2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)
31 July – 2 August 2025, Rangpur, Bangladesh

MRAViT-XAI: A Novel Multi-Resolution Attention Vision Transformer Framework with Explainable AI for Enhanced Lung and Colon Cancer Classification

Md. Abdur Rahman[1], Sabik Aftahee[2], Lamim Zakir Pronay[3],
Md Ashiqur Rahman[4]
[1,4]Department of Computer Science and Engineering, Southeast University,
Dhaka, 1208, Bangladesh
[2]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET),
Chittagong, 4349, Bangladesh
[3]Department of Computer Science and Engineering, National Institute of Technology,
Andhra Pradesh, 534101, India
Email: [1]2021200000025@seu.edu.bd, [2]sabikaftahee.official@gmail.com,
[3]pronayfarab03@gmail.com, [4]ashiqur.rahman@seu.edu.bd

***Abstract*—Early detection of lung cancer and colon cancer is very crucial for successful treatment, with histopathology images serving as the gold standard of detection, despite its reliance on manual expert analysis. In this paper, we develop MRAViT-XAI, a novel Vision Transformer framework enhanced with multi-resolution attention mechanisms for automated histopathological image classification. The proposed architecture is designed to simultaneously capture global contextual information and local tissue morphology through cross-scale attention refinement, enabling precise identification of cancerous patterns at varying magnifications. When tested with the LC25000 dataset, which comprises 25,000 histopathological images in five categories, our proposed method presents state-of-the-art performance with an accuracy of 99.90%, outperforming other recent deep learning approaches. We incorporated Local Interpretable Model-Agnostic Explanations (LIME) visualization technique to improve model interpretability by visualizing diagnostically important image regions, thus promoting clinical trust and transparency. The results of the confusion matrix analysis indicate that the classes are well separated, with a low degree of confusion among the histologically similar subtypes. ROC curve analysis demonstrates perfect AUC scores (1.00) across all classes, indicating the excellent diagnostic performance of the model. This study shows potential for multi-resolution transformer-based architectures to disrupt computerized diagnosis systems for detecting cancer.**

***Index Terms*—Vision Transformer, Multi-Resolution Attention, Histopathology, Cancer Classification, Explainable AI**

I. Introduction

Lung and colon cancers are two of the most widespread and fatal types of cancer worldwide. Approximately 1.8 million deaths due to lung cancer and over 900 thousand colon cancer cases were reported worldwide in 2022 which urges the need for precise early diagnosis to ensure effective treatment [1] [2]. Histopathology imaging is considered the gold standard for cancer diagnosis, but it requires manual analysis which is both tedious, slower and overly reliant on expert pathologists, highlighting the need for automated diagnostic tools that improve efficiency, in the clinical setting [3]. Automated diagnosis systems facilitate dependable classification from histopathology images, making them relevant in case of early detection and treatment of malignancies. In medical imaging, deep learning techniques combined with convolutional networks have accelerated the rate of recognition of global and local features in the images. Recently, Vision Transformers (ViT) have emerged as an alternative which contributed in medical image classification due to their ability to capture intra-image global and local patterns, often outperforming conventional CNN models [4]. In this paper, we propose a new framework MRAViT-XAI, Vision Transformers (ViT) [5] enhanced with multi-resolution attention mechanisms which enables multi-scale detail extraction, focusing on local structures as well as wider patterns. It was evaluated on the LC25000 dataset [6]. The framework helped obtain an accuracy of 99.90% on the dataset. We also integrated Local Interpretable Model-Agnostic Explanations (LIME) [7], to highlight the regions of image that affect the decision making process of the model to enhance clinical trust. The main contributions of this paper are:

- Developing MRAViT-XAI, a novel framework of vision transformer (ViT) featuring multi-resolution attention mechanisms to boost the ViT accuracy to 99.90%.
- Integrating Local Interpretable Model-Agnostic Explanations (LIME), providing transparency and interpretability, crucial for clinical confidence.

