International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 12 No. 1, 2026
https://theaspd.com/index.php

# Improving Non-Small Cell Lung Cancer Classification Through Radiogenomics and Transformer-Based Deep Learning Fusion Strategies

**Moheb R. Girgis [1], Mamdouh M. Gomaa [1 2], Abdel Rahman A. Hashem [1]**
[1]Department of Computer Science, Faculty of Science, Minia University, 61511, Minia, Egypt. [2]Department of Computer Science, Faculty of Information Technology, Amman Arab University, 11953, Amman, Jordan.

***Abstract***
*Getting the right subtype of non–small cell lung cancer (NSCLC) is essential for choosing the best treatment, but it is tough when looking at tissue and genetic data alone. In this study, we set out to improve NSCLC subtype identification—specifically through transformer-based deep learning. We introduce two fusion techniques to fuse CT/PET scans with clinical and genetic data. The first, Intermediate Fusion, combines these data streams partway through the model. The second, Late Fusion, lets each data type run through its own processing pipeline before merging their predictions at the end. Then we compare our approaches with an earlier Intermediate Fusion strategy. Results demonstrated that the proposed Late Fusion achieved superior performance, with 96.12% accuracy, and the proposed Intermediate Fusion achieved 95.64% accuracy, outperforming an earlier Intermediate Fusion which achieved 94.04% accuracy. Both proposed (late fusion, intermediate fusion) methods pick up true cases equally well with 95.5% sensitivity, but Late Fusion's boosting its precision to 96.22% and F1-score to 95.86%. Model interpretability is evaluated using SHAP for tabular data and attention-based analysis for imaging to reveal modality-specific contributions. These findings indicate that late fusion not only improves classification performance but also supports more transparent and clinically meaningful decision-making.*
***Keywords***: *Deep Learning, Fusion Strategies, Model interpretability, Non-small cell lung cancer (NSCLC), Radiogenomics.*

## INTRODUCTION

Non-small cell lung cancer (NSCLC) is the most common cause of cancer-related death worldwide [1]. It is histologically divided into subtypes like adenocarcinoma and squamous cell carcinoma, each of which requires a different approach to treatment [2]. The gold standard for molecular profiling, tissue biopsies, is invasive, risky, and may not be able to detect tumor heterogeneity, while imaging modalities such as CT and PET scans are non-invasive but lack the resolution to accurately predict genomic alterations [3]. Genomic testing, although transformative, is expensive, time consuming, and unavailable in settings with limited resources [4].

Radiogenomics is a potential approach that combines genetics with radiological imaging to find non-invasive biomarkers. Connecting CT patterns to EGFR mutations uses quantitative imaging parameters (such as texture and form) to predict molecular profiles [3, 5]. However, conventional radiogenomic models frequently rely on manually created features or rudimentary machine learning, which restricts their capacity to represent intricate, multimodal interactions and often provide limited insight into the factors driving model predictions, reducing clinical trust and interpretability.

To overcome these constraints, deep learning (DL), specially transformer architecture, provides hitherto unheard-of possibilities. Transformers, known for its self-attention processes, excel at modeling long-range data dependencies, as demonstrated by their performance in natural language processing [6] and medical picture analysis [7]. DL can use three different fusion techniques [8, 9] to combine genomic and imaging data in radiogenomics:

1. Early fusion: Integrating genetic information features with raw images features at the input level.
1. Late fusion: combining outputs at the decision level after processing modalities independently.
2. Intermediate fusion: merges each modality's features at the intermediate fusion layer before processing them in a different model.

Nevertheless, current methods frequently use crude fusion strategies, not taking full advantage of contextual hierarchies or cross-modal interactions [10]. Convolutional neural networks (CNNs), for example, are good at extracting local picture characteristics, but because of their inductive bias toward

hierarchical, translation-invariant patterns, they may miss global tumor context [6, 11]. On the other hand, intermediate fusion leverages intermediate features to improve joint representation learning by combining predictions from multiple modality-specific models at intermediate layers. Despite these advances, limited attention has been given to systematically analyzing how different fusion strategies influence model interpretability and modality-specific contributions in radiogenomic classification.

**Research Gap and Aim**

Despite developments, no research has used transformers to dynamically combine genomic and imaging data at various phases of analysis. Predictive model accuracy is constrained by this disparity. In addition, existing studies rarely examine the transparency of multimodal decision-making, which is critical for clinical adoption.

The aim of this paper is to improve non-small cell lung cancer (NSCLC) classification through radiogenomics and transformer-based deep learning fusion strategies. Accordingly, this paper presents two innovative deep learning fusion strategies—Late Fusion and Intermediate Fusion—designed to enhance the integration of multi-modal data, including fused CT/PET imaging and clinical-genetic tabular data, for distinguishing NSCLC subtypes (adenocarcinoma and squamous cell carcinoma). In situations when DNA testing is not accessible, these methods may enable physicians to customize medicines in a non-invasive manner. Furthermore, the paper compares the proposed Late and Intermediate Fusion strategies with an earlier Intermediate Fusion approach [12]. In the context of NSCLC classification, this comparison aids in determining the best method for integrating multi-modal medical images and tabular data (CT and PET scans, clinical, and genetics). To support reliable deployment, the proposed models are further analyzed using explainability techniques that provide insight into both tabular and imaging contributions to the final predictions.

**Contribution**

Through the investigation of the earlier Intermediate Fusion [12], proposed Intermediate Fusion, and proposed Late Fusion in multi-modal NSCLC classification with image features and tabular features, the work presented in this paper makes the following contributions:

1. Proposed Intermediate Fusion Framework: This is an innovative framework that combines intermediate-layer image features (obtained from BEiT vision transformer pre-training) with tabular features (processed through a fully connected MLP network) in a dynamic way. The combination facilitates cross-modal alignment prior to end-classification for improving feature representation.
2. Proposed Late Fusion Strategy: This is a novel Decision-level fusion strategy in which tabular features (using MLP) and imaging features (using BEiT) are processed separately, and their probabilistic outputs are fused with weights. It retains modality-specific learning while profiting from complementarity.
3. Fusion Strategy Comparison: a thorough analysis of the effects of the proposed intermediate fusion, the proposed late fusion, and the earlier intermediate fusion [12], techniques on classification performance for multi-modal data.
4. Understanding Model Flexibility: illustrating how various fusion techniques can be applied to manage modality-specific data and enhance multi-modal classification tasks in the field of medical image analysis.
5. Analyzing the interpretability of models: a systematic post-hoc analysis for explaining the explainability of models using SHAP-based feature attribution for tabular data, as well as attention analysis for vision transformer models for imaging data.

The paper is organized as follows: Section 2 gives a review of previous research work in the classification of NSCLC. Section 3 shows different fusion strategies and challenges for each one. Section 4 outlines the process used to create and assess the proposed multi-modal NSCLC classification model with the two innovative deep learning fusion strategies, Late Fusion and Intermediate Fusion. Section 5 presents the results of the experiments that we have conducted to evaluate the proposed fusion strategies compared to an earlier Intermediate Fusion strategy [12]. Section 6 presents the model explainability and interpretability analysis, Section 7 discusses the experimental results; and finally, Section 8 concludes the research work presented in this paper.

