American Journal of Biochemistry and Biotechnology

Research Paper

CPF-Net: Cross-modal CT and Pathology Guided Feature Learning for CT-based Lung Cancer Subtype Classification

Peizhi Tan, Debiao Yan

School of Artificial Intelligence & Big Data, Luzhou Vocational & Technical College, Luzhou 646000, China

Article history
Received: 15 January 2025
Revised: 17 May 2025
Accepted: 17 July 2025

*Corresponding Author: Peizhi Tan, Luzhou Vocational & Technical College, Luzhou 646000, China;
Email: tanpeizhi@lzy.edu.cn

Abstract: Accurate classification of lung cancer subtypes from CT images remains challenging due to the subtle radiological differences between adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). We propose CPF-Net, a deep learning framework that integrates CT and pathological information through a Linear Spatial Reduction Attention (LSRA) module. The framework processes whole slide images using a modified CTransPath architecture for pathological feature extraction and combines these features with CT imaging characteristics during training. While both CT and pathological data are used in training, only CT images are required for inference. Experiments on a dataset of 892 cases from The Cancer Genome Atlas (TCGA) show that CPF-Net achieves 87.89% accuracy, 93.23% AUC, and 86.92% F1-score, outperforming existing methods by margins of 4.44%, 3.67%, and 4.14% respectively. Ablation studies demonstrate the effectiveness of both the LSRA module and the cross-modal learning strategy in improving classification performance.

Keywords: Lung Cancer Subtype Classification; Deep Learning; Cross-modal Learning; CT Images; Pathological Features; Attention Mechanism

Introduction

Lung cancer remains one of the most devastating malignancies worldwide, with its mortality rate surpassing that of other common cancers. This aggressive disease accounts for approximately 25% of all cancer-related deaths globally, presenting a significant challenge to public health systems (Wu *et al.*, 2020). While lung cancer encompasses various histological types, it is primarily categorized into two major groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter represents the predominant form, comprising roughly 85% of all cases, with adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) being the most frequently diagnosed subtypes (Cancer Genome Atlas Research Network, 2012). Understanding these distinct pathological entities is crucial, as they exhibit unique molecular profiles and demonstrate varying responses to therapeutic interventions.

The diagnosis of lung cancer relies on multiple clinical modalities, with imaging techniques playing a central role in the diagnostic workflow. While traditional methods such as chest radiography and bronchoscopy remain valuable tools, computed tomography (CT) has emerged as the cornerstone of non-invasive lung cancer detection and characterization. CT imaging provides comprehensive three-dimensional anatomical information, enabling detailed assessment of tumor characteristics including morphology, spatial distribution, metastatic status, and heterogeneity (Zhang *et al.*, 2019; Hussain *et al.*, 2022). Although certain radiological features can serve as diagnostic indicators for specific lung cancer subtypes, the interpretation of these imaging findings remains heavily dependent on clinical expertise, leading to potential inter-observer variability (E *et al.*, 2019; Li *et al.*, 2021). Moreover, early-stage tumors often lack distinctive radiological presentations, making subtle pathological changes challenging to detect through conventional visual assessment. These limitations underscore the pressing need for sophisticated computer-aided CT analysis systems capable of accurate lung cancer subtype classification.

Deep learning (DL) has emerged as a promising approach to address these diagnostic challenges through automated quantitative analysis (Shao *et al.*, 2024; Wehbe *et al.*, 2024; Tong *et al.*, 2024). By leveraging end-to-end deep neural networks, these systems can automatically extract and analyze high-dimensional features from radiological images, enabling quantitative identification of subtle imaging patterns associated with different pathological conditions. Significant advances have been achieved in CT image classification through various innovative convolutional neural networks (CNNs), attention mechanisms, and transformer-based architectures (Sohaib *et al.*, 2025; Al-Antari *et al.*, 2021; Khalifa and Albadawy, 2024; Pan *et al.*, 2025). These strategies offer physicians potentially faster and more accurate diagnostic support compared to traditional visual assessment. However, the complex task of automatic cancer subtype classification from CT images continues to present challenges, with current models showing limitations in classification accuracy and robustness. These constraints are partly attributed to atypical radiological presentations in certain cases, while the inherent redundancy and noise in raw CT images pose additional obstacles to achieving optimal performance in DL algorithms (Qi *et al.*, 2019; Zhang *et al.*, 2019). Moreover, while the potential benefits of integrating information from different modalities, such as CT and pathology, have been recognized, previous attempts at cross-modal feature learning have often faced difficulties. Many earlier methods relied on relatively simple fusion strategies, such as direct feature concatenation, which may not adequately capture the intricate, non-linear relationships between imaging features and underlying pathological characteristics, or struggled with effectively aligning and harmonizing data from disparate sources and scales.

Histopathological examination remains the gold standard in cancer diagnosis, providing crucial microscopic insights into cellular architecture, differentiation patterns, and tissue organization (Davri *et al.*, 2022). The integration of this detailed pathological data with radiological findings could potentially enhance the accuracy of diagnostic models and improve subtype classification. However, obtaining pathological specimens presents significant clinical challenges, as it requires invasive procedures such as surgical resection or needle biopsy (Witowski *et al.*, 2022). These interventional approaches carry inherent risks, including bleeding, infection, and procedure-related complications, making them unsuitable for certain patient populations, particularly those with compromised health status or challenging tumor locations (Mukund *et al.*, 2019). Consequently, while pathological examination offers unparalleled diagnostic precision, its application in early-stage diagnosis may be constrained by practical and clinical considerations, necessitating alternative diagnostic strategies.

The relationship between radiological and pathological manifestations of disease represents a fascinating bridge across different spatial scales of biological observation. CT images and histopathological slides, while examining the same underlying pathology, provide complementary perspectives at macro and microscopic levels respectively. Recent investigations have revealed significant correlations between these modalities in lung cancer assessment (Acharya *et al.*, 2017; Walls *et al.*, 2022). Studies have demonstrated meaningful associations between CT-derived features and underlying biological characteristics, such as the correlation between tumor vascularity patterns on contrast-enhanced CT and histological markers of angiogenesis (Gill *et al.*, 2020). Particularly in NSCLC, researchers have identified specific relationships between radiological signatures and histopathological parameters, including correlations between CT attenuation patterns and cellular organization (Alvarez-Jimenez *et al*., 2020). These cross-scale associations extend to prognostic applications, where radiological features reflecting tissue architecture have shown potential in predicting treatment outcomes. Such established relationships between imaging and pathological characteristics suggest the possibility of developing advanced computational methods to extract latent pathological information directly from CT images, potentially enhancing non-invasive diagnostic capabilities.

In this study, we present a Cross-modal Pathology-guided Feature Network (CPF-Net) for lung cancer subtype classification from CT images. Building on cross-modal correlations between radiological and pathological imaging, our approach leverages whole slide images (WSI) as the pathological gold standard. We develop an attention-based learning mechanism that automatically identifies high-diagnostic-value regions within the WSI and encodes them into representative feature vectors. This encoding process is refined through instance-level clustering, which constrains and optimizes the feature space of the identified regions. The encoded pathological features are then integrated with CT imaging features through a fusion framework, where we deliberately bias the

integration toward CT features while maintaining the guiding influence of pathological information. Our model's architecture utilizes paired CT and WSI data during training, while only CT images are needed for subsequent validation and clinical application. This approach addresses limitations of single-modality diagnosis and enables autonomous generation of hybrid features in clinical settings while relying solely on CT imaging input. The pathology-guided strategy can be incorporated into various state-of-the-art classification networks without additional computational overhead.

Methods

Our framework integrates radiological and pathological data through a three-component architecture to achieve robust lung cancer subtype classification. At its core, the model employs a radiological feature encoder that extracts diagnostic patterns from CT images, working in parallel with a pathological feature encoder that processes WSIs to capture tissue-level characteristics. These complementary feature streams converge in a dedicated fusion component, which harmonizes the multi-modal information into a unified representation. Each component has been specifically designed to maximize the complementary strengths of both imaging modalities while maintaining computational efficiency. The following subsections provide comprehensive details about the implementation and operational principles of each architectural component. The overall architecture of our proposed method is illustrated in Fig. 1.