II. Literature Review

Earlier studies on histopathological image classification in lung tissue have explored various computer-aided diagnosis (CAD) systems. Nishio *et al.* [8] developed a CAD system utilizing homology-based image processing (HI) for feature extraction, demonstrating superior performance over conventional texture analysis methods. Their approach involved calculating Betti numbers to capture topological features of

979-8-3315-9694-1/25/$31.00 ©2025 IEEE

histopathological images, leading to improved classification accuracy. Mangal *et al.* [9] introduced a CAD system employing shallow convolutional neural network (CNN) architectures that preserved image resolution throughout the processing pipeline. This strategy enhanced diagnostic accuracy while reducing computational demands by effectively capturing tissue patterns and connectivity. Merabet and Saighi [10] proposed a hybrid deep learning architecture combining CNN, ResNet50, and InceptionV3 models to improve colon cancer classification. Their integrated approach achieved a predictive accuracy of 99.27%, outperforming standalone models and highlighting the benefits of combined architectures in enhancing diagnostic precision for histopathological images from the LC25000 dataset. Pasha and Ata [11] explored ensemble learning techniques by combining models such as VGG16 + ResNet50, VGG16 + EfficientNetB0, and ResNet50 + EfficientNetB0. By concatenating features from multiple models and applying dense classification layers, they addressed feature redundancy and managed morphological variability in images, demonstrating the efficiency of ensemble learning in histopathological image classification.

III. Methodology

This section presents **MRAViT-XAI**, a novel framework for the automatic classification of lung and colon cancer from histopathology images using Vision Transformers enhanced with multi-resolution attention mechanisms. Our study offers a cross-scale attention refining approach, enabling the model to simultaneously analyze global and local characteristics in histopathology images, which is essential for accurate cancer type classification. Our proposed framework is shown in Figure 1.

A. Dataset Preparation

Our study used the LC25000 dataset [6], consisting of 25,000 H&E-stained histopathology images. It includes five classes: lung adenocarcinoma, lung squamous cell carcinoma, lung benign tissue, colon adenocarcinoma, and colon benign tissue. The original image resolution was 768×768 pixels. The dataset was split into 70% for training, 10% for validation, and 20% for testing using stratified sampling to preserve class balance across all splits.

In preprocessing, All images were resized to 224×224 pixels to align with the ViT backbone's input size. To enhance generalization and robustness, we applied extensive data augmentations that simulated common variations in histopathology imaging. These included: Random Resized Cropping (scale 0.7–1.0) for magnification and framing diversity; Random Horizontal Flips (50%) for orientation invariance; Color Jitter (brightness/contrast up to 30%, saturation 20%, hue 10%) to mimic H&E staining variability; and combined Random Rotations (±20°) with Random Affine transforms (±15° rotation, 10% translation, 90-110% scaling, ±10° shear). These geometric and color-space transformations helped the model focus on diagnostically relevant features. Finally, images were normalized using standard ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to support transfer learning.

B. Proposed Architecture

1) Vision Transformer Backbone: Our backbone feature extractor is the ViT-B/16 architecture [5], pretrained on ImageNet. This model splits input images into 16×16 non-overlapping patches, projecting each patch onto a 768-dimensional embedding space. Twelve transformer encoder blocks process these embeddings together with positional encodings. Each encoder block contains multi-head self-attention (MSA) and multilayer perceptron (MLP) modules with residual connections. Instead of relying solely on the class token output, we preserve spatial information, which is vital for our downstream attention modules by extracting and utilizing the entire feature map from the final transformer layer.

2) Multi-Resolution Attention Module: The main novelty in our methodology is the **Multi-Resolution Attention (MRA)** module. It enhances the model's ability to concurrently focus on features across multiple scales. The MRA module analyzes feature maps at three distinct resolutions, as defined in Equations (1), adapting techniques seen in multi-scale architectures [12].

$$F_i = \mathrm{Pool}_i\big(\,\mathrm{Conv}_i\big(\mathrm{GELU}\big(\mathrm{BN}_i(x)\big)\big)\big)\,, \quad i \in \{1,2,3\} \tag{1}$$

Here, Pool_i denotes adaptive average pooling at scale i with pool dimensions $[7 \times 7,\ 4 \times 4,\ 1 \times 1]$, generating multi-scale feature representations. Each pooled feature map is then processed through a bottleneck architecture [13] with channel reduction and restoration to maintain computational efficiency while capturing scale-specific features.

The importance of each resolution is determined by a learnable scale attention mechanism, similar in principle to attention methods introduced in [14], as shown in Equations (2):

$$\alpha = \mathrm{Softmax}\Big(W_\alpha\,[\,F_1^p,\,F_2^p,\,F_3^p\,]\Big) \tag{2}$$

where F_i^p represents the globally pooled features from each scale. These are concatenated and projected using a weight matrix W_α to generate scale attention weights.