## RELATED WORK

Pathologists typically rely on tissue examinations and scans like CT or PET to sort NSCLC into subtypes, but these approaches can't reveal fine-scale molecular details, are invasive, and may miss regions of the

tumor due to sampling bias [2, 13]. Radiogenomics has emerged to overcome these hurdles by correlating genetic changes—such as EGFR mutations—with measurable imaging traits like texture and shape [3, 5]. However, the intricate, nonlinear genotype–phenotype correlations present in NSCLC could not be well captured by the handmade features and linear models used in early radiogenomic investigations [3, 5, 14].
Medical Imaging with Deep Learning: Deep learning has significantly enhanced the analysis of NSCLC imaging. For tasks like delineating tumors and identifying subtypes, convolutional neural networks (CNNs) have become the gold standard [10, 15]. For example, 3D CNN models consistently outperform classic radiomics approaches in detecting lung nodules [10, 15]. However, because CNNs tend to focus on local texture patterns, they can struggle to capture the global context of tumor's overall context [10].
To address the narrow focus of CNNs, transformer-based models have been adopted in medical imaging. Vision Transformers (ViTs) break each scan into a sequence of patches, enabling them to capture tumor heterogeneity and long-range spatial relationships across the image [6, 7]. Hybrid methods like TransUNet blend convolutional layers with transformer blocks, tapping into both detailed local textures and global contextual cues to improve how accurately images can be segmented and classified [11, 16]. Transformers have also shown real potential for predicting patient survival in NSCLC cases [7].
Deep-learning techniques have become highly effective at using genomic information to distinguish NSCLC subtypes and detect key driver mutations—like EGFR and KRAS—especially when combined with state-of-the-art imaging methods [17]. To better model the long-range relationships within gene expression data, transformer-based architectures such as GeneFormer have also been developed [18]. To further contextualize tumor biology, these genetic models are often created separately and do not integrate with spatial imaging data [17].
Fusion Strategies for Multimodal Integration: There is potential for a more reliable diagnosis of NSCLC by combining complementing data from genetics and imaging. Weak cross-modal interactions and dimensionality mismatches are common problems with early fusion techniques, which merge various modalities at the input level [8, 19, 20]. In contrast, late fusion (or decision-level fusion) addresses these challenges by independently processing each modality and combining their outputs at the final decision stage. In this strategy, dedicated models are built for each modality—such as CNNs for imaging and gradient-boosted trees for genomic data—and their outputs are combined through techniques like weighted averaging, ensemble voting, or a metalearner (for example, logistic regression or an additional neural network) [21, 22]. Late fusion benefits from modality-specific optimization and mitigates dimensionality mismatches but often struggles to exploit fine-grained cross-modal correlations, which are critical for NSCLC where genomic mutations (e.g., EGFR) may correlate with subtle imaging patterns (e.g., spiculated tumor margins). Consequently, late fusion has achieved only modest success in survival prediction, as highlighted in [21].
More recent intermediate fusion algorithms, which align and integrate deep features from CNN and transformer branches using attention mechanisms, have shown superior performance in NSCLC classification [12, 23]. These methods dynamically weight cross-modal interactions, enabling granular fusion of radiogenomic features (e.g., linking tumor texture features to immune gene expression). Datasets enabling radiogenomic analysis, such as The Cancer Imaging Archive (TCIA) paired with genomic data from The Cancer Genome Atlas (TCGA), have been instrumental in developing these models [24–27].
Recent Advances: Significant performance gains in NSCLC diagnosis have been reported by recent studies that combine transformer-based architecture with CT/PET imaging [7, 16]. Notwithstanding these encouraging findings, there are still issues such fragmented model design and underutilized cross-modal learning, and many validations are restricted to small cohorts [21, 23].
In conclusion, our work sits at the nexus of multiple important research directions in the fields of cancer diagnostics and medical imaging. Our work intends to address some of the limitations noted in earlier research by combining and expanding upon these various approaches in order to provide a novel approach to the early detection and categorization of NSCLC.

**Fusion Strategies Comparison**

One of the key contributions of this study is the comparative analysis of primary fusion strategies: early, intermediate, and late fusion. These approaches dictate how multimodal data—such as CT/PET scans, clinical records, and genetic information—are combined and processed before classification. The main

objective is to assess how each method affects classification performance, especially in terms of capturing the complex relationships among the different data modalities.

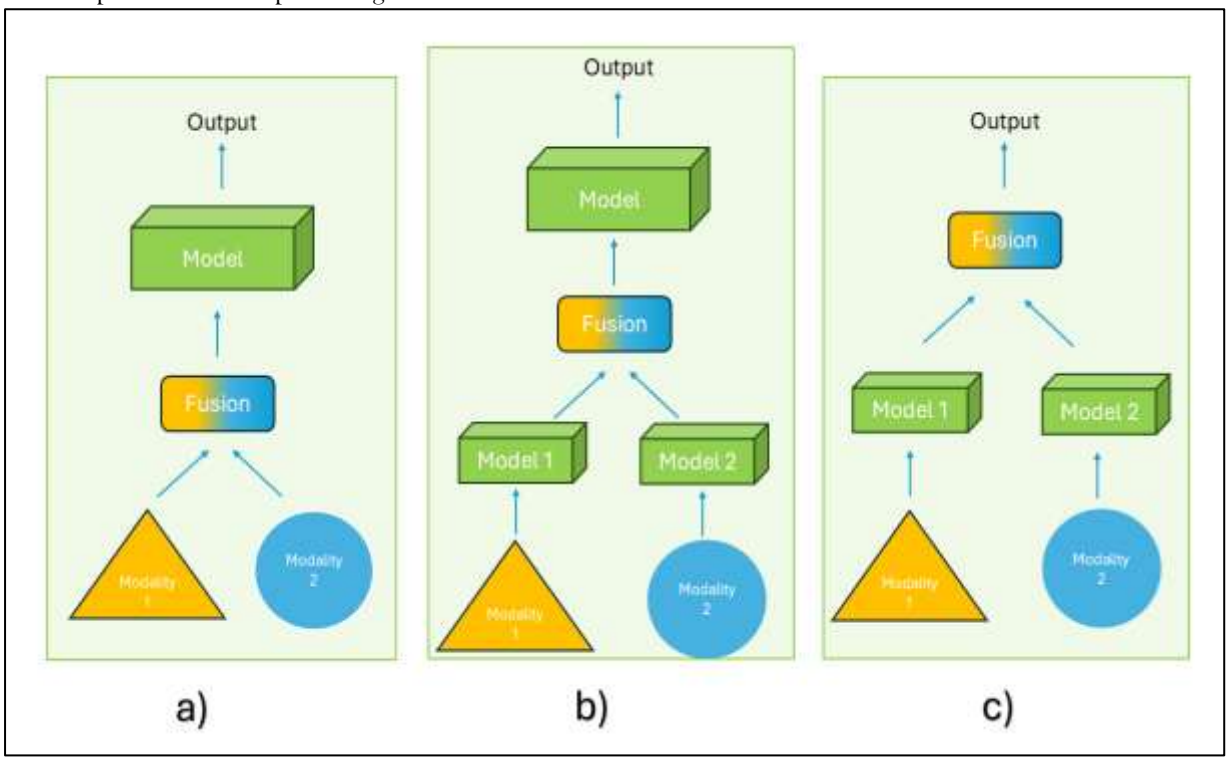



**Fig. 1**: Fusion techniques based on DL. (a) In early fusion techniques, combined vectors are used as input. There is no learning of marginal representations. (b) Intermediate fusion techniques fuse marginal representations inside another model after initially learning them. This can happen gradually or all at once. (c) Late fusion strategies aggregate each modality's decisions via sub-models.

