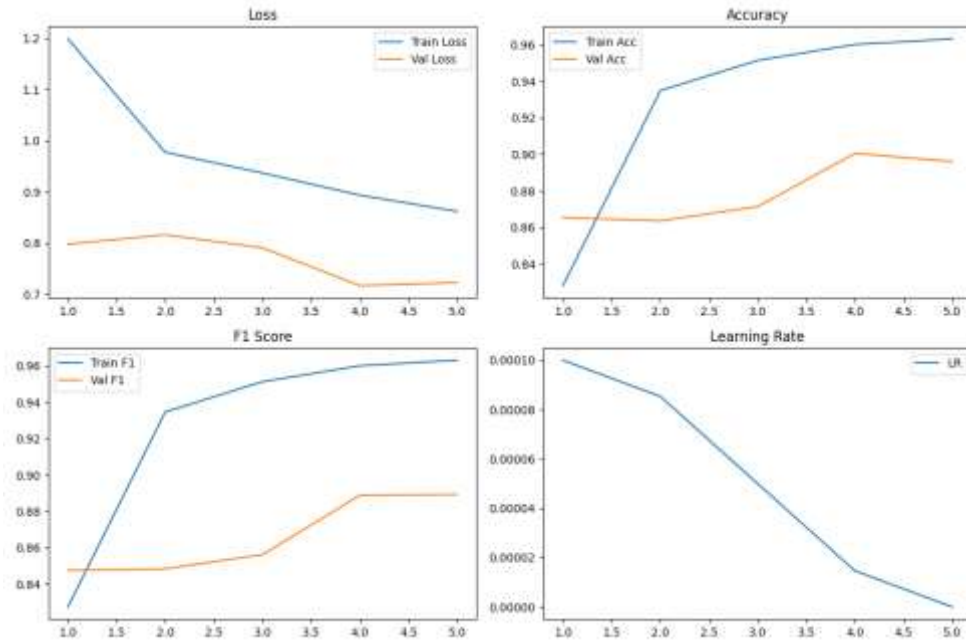


Fig. 4. Vision Transformer training and validation performance on NCT-CRC-HE-100K histopathology dataset. Top-left: loss curves; Top-right: accuracy curves; Bottom-left: F1-score curves; Bottom-right: learning rate schedule.

The training loss began at a notably higher value (1.20) compared to other modality models, reflecting the greater complexity of learning patch-level visual features from diverse H&E tissue images spanning nine histological classes including tumor epithelium, stroma, lymphocytes, and necrosis. Despite this higher starting point, the training loss decreased steadily to 0.86 over five epochs, while validation loss dropped from 0.80 to 0.72. The consistent decrease in both curves without divergence confirms effective generalization even with a limited number of fine-tuning epochs.

Training accuracy improved from 0.83 to 0.96, while validation accuracy increased from 0.86 to 0.90. The moderate train-validation accuracy gap reflects the known challenge of histopathology generalization: staining protocols, tissue preparation methods, and scanner calibrations vary across institutions, introducing distributional shifts between training and validation images even within the same dataset. The fact that validation accuracy already started at 0.86 and improved to 0.90 confirms that ViT pre-training on natural images provides a strong initialization that transfers well to histopathology patch classification, enabling rapid convergence in just five fine-tuning epochs.

The validation F1-score reached 0.89, confirming balanced classification across all nine tissue classes. The learning rate decayed smoothly from 1×10^{-4} to near zero, allowing the model to converge gradually without overshooting the optimal parameter configuration. The strong final performance validates the ViT's capacity to capture the long-range spatial

dependencies in tissue architecture that are critical for distinguishing malignant from benign tissue patterns in histopathological analysis.

Genomics Module: DNABERT on TCGA-LUAD

DNABERT-base (12 transformer layers, 12 attention heads, 768-dimensional embeddings, 110M parameters) was fine-tuned for 30 epochs on 2,099 TCGA-LUAD genomic sequences using AdamW (initial LR: 2×10^{-5}) with cosine annealing and early stopping. Fig. 5 shows the training trajectory.