The final enhanced feature representation is given in Equations (3):

$$F_{\text{enhanced}} = \sum_{i=1}^{3} \alpha_i\, F_i^{\text{resized}} + x \tag{3}$$

where F_i^{resized} denotes the upsampled features aligned to the original spatial dimensions, and the original input x is added as a residual connection [13].

3) Cross-Scale Attention Refinement: After multi-resolution feature extraction, we employed a **Cross-Scale Attention Refinement (CSAR)** approach to facilitate interaction between features at multiple scales. This module applies self-attention [15] over the spatial dimensions, treating the enhanced spatial features as a sequence, as shown in Equation (4).

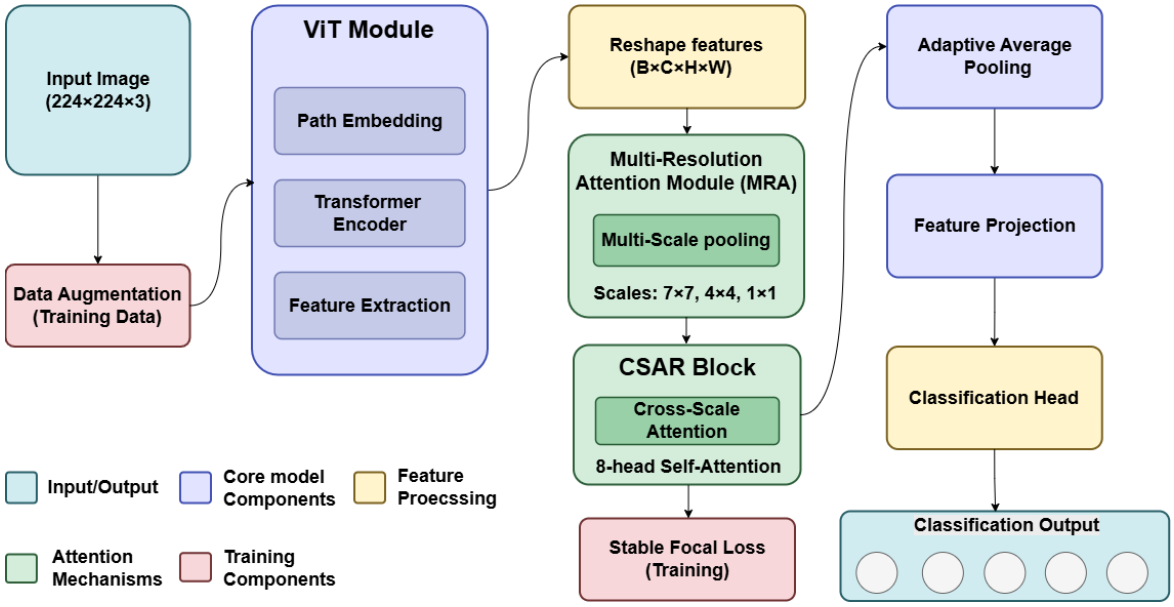

Fig. 1: MRAViT-XAI Framework

$$\hat{F} = \mathrm{Norm}\big(F_{\mathrm{flat}}\big), \quad F_{\mathrm{refined}} = \hat{F} + \mathrm{MHA}\big(\hat{F},\, \hat{F},\, \hat{F}\big) \quad (4)$$

In this equation, F_{flat} denotes the spatially flattened form of F_{enhanced}, and MHA represents multi-head attention [15] with 8 attention heads. The output F_{refined} is the refined feature representation.

This operation enables the model to capture long-range dependencies across spatial locations and resolutions by effectively integrating both local and global contextual information.

4) Classification Head: The refined features from the CSAR module are globally pooled and passed through a feature projection module consisting of layer normalization [16], dimensionality reduction to 512 features, GELU activation [17], and dropout regularization (rate = 0.2) [18]. The final classification layer outputs logits corresponding to the five cancer classes, as shown in Equations (5)

$$y = W_{\mathrm{cls}} \cdot \mathrm{Dropout}(\mathrm{GELU}(\mathrm{Linear}(\mathrm{LayerNorm}(F_{\mathrm{pool}})))) \quad (5)$$

We employ a stable focal loss function [19] to handle class imbalance and focus learning on harder examples. The loss function is defined in Equations (6)

$$\mathcal{L}_{\mathrm{focal}} = -\,\alpha_t\,(1 - p_t)^{\gamma}\,\log\big(p_t\big) \quad (6)$$

Here, p_t denotes the predicted probability for the true class, α_t is a weighting factor used to balance class frequencies, and $\gamma = 2$ is the focusing parameter that down-weights easy examples and emphasizes hard ones.