**Early Fusion**

In early fusion, different types of input data—such as genetic profiles, clinical records, and fused medical images—are combined into a single feature set right from the start, as illustrated in Fig. 1(a). This unified input is then fed into a deep learning model that learns shared representations across all modalities without treating them separately [8, 9].

**Advantages:** Early fusion is easy to implement and efficient in terms of computation since it avoids the complexity of building separate models for each data type. It is particularly effective at capturing straightforward cross-modal relationships from the raw, low-level features [8, 9].

**Disadvantages:** Best Suited for Similar Data Types: This approach has difficulty handling highly diverse data formats—like combining images with text—because it requires uniform feature representation, which often demands heavy preprocessing.

Misses Deeper Interactions: By merging data at the input stage, early fusion can obscure more complex relationships that tend to emerge in deeper layers of modality specific models, such as semantic content or spatial structures. Sensitive to Data Imbalance and Noise: Variations in feature scale or data quality—for example, high dimensional genomic data versus low-dimensional clinical metrics—can harm model performance, as early fusion doesn't dynamically adapt to such differences [8].

**Intermediate Fusion**

Intermediate fusion starts with processing individual modalities of data, such as images, clinical records, or genetics. This allows the model to learn features for individual modalities first before finally fusing them within the same representation, as shown in Fig. 1(b). The process of fusing can either utilize standardized networks or modality-specific networks. There are two main flavors of such an approach: Marginal Intermediate Fusion combines independent representation of the individual data and then passes this combined information through a classifier in order to make predictions. Joint Intermediate

Fusion goes even further in enabling the model to learn even more integrated and complex representations from the higher-level features of both the modalities.

**Advantages:** It enables the model to uncover complex patterns between various types of data. Knowing what is unique about every type of data, then it is capable of uncovering richer patterns, which is especially useful for multi-type data classification problems.

**Disadvantages:** It is an involved process in deciding precisely the right depth and order for combining it all together and can involve extensive experimentation in order to provide the maximum performance.

**Late Fusion**

Individual submodels are trained for individual modalities (Fused images, Clinical, Genetics) in late fusion, with their individual predictions being fused in an attempt to obtain the ultimate decision as shown in Fig. 1(c). Averaging, weighted averaging, or meta-learning methods can be employed where a meta-model is trained for combining the prediction probabilities of individual sub-models [8].

**Advantages:** Each modality can be processed individually in late fusion, which can enhance the performance of each modality. The overall model's robustness is increased regarding the errors of the submodels combined, with the possibility of the submodels errors being uncorrelated as well.

**Disadvantages:** Late fusion is not capable of learning the feature-level interaction of the modalities as it misses the useful multimodal information.

## METHODOLOGY

In this section, the approach used in developing and comparing the proposed multimodal NSCLC classifiers is described. It includes the selection of the dataset, pre-processing methods, the techniques for fusion, as well as the classifiers. Comparison of the three fusion models (earlier intermediate [12], proposed intermediate, proposed late fusion) is another contribution of the current work. In the classification of NSCLC, this comparison serves to identify the best method of fusing multi-modal image scans (PET, CT) with tabular features (clinical, genetic).

**Datasets**

The source of the datasets is the public available large-scale multi-modal datasets used in the field of NSCLC. The datasets used:

- Radiogenomics Dataset for NSCLC (primary dataset): The dataset comprises 285,411 images from 303 studies of 211 NSCLC patients, in the form of CT as well as PET scans of the lungs, procured from the Cancer Imaging Archive (TCIA) [24]. Also, included in the dataset is parallel RNA sequencing of tumor tissues from biopsy, as well as clinical variables including age, smoking, histology, treatment, and recurrence of cancer.
- NSCLC Radiomics Dataset [25]: This dataset is used in augmenting the primary dataset with images for 422 NSCLC patients. The dataset is particularly useful with the acknowledgment of class imbalance in the primary dataset (where the Adenocarcinoma class overwhelmed the dataset). Both datasets were created from the same scanner, therefore the quality of the datasets is consistent. For training the models, we use the NSCLC Radiomics and NSCLC Radiomics-Genomics datasets. The NSCLC Radiomics-Genomics dataset [27] features images of 89 NSCLC patients who have undergone surgery, the pre-treatment CT scans, gene expressions, as well as clinical characteristics of the patients.
- De-noised PET scans from large-scale CT and PET database [26] are used for training the deep convolutional autoencoder for PET image de-noising. PET scans are then de-noised with the help of CNN Deep model and input data quality is enhanced for further analysis from the original database.

**Data Preprocessing**

The multi-modal data is in the correct representation for classification as well as fusion owing to the preparation pipeline. As the dataset is multi-modal in nature, some techniques from prior work [12] are adopted for preparation:

- Preprocessing of clinical and genetic data: Both clinical tabular data as well as RNA sequencing data are put through the routine preprocessing steps such as imputing missing values, encoding categorical variables, and normalization. Clinical and genetic variables were selected based on prior evidence of prognostic and biological relevance [28], which investigated associations between imaging phenotypes, clinical characteristics, gene expression profiles, and tumor histological subtypes in lung cancer. The selected clinical features included demographic variables (age at histological diagnosis, ethnicity, weight),

smoking-related factors (smoking status, pack years), treatment-related variables (chemotherapy, radiation, adjuvant treatment), pathological staging parameters (pathological T, N, and M stages, histopathological grade, lymphovascular invasion, pleural invasion), molecular alteration status (EGFR, KRAS, and ALK), tumor location indicators, and (recurrence, recurrence location, survival status, and time to death). The genetic features comprised a curated set of gene expression markers associated with tumor biology, immune regulation, extracellular matrix remodeling, hypoxia, and epithelial–mesenchymal transition. These included BGN, CD37, CD4, CD44, CD48, CDH2, COL4A1, COL5A1, COL5A2, EGR2, GDF15, HPGD, LMO2, LRIG1, LYL1, PDGFRA, POSTN, SPI1, VCAM1, VIM, and VCAN.

- Imaging Data Preprocessing

o PET scan De-noising: Deep CNN autoencoder trained from the de-noised PET dataset is utilized for PET scans' noise reduction. Hence, the images are of superior quality and suitable for the classification as well as the image fusion operation.

- Preprocessing of CT scan: The CT scans are normalized and contrast-enhanced for the purpose of bringing out notable anatomical features of importance in the classification model.
- Image Fusion: The CT is not metabolic in character but provides clear anatomical characteristics of the lung tissues and their environment. The PET is not anatomically accurate but indicates clearly regions of increased metabolic activity characteristic of tumor presence. The combined image, which outlines regions of increased metabolic activity as well as the fine texture of the lung tissues, not only overcomes the limitations of both modalities, but also offers an overview most sought after in medical imaging [12]. We employ the Image Fusion method in which the advantages of CT scans and PET scans are combined, as described in [12] (Fig. 2). Once the scans were preprocessed, the filtered input is supplied to the VGG19 fusion model. The algorithm makes use of discrete wavelet transformation for decomposing the CT scans and PET scans in terms of the coefficient LL1 and the detail coefficients LH1 (horizontal), LV1 (vertical), and LD1 (diagonal). The image fusion is produced by applying the inverse wavelet transformation in the four bands post the fusion of the four couples.