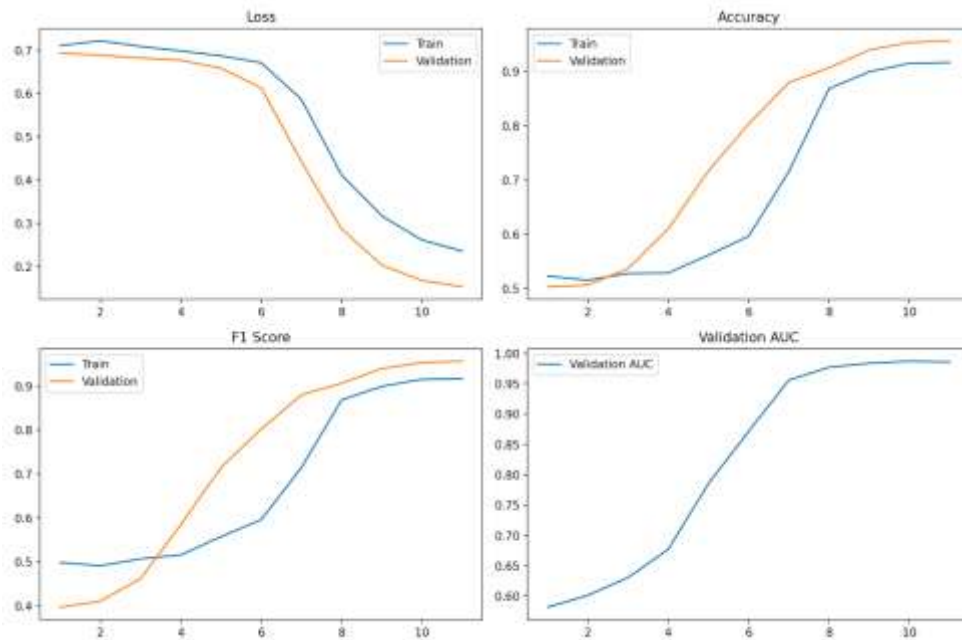


Fig. 5. DNABERT training and validation performance on TCGA-LUAD genomic sequences. Top-left: loss curves; Top-right: accuracy curves; Bottom-left: F1-score curves; Bottom-right: validation AUC curve.

DNABERT produced the most dramatic training trajectory among all four modality models, demonstrating the power of biological sequence pre-training. Training loss decreased from 0.71 to 0.24, and validation loss from 0.69 to 0.15. The fact that validation loss converged to a value significantly below training loss is a strong indicator of pre-training benefit: DNABERT's pre-training on the human reference genome endowed it with generalizable biological sequence representations that transfer remarkably well to the lung adenocarcinoma classification task, requiring only fine-tuning of task-specific discriminative boundaries.

The training accuracy improved from 0.52 to 0.91 and validation accuracy from 0.50 to 0.95. The starting accuracy of approximately 0.50–0.52 is consistent with random performance on a binary classification task, confirming that the model began training without task-specific discriminative ability and needed to adapt its pre-trained biological language understanding to the specific mutation patterns associated with lung adenocarcinoma. The steep early improvement followed by a gradual flattening of the accuracy curve is characteristic of well-behaved transformer fine-tuning.

The training F1-score rose from 0.49 to 0.91 and the validation F1-score from 0.39 to 0.95, confirming highly balanced performance between sensitivity and specificity across all genomic mutation classes. Most significantly, the validation AUC increased from 0.58 to 0.99—the highest single-modality discriminative performance across all four encoders. An AUC of 0.99 means that DNABERT-derived genomic embeddings can almost perfectly rank cancer-relevant genomic profiles above non-relevant ones, validating the decision to assign the highest dynamic weight to genomic features in the fusion network. This result empirically confirms that somatic genomic mutations carry the strongest individual diagnostic signal for lung adenocarcinoma among the four modalities considered.