C. Explainable AI with LIME

Our study employed **Local Interpretable Model-Agnostic Explanations (LIME)** [7] to enhance clinical trustworthiness and provide interpretability for our model's predictions. LIME produces locally faithful explanations by approximating the complex model with an interpretable one in the neighborhood of a specific prediction. The process involves segmenting test images into superpixels, generating perturbed samples by randomly masking these components, and obtaining predictions from the model for each variation. A weighted linear model is then trained on these samples to locally approximate the original model's behavior. This approach highlights the image regions that most influence classification decisions. The explanations are visualized as heatmaps overlaid on the original images, where green indicates a positive contribution to the prediction and red signifies a negative one. These visual cues allow clinicians to verify whether the model's focus aligns with medically relevant features, thereby increasing trust in its diagnostic outputs.

IV. Results and Discussions

This section describes the experimental setup, presents both quantitative and qualitative results of our proposed Multi-Resolution Attention Vision Transformer (MRAViT-XAI) framework on the LC25000 dataset, compares it against state-of-the-art methods, and discusses our findings, including insights from explainable artificial intelligence.

A. Implementation Details

1) Environment Setup: Experiments were conducted on the Kaggle platform, utilizing cloud computing resources. The setup included a 2-core Intel Xeon CPU with 29 GB of RAM and a 16 GB VRAM accelerating NVIDIA Tesla P100 GPU. We used PyTorch as the deep learning framework, along with libraries such as torchvision, scikit-learn, and LIME.

2) Hyperparameters and Training: The MRAViT-XAI model was trained using a weight decay of 0.01 and the AdamW optimizer [20]. We used a OneCycle learning rate scheduler [21] with a maximum learning rate of 3e-5 over 10 epochs. To address potential class imbalance and focus on harder instances, we employed Stable Focal Loss with $\gamma = 2.0$ and equal class weighting $\alpha = 0.2$ for the 5 classes. A batch size of 64 was used for training, while a batch size of 128 was used for validation and testing. On the GPU, mixed-precision training (AMP) was activated. To minimize overfitting and select the best-performing model based on validation accuracy, early stopping was employed with a patience of three epochs and a minimum validation accuracy improvement delta of 0.001.

B. Training Dynamics

Fig. 2 shows the training and validation performance curves. Within the first epoch, the training loss shows a significant initial drop, and then across later epochs, it gradually converges towards zero. Reflecting this pattern, the validation loss stabilizes at a relatively low value early on, suggesting fast learning and high generalization. In the training and validation accuracy curve, we can observe a similar rapid increase, surpassing 95% after the first epoch, and a plateau near 99% by the third epoch. The small gap between training and validation measurements points to the efficient reduction of overfitting through the augmentations and regularization methods used. Training was terminated early due to the early stopping criteria being met at 9 epochs, as performance saturated after 6 epochs. In terms of computational cost, the training for these 9 epochs completed in 51 minutes, averaging 5.6 minutes per epoch.

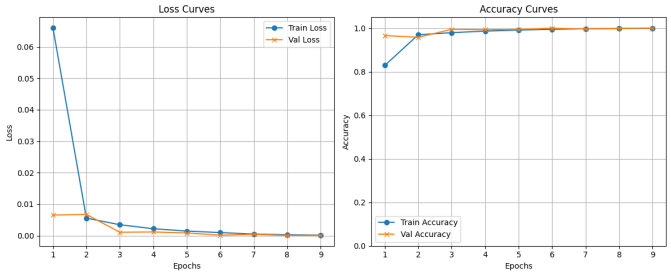

Fig. 2: Training and validation loss and accuracy curves over epochs.

1) Overall Performance: Our final model achieved an overall performance of 99.90%. The class-wise proprietary performance is shown in the classification report in Table I. For the macro and weighted average, precision, recall, and F1-score are all 0.9990, reflecting the uniformly high performance across all five classes.

