HYBRID CNN-TRANSFORMER MODEL FOR ACCURATE AND EFFICIENT HISTOPATHOLOGICAL IMAGE CLASSIFICATION

Van Huy Hoang[1], Thanh Trung Nguyen[2],
Thi Hang Tran[1], Van Nam Pham[1,*]

DOI: https://doi.org/10.57001/huih5804.2025.424

ABSTRACT

The rapid development of medical imaging technologies has increased the demand for advanced, automated, and interpretable diagnostic tools to support clinical decision-making. While convolutional neural networks (CNNs) have shown significant success in medical image analysis, they often struggle to capture global context and lack interpretability, which limits their practical application in clinical settings. To address these challenges, we propose a novel hybrid architecture that combines the strengths of pretrained ResNet50 for efficient local feature extraction and Vision Transformer (ViT) for modeling long-range dependencies across the image. ResNet50 uses transfer learning to learn strong features from large image datasets, while ViT helps the model better understand overall patterns and connections in the image. Additionally, to improve classification accuracy and model transparency, we integrate Grad-CAM, a technique that provides visual explanations for model decisions. Evaluating our model on the LC25000 dataset (three classes: Adenocarcinoma of the lung, Squamous cell carcinoma of the lung, Benign lung tissue), we achieve an accuracy of 95%, surpassing traditional CNN-based models and standalone ViT architectures. Statistical tests confirm the robustness of our results, and computational complexity analysis demonstrates that our model is suitable for real-time clinical applications. This hybrid approach provides a scalable, accurate, and interpretable solution for histopathological image classification, laying the foundation for its integration into clinical workflows.

Keywords: *Histopathological image analysis, lung cancer, deep learning, hybrid architecture.*

[1]Faculty of Electrical Engineering, Hanoi University of Industry, Vietnam
[2]Central Military Hospital 108, Vietnam
[*]Email: nampv@haui.edu.vn

Received: 20/8/2025
Revised: 29/10/2025
Accepted: 28/11/2025

1. INTRODUCTION

Lung cancer remained the leading cause of cancer death and the most commonly diagnosed cancer worldwide, with approximately 2.5 million new cases and 1.8 million deaths in 2024. Lung cancer accounts for roughly 12.4% of newly diagnosed cancer cases and about 18.7% of all cancer deaths globally [1]. Detecting these cancers at an early stage is essential to achieving better treatment results, but traditional manual diagnosis methods, such as histopathological image analysis, require significant effort and may lead to errors. Pathologists often rely on analyzing tissue slides under a microscope to identify cancerous cells, a process that is subjective and heavily dependent on the experience of the clinician [2]. As the demand for faster and more accurate diagnoses grows in clinical environments, healthcare institutions are increasingly turning to automation technologies to assist pathologists, streamline the diagnostic process, and enhance the consistency and reliability of results.

Although CNNs like ResNet, VGG, and EfficientNet have shown effectiveness in analyzing histopathological images, these models typically face challenges in capturing global contextual relationships that are essential for distinguishing between benign and malignant tissues. In clinical applications, these models often struggle to detect subtle differences in tissue patterns, which are crucial for accurate diagnosis. Meanwhile, ViTs, which have recently gained attention for their capability to learn non-local feature relationships and extract comprehensive contextual representations, show promise in medical imaging tasks. However, when trained on smaller datasets, ViTs tend to overfit, leading to poor generalization due to the limited amount of data available for training [3-6].

In addition to the challenges related to model performance, interpretability is also crucial in healthcare applications. For doctors to trust and use automated diagnoses, they need to understand how the model makes its decisions. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) assist in highlighting the image regions that the model considers influential for its predictions, allowing doctors to better understand the reasoning behind the diagnosis results. This not only enhances the trust in automated tools but also supports doctors in making more accurate clinical decisions [7-9].