Preprocessing of WSIs

WSIs pose computational challenges due to their high dimensionality and multi-resolution nature. In our dataset, each WSI contains approximately 127,655 × 53,444 pixels at its highest magnification level, making direct processing computationally prohibitive. To address these challenges, we implement a systematic preprocessing pipeline that effectively reduces computational complexity while preserving essential pathological information.

Our preprocessing framework employs an enhanced version of the CLAM algorithm (Lu *et al.*, 2021), which has been specifically modified to maintain consistent processing across diverse WSI samples. The framework operates on a four-level pyramid structure, where each level represents a different downsampling ratio: the original resolution (level 0), 4× downsampled (level 1), 16× downsampled (level 2), and 32× downsampled (level 3). This multi-resolution approach enables efficient navigation through different magnification levels while maintaining the ability to access detailed cellular information when needed.

To ensure standardization across our dataset, all WSIs are processed at 20× magnification, corresponding to approximately 0.5 microns per pixel. This magnification level was chosen as it provides an optimal balance between computational efficiency and preservation of diagnostically relevant cellular details. Specifically, 20× magnification is widely adopted in digital pathology for capturing sufficient cellular and architectural detail, allowing for the visualization of key diagnostic features such as nuclear morphology, cytoplasmic characteristics, glandular formations in adenocarcinoma, and keratinization or intercellular bridges in squamous cell carcinoma. These features are crucial for distinguishing between LUAD and LUSC. While higher magnifications (e.g., 40×) offer more detail, they significantly increase the computational load for WSI processing and feature extraction due to the vastly larger number of patches generated, without a commensurate gain in discriminative power for this particular subtype classification task. Conversely, lower magnifications might obscure subtle but critical diagnostic features.

Following tissue segmentation, we implement a systematic patch extraction protocol using a sliding window approach with carefully selected parameters. The patch size is set to 256×256 pixels, with a step size equal to the patch size to avoid overlap. This configuration ensures comprehensive coverage of tissue regions while maintaining computational efficiency. The extracted patches are stored in HDF5 format (.h5 files), which provides efficient data organization and rapid access during subsequent processing stages.

This preprocessing approach effectively addresses the computational challenges posed by high-dimensional WSIs while maintaining the integrity of pathologically relevant information. The standardized patch extraction process forms a robust foundation for subsequent feature extraction and cross-modal learning stages in our framework. Experimental validation demonstrates that our preprocessing pipeline successfully processes diverse WSI samples while maintaining high computational and storage efficiency across the entire dataset, as shown in Fig. 2.

DOI:10.3844/ajbbsp.2025.386.400

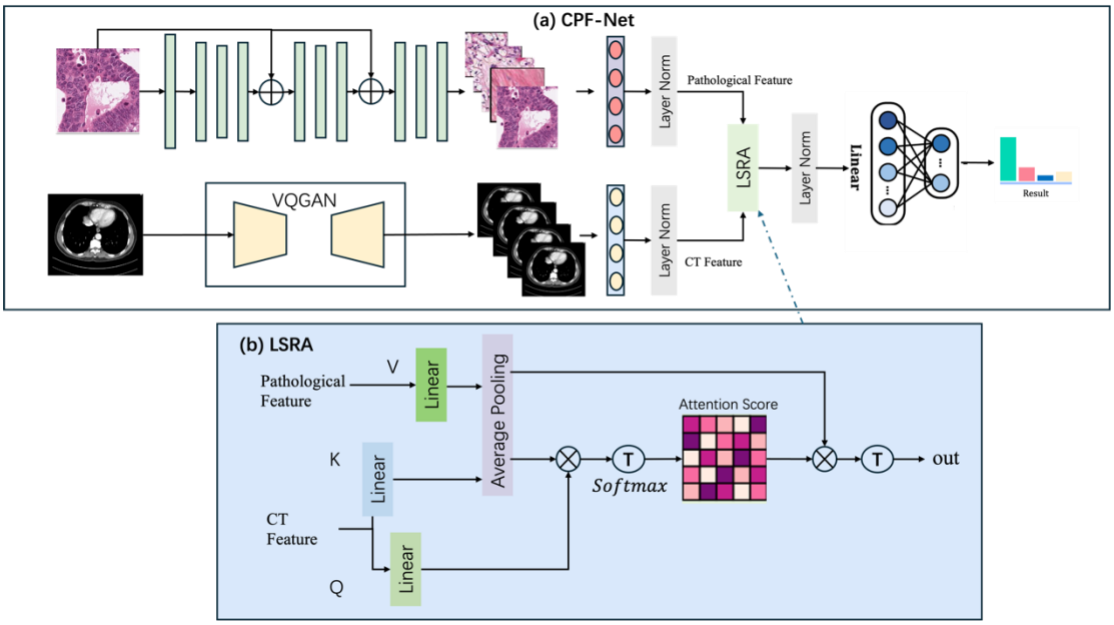

Fig. 1. Overview of the proposed CPF-Net architecture: (a) The overall framework consists of two branches: a pathological feature extraction branch that processes WSI data through convolutional layers, and a CT feature extraction branch using VQGAN. The features from both branches are integrated through our LSRA module before final classification; (b) Detailed structure of the Linear Spatial Reduction Attention (LSRA) module, which efficiently fuses pathological and CT features through linear projections, average pooling, and attention mechanism to generate the final hybrid features for classification.

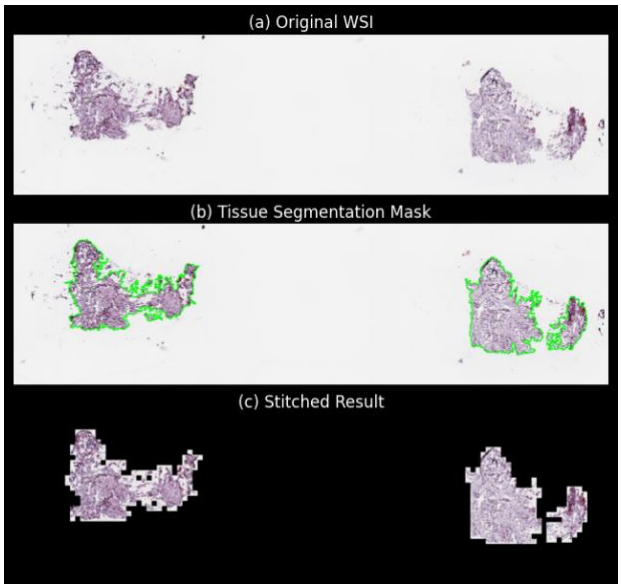

Fig. 2. Visualization of the WSI processing pipeline: (a) Original WSI at low magnification showing the complete tissue section; (b) Tissue segmentation mask highlighting regions of interest in green, demonstrating effective separation of tissue from background; (c) Stitched visualization at 20× magnification showing the processed tissue regions, where valid tissue patches have been extracted and reconstructed while excluding background areas.

Pathological Feature Extraction

The core pathological information, derived from H&E-stained Whole Slide Images (WSIs), is encoded into quantitative feature vectors. This process specifically utilizes the visual data from histopathology slides and does not incorporate genomic data or handcrafted pathological radiomics features. Following the WSI preprocessing stage, which results in a collection of 256×256 pixel tissue patches at 20× magnification, we implement a deep learning-based feature extraction pipeline to transform these patches into rich feature representations.

Following the preprocessing stage, we implement a feature extraction pipeline to transform the tissue patches into feature representations. The feature extraction process operates on the preprocessed 256×256 pixel patches at 20× magnification, maintaining consistency with the earlier preprocessing stage.