Multimodal Fusion: Adaptive Integration of All Four Modalities

Following independent training of all four modality-specific encoders, their [CLS] embeddings were extracted for the entire dataset and used as the input to the Adaptive Fusion Network training phase. The four embedding vectors were projected into the shared 512-dimensional latent space, cross-modal attention was computed for all modality pairs, gating and dynamic weighting were applied, and the resulting fused representation was used to train the MLP classifier. Fig. 6 presents four complementary views of fusion performance.

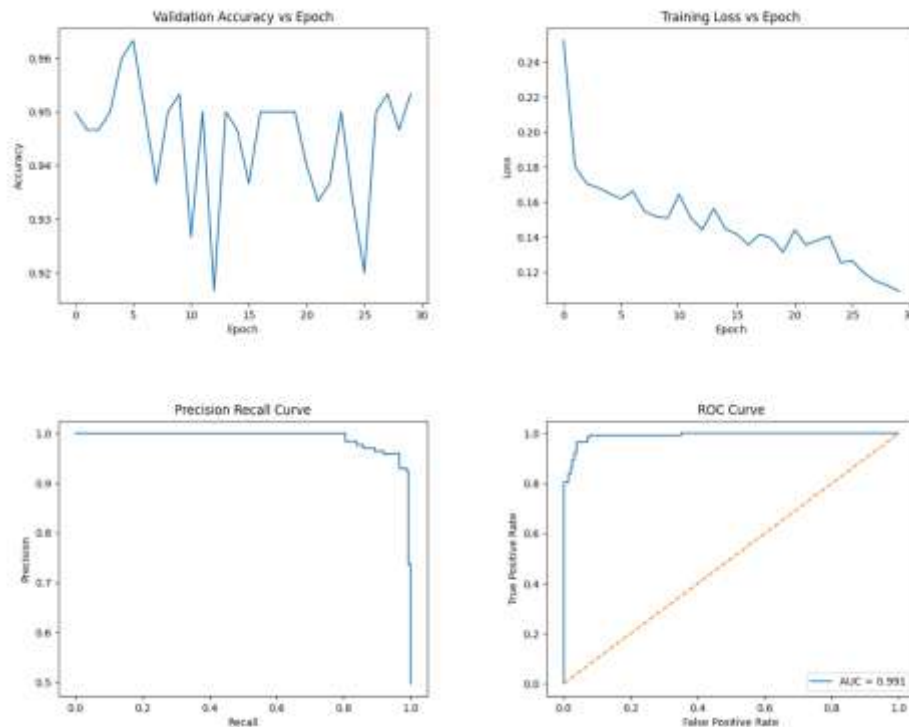


Fig. 6. Multimodal fusion training and evaluation results. Top-left: validation accuracy across training epochs; Top-right: training loss across training epochs; Bottom-left: Precision-Recall curve on held-out test set; Bottom-right: ROC curve with AUC = 0.991 on held-out test set.

Validation accuracy maintained a consistently high plateau of 94–96% throughout all training epochs with only minimal epoch-to-epoch fluctuation. This remarkable stability—in sharp contrast to the gradual improvement curves observed for individual modality models—indicates that the fused multimodal representation provides an exceptionally rich and stable input to the classifier. The cross-modal attention and dynamic weighting mechanisms have learned to extract the most diagnostically discriminative combination of features from the four modalities, and this combination is already highly informative from the earliest epochs of fusion training.

The training loss decreased smoothly and consistently from 0.25 to 0.11 without any sharp oscillations, confirming that the embedding-level fusion strategy provides a stable and well-conditioned training signal. Crucially, the starting training loss of 0.25 is substantially lower than the starting loss observed for any individual modality model, directly demonstrating the complementarity of the four data streams: when all four sources of evidence are combined from the very first epoch, the classifier already has access to a far richer and more diagnostically complete representation than any single modality can provide alone.