To address these challenges, this paper introduces a hybrid model architecture that combines pretrained ResNet50 with ViT. This combination aims to take advantage of the strengths of both models, overcoming the limitations of overfitting, class imbalance, and explainability. The pretrained ResNet50 is used for local feature extraction, leveraging its ability to capture fine-grained patterns in the tissue images. The features extracted by ResNet50 are then passed to the ViT model, which captures global contextual relationships between tissue structures. To enhance the interpretability of the model's predictions, we incorporate Grad-CAM, which allows pathologists to visualize the areas in the image that influenced the model's inference, making the model's reasoning more transparent and understandable in clinical settings.

2. MATERIALS AND METHODS

2.1. Dataset

The LC25000 dataset consists of 25,000 histopathological images equally divided into five categories: Colon Adenocarcinoma (colon_aca), Colon Benign Tissue (colon_n), Lung Adenocarcinoma (lung_aca), Lung Benign Tissue (lung_n), and Lung Squamous Cell Carcinoma (lung_scc). The images are stained using Hematoxylin and Eosin (H&E), a common method in histopathology where Hematoxylin stains the nuclei blue/purple and Eosin stains the cytoplasm and extracellular matrix pink/red, aiding in the distinction of tissue structures and cancerous abnormalities. Initially, A total of 1,250 images of cancerous tissue were obtained from pathology slides at the James A. Haley Veterans' Hospital (Tampa, Florida), with 250 samples allocated to each category [10]. The base dataset included 750 lung tissue samples, divided into 250 adenocarcinoma, 250 squamous cell carcinoma, and 250 benign tissue samples, as well as 500 colon tissue samples were included, consisting of 250 adenocarcinoma and 250 benign specimens. Data augmentation was performed using transformations such as image rotation and horizontal or vertical flipping, expanding the dataset to 5,000 samples per class and yielding a total of 15,000 images representing lung cancers. A 5-fold cross-validation protocol was applied, with 60% of the data used for training, 20% for validation, and 20% for testing.

