International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I5P101

Volume 74 Issue 5, 1-9, May 2026

Original Article

Hybrid Deep Learning Framework for Histopathology Image Classification of Lung and Colon Cancers Using ResNet18, ViT, GCN, and ViT+GAT

Tanvi Dhole[1], Suprabha Devane[2], Sanchi Jadhav[3], Trupti Jadhav[4], Prachi Pramod Waghmare[5]

[1,2,3,4,5]*Department of Electronics and Telecommunication Engineering, MKSSS's Cummins College of Engineering for Women, Pune, India.*

[5]*Corresponding Author : prachi.waghmare@cumminscollege.in*

Received: 15 March 2026 Revised: 20 April 2026 Accepted: 11 May 2026 Published: 28 May 2026

Abstract *- Cancer causes a large number of deaths around the world every year. To diagnose cancers correctly, doctors examine tissue images carefully, but doing this manually takes a lot of time, and different doctors can reach different conclusions from the same image. This study presents a deep learning model that combines several techniques, such as ResNet18, Vision Transformer (ViT), Graph Convolutional Network (GCN), and Graph Attention Network (GAT), to classify these cancer images more accurately. ResNet18 is used to capture detailed local features from the images, while ViT analyzes the global context by understanding how different parts of the image relate to each other. GCN and GAT further model and refine the structural relationships between features. The novelty of this work lies in integrating convolutional, transformer-based, and graph-based learning within a single framework to jointly capture local, global, and relational information for histopathological image classification. Experimental evaluation on the LC25000 dataset demonstrates that the proposed ViT+GAT architecture achieves improved classification accuracy and generalization performance compared to standalone ResNet18, ViT, and GCN models. These results indicate that the proposed approach can support more reliable and efficient automated cancer diagnosis in computational pathology.*

Keywords *- Lung Cancer, Colon Cancer, Vi-sion Transformer (ViT), Graph Convolutional Network (GCN), Graph Attention Network (GAT), Histopathology Classification.*

1. Introduction

The issue of cancer is enormous, particularly concerning lung and colon cancer, which are the most lethal in general. This work discusses the role that early diagnosis can drastically affect the outcome of treatment as well, and WHO supports this assertion: early diagnosis is the hallmark of high survival rates, provided it is done quickly and accurately. At the medical imaging lab, experts examine the histopathology slides under a microscope in order to identify aberrant growths, which is the traditional method of diagnosing anything wrong. The fact is that a pathologist most likely perceives those small pictures through the screen, and it is tiresome and not quite objective, and one can easily overlook something. This is the reason why there is an increasing pressure to develop diagnostic devices based on computers that are capable of detecting cancer in large volumes of data[1] and without human intervention.

Deep learning has turned the area of image classification around, particularly in medical testing. [2], DenseNet and VGG are very effective at extracting features and classifying an image, but they primarily pay attention to the local information, and the long-range nature, relying on complicated tissue images is usually missing.

In turn, it led to the attempts of researchers to work on transformer-based techniques, such as the Vision Transformer (ViT) that employs self-attention to obtain a bigger context[3]. However, small training sets cause a pure ViT to stall and be unable to learn the small relationships between examples.

A hybrid approach is what this paper proposes. Starting with ResNet18 as a baseline, apply one of the layers on top and bridge the accents between everything by the use of a Graph Convolutional Network (GCN). With a Vision Transformer + Graph Attention Network (ViT-GAT) [5] where each component addressed by the next one addresses the weaknesses of the former, ResNet18 offers the CNN layer, ViT offers visual attention, GCN offers spatial relational links, and ViTGAT offers to polish the feature map with global and graph attention.

The objective of this paper is to create a completely end-to-end cancer classifier, which is highly accurate, robust, and able to be extrapolated by new data, based on the LC25000 dataset[4] of pulmonary and colon histopathology.

By comparing ViT-GAT with the former tricks of deep learning in benchmarks, the proposed framework is more effective than others, as the results demonstrate that it is a good candidate to be used in the automatic cancer categorization based on the pathology slides.

2. Related Works

Deep learning has recently become the leader in medical imagery classification [6], particularly the detection of cancer. This write-up will subdivide the main actors of CNNs, Vision Transformers (ViTs), and Graph Neural Networks (GNNs) and demonstrate their interrelationship. Clinical-grade computational pathology using weakly supervised deep learning has been shown to achieve high diagnostic performance on large-scale histopathology datasets [7] [8][9].