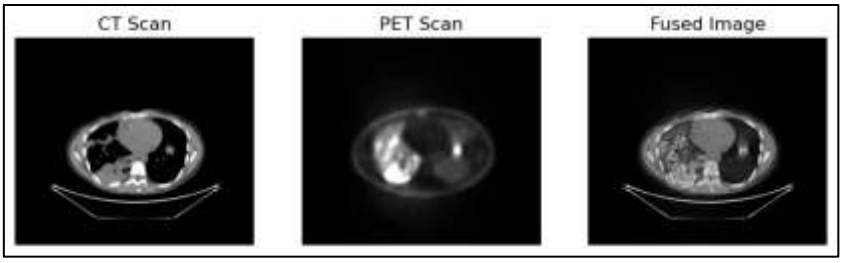



**Fig. 2**: Fusion of CT and PET [12]

- Data Augmentation: Several techniques of data augmentation such as random rotations, gaussian blur, random affine, flipping, as well as scaling are used for image data in order to improve the generalization of the model as well as avoid overfitting.

**Multi-modal Classification Model**

As mentioned above, our aim is to compare the three modes of fusion—Intermediate Fusion [12], Proposed Intermediate Fusion, and Proposed Late Fusion—for multimodal classification. All three modes of processing integrate tabular data (clinical and genetic data) and imaging data (CT and PET scans) in varied phases of the processing pipeline. We employ varying network structures, some of which use MLP based structures for processing tabular data and multi-modal feature integration, and others use the BEiT (Bidirectional Encoder Representation from Image Transformers) [29], which has 12 Transformer layers, for feature extraction for images.

**Intermediate Fusion Strategy**

The earlier intermediate fusion architecture, proposed by Hassan et al. [12], integrates image features and tabular features at the feature-level for the purpose of classification (Fig. 3). Image features are derived using pretrained Vision Transformer (BEiT) that generates an image feature of 768 dimensions from the last transformer layer. For tabular features, the input is projected through the feedforward network with

components including linear layer, ReLU, InstanceNorm1d, and dropout, resulting in an embedding of 64 dimensions. The combined 768 + 64 dimensions are utilized by a network of

1. Linear layer with GELU activation
2. Dropout
3. Feature stabilization with LayerNorm
4. A final linear layer which creates classification logit

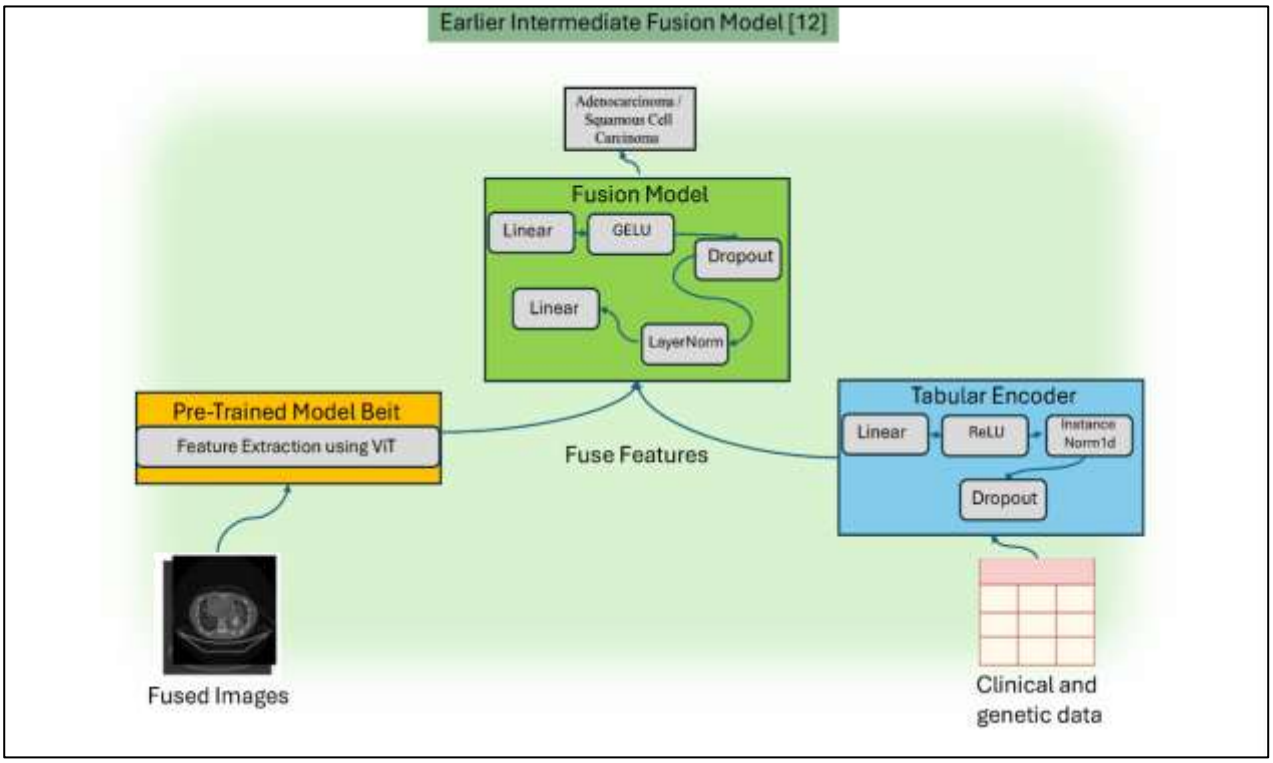



**Fig. 3**: Earlier Intermediate Fusion Architecture [12]: Image features (extracted via pretrained Vision Transformer) and tabular embeddings (encoded with feedforward networks) are concatenated and processed through a fusion network (GELU activation, dropout, LayerNorm) prior to classification.

**Proposed Intermediate Fusion**

Our intermediate fusion combines image and tabular features after modality-specific processing (Fig. 4). The Vision Transformer (ViT) extracts final-layer [CLS] token features (768-D), which are projected to 128-D through a GELU-activated linear layer with dropout. Tabular data is encoded to 64-D via ReLU-activated feedforward networks with InstanceNorm and dropout, then projected to 128-D. The concatenated 256-D feature vector (image + tabular features) goes through

1. Linear projection to 128-D with GELU
2. LayerNorm
3. Dropout regularization
4. Last classification projection (128-D → 2)

**Rationale** This model handles modalities separately before fusion to preserve individual feature hierarchies, then projects both modalities onto the same dimensions (128-D) prior to concatenation for dealing with embedding space imbalance, LayerNorm and dropout encourage generalization despite high-dimensional integration (256-D), and GELU activation allows non-linear cross-modal interaction modeling.