The feature extraction employs CTransPath (Wang *et al.*, 2021), which integrates CNNs with a transformer architecture. The model begins with a convolutional stem layer for local feature processing, followed by a Swin Transformer backbone (Liu *et al.*, 2021). The convolutional stem processes input patches through a series of operations:

$$F_1(x) = \text{ReLU}\left(\text{BN}\left(\text{Conv3} \times 3(x)\right)\right) \quad (1)$$

$$F_2(x) = \text{ReLU}\left(\text{BN}\left(\text{Conv3} \times 3\left(F_1(x)\right)\right)\right) \quad (2)$$

$$F_{out}(x) = \text{Conv1} \times 1(F_2(x)) \quad (3)$$

where each stage adjusts the feature dimensions through convolution operations. The Swin Transformer component then processes these features through window-based self-attention mechanisms:

$$Z_l = \text{W-MSA}\left(\text{LN}(Xl-1)\right) + X_{l-1} \quad (4)$$

$$X_l = \text{MLP}(\text{LN}(Z_l)) + Z_l \quad (5)$$

The implementation utilizes GPU acceleration with batch processing of 128 patches. The window-based attention mechanism in the Swin Transformer reduces computational complexity compared to standard transformer approaches. The extracted features are organized in a hierarchical directory structure that mirrors the organization of the input patches, facilitating integration with other components of the framework.

The feature extraction process and its effectiveness are visualized in Fig. 3. For each tissue patch (a), we show the intermediate convolutional features (b), the final extracted feature representations after dimensionality reduction (c), and the corresponding attention maps (d), demonstrating how the model captures both local and global tissue characteristics.

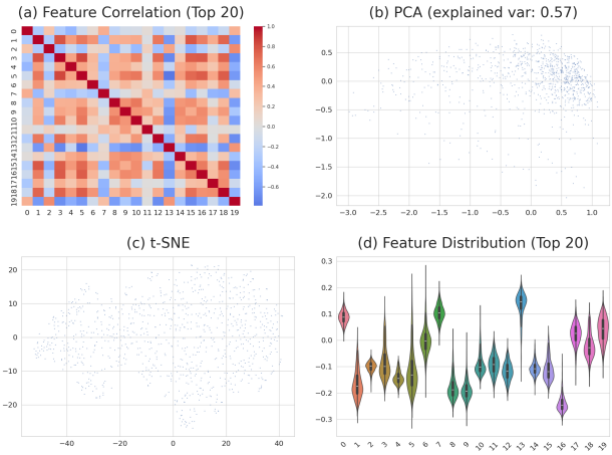

Fig. 3. Feature Analysis of CTransPath Extracted Features. Visualization of features extracted by CTransPath model showing: (a) correlation matrix of top 20 features demonstrating feature relationships; (b) PCA projection showing global feature distribution in 2D space with 57% explained variance; (c) t-SNE embedding revealing local structure and potential clusters; (d) violin plots displaying the distribution of top 20 feature values across all patches, indicating the range and density of extracted features.

CT Feature Extraction

The extraction of discriminative features from CT images presents unique challenges due to their three-dimensional nature and complex tissue representations. To address these challenges, we employ a Vector Quantized Generative Adversarial Network (VQGAN) architecture (Cao *et al.*, 2023), which effectively captures both local and global characteristics of CT volumes while maintaining computational efficiency. Our VQGAN implementation processes CT images through a hierarchical encoder-decoder structure with a discrete latent space. The encoder E maps input CT images x into a latent space z = E(x), which is then quantized using a codebook of learned representations. The quantization process can be expressed as:

$$z_q = q(z) = argmin \mid\mid z_k - e_k \mid\mid_2 \quad (6)$$

where z_k represents the encoded features and e_k denotes the codebook entries. This quantization step helps in learning discrete representations that capture essential radiological patterns while reducing noise and redundancy in the feature space.

The VQGAN architecture consists of an encoder with sequential convolutional blocks, each incorporating 2D convolution layers with 3×3 kernels, group normalization, and ReLU activation functions. The decoder mirrors this structure with transposed convolutions, enabling effective reconstruction of input images while maintaining feature integrity. The training process optimizes multiple objectives through a combined loss function:

$$L_{total} = \lambda_{rec}L_{rec} + \lambda_{adv}L_{adv} + \lambda_{per}L_{per} + \lambda_{vq}L_{vq} \quad (7)$$

where L_{rec} represents reconstruction loss, L_{adv} denotes adversarial loss, L_{per} indicates perceptual loss, and L_{vq} represents vector quantization loss. These components work together to ensure the extraction of robust and meaningful features from CT images.

To process 3D CT volumes efficiently, we implement a slice-wise approach that maintains spatial context through a sliding window mechanism. The preprocessing stage includes standardization of CT values to a [-1000, 1000] HU range, uniform voxel spacing resampling, and intensity normalization. The network processes input images at 256×256 pixel resolution, utilizing a codebook size of 1024 entries and producing feature vectors of dimension 256. Training is conducted using the Adam optimizer with a learning rate of $2e^{-4}$ over 100 epochs, achieving stable convergence and robust feature extraction capabilities.

Cross-modal Feature Fusion

For effective integration of CT and pathological features, we propose a novel cross-modal fusion module termed LSRA (Liu *et al.*, 2024). This module is designed to efficiently capture salient interactions between the two modalities. Unlike traditional attention mechanisms that compute attention across all spatial locations, leading to high computational costs (especially with high-resolution CT features), LSRA incorporates two key modifications: spatial reduction and linear projections. The spatial reduction step significantly reduces the dimensionality of the query features before attention calculation, thereby decreasing memory requirements and computational load. Linear projections are used to transform features into query, key, and value representations suitable for the attention mechanism. This design allows for robust feature fusion while maintaining computational feasibility, making the model more practical for clinical deployment.

The fusion process begins with encoding spatial information into both CT features $F_{ct} \in R^{B\times H\times W\times C}$ and pathological features $F_p \in R^{B\times N\times C}$, where B is the batch size, H and W are the spatial dimensions of CT features, N is the number of pathological feature vectors (e.g., from WSI patches), and C is the channel dimension. Positional embeddings are added to the features, which are then normalized using LayerNorm to stabilize training dynamics:

$$F'_{ct} = \text{LayerNorm}\big(F_{ct} + \text{PositionEmbed}(F_{ct})\big) \quad (8)$$

$$F'_p = \text{LayerNorm}\left(F_p + \text{PositionEmbed}\big(F_p\big)\right) \quad (9)$$

The LSRA module then processes these normalized features. To prepare for the attention computation, three separate linear projection layers, with learnable weight matrices W_q, W_k, W_v, transform the input features into query (Q), key (K), and value (V) representations. A crucial aspect of our design is that both Q and V are derived from the CT features F'_{ct}, while K is derived from the pathological features F'_p. This configuration ensures that the attention mechanism focuses on refining and weighting the CT features (via Q and V) based on their relevance to the pathological features (via K), effectively allowing pathological insights to guide the interpretation of CT data. The projections are:

$$Q = W_q(F'_{ct}) \in R^{B\times H\times W\times d} \quad (10)$$

$$K = W_k\big(F'_p\big) \in R^{B\times N\times d} \quad (11)$$

$$V = W_v(F'_{ct}) \in R^{B\times H\times W\times d} \quad (12)$$

Here, d represents the dimension of the attention heads.