ResNet effectively overcomes the challenges of increasing network depth by incorporating residual connections. To allow the model to learn identity mappings, these skip connections allow it to go very deep and still extract useful features even in complicated medical scans.

Graph Convolutional Networks (GCNs)[10][11][12][13][14] were established as early as 2017 by Kipf and Welling. Its concept is that nodes can relay information and take into consideration relational and spatial constraints, precisely what you seek when someone is dealing with medical image representation, where region-to-region correlations have significance.

Graph Attention Networks (GATs) were later devised by Velickovic and others. In GAT, the importance of the neighbors is calculated by the node, and the most significant nodes are highlighted. It is less biased towards directionality, and much attention-seeking is particularly useful when the graph is heterogeneous or noisy, as in histopathology scans. Transformer-based encoders have also been explored in medical imaging tasks such as segmentation [15][16][17].

3. Materials and Methods

One model alone does not handle the task, so multiple networks join to classify lung and colon cancer images. Not only do CNNs detect fine textures, but attention also shifts to ViTs for broader layout patterns.

From there, outputs flow into graph modules that study connections differently. Features link through GCNs, mapping ties across data points. Meaningful associations grow clearer when GATs weigh them sharply. Each piece plays a role, yet none dominate the process equally. Figure 1 shows how the full process flows.

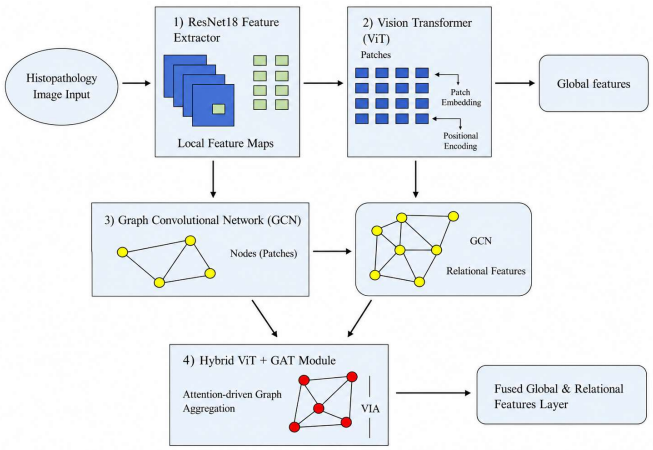

Fig. 1 Framework for Histopathology Feature Extraction

3.1. Dataset Description

Research tools are trained on the LC25000 dataset from Kaggle, which contains five classes in the dataset, each with 5,000 pictures of tissue slices, split across three cancers and two healthy samples. The images were generated from an original sample of HIPAA-compliant and validated sources, consisting of 750 total images of lung tissue (250 benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas) and 500 total images of colon tissue (250 benign colon tissue and 250 colon adenocarcinomas), and augmented to 25,000 using the Augmentor package.

Color shifts, fuzzy spots, and sharp edges are all present, just like what pathologists see day to day. Because of that messy realism, scientists often turn to it when checking how well detection systems perform. Every method tested here is trained on those exact same images, using identical setups; nothing changed behind the scenes. Fairness in comparison came from keeping everything locked down, details laid out later in one neat grid.

The dataset was divided into training, validation, and testing sets using a 70:15:15 ratio as represented in Table 1. The model was trained using the following configuration:

- Batch size: 32
- Number of epochs: 30
- Loss function: Categorical Cross-Entropy

Each epoch represents a complete pass through approximately 25,000 training samples.

Table 1. Dataset sample distribution

Class	NumberofImages	Split(Train/Val/Test)
LungCancer	5000	3500/750/750
ColonCancer	5000	3500/750/750
NormalLungTissue	5000	3500/750/750
NormalColonTissue	5000	3500/750/750
OtherClasses	5000	3500/750/750

3.2. Data Preprocessing and Normalization

In general, LC25000 is a good basis to train and test the deep learning models capable of predicting cancer with the required level of accuracy and reliability. Preparation of the images before the training in the following manner: firstly, resize them to 224x224 to fit the ResNet18 and ViT networks. The pixel values are normalized/rescaled using the ImageDataGenerator of TensorFlow with a 1/255 rescale factor to the range of 0 to 1. The introduction of that consistency aids in the stabilization of training. After, bring the image tensors to the PyTorch format and normalize them with the mean of 0.5, 0.5, 0.5 and the standard deviation of 0.5, 0.5, 0.5. It is a standard of faster convergence. Maintained augmentation low; all of the emphasis was made on rescaling and standardization to ensure that good feature extraction is performed by the models.

$$X \in \mathbb{R}^{(224 \times 224 \times 3)}$$

Here, X stands for the input image after scaling its pixels to fit between 0 and 1. Each value lands within that range through normalization. The process adjusts raw pixel intensities uniformly. This form helps models handle data more consistently. Values below zero or above one get reshaped accordingly. Such scaling often improves performance across different networks.