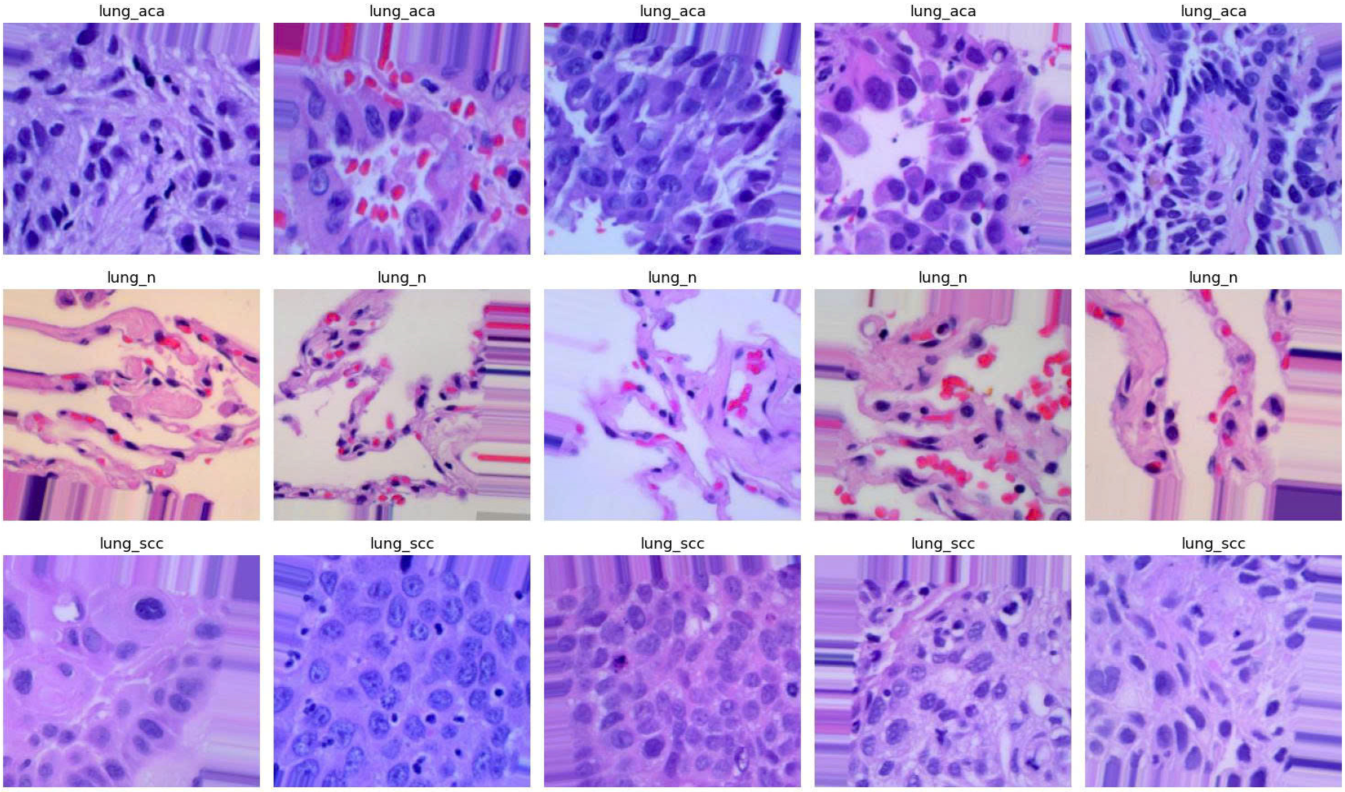

Figure 1. Sample images of LC25000 dataset

The above figure presents sample images from the LC25000 dataset, showing different classes of tissue stained using Hematoxylin and Eosin (H&E). The H&E staining enhances the visibility of tissue components, with the blue/purple areas representing cell nuclei and the pink/red areas corresponding to the cytoplasm and extracellular matrix. These images help highlight the distinct characteristics of benign and malignant tissues. The dataset serves as a valuable resource for deep learning models trained to classify various types of cancer, including colon and lung cancers.

2.2. Preprocessing

To ensure optimal performance, we applied several preprocessing techniques to the images:

Resize and Crop: All images were resized and cropped to a standardized resolution of 768 x 768 pixels. Initially, the images had a resolution of 1024 x 768 pixels, but they were cropped to a square shape (768 x 768 pixels) to maintain consistency in size across all images. This step helps ensure that all input images have the same dimensions, which is crucial for deep learning models, as varying image sizes can lead to inefficient processing and training.