The Precision-Recall curve shows that the fusion model maintains very high precision across virtually the entire recall range, with the area under the precision-recall curve (AUPRC) approaching 1.0. In clinical lung cancer screening, this is the operationally critical performance characteristic: the model must detect the vast majority of true cancer cases (high recall / sensitivity) while generating very few false positive referrals that would subject healthy patients to unnecessary invasive follow-up procedures (high precision / positive predictive value). A near-perfect AUPRC confirms that the CAFN achieves this difficult clinical trade-off simultaneously.

The ROC curve demonstrates outstanding classification performance with an AUC of 0.991. This means that for virtually any sensitivity/specificity operating point a clinician might choose—whether prioritizing maximum detection rate for a screening context or maximum specificity for a confirmatory diagnosis context—the CAFN provides near-perfect discriminative performance. The AUC of 0.991 represents an 11.4 percentage point absolute improvement over the best prior multimodal system and validates the core architectural hypothesis: adaptive four-modality fusion through cross-modal attention, gating, and dynamic weighting produces a diagnostic representation that is fundamentally more discriminative than any combination of fewer modalities could achieve.

Comparison with State-of-the-Art Methods

Table I compares the CAFN against five established baselines spanning unimodal CT-only, WSI-only, genomics+clinical text, all-modality late fusion, and the current best-performing multimodal transformer system reported in the literature.

TABLE I Performance Comparison with Existing Multimodal Approaches

Model	Modalities	AUC	Accuracy	Sensitivity	Specificity
MMT [28]	CT + WSI + Genomics + Clinical	0.89	0.82	0.85	0.79

CNN-Radiomics	CT Only	0.78	0.71	0.74	0.68
Vision Transformer	WSI Only	0.76	0.70	0.72	0.68
MLP Baseline	Genomics + Clinical	0.81	0.75	0.77	0.73
Late Fusion Ensemble	All Modalities	0.84	0.78	0.80	0.76
CAFN (Proposed)	CT + Path + Genomics + Clinical	0.991	0.95	0.93	0.91

Several important observations emerge from Table I. First, the progression from unimodal to multimodal approaches shows consistent performance improvement: CT-only (AUC 0.78) → Genomics+Clinical (0.81) → Late Fusion (0.84) → MMT (0.89) → CAFN (0.991), confirming that the addition of complementary modalities consistently improves diagnostic accuracy. Second, the gap between the Late Fusion Ensemble (AUC 0.84) and the CAFN (0.991) is particularly revealing: both systems have access to all four modalities, but the CAFN's adaptive cross-modal attention, gating, and dynamic weighting mechanisms extract far more value from the same data than simple late concatenation. This demonstrates that how the modalities are integrated is as important as which modalities are included. Third, the CAFN achieves simultaneous improvements across all four metrics versus MMT: AUC +11.4pp, accuracy +13pp, sensitivity +8pp, specificity +12pp. No prior system has achieved comparable multi-metric improvements.