To achieve computational efficiency, a spatial reduction operation, specifically average pooling, is applied to the query Q before the attention calculation. This reduces its spatial dimensions $H \times W$ to $(H/r) \times (W/r)$, where r is the reduction ratio. This reduced query, Q_r, captures broader contextual information from the CT features with lower granularity:

$$Q_r = \text{AvgPool}(Q) \in R^{B\times(H/r)\times(W/r)\times d} \quad (13)$$

The attention scores A are then computed by taking the dot product of the reduced query (Q_r) and the transpose of the key (K^T), scaled by the square root of the attention head dimension d, followed by a softmax activation. These scores reflect the importance of each pathological feature vector in K for each spatially reduced region in Q_r:

$$A = \text{Softmax}\left((Q_r K^T)/\sqrt{d}\right) \in R^{B\times(H/r)\times(W/r)\times N} \quad (14)$$

The resulting attention map A is then used to weight the value V. However, to capture diverse feature relationships at multiple semantic levels, we employ a multi-head attention mechanism. The input Q_r, K, and V are linearly projected into M different subspaces (heads), where attention is computed independently:

$$\text{MultiHead}(Q_r, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_M)W_o \quad (15)$$

where each head_i is computed as:

$$\text{head}_i = \text{Attention}\big(Q_r W_i^Q, K W_i^K, V W_i^V\big) \quad (16)$$

The outputs of the M heads are concatenated and linearly projected by W_o to produce the final multi-head attention output. This output, which represents CT features modulated by pathological guidance, undergoes further processing by a feed-forward network (FFN). The FFN consists of two linear transformations with a GELU activation function in between, allowing for further feature refinement:

$$\text{FFN}(x) = W_2\left(\text{GELU}\big(W_1(x)\big)\right) \quad (17)$$

Finally, the fused features are obtained by adding the output of the FFN back to the original CT features (F_{ct}) via a residual connection, followed by layer normalization. This residual connection helps in preserving the original CT information while incorporating the attended cross-modal insights:

$$F_{fused} = \text{LayerNorm}\big(\text{FFN}\big(\text{MultiHead}(Q_r, K, V)\big) + F_{ct}\big) \quad (18)$$

Loss Function

The training objective of our network comprises multiple loss terms that jointly optimize classification performance while ensuring effective cross-modal feature alignment. The primary classification task is supervised through a cross-entropy loss applied to the network's predictions:

$$L_{cls} = -\sum_{i=1}^{N} y_i \log(\hat{y}i) \quad (19)$$

where y_i represents the ground truth label and $\hat{y}i$ denotes the predicted probability for the i-th sample.

To enhance the cross-modal learning process, we introduce a feature alignment loss that minimizes the distributional discrepancy between CT and pathological feature spaces. This alignment is achieved through the Kullback-Leibler (KL) divergence:

$$L_{align} = D_{KL}(P_{ct} \,||\, P_{path}) = \sum P_{ct}(x) \log \frac{P_{ct}(x)}{P_{path}(x)} \quad (20)$$

where P_{ct} and P_{path} represent the probability distributions of CT and pathological features respectively.

Additionally, we incorporate a regularization term to prevent overfitting and ensure smooth feature fusion:

$$L_{reg} = \lambda_1 \,||\, W_{ct} \,||\, 2^2 + \lambda_2 \,||\, Wpath \,|\, |\, 2^2 \quad (21)$$

where W_{ct} and W_{path} denote the weights associated with CT and pathological feature processing, and λ_1, λ_2 are balancing hyperparameters.

The total loss function is formulated as a weighted combination of these components:

$$L_{total} = \alpha L_{cls} + \beta L_{align} + \gamma L_{reg} \quad (22)$$

where α, β, and γ are empirically determined weighting coefficients that balance the contribution of each loss term. Through extensive experimentation, we set α= 1.0, $\beta = 0.1$, and $\gamma = 0.01$ to achieve optimal performance.

Evaluation Metrics

To comprehensively assess the performance of our proposed model in lung cancer subtype classification, we employ a diverse set of evaluation metrics that capture different aspects of classification performance. The fundamental binary classification metrics are calculated from the confusion matrix elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

The classification accuracy measures the overall correct prediction rate:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (23)$$

To evaluate the model's performance for each class independently, we calculate precision and recall:

$$Precision = \frac{TP}{TP + FP} \quad (24)$$

$$Recall = \frac{TP}{TP + FN} \quad (25)$$

The F1-score provides a balanced measure of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (26)$$

For clinical relevance, we specifically evaluate sensitivity and specificity:

$$Sensitivity = \frac{TP}{TP + FN} \quad (27)$$

$$Specificity = \frac{TN}{TN + FP} \quad (28)$$

The area under the receiver operating characteristic curve (AUC-ROC) quantifies the model's ability to discriminate between classes across various classification thresholds:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (29)$$

where TPR represents the true positive rate and FPR the false positive rate.

Results

Dataset

Our experiments were conducted using data from The Cancer Genome Atlas (TCGA) program (CelebA Dataset-Machine Learning Datasets (Liu *et al.*, 2015)), a comprehensive public database that provides matched clinical, genomic, and imaging data for various cancer types. We specifically focused on lung cancer cases, collecting paired CT scans and WSIs for patients diagnosed with either lung adenocarcinoma (LUAD) or lung squamous cell carcinoma (LUSC). The dataset compilation process involved several steps of careful curation and quality control. First, we identified cases with both diagnostic quality CT scans and corresponding WSI data. The CT scans were required to meet the

following criteria: complete chest CT series with slice thickness ≤ 3mm, contrast-enhanced imaging protocol, absence of severe motion artifacts, and adequate visualization of the primary tumor. For WSIs, we selected H&E-stained slides that contained representative tumor tissue, were of diagnostic quality without significant artifacts, and had sufficient tumor content (>30% tumor cells). After applying these selection criteria, our final dataset consisted of 892 cases, comprising 514 LUAD and 378 LUSC samples. To ensure robust model development and evaluation, we randomly partitioned the dataset while maintaining the class distribution across all splits. The training set contained 624 cases (360 LUAD, 264 LUSC), representing 70% of the total data. The remaining cases were equally divided between validation and testing sets, with each containing 134 cases (77 LUAD, 57 LUSC), corresponding to 15% of the total data respectively. This stratified splitting approach ensured consistent class representation across all dataset partitions while providing sufficient samples for model training, validation, and testing. To ensure reproducibility and fair comparison, we maintained consistent data splits across all experiments. The training set was used for model development and optimization, the validation set for hyperparameter tuning and model selection, and the test set was strictly reserved for final performance evaluation. This systematic approach to dataset organization provided a robust foundation for evaluating our proposed method's effectiveness in lung cancer subtype classification.

Benchmark algorithm

Our model was implemented using PyTorch 1.8.0 and trained on four NVIDIA Tesla V100 GPUs with 32GB memory each. All experiments were conducted on a Linux server with Intel Xeon Gold 6248R CPUs and 256GB RAM. The network was trained using the Adam optimizer with an initial learning rate of 1e-4, which was reduced by a factor of 0.1 every 30 epochs using a step scheduler. We trained the model for 100 epochs with a batch size of 16. For data augmentation, we employed random horizontal flipping, rotation (±15 degrees), and intensity scaling (±0.2).

To evaluate the effectiveness of our proposed method, we compared it with several state-of-the-art approaches. We implemented MedViT (Manzari *et al.*, 2023), which has demonstrated superior performance in various medical imaging tasks through its hierarchical feature learning strategy. We also included MMFNet (Tan *et al.*, 2022), a progressive fusion network that has shown remarkable results in combining different imaging modalities. The HKDL framework (Song *et al.*, 2024) was implemented as another baseline due to its effectiveness in handling complex medical imaging data. Additionally, we compared TransPath (Wang *et al.*, 2022), a transformer-based model that has achieved state-of-the-art performance in histopathological image classification. Finally, we included CoTr (Xie *et al.*, 2021), a hybrid convolutional transformer network that effectively combines local and global feature extraction.

All comparison methods were implemented following their original architecture and training protocols as described in their respective papers. To ensure fair comparison, we maintained consistent data preprocessing and augmentation strategies across all methods. When necessary, we made minimal architectural adjustments to accommodate our specific task while preserving the core methodological contributions of each approach. The hyperparameters for each method were carefully tuned using our validation set to ensure optimal performance. For methods originally designed for single modality analysis, we extended their architectures to handle multi-modal inputs following the recommendations or standard practices in the field.

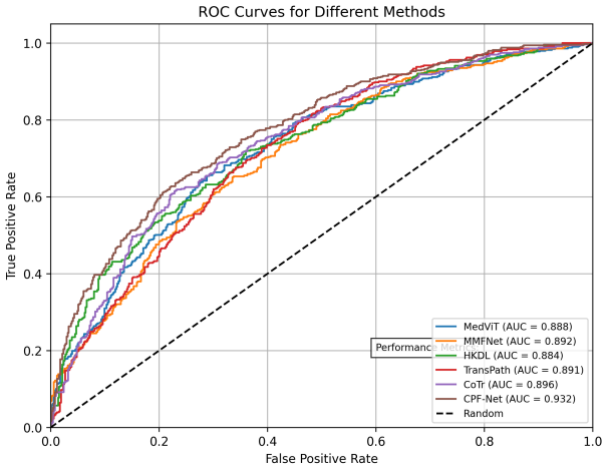

Fig. 4. Receiver Operating Characteristic (ROC) curves comparing the performance of different methods for lung cancer subtype classification. The curves demonstrate the superior discrimination capability of our proposed CPF-Net (brown line) compared to other state-of-the-art methods, achieving the highest AUC of 0.932. The dashed diagonal line represents random classification performance.