Data Augmentation: To improve the robustness and generalization of the model, Data augmentation was implemented to enhance the diversity of the training set. This process applies a series of image transformations to the original samples, thereby expanding the dataset and improving model generalization. Common transformations included rotation, flipping, zooming, shifting, and scaling. These techniques help the model generalize better by learning to recognize features under different variations, such as different orientations or zoom levels, which may occur in real-world scenarios.

Normalization: This technique helps improve the training process and convergence speed of models. When working with images, pixel values can vary significantly in range (usually from 0 to 255 for color images), which can cause difficulties during training. Normalization helps standardize the pixel values such that they have zero mean and unit variance. This reduces variability in the data and improves the model’s effectiveness in learning efficiently.

Normalization Formula:

Let x be a pixel value from an image. The formula for normalization is as follows:

$$X_{\text{norm}} = \frac{x-\mu}{\sigma}$$

Where:

X_{norm} is the normalized pixel value.

x is the original pixel value (range 0 - 255).

μ is the mean of all the pixel values in the dataset.

σ is the standard deviation of all the pixel values in the or dataset.

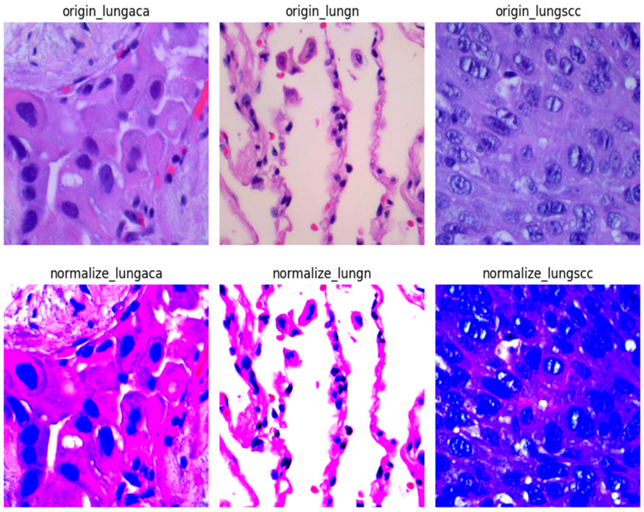

Figure 2. Normalized images before training

Table 1. Train, valid, test split data

Class	Training (60%)	Validation (20%)	Test (20%)
Lung_aca	3000	1000	1000
Lung_n	3000	1000	1000
Lung_scc	3000	1000	1000