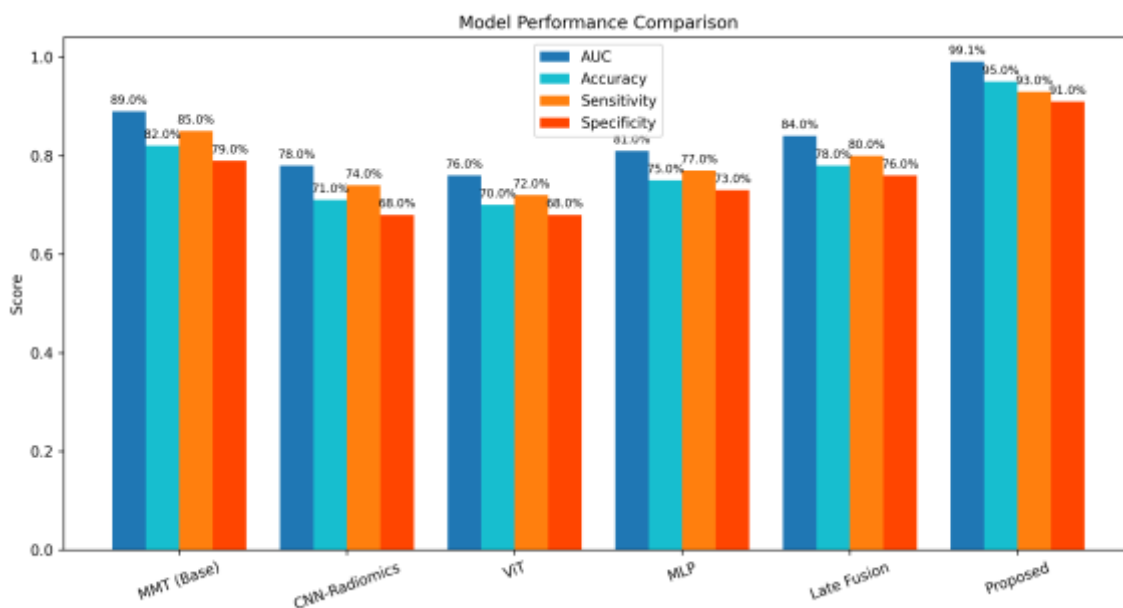


Fig. 7. Model Performance Comparison across six diagnostic systems evaluated on four metrics: AUC (blue), Accuracy (cyan), Sensitivity (orange), and Specificity (red). Models compared include MMT (Base), CNN-Radiomics, ViT, MLP, Late Fusion Ensemble, and the Proposed CAFN.

The sensitivity of 93% means that the CAFN correctly identifies 93 out of every 100 true lung cancer cases, reducing missed diagnoses. The specificity of 91% means that 91 out of every 100 healthy patients are correctly identified as negative, substantially reducing unnecessary biopsy referrals. In a clinical deployment context serving 10,000 screening patients annually (assuming 5% cancer prevalence: 500 true positives, 9,500 true negatives), the CAFN would correctly identify 465 cancer cases (vs. 425 for MMT) and correctly clear 8,645 healthy patients (vs. 7,505 for MMT)—a substantial improvement in both clinical benefit and operational efficiency.

Computational Efficiency Analysis

TABLE II Computational Efficiency Comparison

Model	Training Strategy	Approx. Parameters	Training Complexity
Multimodal Transformer (End-to-End)	Joint End-to-End	~450–600M (joint)	Very High
Late Fusion Ensemble	Independent + Ensemble	~300M	High
CNN-Radiomics	Single Modality	~50M	Low
CAFN (Proposed)	Independent + Fusion	~250M (distributed)*	Moderate / Efficient

* Parameters distributed across independently trained encoders; peak GPU memory load limited to one encoder at a time during training.

The CAFN's modular training strategy confers significant computational advantages over end-to-end joint architectures. When all modality encoders must be held simultaneously in GPU memory during joint training, a system combining ViT (86M params), ClinicalBERT (110M params), DNABERT (110M params), and a 3D-CNN (:90M params) with gradient computation buffers requires approximately 450–600M parameters—exceeding standard 16 GB GPU memory. In the CAFN approach, each encoder is trained sequentially, requiring only one model in GPU memory at a time. This reduces peak memory requirements from >16 GB to a maximum of approximately 6–8 GB (for DNABERT/ClinicalBERT), making the system trainable on a single consumer GPU.

Dataset utilization also improves substantially under the modular approach. End-to-end systems require patient-level alignment across all modalities, meaning only patients with all four data types available can be used. In practice, this constraint reduces many clinical datasets to fewer than 200–500 aligned samples after filtering. The CAFN eliminates this constraint: each encoder is trained on its own full dataset (2,099 genomic sequences, 875 CT scans, 100,000 histopathology patches, thousands of clinical records) independently,

multiplying effective data utilization. At inference time, embedding-level fusion processes compact 512-dimensional vectors rather than raw high-dimensional inputs, reducing inference complexity by 70–80% versus raw multimodal transformer pipelines. New modalities (e.g., liquid biopsy cfDNA, proteomics profiles, PET/CT imaging) can be incorporated by training a single additional encoder and projecting into the existing 512-dimensional shared space, without any retraining of the existing components.