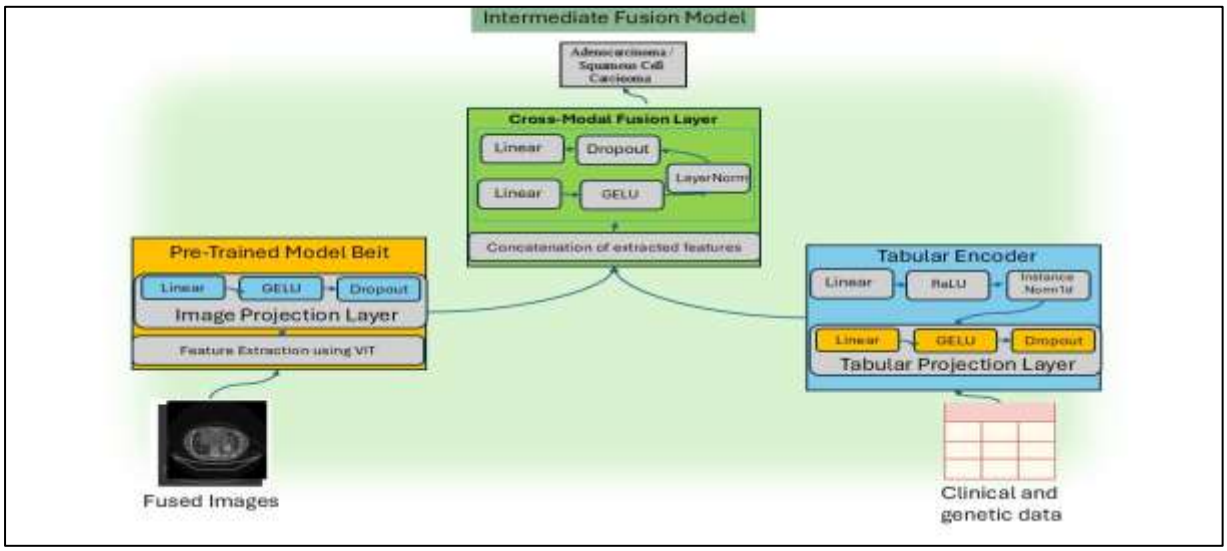



**Fig. 4**: Proposed Intermediate Fusion Model – This method extracts feature representations from both image and tabular data before fusing them at an intermediate stage.

**Proposed Late Fusion with Weighted Attention**

Our late fusion approach sends image and tabular data through separate pipelines before the predictions are fused (Fig. 5). The vision pipeline utilizes pre-trained ViT (BEiT) with: **Feature regularization:** LayerNorm and Dropout of [CLS] token features Image classifier (MLP with GELU activation, LayerNorm, and dropout). **The tabular branch:** Feature Encoder (MLP with LeakyReLU activation, LayerNorm, drop-out), Tabular classifier (GELU-activated MLP with LayerNorm + dropout.

**Fusion Mechanism:**

1. Modality attention: The networks compute image/tabular attention scores through Tanh-activated MLP
2. Feature projection: Image (768→128) and tabular (64→128) projected in common space.
3. Weighted combination: Weighted-sum of projected features is combined with softmax-normalized.
4. Final classifier: MLP with GELU, LayerNorm, and drop-out.

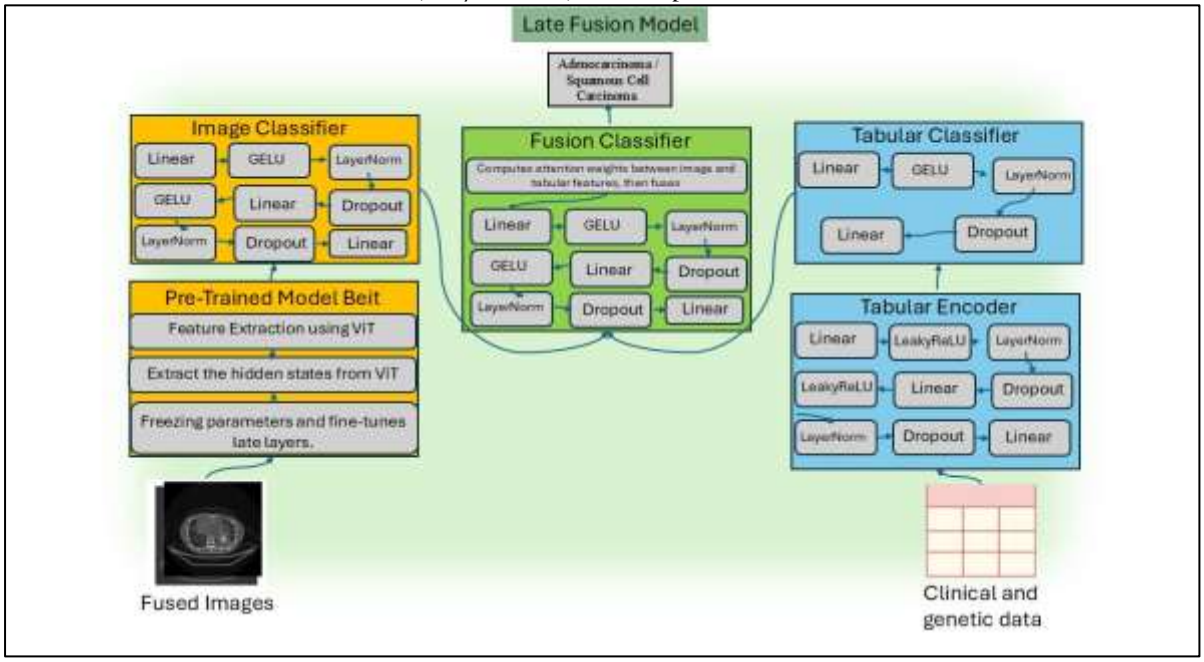



**Fig. 5**: Proposed Late Fusion Model – This approach processes image and tabular data independently, generating separate predictions for each modality. A weighted fusion mechanism is then applied to combine the predictions before the final classification.

**Model Training and Evaluation**

For the performance evaluation of the three fusion models, the dataset in this work is partitioned into training (60%), validation (15%), and testing (25%) sets. The models are trained using cross-entropy loss, and Adam optimizer as well as with learning rate scheduler. The models utilize dropout and instance normalization for enhancing the models' generalization and avoiding overfitting. The models are protected from overtraining with the use of early stopping, so that they possess sufficient generalization when used on unseen inputs. The performance of every fusion method is tested with the performance measures: Accuracy, Precision, Recall (Sensitivity), Specificity and F1-score.

**Experimental Results**

The performance of the three fusion strategies—proposed Late Fusion, proposed Intermediate Fusion, and earlier Intermediate Fusion [12]—was rigorously evaluated for NSCLC subtype classification (adenocarcinoma and squamous cell carcinoma) using fused CT/PET imaging and tabular data (clinical and genetic features). Table 1 summarizes the performance comparison in terms of accuracy, precision, recall/sensitivity, specificity, and F1-score (Fig. 6).

**Table 1**: Performance comparison of fusion strategies for NSCLC subtype classification (all values in %).

| Metric | Proposed Late Fusion | Proposed Intermediate Fusion | Intermediate Fusion [12] |
|---|---|---|---|
| Accuracy | 96.12 | 95.64 | 94.04 |
| Precision | 96.22 | 95.26 | 96.06 |
| Sensitivity | 95.5 | 95.5 | 91.13 |
| Specificity | 96.65 | 95.76 | 96.65 |
| F1-score | 95.86 | 95.38 | 93.52 |