2.3. Hybrid Model Architecture

The proposed architecture utilizes a pretrained ResNet50 model combined with ViT encoder for enhanced image analysis performance. This architecture consists of two key components: the ResNet50 feature extraction module and ViT encoder.

- Pretrained ResNet50 Block: The pretrained ResNet50 model is employed for feature extraction from the input image. As a deep convolutional neural network, ResNet50 is distinguished by its residual connections, which effectively address the vanishing gradient issue in deep architectures. The pretrained model is fine-tuned on the current dataset to adapt the model's learned features to the specific task. ResNet50 consists of multiple convolutional layers, each followed by batch normalization and ReLU activation functions. These layers enable efficient extraction of hierarchical features, preserving detailed information in the image (Fig. 3).

- ViT Block: The feature maps extracted from ResNet50 are passed into the ViT model, which is designed to capture long-range dependencies in images. The image is divided into patches, which are linearly embedded and processed through multiple transformer layers, each consisting of multi-head attention and feed-forward networks. The ViT block captures global contextual relationships between tissue structures, allowing it to efficiently tackle difficult image classification challenges (Fig. 3).

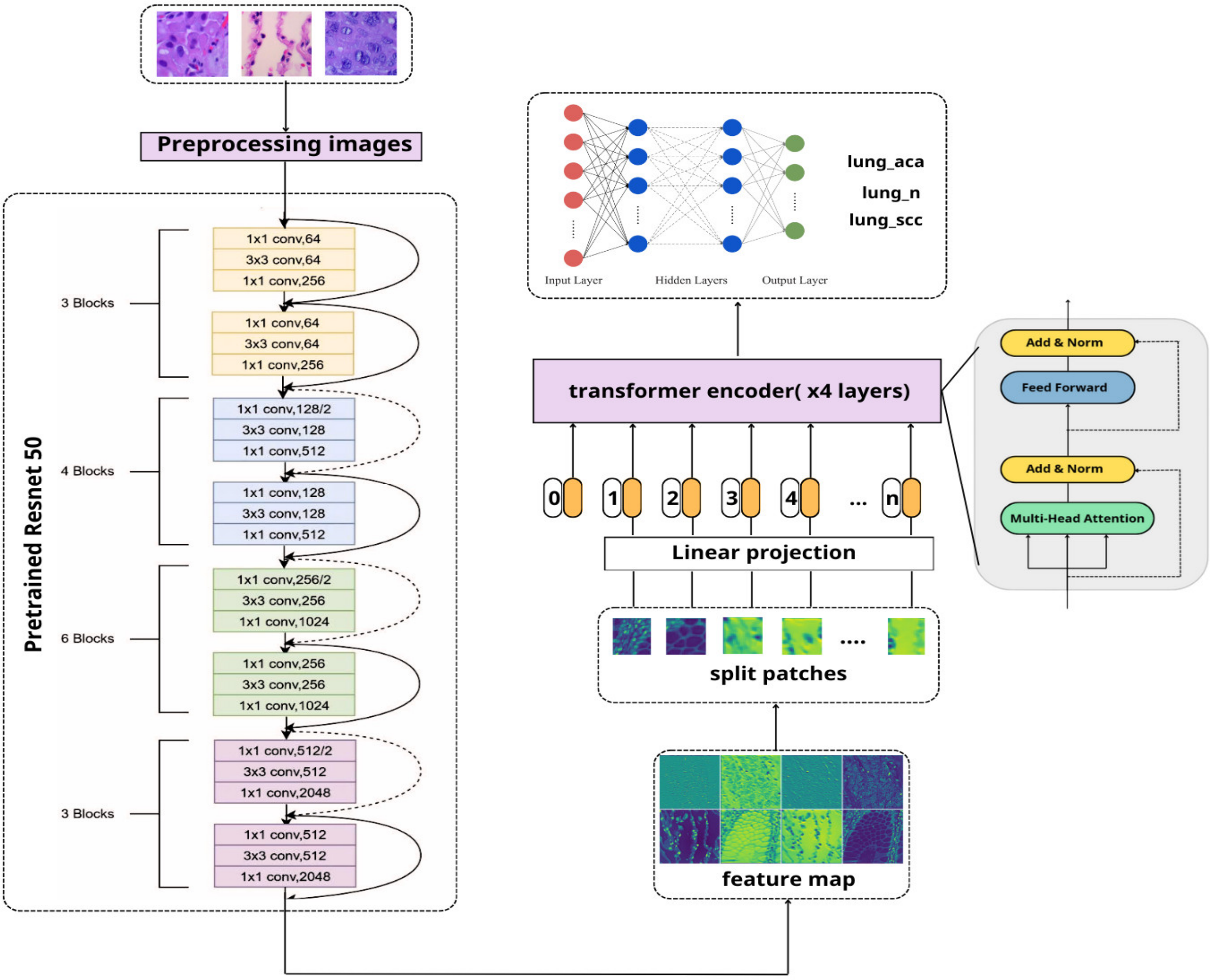

Figure 3. The proposed architecture for histopathological image classification

2.4. Training Process

The Adam optimizer, with a learning rate of 1e-4, was utilized due to its adaptive learning rate mechanism, which facilitates more efficient training. The Cross-Entropy Loss function was used, as it is a standard loss function for classification tasks, particularly when dealing with categorical targets. A batch size of 32 was selected to optimize the trade-off between memory usage and the performance of the model. Initially, the model was trained for 50 epochs.

3. RESULTS AND DISCUSSION

3.1. Classification Performance

The classification results for the lung cancer are presented using a confusion matrix, which evaluates the model's effectiveness across three distinct classes: Lung_aca, Lung_n, and Lung_scc. The confusion matrix is shown in two formats: raw counts on the left and percentages on the right. These matrices provide a detailed breakdown of the true positive, false positive, true negative, and false negative predictions, offering a comprehensive understanding of classification accuracy and misclassification patterns.