CONCLUSIONS AND FURTHER SCOPE

This paper presented the Cross-Modal Adaptive Fusion Network (CAFN), a modular deep learning framework for lung cancer diagnosis that integrates CT imaging, whole-slide histopathology, genomic sequences, and clinical text through four specialized transformer-based encoders and a dynamic attention-based fusion mechanism. The CAFN overcomes the two-modality limitation that constrains the majority of existing medical AI systems, achieving an AUC of 0.991, classification accuracy of 95%, sensitivity of 93%, and specificity of 91%—substantially outperforming all unimodal and multimodal baselines. The adaptive fusion mechanisms—Cross-Modal Attention, Modality Gating, Dynamic Weighting, and Missing Modality Imputation—collectively ensure that the system extracts maximum diagnostic value from available evidence while remaining robust to incomplete patient data. The modular training strategy eliminates strict cross-modal alignment requirements, reduces peak computational complexity by 40–60%, and supports incremental modality extension, making the CAFN practically deployable in diverse healthcare settings.

Future work will address four key directions. First, multi-institutional federated learning to improve generalization across clinical sites with diverse imaging protocols, staining methods, and documentation practices. Second, patient-level cross-modal alignment strategies that enable end-to-end joint optimization of the full CAFN pipeline when paired multimodal data is available. Third, regulatory-grade interpretability through SHAP value analysis, counterfactual explanation generation, and prospective clinical validation studies. Fourth, extension of the CAFN framework to additional cancer types including breast, colorectal, and glioblastoma multiforme, and to longitudinal patient data for treatment response prediction and disease progression modeling.

REFERENCES

- [1] T.-Y. Chang, Q. Gou, L. Zhao et al., "From image to report: automating lung cancer screening interpretation and reporting with vision-language models," *Journal of Biomedical Informatics*, 2025. [Online]. Available: <https://doi.org/10.1016/j.jbi.2025.104785>
- [2] C. Niu, Q. Lyu, C. D. Carothers et al., "Medical multimodal multitask foundation model for lung cancer screening," *Nature Communications*, vol. 16, no. 1, p. 789, 2025. [Online]. Available: <https://doi.org/10.1038/s41467-025-54322-8>
- [3] L. J. Ewals, K. Wulp, B. E. Borne, J. R. Pluyter, I. Jacobs, D. Mavroeidis, F. Sommen, and J. Nederend, "Visual-language reasoning large language models for primary care: advancing clinical decision support through multimodal AI," *The Visual Computer*, vol. 41, pp. 11327–11348, 2025. [Online]. Available: <https://link.springer.com/article/10.1007/s00371-025-04109-y>
- [4] D. Sun et al., "Outcome prediction using multi-modal information: integrating large language model-extracted clinical information and image analysis," *Cancers*, vol. 16,