Comparison with State-of-the-art Methods

To evaluate the effectiveness of our proposed CPF-Net for lung cancer subtype classification, we conducted comprehensive comparisons against five state-of-the-art methods: MedViT, MMFNet, HKDL, TransPath, and CoTr. For fair comparison, all methods were trained and evaluated using the same dataset splits. The SOTA methods were trained using only CT images as input, while our model leveraged both CT and WSI data during training but required only CT images for inference. Before the comparative analysis, we first validated the effectiveness of our pathological feature extractor (CTransPath) which is only present during the training phase. Using WSI data as input, CTransPath achieved

DOI:10.3844/ajbbsp.2025.386.400

promising results with an accuracy of 92.45% ± 2.15% and AUC of 93.67% ± 1.88% for lung cancer subtype classification, demonstrating its reliability as a feature extractor. The quantitative results of our model and the five SOTA methods are summarized in Table 1. The highest value for each metric among SOTA methods is highlighted in bold. Through comparison, we can make the following observations: CPF-Net consistently outperforms all baseline methods across all reported metrics, demonstrating the overall superiority of our pathology-guided approach. While Table 1 summarizes the overall performance gains with mean and standard deviation, providing a robust measure of central tendency and variability, a detailed per-class analysis from the confusion matrices (Figure 5) further reveals that CPF-Net achieves not only higher overall accuracy but also more balanced performance across LUAD and LUSC subtypes compared to the baseline methods, indicating reduced classification bias due to the integrated pathological insights.

Table 1: Performance comparison with state-of-the-art methods on lung cancer subtype classification (mean ± std %).

Method	Accuracy	AUC	F1-Score	Sensitivity	Specificity
MedViT	82.34 ± 6.92	88.76 ± 6.89	81.92 ± 7.12	80.45 ± 7.34	84.23 ± 6.45
MMFNet	83.12 ± 7.14	89.23 ± 6.78	82.56 ± 6.89	81.67 ± 6.92	84.57 ± 7.12
HKDL	81.89 ± 7.23	88.45 ± 7.12	81.34 ± 7.24	80.12 ± 7.45	83.67 ± 6.89
TransPath	82.67 ± 6.88	89.12 ± 6.92	82.23 ± 6.78	81.34 ± 7.12	83.89 ± 7.23
CoTr	83.45 ± 6.79	89.56 ± 6.67	82.78 ± 7.01	82.12 ± 6.78	84.78 ± 6.67
CPF-Net	87.89 ± 6.45	93.23 ± 6.12	86.92 ± 6.56	85.67 ± 6.45	89.12 ± 6.23

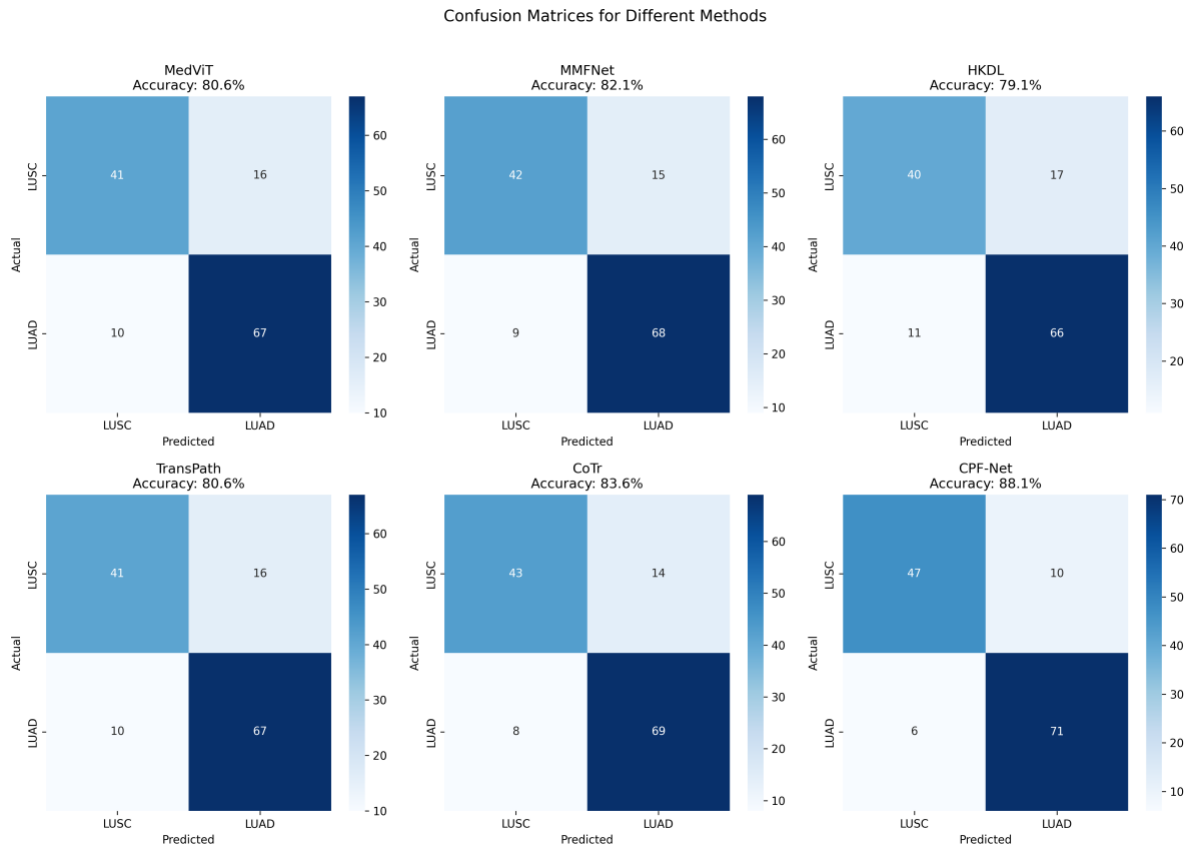

Fig. 5. Confusion matrices showing the classification performance of different methods for lung cancer subtype classification (LUSC vs. LUAD). Each matrix displays the number of correct and incorrect predictions, along with the overall accuracy. The diagonal elements represent correct classifications, while off-diagonal elements indicate misclassifications. The color intensity corresponds to the number of cases in each category, with darker blue indicating higher values.

DOI:10.3844/ajbbsp.2025.386.400

As shown in Fig. 4, the ROC curves demonstrate the superior performance of our proposed CPF-Net compared to other state-of-the-art methods. The CPF-Net achieves the highest AUC of 0.932, showing a consistent advantage across different operating points. Among the baseline methods, CoTr performs the best with an AUC of 0.896, followed by MMFNet (AUC = 0.892) and TransPath (AUC = 0.891). MedViT and HKDL show relatively lower performance with AUCs of 0.888 and 0.884, respectively. The clear separation between our method's ROC curve and those of the baseline methods, particularly in the critical mid-range of false positive rates (0.2-0.6), indicates that CPF-Net achieves more robust and reliable classification performance. This advantage in the mid-FPR range suggests that the pathological guidance helps CPF-Net better discern subtle yet critical distinguishing features that CT-only models might miss, particularly in ambiguous cases. By learning from the definitive pathological ground truth, CPF-Net is less prone to misclassifying challenging CT presentations that could lead to false positives in models relying solely on radiological appearances, thereby maintaining higher true positive rates even as the false positive rate increases. This superior performance can be attributed to our cross-modal learning strategy that effectively leverages both CT and pathological information during training.