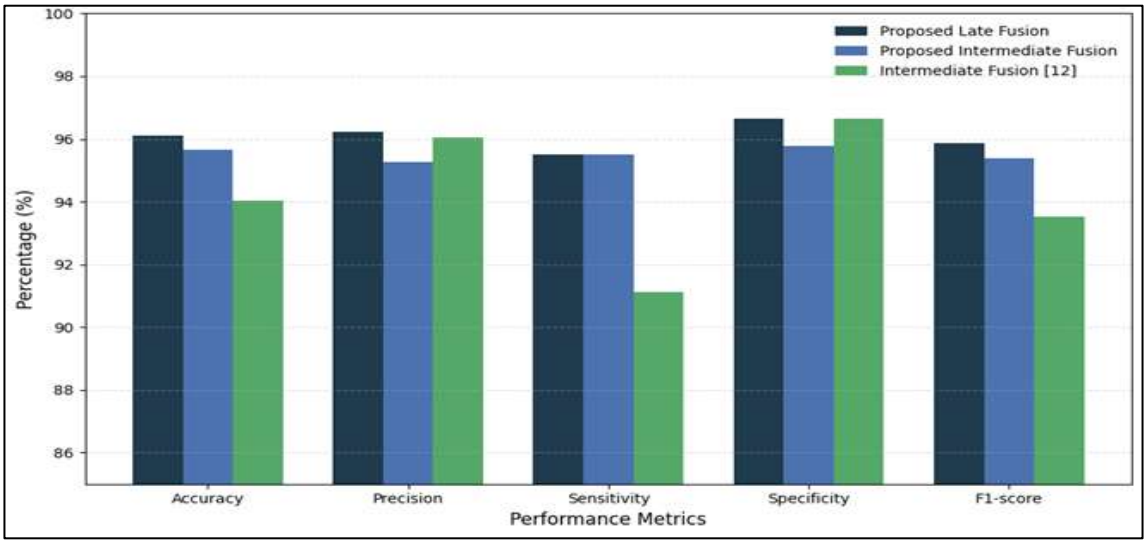



**Fig. 6**: Performance comparison of different fusion strategies for NSCLC subtype classification.

**Accuracy**

The proposed Late Fusion had the highest accuracy of 96.12%, with the proposed Intermediate Fusion (95.64%) ranking as the closest performing, while the Intermediate Fusion [12] had the lowest accuracy

with 94.04%. This indicates the superior ability of Late Fusion in combining multi-modal (image + tabular) input for subtyping NSCLC while not sacrificing discriminative features.

**Precision and Recall/Sensitivity**

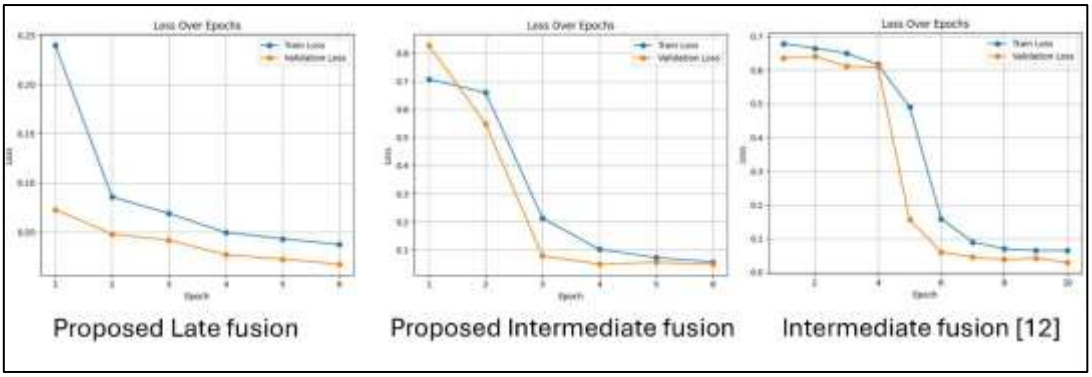
The proposed Late Fusion recorded the highest precision (96.22%) in eliminating false positives for subtype detection. The proposed Intermediate Fusion recorded slightly weaker precision (95.26%) but remained competitive. The Intermediate Fusion [12] recorded the lowest precision (96.04%). Both the proposed methods recorded the same recall/sensitivity (95.5%), which is an indication of robust detection of the positive samples, while Intermediate Fusion [12] recorded the lowest recall/sensitivity (91.13%).

**Specificity**

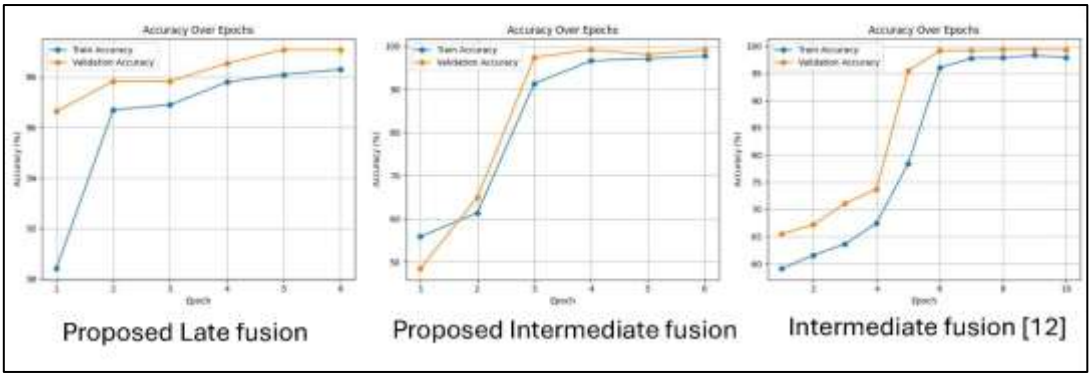
The proposed Late Fusion and the earlier Intermediate Fusion [12] ranked highest in specificity (96.65%), being critical in distinguishing squamous cell carcinoma from adenocarcinoma, and the proposed Intermediate Fusion get specificity (95.76%), with the low specificity in excluding false positives for histologically comparable subtypes.

**F1-Score**

The proposed Late Fusion and the proposed Intermediate Fusion achieved close F1- scores (95.86%, 95.38%), reflecting balanced precision-recall trade-offs. Earlier Intermediate Fusion [12] trailed with 93.52%, emphasizing its limitations in harmonizing multi-modal data.




(a)




(b)

**Fig. 7**: The validation and training (a) loss and (b) accuracy curves for the three fusion strategies.

**Model Explainability and Comparative Analysis Across Fusion Strategies**

To better understand how the proposed multimodal models arrive at their predictions, we conducted a post-hoc explainability analysis across three fusion strategies: Proposed Late Fusion, Proposed Intermediate Fusion, and the baseline Intermediate Fusion model reported in [12]. These architectures differ in the stage at which image and tabular information are integrated, which has implications for predictive behavior and interpretability. All models were implemented in PyTorch and evaluated in

inference mode using fixed, trained weights. Explainability analyses were performed on the test set using methods appropriate to the computational structure of each fusion strategy.

**Tabular Feature Attribution**

**Proposed Late Fusion Model**

In the proposed late-fusion model, image and tabular inputs are processed independently and combined only at the decision stage through learned fusion weights. This architectural distinction allows for the analysis of each modality separately without affecting the internal representation of the other modality. For analyzing the role of tabular features, the input to the network was passed through the tabular encoder and classification head, obtaining predictions solely on the basis of structured inputs. SHAP Gradient Explainer was used with a randomly chosen background dataset containing 100 samples from the training set. Feature importance on the test set was computed by aggregating mean absolute SHAP values across samples and output classes. The resulting attributions provide consistent and well-defined measures of tabular feature influence and are consistent with the strong predictive performance of the late-fusion model.