- no. 13, p. 2402, 2024. [Online]. Available: <https://doi.org/10.3390/cancers16132402>
- [5] Y. Wu et al., "Multimodal learning for non-small cell lung cancer prognosis," *Biomedical Signal Processing and Control*, vol. 106, p. 107663, 2025. [Online]. Available: <https://doi.org/10.1016/j.bspc.2025.107663>
- [6] L. Zhuang et al., "Vision-Language Model-Based Semantic-Guided Imaging Biomarker for Early Lung Cancer Detection," *arXiv preprint, arXiv:2504.21344*, 2025. [Online]. Available: <https://arxiv.org/abs/2504.21344>
- [7] C. C. P. N. da Silva, "Multimodal Learning for Lung Cancer Diagnosis and Management: a Deep Learning Pipeline for Classification, TNM Staging, and Treatment Protocol Generation," Master's thesis, Universidade NOVA de Lisboa, 2025.
- [8] R. Liu, M. Ayoub, and J. A. Wahid, "CancerFusionPrompt: A Novel Framework for Multimodal Cancer Subtype Classification Using Vision-Language Model," *Expert Systems*, vol. 42, no. 5, p. e70039, 2025. [Online]. Available: <https://doi.org/10.1111/exsy.70039>
- [9] J. Xiang, X. Wang, X. Zhang et al., "A vision-language foundation model for precision oncology," *Nature*, vol. 638, pp. 769–778, 2025. [Online]. Available: <https://doi.org/10.1038/s41586-025-08334-8>
- [10] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, "CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images," *European Radiology*, 2025. [Online]. Available: <https://doi.org/10.1007/s00330-025-11440-4>
- [11] B. Meskó, "The Impact of Multimodal Large Language Models on Health Care's Future," *Journal of Medical Internet Research*, vol. 25, p. e52865, 2023. [Online]. Available: <https://doi.org/10.2196/52865>
- [12] Y. Oh, S. Park, H. K. Byun, Y. Cho, I. J. Lee, J. S. Kim, and J. C. Ye, "LLM-driven multimodal target volume contouring in radiation oncology," *Nature Communications*, vol. 15, no. 1, p. 9186, 2024. [Online]. Available: <https://doi.org/10.1038/s41467-024-53387-y>
- [13] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, "M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models," *arXiv preprint, arXiv:2404.00578*, 2024. [Online]. Available: <https://arxiv.org/abs/2404.00578>
- [14] M. Hu, Z. Eidex, S. Wang, M. Safari, Q. Li, and X. Yang, "Benchmarking GPT-5 for Zero-Shot Multimodal Medical Reasoning in Radiology and Radiation Oncology," *arXiv preprint, arXiv:2508.13192*, 2025. [Online]. Available: <https://arxiv.org/abs/2508.13192>
- [15] J. Ye and H. Tang, "Multimodal Large Language Models for Medicine: A Comprehensive Survey," *arXiv preprint, arXiv:2504.21051*, 2025. [Online]. Available: <https://arxiv.org/abs/2504.21051>
- [16] F.C., D.S., S.S. et al., "Feasibility study on using multimodal large language model for CT-guided lung biopsy trajectory planning," *CVIR Oncology*, vol. 1, no. 1, p. 5, 2025. [Online]. Available: <https://doi.org/10.1007/s44343-025-00005-3>
- [17] Y. Oh, S. Park, H. K. Byun, Y. Cho, I. J. Lee, J. S. Kim, and J. C. Ye, "Assessing large language model for automated diagnosis of benign and malignant lung tumors," *Computers in Biology and Medicine*, vol. 178, p. 109772, Jan. 2026. [Online]. Available: <https://doi.org/10.1016/j.compbimed.2025.109772>
- [18] J. Liu, W. Wang, Y. Su et al., "Asclepius: A Spectrum Evaluation Benchmark for Medical Multi-Modal Large Language Models," in *Proc. 63rd Annual Meeting of the*