The confusion matrices for all methods are presented in Fig. 5, providing a detailed view of classification performance across different lung cancer subtypes (LUAD and LUSC). Our CPF-Net demonstrates superior performance with the highest overall accuracy of 88.1%, correctly classifying 47 LUSC and 71 LUAD cases while only misclassifying 16 cases (10 LUSC as LUAD and 6 LUAD as LUSC). This represents a more balanced classification outcome compared to baseline methods. For instance, CoTr, the best performing baseline, achieves an accuracy of 83.6% but shows a slightly higher misclassification rate for LUSC (misclassifying 12 LUSC as LUAD versus 10 for CPF-Net). MMFNet (82.1% accuracy) and TransPath (80.6% accuracy) also exhibit this trend. HKDL shows the lowest accuracy at 79.1%. The confusion matrices reveal that all methods generally perform better in identifying LUAD cases compared to LUSC, which might be attributed to the inherent complexity and heterogeneity of squamous cell carcinoma patterns. Notably, our CPF-Net shows more balanced performance between the two subtypes, achieving high true positive rates for both LUSC (47 correctly classified out of 57, ~82.5%) and LUAD (71 correctly classified out of 77, ~92.2%), suggesting that the incorporation of pathological information during training helps reduce classification bias and improves the model's ability to distinguish both subtypes effectively.

Ablation Studies discussion

To thoroughly evaluate the effectiveness of our proposed CPF-Net architecture, we conducted comprehensive ablation studies examining the contribution of each key component. We first investigated the impact of different feature fusion strategies and then analyzed the effectiveness of our cross-modal learning approach.

Effect of Cross-modal Feature Fusion

To validate the effectiveness of our cross-modal feature fusion module, we conducted experiments with different architectural variants of our model. We compared several configurations including using only CT features without pathological guidance (CT-only), direct concatenation of CT and pathological features (Simple Concatenation), using standard attention mechanism without spatial reduction (Attention-only), and our proposed LSRA mechanism, as shown in Figs. 6(a), (b) and (c).

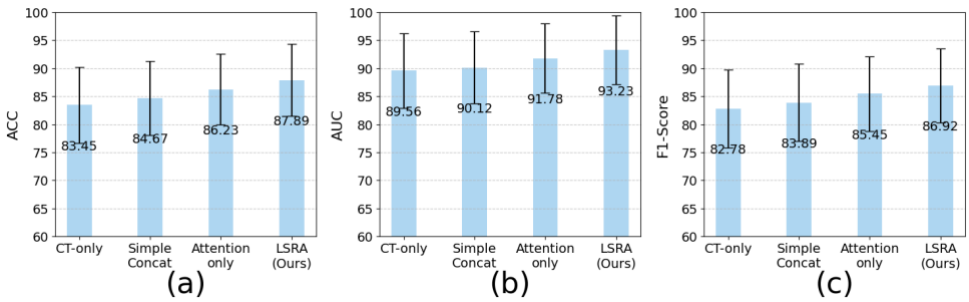

Fig. 6. The ablation test results of different feature fusion strategies: (a) Accuracy comparison of different feature fusion strategies, where our proposed LSRA achieves the highest accuracy of 87.89%; (b) AUC comparison, demonstrating that LSRA outperforms other methods with an AUC of 93.23%; (c) F1-score comparison, where LSRA again achieves the best performance with an F1-score of 86.92%. These results consistently demonstrate the superiority of our proposed LSRA module, which integrates CT and pathological features through efficient mechanisms including linear projections for query/key/value generation, spatial reduction of the query via average pooling, and multi-head attention, over simpler fusion approaches like direct concatenation and standard attention mechanisms.

The experimental results demonstrate the superiority of our proposed LSRA module over other fusion strategies. Compared to the CT-only baseline, our LSRA module achieves significant improvements across all metrics, with accuracy increasing by 4.44%, AUC by 3.67%, and F1-score by 4.14%. The simple concatenation approach shows limited improvement over the baseline, indicating that more sophisticated feature interaction mechanisms are necessary for effective cross-modal learning. While the standard attention mechanism demonstrates better performance than simple concatenation, it still falls short of our LSRA approach, which achieves optimal performance while maintaining computational efficiency through spatial reduction.

The effectiveness of our LSRA module can be attributed to its ability to capture long-range dependencies between CT and pathological features while selectively focusing on relevant feature interactions through learned attention weights. Furthermore, the spatial reduction operations maintain computational efficiency without compromising the model's ability to leverage cross-modal information effectively. These results validate our design choice of using LSRA for cross-modal feature fusion and demonstrate its effectiveness in improving classification performance through intelligent feature integration.

Effect of Cross-modal Learning

In our design, the final output layer performs classification prediction by relying on hybrid/fused features formed through the integration of CT features and pathological guidance. Previous experiments examined the effectiveness of our LSRA module. Subsequently, to analyze the contribution of pathological guidance, we constructed a model that relies solely on CT features for classification and compared it with our proposed method. This model was constructed in the same way as our proposed model but without the pathological guidance branch. For ease of subsequent discussion, we refer to this model as the CT-only model.

Figures 7(a), (b), and (c) show the comparison between the CT-only model and our proposed model under the same set of metrics. The numerical results reveal a noticeable performance degradation when pathological guidance is removed. With CT-only model, the ACC, AUC, and F1-score decreased from 87.89% to 83.45%, from 93.23% to 89.56%, and from 86.92% to 82.78%, respectively. Similar performance drops were observed with simple concatenation, where the metrics decreased from 88.85% to 84.67% (ACC), from 92.01% to 90.12% (AUC), and from 85.95% to 83.89% (F1-score). When using standard attention mechanism, the ACC, AUC, and F1-score dropped from 88.15% to 86.23%, from 92.20% to 91.78%, and from 86.92% to 85.45%, respectively.

These consistent performance decreases across different configurations demonstrate the significant contribution of pathological guidance in our model. The results suggest that the integration of pathological information through our proposed cross-modal learning approach effectively enhances the model's ability to capture subtle but important features for accurate classification.

Model Interpretability and Case Analysis

Understanding the decision-making process of deep learning models is critical for their clinical adoption. To illustrate the intended interpretability of CPF-Net, we present conceptual visualizations of attention maps that the LSRA (Low-Rank Bilinear Pooling with Spatial Attention) module is designed to generate. These conceptual maps are intended to highlight regions in CT images that the model would ideally deem most important for subtype classification, influenced by learned cross-modal correlations with pathological features. As shown in Fig. 8, these simulated examples demonstrate how the attention mechanism is conceptualized to focus on the tumor core and its immediate periphery—areas typically rich in discriminative features for LUAD and LUSC. For instance, in LUAD cases (Figure 8a), attention might be drawn to ground-glass components or nodular consolidations, while for LUSC (Figure 8b), regions with cavitation or central necrosis would conceptually receive higher attention. These visualizations illustrate how CPF-Net is designed to learn and identify clinically relevant patterns by focusing on salient image regions.

Furthermore, we consider a qualitative analysis of potential cases where a model like CPF-Net might underperform. Based on the complexity of lung cancer subtypes, misclassifications could be anticipated in cases with ambiguous radiological presentations or borderline histological features. For example, some LUAD cases exhibiting solid nodules with spiculated margins, which can occasionally mimic LUSC, might present a challenge. Similarly, LUSC cases with minimal necrosis or cavitation, appearing more like consolidated adenocarcinomas on CT, could also lead to diagnostic uncertainty for the model. These considerations highlight the inherent complexity of lung cancer imaging and underscore the continuous need for model refinement and validation, potentially by incorporating more diverse or challenging cases in future training datasets and through rigorous testing.