2) Confusion Matrix Analysis: The confusion matrix shown in Fig. 3 visually verifies the good performance of the classification. The matrix has very strong diagonal dominance because test images were classified correctly in almost all samples. No substantial confusion was observed, except for 3 Lung Adenocarcinoma cases that were mislabeled as Lung Squamous Cell Carcinoma and 2 Lung Squamous Cell Carcinoma cases that were labeled as Lung Adenocarcinoma. This small mix-up of histological lung cancer subtypes is reasonable and again stresses how difficult it is to distinguish these particular categories, even for automated tools. All other classes were classified perfectly.

TABLE I: Classification Report on the Test Set

Class	Precision	Recall	F1-score
colon_aca	1.0000	1.0000	1.0000
colon_n	1.0000	1.0000	1.0000
lung_aca	0.9980	0.9970	0.9975
lung_n	1.0000	1.0000	1.0000
lung_scc	0.9970	0.9980	0.9975
Accuracy		**0.9990**	
Macro avg	**0.9990**	**0.9990**	**0.9990**
Weighted avg	**0.9990**	**0.9990**	**0.9990**

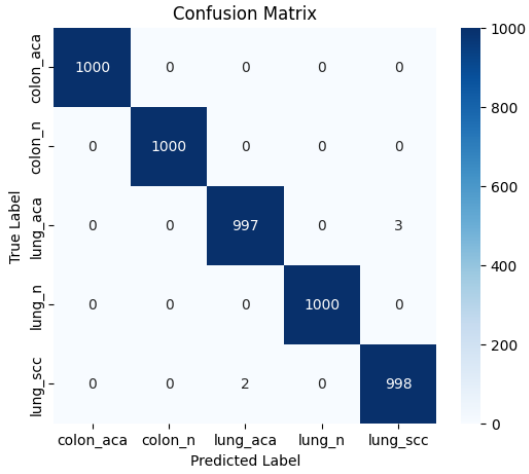

Fig. 3: Confusion matrix for the 5-class classification on the LC25000 testset.

3) ROC Curve Analysis: Receiver Operating Characteristic (ROC) curves and AUC values also support the discriminative capability of the model as shown in Fig. 4. The AUC scores of each individual class ('colon_aca', 'colon_n', 'lung_aca', 'lung_n', 'lung_scc') were 1.0000. The micro-average AUC and macro-average AUC scores were, as a result, also 1.0000. The ideal AUC score value indicates that the model has excellent identifiability for all the classes over all the decision thresholds.

4) Comparison with State-of-the-Art: We benchmarked our MRAVIT-XAI framework with recently published SOTA methods on the same dataset, LC25000. As presented in Table II, our proposed method obtains an accuracy of 99.90%, which is very competitive and superior to the performance of many recent studies, such as DenseNet+RF, VGG16+CLAHE, ColonNet, and CNN+VGG19, as well as the best algorithm Ensemble DL and Self-ONN. Performance differences across studies relate to varied preprocessing: Kumar et al. [22] used resizing, Omar et al. [23] resizing with augmentation and cropping, Hadiyoso et al. [24] CLAHE with resizing, Iqbal

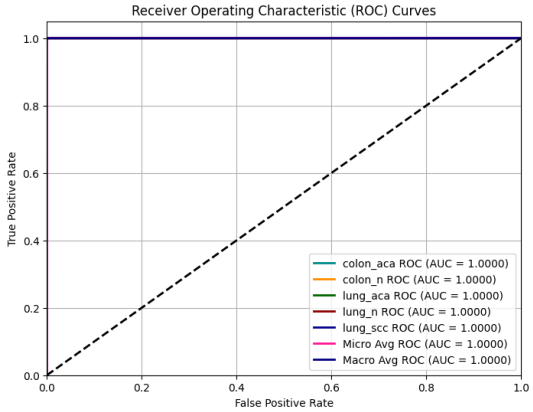

Fig. 4: Receiver Operating Characteristic (ROC) curves.

et al. [25] stain normalization and augmentation, Said et al. [26] pixel normalization with resizing, and Rawashdeh et al. [27] resizing, normalization and augmentation, differing from our augmentation, resizing and normalization approach. Our method demonstrates consistent superiority across this diverse methodological landscape. This demonstrates the success of using the global receptive field of ViT jointly with our multi-resolution attention mechanisms in such a histopathological classification task.