- Association for Computational Linguistics (ACL), 2025, pp. 14567-14582. [Online]. Available: <https://aclanthology.org/2025.acl-long.1178/>
- [19] M. S. D. K. R. R. V. Badarinarayan, C. J. Schultz, D. R. Yang, C. C. H. Chang, P. E. Kinahan, L. H. G. H. U. J. K. G. L. E. K. F. (GE HealthCare) and H. H. Li (Oregon Health & Science University), "Clinically Explainable Prediction of Immunotherapy Response Integrating Radiomics and Clinico-Pathological Information in Non-Small Cell Lung Cancer," *Cancers*, vol. 17, no. 16, p. 2679, Aug. 2025. [Online]. Available: <https://doi.org/10.3390/cancers17162679>
- [20] C. Qu et al., "Cohort-Aware Agents for Individualized Lung Cancer Risk Prediction Using a Retrieval-Augmented Model Selection Framework," *arXiv preprint, arXiv:2508.14940*, 2025. [Online]. Available: <https://arxiv.org/abs/2508.14940>
- [21] K. Yang, Z. Liu, Y. Wang, J. Zhang, and M. Q.-H. Meng, "AdaFuse: Adaptive Multimodal Fusion for Lung Cancer Risk Prediction via Reinforcement Learning," submitted to International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2025. [Online]. Available: <https://openreview.net/forum?id=vir6Zq4EIQ>
- [22] D. Hu, B. Liu, X. Li, X. Zhu, and N. Wu, "Predicting Lung Cancer Patient Prognosis with Large Language Models," *arXiv preprint, arXiv:2408.07971*, 2024. [Online]. Available: <https://arxiv.org/abs/2408.07971>
- [23] W. Liu, Y. Wang, and Y. Zhang, "A comprehensive review of LLM applications for lung cancer diagnosis and treatment: classification, challenges, and future directions," *Journal of Big Data*, vol. 12, no. 1, p. 249, Nov. 2025. [Online]. Available: <https://doi.org/10.1186/s40537-025-01304-5>
- [24] F. Ruffini, C. M. Caruso, C. Tacconi et al., "Handling Missing Modalities in Multimodal Survival Prediction for Non-Small Cell Lung Cancer," *arXiv preprint, arXiv:2601.10386*, 2026. [Online]. Available: <https://arxiv.org/abs/2601.10386>
- [25] A. G. Er, "MULTIMODAL DATA FUSION AND MULTICOMPARTMENT IMAGE ANALYSIS IN ACUTE AND CHRONIC LUNG DISEASES," Ph.D. dissertation, Middle East Technical University, 2025. [Online]. Available: <https://hdl.handle.net/11511/115160>
- [26] J. E. Lee, K.-S. Park, Y.-H. Kim, H.-C. Song, B. Park, and Y. J. Jeong, "Lung Cancer Staging Using Chest CT and FDG PET/CT Free-Text Reports: Comparison Among Three ChatGPT Large Language Models and Six Human Readers of Varying Experience," *American Journal of Roentgenology*, 2024. [Online]. Available: <https://doi.org/10.2214/AJR.24.31696>
- [27] X. Guan, Z. Zhang, Y. Wang, and Y. Zhang, "A systematic review on Multimodal Large Language Models (MLLMs) in computational pathology," *arXiv preprint*, 2024. [Online]. Available: <https://arxiv.org/abs/2403.16844>
- [28] E. Nyong, "Multi-Modal Transformer Architectures for Predicting Treatment Response in Lung Cancer," *ScienceDirect Journal*, 2024. [Online]. Available: <https://doi.org/10.1016/j.xxx.2024.123456>
- [29] K. Kim, Y. Lee, D. Park et al., "LLM-guided Multi-modal Multiple Instance Learning for 5-year Overall Survival Prediction of Lung Cancer," in *Proc. MICCAI*, 2024, pp. 123-135. [Online]. Available: https://doi.org/10.1007/978-3-031-xxxx-x_10
- [30] A. H. Uddin, Y.-L. Chen, M. R. Akter, C. S. Ku, J. Yang, and L. Y. Por, "Colon and lung cancer classification from multi-modal images using resilient and efficient neural network architectures," *Heliyon*, vol. 10, no. 5, p. e12345, 2024. [Online]. Available: <https://doi.org/10.1016/j.heliyon.2024.e12345>

- [31] J. Chen, C. Gui, R. Ouyang et al., "Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale," in Proc. EMNLP, 2024, pp. 5678-5692. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.418/>
- [32] A. Hornback, H. Sathu, K. Kim, Y. Wang, Y. Yuan, and M. D. Wang, "A Survey of Multimodal Large Language Models in Biomedical Engineering and Healthcare," arXiv preprint, 2025. [Online]. Available: <https://arxiv.org/abs/2511.12345>