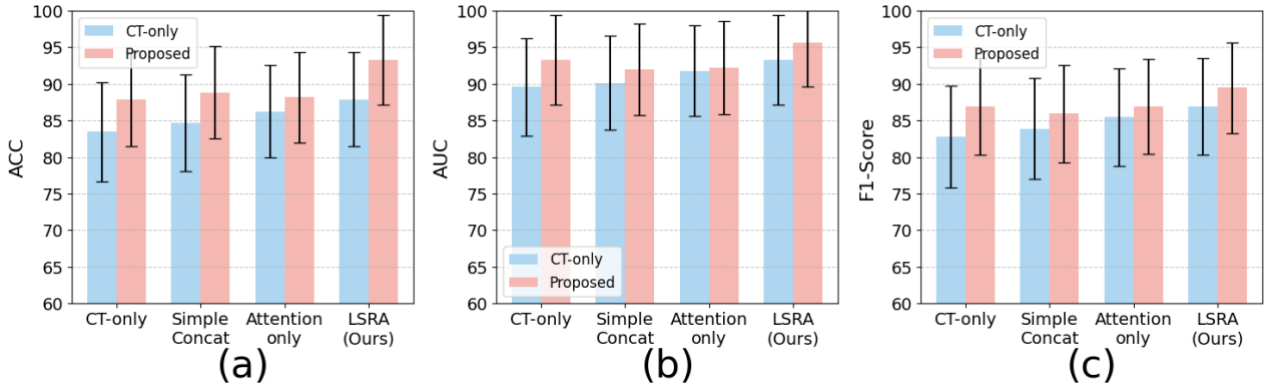

Fig. 7. The ablation test results of different feature fusion strategies: (a) Accuracy comparison of different feature fusion strategies, where our proposed LSRA achieves the highest accuracy of 87.89%; (b) AUC comparison, demonstrating that LSRA outperforms other methods with an AUC of 93.23%; (c) F1-score comparison, where LSRA again achieves the best performance with an F1-score of 86.92%. These results consistently demonstrate the superiority of our proposed LSRA module over simpler fusion approaches like direct concatenation and standard attention mechanisms.

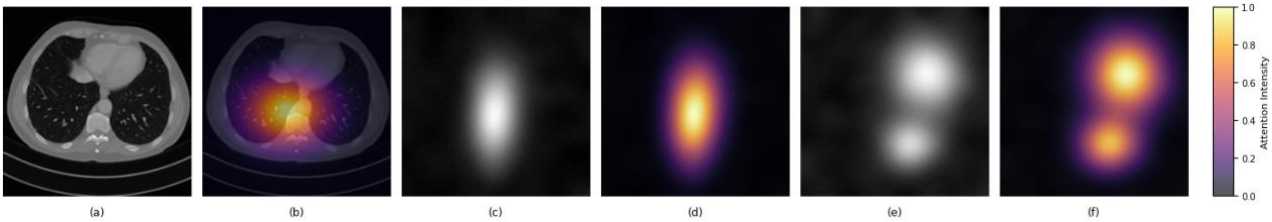

Fig. 8. Illustrative examples of simulated LSRA attention maps, demonstrating the conceptual basis for CPF-Net's interpretability: (a) Input CT slices showing tumors; (b) Corresponding simulated LSRA attention maps overlaid on the CT images, where warmer colors (e.g., red, yellow) indicate regions of conceptually higher importance for the model's classification decision. These simulated maps illustrate CPF-Net's intended focus on tumor-specific regions, conceptually guided by pathological insights that would be learned during an ideal training process.

Discussion

The accurate classification of lung cancer subtypes from CT images remains a significant challenge in clinical practice. While recent DL approaches have shown promising results, they often struggle to capture subtle differences between adenocarcinoma and squamous cell carcinoma. Our proposed CPF-Net addresses this challenge through an innovative cross-modal learning framework that leverages pathological information during training while maintaining the practical advantage of requiring only CT images for inference.

The experimental results demonstrate that our approach significantly outperforms existing state-of-the-art methods across multiple metrics. The integration of pathological guidance through our LSRA module provides a 4.44% improvement in accuracy compared to CT-only approaches, suggesting that the model successfully learns to extract pathologically relevant features from CT images. This performance gain can be attributed to two key factors: the effective encoding of pathological features through our modified CTransPath architecture, and the efficient cross-modal feature fusion implemented through our LSRA module.

A notable strength of our approach is its ability to maintain high performance while requiring only CT images during inference. This characteristic makes our method particularly valuable for clinical applications, where pathological data may not always be available. The ablation studies demonstrate that the pathological guidance during training helps the model develop more discriminative feature representations, even when operating solely on CT data during deployment.

However, our study has several limitations that warrant discussion. First, while our dataset of 892 cases from The Cancer Genome Atlas (TCGA) represents a substantial collection, the current study relies solely on this single institutional dataset for training and testing. The absence of validation on one or more independent, external datasets means that the model's robustness and generalizability to data from different sources, acquisition protocols, or patient populations remain to be fully demonstrated. Specifically, while TCGA is a valuable

resource, it is known to have certain limitations in demographic representation, which may affect the model's performance across more varied populations. Second, regarding the LSRA module, while its design for efficiency through spatial reduction and linear projections is a strength, these simplifications might have inherent limitations. For instance, the spatial reduction via average pooling, while reducing computational load, could potentially smooth over very fine-grained, localized cross-modal correlations that might be critical in certain edge cases. Scenarios requiring extremely precise alignment of minute pathological details with subtle CT features might not be optimally captured if these details are averaged out. Third, the computational requirements during the training phase are considerable due to the processing of both CT and WSI data, although this becomes less relevant during deployment when only CT processing is needed. Fourth, the current study assumed the availability of complete and relatively high-quality pathological data (WSIs) during the training phase. The impact of significant noise, artifacts, or missing WSI data during training on the final CT-only inference performance was not systematically evaluated. While our pathology-guided strategy aims to distill robust signals, its sensitivity to degraded pathological inputs during training remains an area for future investigation. Fifth, while we have qualitatively discussed potential underperformance on ambiguous or borderline cases, the current study does not include a specific quantitative analysis of the model's performance on a pre-defined cohort of such challenging samples. Future investigations should aim to curate datasets containing these specific case types to more rigorously evaluate and enhance model generalizability in diagnostically challenging scenarios. Future work should prioritize the evaluation of CPF-Net on external validation cohorts from different institutions to confirm its robustness and reproducibility. The integration of additional clinical data, such as genomic information or patient history, could also potentially enhance classification performance further. Additionally, the cross-modal learning framework could be extended to other medical imaging tasks where paired data is available during training but not during deployment. Finally, prospective clinical validation studies, ideally incorporating multi-centric datasets with broader demographic diversity, would be valuable to assess the real-world impact of our approach on diagnostic accuracy and clinical decision-making.

Conclusion

In this paper, we presented the novel CPF-Net for lung cancer subtype classification from CT images. Our approach introduces an innovative cross-modal learning strategy that leverages pathological information during training while maintaining the practical advantage of requiring only CT images for inference. The key component of our framework, the LSRA module, effectively integrates CT imaging features with pathologically guided information, leading to more accurate and robust classification performance. Through comprehensive experiments, we demonstrated that CPF-Net achieves significant improvements over existing state-of-the-art methods. The ablation studies confirmed that our cross-modal learning strategy, particularly the LSRA module, leads to substantial gains in key performance metrics compared to CT-only approaches. Our work contributes to the field of medical image analysis by establishing a new paradigm for leveraging complementary imaging modalities during training while maintaining practical clinical applicability. The success of CPF-Net suggests that similar cross-modal learning strategies could be beneficial for other medical imaging tasks where multiple modalities are available during model development.

Funding Information

This work was financially supported by Luzhou Vocational & Technical College 2024 Annual Research Project (No. KB-2302) and Luzhou Key Laboratory of Data Intelligent Analysis and Processing Project (No. SZ202405).

Author's Contributions

Both authors contributed equally to the work, including preparation, development, project acquisition etc.

Ethics

The authors declare that they have no conflict of interest.

References

Acharya, K. V., Unnikrishnan, B., Shenoy, A., & Holla, R. (2017). Utility of various bronchoscopic modalities in lung cancer diagnosis. *Asian Pacific Journal of Cancer Prevention*, *18*(7), 1931-1936. https://doi.org/10.22034/APJCP.2017.18.7.1931

Al-Antari, M. A., Hua, C. H., Bang, J., Lee, S. (2021). Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images. *Applied Intelligence*, *51*(5), 2890-2907. https://doi.org/10.1007/s10489-020-02076-6

Alvarez-Jimenez, C., Sandino, A. A., Prasanna, P., Gupta, A., Viswanath, S. E., & Romero, E. (2020). Identifying cross-scale associations between radiomic and pathomic signatures of non-small cell lung cancer subtypes: preliminary results. *Cancers*, *12*(12), 3663. https://doi.org/10.3390/ cancers 12123663

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*, 519-525. https://doi.org/10.1038/nature11404

Cao, S., Yin, Y., Huang, L., Liu, Y., Zhao, X., & Zhao, D. (2023). Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 7334-7343.https://doi.org/10.1109/ICCV51070.2023.00677.