**Proposed Intermediate Fusion Model**

The proposed intermediate-fusion model integrates image and tabular representations at a shared latent level through a cross-fusion module, allowing interactions between modalities. Due to the coupled nature of this architecture, tabular feature attribution was performed with both image and tabular inputs present, ensuring that explanations reflect the model's true inference behavior. SHAP Gradient Explainer was applied to tabular inputs while maintaining the full multimodal forward pass. The same background sampling strategy used in the late-fusion model was adopted to ensure comparability. Global feature importance was obtained by aggregating mean absolute SHAP values across samples. In addition, permutation-based feature importance was used to assess the sensitivity of model predictions to individual tabular variables. In comparison to the late fusion method, the attributions presented in this work are less isolated. It is able to demonstrate the effect of cross-modal interaction that results from intermediate fusion, which corresponds with differences in performance levels achievable by both fusion techniques.

**Intermediate Fusion Model [12]**

The baseline intermediate-fusion model [12] performs feature-level fusion by concatenating encoded image and tabular representations at an intermediate stage, followed by joint classification. Tabular feature attribution was conducted using SHAP Gradient Explainer with both modalities present during inference, ensuring that explanations are consistent with the model's operational pathway. Feature importance was computed by aggregating mean absolute SHAP values across samples and output classes. Due to intermediate fusion and strong modality entanglement, tabular attributions in this model are more diffuse, indicating reduced separability between image and structured feature contributions. This characteristic is associated with the comparatively lower predictive performance of the baseline model.

**Image Explainability Using Vision Transformer Attention**

**Proposed Late Fusion Model**

Image explainability for the proposed late-fusion model was derived from Vision Transformer self-attention. Attention weights were extracted from the final transformer layer and averaged across heads to obtain CLS-to-patch attention scores. To reflect the contribution of the image modality to the final prediction, attention maps were scaled by the learned fusion weight corresponding to the image branch for the predicted class. This scaling is intended as a qualitative visualization aid rather than a causal attribution mechanism. This approach incorporates both spatial attention and modality relevance. The resulting attention maps are spatially localized and highlight regions consistent with clinically relevant image features.

**Proposed Intermediate Fusion Model**

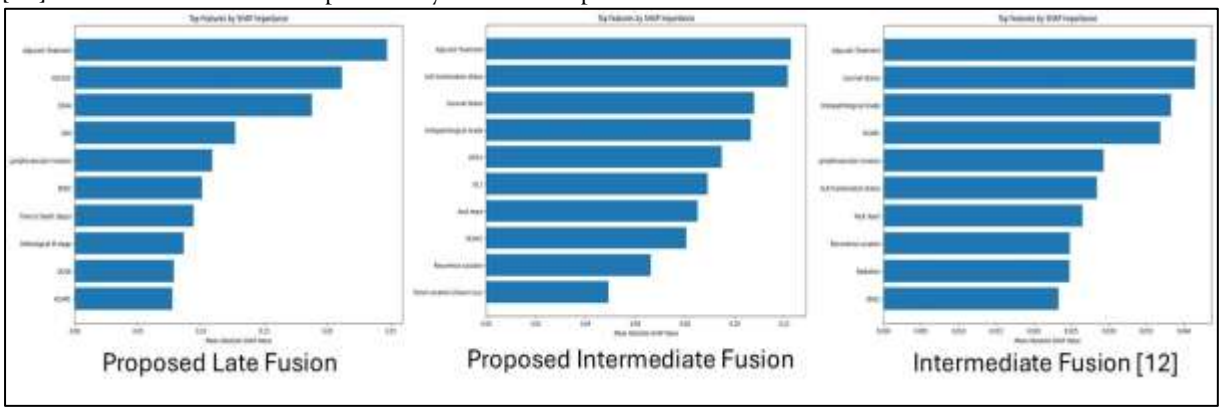
For the proposed intermediate-fusion model, image explanations were derived from Vision Transformer self-attention extracted from the final transformer layer. Because image representations interact with tabular features at the fusion stage, the resulting attention maps exhibit broader spatial patterns compared to the late-fusion model.

**Intermediate Fusion Model [12]**

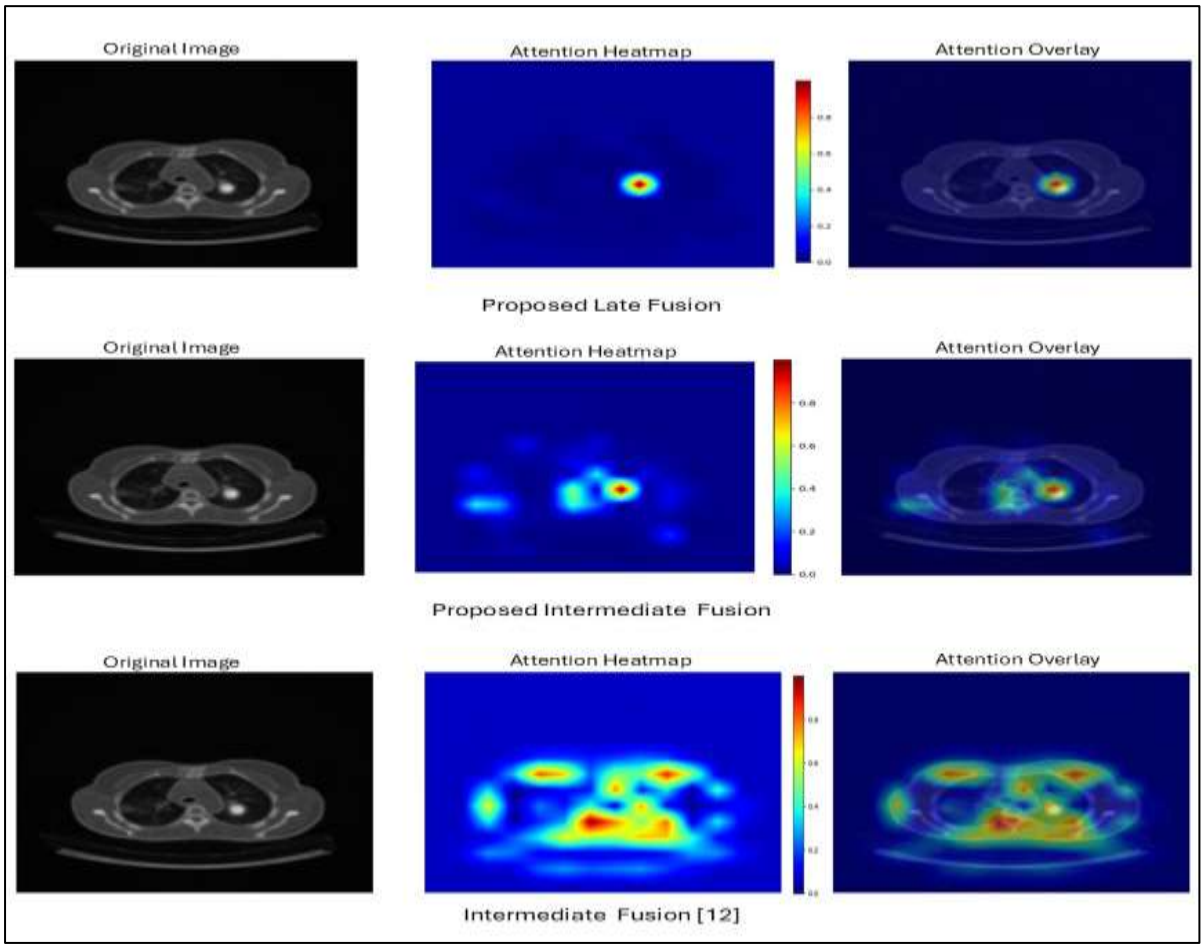
In the baseline intermediate-fusion model [12], image explanations were obtained from CLS-to-patch attention maps derived from the Vision Transformer encoder.