TABLE II: Performance comparison of the proposed MRAViT-XAI framework against state-of-the-art methods

Reference	Year	Method	Dataset	Accuracy (%)
Kumar et al. [22]	2022	DenseNet121+RF	LC25000	98.60
Omar et al. [23]	2023	Ensemble DL	LC25000	99.44
Hadiyoso et al. [24]	2023	VGG16+CLAHE	LC25000	98.96
Iqbal et al. [25]	2023	ColonNet	LC25000	96.31
Said et al. [26]	2024	Self-ONN	LC25000	99.74
Rawashdeh et al. [27]	2024	CNN+VGG19	LC25000	99.04
Proposed	2025	**MRAViT-XAI**	LC25000	**99.90**

C. Ablation Study

To validate our architectural contributions, we conducted an ablation study as shown in Table III. The findings show that the performance over ViT baseline is confirmed with all components added, with each step adding on MRA and CSAR further improving performance.

TABLE III: Ablation study of proposed MRAViT-XAI variants

Configuration	Accuracy	Weighted F1-Score
ViT Baseline	0.9604	0.9604
ViT + MRA only	0.9736	0.9736
MRAViT-XAI	**0.9990**	**0.9990**

D. Qualitative Analysis and Explainability (XAI)

To interpret the decision-making mechanism of the model and improve its reliability, we used LIME [7] to produce visual explanations on test images. Fig. 5 shows the representative samples for various categories.

The left column shows the original image along with its true and predicted labels. The right column shows the LIME explanation, where green regions are the super-pixels that positively contribute to the predicted class, red regions contribute negatively, and yellow contours represent superpixels considered by LIME.

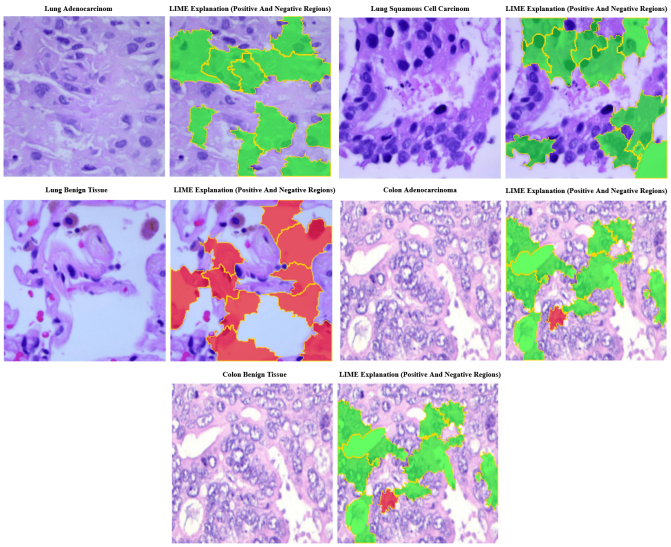

Fig. 5: LIME explainability results for representative test images. (a) Lung Adenocarcinoma, (b) Lung Squamous Cell Carcinoma, (c) Lung Benign Tissue, (d) Colon Adenocarcinoma, (e) Colon Benign Tissue.

Observations from Fig. 5 indicate that the model tends to attend to salient diagnostic features. The green highlighted areas in cancerous tissues like Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, and Colon Adenocarcinoma imply high cellular density, atypical nuclei, or glandular structures typically associated with malignancies. For normal tissues (Colon Benign Tissue, Lung Benign Tissue), the model accurately identifies positive contributions from normal structures (green in Colon Benign Tissue) or highlights uncertain regions that detract from cancer appearance (red in Lung Benign Tissue). These qualitative results increase confidence in the model's alignment with known pathology. While these visualizations are promising, a formal quantitative evaluation through a user study with pathologists would be necessary to confirm their clinical utility, which remains a key direction for future work.

V. Discussion

Experimental results confirm that the proposed MRAViT-XAI framework is effective for multi-class classification of lung and colon cancer histopathology. The ViT backbone's ability to model global dependencies, combined with MRA-CSAR modules that integrate features across spatial scales, leads to state-of-the-art performance (99.90% accuracy) on the LC25000 dataset. Quantitatively, the model demonstrates high precision, recall, F1-scores, and perfect AUC values, reflecting robustness and discrimination power. Compared with SOTA models, our approach is highly competitive. Qualitatively, LIME explanations indicate that the model focuses on

diagnostically meaningful features, enhancing interpretability, and potential for clinical trust. Despite its strengths, limitations include the computational cost inherent to ViT-based models, which directly impacts its scalability for real-time clinical applications like high-throughput whole-slide image analysis. The localized nature of LIME explanations also presents a limitation. Furthermore, while we employed extensive data augmentation and regularization techniques to mitigate overfitting, the model's high accuracy on a single dataset warrants further validation. Future work should therefore focus on evaluating the MRAVIT-XAI framework on external datasets, exploring model compression and optimization techniques to enable clinical scalability, and quantitatively assessing the clinical utility of the XAI features through user studies with pathologists.