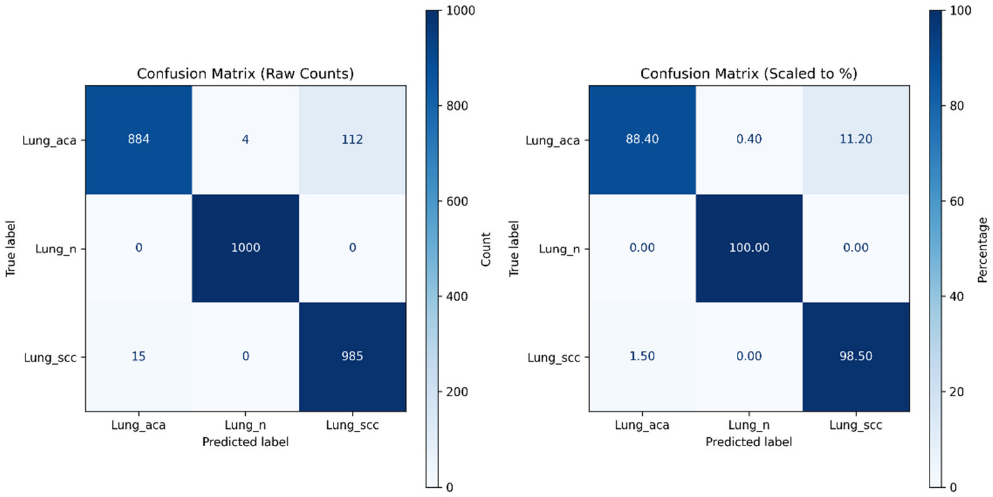

Figure 4. Confusion matrix of test set

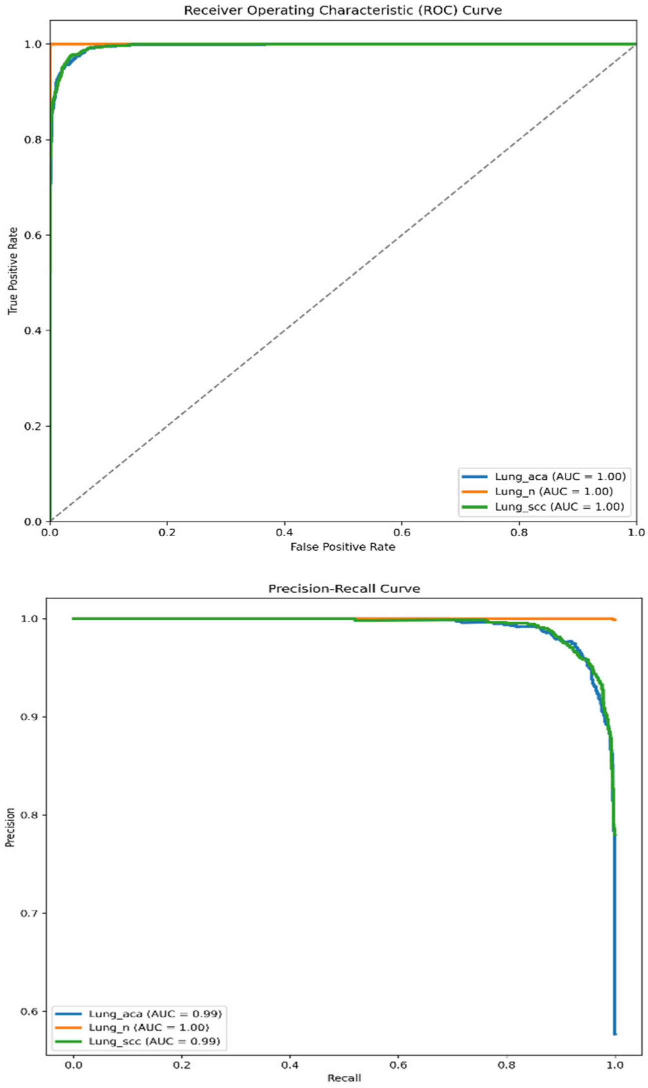

Figure 5. Roc and precision-recall curves

From the confusion matrix, it is evident that the model performs exceptionally well in identifying Lung_n samples, achieving a perfect classification rate of 100%. Strong performance is also observed for Lung_scc, with a high accuracy of 98.50%. However, a small percentage of Lung_scc instances were misclassified as Lung_aca. The Lung_aca class, on the other hand, shows a slightly higher misclassification rate of 11.20%.

The ROC curve illustrates the trade-off between the True Positive Rate and False Positive Rate, while the Precision-Recall curve demonstrates the relationship between Precision and Recall.

Fig. 6 illustrates the application of Grad-CAM visualizations on lung cancer histopathological images, processed by a model combining pretrained ResNet50 for feature extraction and ViT for capturing long-range dependencies. The Grad-CAM heatmaps highlight the regions of the images deemed most significant by the model in making its predictions, providing insights into the decision-making process of the deep learning model.

3.2. Comparison with Baseline Models

The following table shows an evaluation comparing our hybrid model to the baseline model CNN models (ResNet-50, DenseNet-121) on various performance metrics. Our hybrid architecture consistently outperforms all baselines in terms of accuracy, precision, and recall, especially in the Lung Squamous Cell Carcinoma (Lung_scc) class.