**Comparative Analysis Across Fusion Strategies**

Across all experiments, consistent trends were observed between fusion strategy, interpretability, and predictive performance (Fig. (8a, 8b)). The proposed late-fusion model achieves the highest classification accuracy and provides the most localized and stable explanations for both tabular and image modalities. The proposed model of intermediate fusion has good interaction between the modalities but only has moderate interpretability and performance results. In contrast, the baseline model of intermediate fusion [12] has lower levels of interpretability and model performance.




(a)




(b)

**Fig. 8**: (a) SHAP-based tabular feature importance and (b) Vision Transformer attention visualizations for the proposed late fusion, proposed intermediate fusion, and baseline intermediate fusion models, illustrating how fusion strategy affects feature attribution clarity and spatial localization of model attention.

## DISCUSSION

**The proposed Late Fusion** produced the highest accuracy (96.12%) and specificity (96.65%) as the superior method for subtype classification of NSCLC due to the processing of imaging (CT/PET) and tabular (clinical/genetic) modalities separately before their late-stage combination.

**The proposed Intermediate Fusion** (with accuracy 95.64%) worked almost as effectively as Late Fusion, indicating that partial cross-modal interaction in mid network processing can retain discriminative power. Its slightly lower specificity (95.76%) compared to Late Fusion, however, indicates subtle trade-offs in feature integration.

**The Earlier Intermediate Fusion [12]** got the lowest performance among other approaches in accuracy (94.04%), recall (91.13%), and F1-score (93.52%).

**Conclusion**

This work introduced two novel deep learning fusion methods-Late Fusion and Intermediate Fusion-to improve the integration of multi-modal data like fused CT/PET imaging and clinical-genetic tabular data, for subtyping of NSCLC (squamous cell carcinoma and adenocarcinoma). The work also compares the proposed Late and Intermediate Fusion methods with an earlier Intermediate Fusion method [12]. Experimentally, the work demonstrated that Late Fusion is superior in performance compared to other methods with accuracy 96.12% and specificity 96.65% maintaining modality-specific discriminative features. The proposed Intermediate Fusion offers an acceptable substitute with balanced performance (with accuracy 95.64%), whereas the earlier Intermediate Fusion [12] is not as suitable for precision-critical applications. In addition to performance gains, this work underlines the importance of explainability in multimodal clinical decision support systems as a whole. Through the combination of post-hoc explainability methods such as SHAP feature attribution for tabular data and Vision Transformer attention analysis for imaging data, the proposed system provides an indication of the influence of different fusion strategies on decision-making processes. The work emphasizes the importance of the integration timing of multimodal learning and recommends Late Fusion in clinical decision-support systems for enhancing diagnostic accuracy while preserving interpretable, ultimately supporting improved patient outcomes.

Further studies are necessary for developing data-efficient, explainable hybrid models that can successfully integrate imaging and genetic data for improving the diagnosis and prognostication of NSCLC, as well as for validating explainability findings in prospective clinical settings.

**REFERENCES**

[1] Society, A.C.: Key Statistics for Lung Cancer. Accessed: 2023-10-16 (2023)

[2] Travis, W.D., Brambilla, E., Nicholson, A.G., Yatabe, Y., Austin, J.H., Beasley, M.B., Wistuba, I.: The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. Journal of Thoracic Oncology **10**(9), 1243–1260 (2015)

[3] Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Walsh, S.: Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology **14**(12), 749–762 (2017)

[4] Pennell, N.A., Mutebi, A., Zhou, Z.Y., Ricculli, M.L., Tang, W., Wang, H., Otterson, G.A.: Economic impact of next-generation sequencing versus single-gene testing to detect genomic alterations in metastatic non–small-cell lung cancer using a decision analytic model. JCO Precision Oncology **3**, 1–9 (2019)

[5] Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Lambin, P.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**(1), 4006 (2014)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[7] Henry, E.U., Emebob, O., Omonhinmin, C.A.: Vision transformers in medical imaging: A review. arXiv preprint arXiv:2211.10043 (2022)

[8] Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J.: Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics **23**(2), 569 (2022)

[9] Kumar, S., Sharma, S., Megra, K.T.: Transformer enabled multi-modal medical diagnosis for tuberculosis classification. Journal of Big Data **12**(5) (2025) https: //doi.org/10.1186/s40537-024-01054-w

[10] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., S´anchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)

[11] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[12] Hassan, S., Al Hammadi, H., Mohammed, I., Khan, M.H.: Multi-modal medical image fusion for non-small cell lung cancer classification. In: 2024 IEEE International Conference on Image Processing (ICIP), pp. 3091–3097 (2024). IEEE

[13] (NCCN), N.C.C.N.: Non-Small Cell Lung Cancer. Accessed: 2023-10-16 (2023)

[14] Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)
[15] Hosny, A., Parmar, C., Coroller, T.P., Grossmann, P., Zeleznik, R., Kumar, A., Aerts, H.J.: Deep learning for lung cancer prognostication: a retrospective multicohort radiomics study. PLOS Medicine **15**(11), 1002711 (2018)
[16] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical selfsupervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)
[17] Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B., Veeriah, S., Swanton, C.: Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine **376**(22), 2109–2121 (2017)
[18] Liu, J., Yang, M., Yu, Y., Xu, H., Li, K., Zhou, X.: Large language models in bioinformatics: applications and perspectives. arXiv (2024). arXiv:2401
[19] Baltru˘saitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018)
[20] Park, C., Ha, J., Park, S.: Prediction of alzheimer's disease based on deep neural network by integrating gene expression and dna methylation dataset. Expert Systems with Applications **140**, 112873 (2020)
[21] Deng, R., Shaikh, N., Shannon, G., Nie, Y.: Cross-modality attention-based multimodal fusion for non-small cell lung cancer (nsclc) patient survival prediction. In: Medical Imaging 2024: Digital and Computational Pathology, vol. 12933, pp. 46–50. SPIE, Bellingham, WA, USA (2024). https://doi.org/10.1117/12.3006036
[22] Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017)
[23] Aksu, F., Gelardi, F., Chiti, A., Soda, P.: Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from ct and pet. arXiv preprint arXiv:2501.12425 (2025)
[24] Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Napel, S.: A radiogenomic dataset of non-small cell lung cancer. https://doi.org/10.7937/ K9/TCIA.2017.7hs46erv (2018)
[25] Aerts, H.J.W.L., Wee, L., Rios Velazquez, E., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P.: Data From NSCLC-Radiomics (version 4). https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI. The Cancer Imaging Archive. Data set (2014)
[26] Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., Wang, D.: A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Version 5) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.NNC2-0461 (2020)
[27] Aerts, H.J.W.L., Rios Velazquez, E., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P.: Data From NSCLC-Radiomics-Genomics. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z (2015)
[28] Zhou, M., Leung, A., Echegaray, S., Gentles, A., Shrager, J.B., Jensen, K.C., Berry, G.J., Plevritis, S.K., Rubin, D.L., Napel, S., Gevaert, O.: Non–small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology **286**(1), 307–315 (2018) https://doi.org/10.1148/radiol.2017161845
[29] Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)