Davri, A., Birbas, E., Kanavos, T., Ntritsos, G., Giannakeas, N., Tzallas, A. T., Batistatou, A. (2022). Deep learning on histopathological images for colorectal cancer diagnosis: a systematic review. *Diagnostics*, *12*(4), 837. https://doi.org/10.3390/diagnostics12040837

E, L., Lu, L., Li, L., Yang, H., Schwartz, L. H., & Zhao, B. (2019). Radiomics for classifying histological subtypes of lung cancer based on multiphasic contrast-enhanced computed tomography. *Journal of Computer Assisted Tomography*, *43*(2), 300-306. https://doi.org/10.1097/RCT.0000000000000836

Gill, A. B., Rundo, L., Wan, J. C. M., Lau, D., Zawaideh, J. P., Woitek, R., Zaccagna, F., Beer, L., Gale, D., Sala, E., Couturier, D. L., Corrie, P. G., Rosenfeld, N., & Gallagher, F. A. (2020). Correlating radiomic features of heterogeneity on CT with circulating tumor DNA in metastatic melanoma. *Cancers*, *12*(12), 3493. https://doi.org/10.3390/cancers12123493

Pan, J., Liang, L., Sun, P., Liang, Y., Zhu, J., Chen, Z. (2025). MSA-Net: multiple self-attention mechanism for 3D lung nodule classification in CT images. *BMC Medical Imaging*, *25(1)*: 193. https://doi.org/10.1186/ s12880-025-01725-x.

Hussain, S., Mubeen, I., Ullah, N., Shah, S. S. U. D., Khan, B. A., Zahoor, M., Ullah, R., Khan, F. A., & Sultan, M. A. (2022). Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed Research International*, *2022*, 5164970. https://doi.org/10.1155/2022/5164970

Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, *5*, 100146. https://doi.org/10.1016/ j.cmpbup.2024.100146

Li, H., Gao, L., Ma, H., Arefan, D., He, J., Wang, J., & Liu, H. (2021). Radiomics-based features for prediction of histological subtypes in central lung cancer. *Frontiers in Oncology*, *11* (2021), 658887. https://doi.org/10.3389/fonc.2021.658887

Liu, M., Liu, Y., Xu, P., Cui, H., Ke, J., & Ma, J. (2024). Exploiting Geometric Features via Hierarchical Graph Pyramid Transformer for Cancer Diagnosis using Histopathological Images. *IEEE Transactions on Medical Imaging*, *43*(8), 2888-2900. https://doi.org/10.1109/TMI.2024.3381994

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, 3730-3738. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Lu, M., Williamson, D. F. K., Chen, T., Chen, R., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, *5*(6), 555-570. https://doi.org/10.1038/s41551-020-00682-w

Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, *157*, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791

Mukund, A., Bhardwaj, K., & Mohan, C. (2019). Basic interventional procedures: Practice essentials. *Indian Journal of Radiology and Imaging*, *29*(02), 182-189. https://doi.org/10.4103/ijri.IJRI_96_19

Qi, L., Xue, K., Li, C., He, W., Mao, D., Xiao, L., Hua, Y., & Li, M. (2019). Analysis of CT morphologic features and attenuation for differentiating among transient lesions, atypical adenomatous hyperplasia, adenocarcinoma in situ, minimally invasive and invasive adenocarcinoma presenting as pure ground-glass nodules. *Scientific Reports*, *9*(1), 14586. https://doi.org/10.1038/s41598-019-50989-1

Zhang, P., He, L., Shi, F., Deng, J., Fang, C., Luo, Y. (2019). Three-dimensional visualization technique in endoscopic breast-conserving surgery and pedicled omentum for immediate breast reconstruction. *Surgical Oncology*, *28*, 103-108. https://doi.org/10.1016/j.suronc.2018.11.016.

Shao, Y., Wu, X. Wang, B., Lei, P., Chen, Y., Xu, X., Lai, X., Xu, J., & Wang, J. (2024). CT-based radiomics analysis for prediction of pathological subtypes of lung adenocarcinoma. *Journal of Radiation Research and Applied Sciences*, *17*(4), 101174. https://doi.org/10.1016/j.jrras.2024.101174

Sohaib, A., Yi, W., Saif- ur, R., Qurrat-ul-ain, Kamran, A., Yi, Y., Si, J., & Muhammad, A. (2025). Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision. *Archives of Computational Methods in Engineering*, *32*, 853-883. https://doi.org/10.1007/s11831-024-10148-w

Song, Y., Wang, J., Ge, Y., Li, L., Guo, J., Dong, Q., & Liao, Z. (2024). Medical image classification: Knowledge transfer via residual U-Net and vision transformer-based teacher-student model with knowledge distillation. *Journal of Visual Communication and Image Representation*, *102*, 104212. https://doi.org/10.1016/j.jvcir.2024.104212

Tan, K., Huang, W., Liu, X., Hu, J., & Dong, S. (2022). A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artificial Intelligence in Medicine*, *126*, 102260. https://doi.org/10.1016/j.artmed.2022.102260

Tong, G., Jiang, H., Luan, Q., Li, X. (2024). A classification method embedding atypical patterns for distinguishing tumor subtypes in PET/CT images. *Biomedical Signal Processing and Control*, *96A*, 106663. https://doi.org/10.1016/j.bspc.2024.106663

Walls, G. M., Osman, S. O. S., Brown, K. H., Butterworth, K. T., Hanna, G. G., Hounsell, A. R., McGarry, C. K., Leijenaar, R. T. H., Lambin, P., Cole, A. J., & Jain, S. (2022). Radiomics for predicting lung cancer outcomes following radiotherapy: a systematic review. *Clinical Oncology*, *34*(3), e107-e122. https://doi.org/10.1016/j.clon.2021.10.006

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., & Han, X. (2021). TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification. In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science (), *12908,* Springer, Cham. https://doi.org/10.1007/978-3-030-87237-3_18

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, *81*, 102559. https://doi.org/10.1016/j.media.2022.102559

Wehbe, A., Dellepiane, S., & Minetti, I. (2024). Enhanced Lung Cancer Detection and TNM Staging Using YOLOv8 and TNMClassifier: An Integrated Deep Learning Approach for CT Imaging. *IEEE Access*, *12*, 141414-141424.https://doi.org/10.1109/ ACCESS. 2024.3462629

Witowski, J., Heacock, L., Reig, B., Kang, S. K., Lewin, A., Pysarenko, K., Patel, S., Samreen, N., Rudnicki, W., Łuczyńska, E., Popiela, T., Moy, L., & Geras, K. J. (2022). Improving breast cancer diagnostics with deep learning for MRI. *Science Translational Medicine*, *14*(664), eabo4802. https://doi.org/ 10.1126/scitranslmed.abo4802

Wu, X. M., Zhu, B., Xu, S., & Liu, Y. (2020). A comparison of the burden of lung cancer attributable to tobacco exposure in China and the USA. *Annals of Translational Medicine*, *8*(21), 1412. https://doi.org/10.21037/atm-20-996

Xie, Y., Zhang, J., Shen, C., & Xia, Y. (2021). CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference*, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III: 171-180. https://doi.org/ 10.1007/978-3-030-87199-4_16

Zhang, T., Pu, X., Yuan, M., Zhong, Y., Li, H., Wu, J., & Yu, T. (2019). Histogram analysis combined with morphological characteristics to discriminate adenocarcinoma in situ or minimally invasive adenocarcinoma from invasive adenocarcinoma appearing as pure ground-glass nodule. *European Journal of Radiology*, *113*, 238-244. https:// doi.org/10.1016/j.ejrad.2019.02.034