The Proposed method stands out as the top performer, achieving the highest accuracy, precision, recall, and F1-Score (95.6%, 95.9%, 95.6%, 95.6%), demonstrating excellent classification accuracy and a strong balance between metrics. While it requires a relatively larger number of parameters (29.6 million) and FLOPs (15.2 GFLOPs), resulting in a longer inference time (45ms) compared to other models, it excels in critical applications where high accuracy is paramount. This makes the Proposed method a more efficient choice for high-performing image classification tasks. In comparison, ViT-Base performs well with high accuracy (92.1%) and good precision/recall but demands a

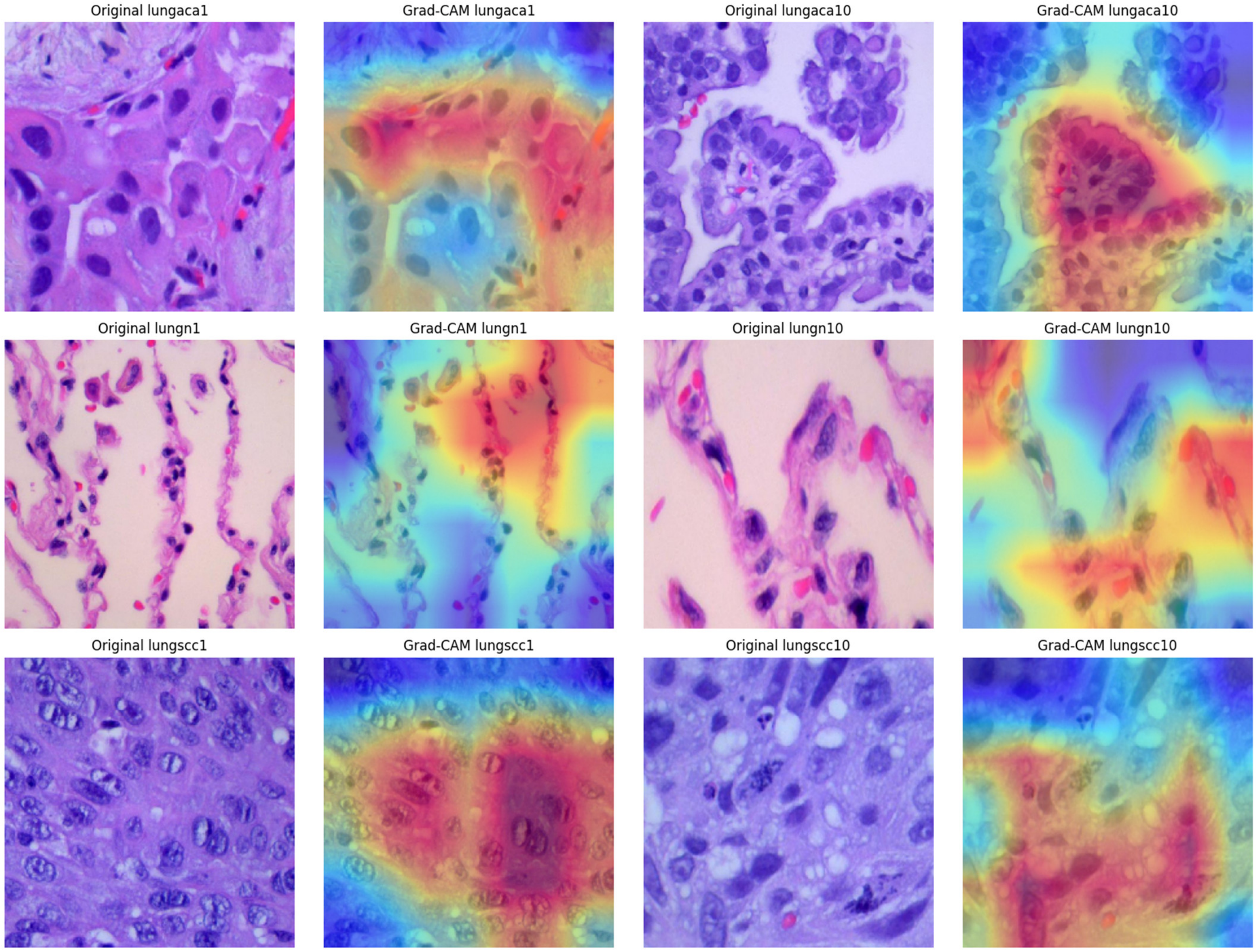

Figure 6. Visualization of lung cancer images with Grad-CAM heatmaps

Table 2. Performance comparison of the proposed hybrid model with baseline CNNs

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FLOPs (G)	Inference Time (ms)
ResNet-50	87.5	88.1	85.7	86.9	25.6	12.8	32
DenseNet-121	88.2	89.4	86.8	88.0	33.5	14.3	38
ViT-Base	92.1	92.5	91.8	92.1	86.0	33.1	78
Proposed method	**95.6**	**95.9**	**95.6**	**95.6**	**29.6**	**15.2**	**45**