VI. Conclusion

This paper proposes **MRAViT-XAI**, a new Vision Transformer model with a multi-resolution attention mechanism for classifying lung and colon cancer. On the LC25000 dataset, our method outperformed prior approaches with a high accuracy of **99.90%**. The use of Local Interpretable Model-agnostic Explanations (LIME) for model explainability provides visual insights to enhance clinical trust in the model's decision-making process. Our results highlight the promise of MRAViT-XAI as a powerful yet interpretable tool for automated histopathological diagnosis, supporting the field of computational pathology.

References

[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

[2] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(1):7–33, 2022.

[3] Weiming Hu, Xintong Li, Chen Li, Rui Li, Tao Jiang, Hongzan Sun, Xinyu Huang, Marcin Grzegorzek, and Xiaoyan Li. A state-of-the-art survey of artificial neural networks for whole-slide image analysis: from popular convolutional neural networks to potential visual transformers. *Computers in Biology and Medicine*, 161:107034, 2023.

[4] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif, and Umair Farooq. A survey of the vision transformers and their cnn-transformer based variants. *Artificial Intelligence Review*, 56(Suppl 3):2917–2970, 2023.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[8] Mizuho Nishio, Mari Nishio, Naoe Jimbo, and Kazuaki Nakane. Homology-based image processing for automatic classification of histopathological images of lung tissue. *Cancers*, 13(6):1192, 2021.

[9] Sanidhya Mangal, Aanchal Chaurasia, and Ayush Khajanchi. Convolution neural networks for diagnosing colon and lung cancer histopathological images. *arXiv preprint arXiv:2009.03878*, 2020.

[10] Asma Merabet, Asma Saighi, Mohamed Abderraouf Ferradji, and Zakaria Laboudi. Enhancing colon cancer prediction in histopathology with integrated deep learning models: A comparative study on the lc25000 dataset. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–7. IEEE, 2024.

[11] Masthan Pasha, Kishore Kumar ATA, and V Vijaya Kishore. Optimized ensemble learning for lung and colon cancer classification using histopathology images from lc25000 dataset. In *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1874–1879. IEEE, 2025.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[21] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[22] Naresh Kumar, Manoj Sharma, Vijay Pal Singh, Charanjeet Madan, and Seema Mehandia. An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. *Biomedical Signal Processing and Control*, 75:103596, 2022.

[23] Lava Th Omar, Judy M Hussein, Lava F Omer, Abdalbasit Mohammed Qadir, and Mazen Ismaeel Ghareb. Lung and colon cancer detection using weighted average ensemble transfer learning. In *2023 11th international symposium on digital forensics and security (ISDFS)*, pages 1–7. IEEE, 2023.

[24] Sugondo Hadiyoso, Suci Aulia, Indrarini Dyah Irawati, et al. Diagnosis of lung and colon cancer based on clinical pathology images using convolutional neural network and clahe framework. *Int. J. Appl. Sci. Eng*, 20(1):1–7, 2023.

[25] Saeed Iqbal, Adnan N Qureshi, Musaed Alhussein, Khursheed Aurangzeb, and Seifedine Kadry. A novel heteromorphous convolutional neural network for automated assessment of tumors in colon and lung histopathology images. *Biomimetics*, 8(4):370, 2023.

[26] Menatalla MR Said, Md Sakib Bin Islam, Md Shaheenur Islam Sumon, Semir Vranic, Rafif Mahmood Al Saady, Abdulrahman Alqahtani, Muhammad EH Chowdhury, and Shona Pedersen. Innovative deep learning architecture for the classification of lung and colon cancer from histopathology images. *Applied Computational Intelligence and Soft Computing*, 2024(1):5562890, 2024.

[27] Majdi Rawashdeh, Muath A Obaidat, Meryem Abouali, and Kutub Thakur. A deep learning-driven approach for detecting lung and colon cancer using pre-trained neural networks. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 183–188. IEEE, 2024.