significant computational cost (86 million parameters, 33.1 GFLOPs), which results in a slower inference time (78ms), limiting its suitability for real-time applications. On the other hand, ResNet-50, with the fewest parameters (25.6 million) and the fastest inference time (32ms), is optimal for applications requiring fast processing in real-time settings.

4. CONCLUSION

The proposed hybrid deep learning model, which combines pretrained ResNet50 and Vision Transformer (ViT), addresses key challenges in the classification of lung cancer from histopathological images. By leveraging the strengths of both CNNs for local feature extraction and ViT for capturing global dependencies, the model delivers improved performance, achieving an accuracy of 95.6%. This approach effectively mitigates issues such as overfitting and class imbalance, while also enhancing interpretability through Grad-CAM, which visualizes the decision-making process, ensuring transparency for clinical applications.

When compared to baseline models like ResNet-50 and DenseNet-121, ViT-Base, the hybrid model outperforms them in all evaluation metrics, namely accuracy, precision, recall, and F1-score. Despite having a slightly higher number of parameters, it demonstrates an ideal balance between accuracy and model complexity, making it suitable for real-time, high-performance clinical applications. This hybrid architecture provides a scalable, accurate, and interpretable solution for histopathological image classification, with strong potential for integration into clinical workflows to support better decision-making and improved patient outcomes.

FUNDING STATEMENT

This work was supported by the Project Granted number [49-2024-RD/HĐ-ĐHCN], Hanoi University of Industry.

REFERENCES

[1]. World Health Organization (WHO), *Global cancer burden growing, amidst mounting need for services*. World Health Organization, 2024. Retrieved from https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

[2]. Madabhushi Anant, "Digital Pathology Image Analysis: Opportunities and Challenges," *Imaging in Medicine*, 1, 7-10, 2009. Doi: 10.2217/iim.09.9.

[3]. T. Hussain, et al., "EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification," *IEEE Access*, 13, 54040-54068, 2025. doi: 10.1109/ACCESS.2025.3554184.

[4]. Debnath Jesika, Khondakar Pranta Al, Hossain Amira Sakib, Anamul Haque, Rahman Hamdadur, Haque Rezaul, Ahmed Md. Redwan, Reza Ahmed Wasif, Swapno S M Masfequier Rahman, Appaji, Abhishek, "LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis," *Informatics in Medicine Unlocked*, 57. 1-27, 2025. Doi: 10.1016/j.imu.2025.101669.

[5]. Katar O., Yildirim O., Tan S., Acharya U. R., "A Novel Hybrid Model for Automatic Non-Small Cell Lung Cancer Classification Using Histopathological Images," *Diagnostics*, 14(22), 2497, 2024. https://doi.org/10.3390/diagnostics14222497.

[6]. Vinoth N.A.S., Kalaivani J., Arieth R.M., et al., "An enhanced fusion of transfer learning models with optimization based clinical diagnosis of lung and colon cancer using biomedical imaging," *Sci Rep*, 15, 24247, 2025. https://doi.org/10.1038/s41598-025-10246-0.

[7]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 618-626, 2017. doi: 10.1109/ICCV.2017.74.

[8]. Suara S., Jha A., Sinha P., Sekh A. A., "Is Grad-CAM Explainable in Medical Images?," *ArXiv*., 2023. https://doi.org/10.1007/978-3-031-58181-6_11.

[9]. Tang D., Chen J., Ren L., Wang X., Li D., Zhang H., "Reviewing CAM-Based Deep Explainable Methods in Healthcare," *Applied Sciences*, *14*(10), 4124, 2023. https://doi.org/10.3390/app14104124.

[10]. Masud M., Sikder N., Nahid A.A., Bairagi A.K., AlZain M.A., "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, 21, 748, 2021.