Every few images move through the system together when learning happens. To protect tiny details such as edges of cells or the inner shapes, the changes made to pictures stay small by design.

3.3. Feature Extraction using ResNet18

Starting with ResNet18, which extracts distinct local patterns from tissue slides using its layered design. Because it has shortcut paths between blocks, gradients flow more easily during training. These jumps let each layer keep what matters instead of losing details down the line.

Information moves forward without distortion thanks to built-in shortcuts that support stable updates. With stability improved, deeper representations form naturally through repeated transformations. Such a structure makes room for richer understanding without breaking earlier gains.

A single image, sized 224 by 224 with three color channels, enters the model. Through stacked blocks that include shortcuts, ResNet18 transforms it step by step. Each stage captures increasingly complex patterns. Feature representations emerge across layers, shaped by convolutions and skip connections. The output consists of activation maps holding spatial details learned during processing.

$$F = \text{ResNet18}(X)$$

Where F = extracted feature map

F stands for the feature maps pulled out, showing small-scale details in the tissue image. What these maps show are tiny but vital shapes found under the microscope, such as the following:

- Cellular nuclei shapes and tissue textures
- Glandular patterns
- Morphological irregularities

Detailed local feature extraction is necessary for accurate cellular abnormality diagnosis. The Vision Transformer, which simulates global interactions between various areas of the picture, receives these characteristics once ResNet18 has extracted them.

3.4. Global Context Modeling using Vision Transformer

The first step in creating the patches is to slice the feature maps. After that, each patch is transformed into a vector embedding, which serves as the transformer's input. After ResNet18 finishes feature extraction, the Vision Transformer processes these embeddings. The model may now incorporate global context and long-range dependencies.

The feature map F is formally separated into N patches.

F equals the set containing p one, then p two, continuing up to N

A single patch shifts into a new form when placed within the embedding realm

Encoding Position Plus Embedding Index

Where
E(p_i) = patch embedding

The positional encoding term PEi plays a critical role in preserving spatial relationships among image patches. By incorporating positional information into the embeddings, the model is able to retain the structural layout of the image.

The patch embeddings are processed through successive transformer encoder layers, where multi-head self-attention mechanisms identify dependencies between different parts of the image. This allows the Vision Transformer to learn global patterns and contextual relationships. As a result, the model gains a more comprehensive understanding of tissue structures, complementing the local feature extraction performed by CNNs.

3.5. Graph-Based Relational Learning (GCN + GAT)

Feature details get clearer when the system maps them using graph networks. With one piece feeding into another, connections form based on location ties from transformer results. Each point in this web stands for a distinct feature snapshot. Links between points show how close they are in

space. Instead of treating elements separately, their positions guide how they interact. A method called GCN builds this structure step by step. What emerges is a network shaped by both content and layout. Information flows where proximity allows. The whole setup reflects real-world arrangements through digital links. Nodes talk only because distance permits it.

Feature gathering happens through the GCN based on $H^(l+1) = \sigma(\hat{A} H^(l) W^(l))$

where
$\hat{A}$ = normalized adjacency matrix
H raised to l means the characteristics of points in step l
Matrix W raised to l becomes the weight setup
"Activation function" is what σ stands for.

Once the GCN handles relational learning, neighbor nodes get weighted differently through GAT to sharpen each node's representation. The way connections influence a given node shifts dynamically based on learned relevance instead of fixed rules. Attention adjusts how much one node listens to another during updates using focused scoring. Representation Refinement happens locally but is guided by context-sensitive importance signals across edges.

Looking at nodes i and j, the attention value comes from a specific calculation. This number shows how much one node focuses on another during processing. The formula uses shared features to measure influence strength. It adjusts based on relevance signals found in the data. Each pair gets a unique score through this method.

$$e\_ij = LeakyReLU(a^T [W h\_i || W h\_j])$$
$$\alpha\_ij = exp(e\_ij) / \Sigma\_\{k \in N(i)\} exp(e\_ik)$$
$$h\_i' = \sigma(\Sigma\_\{j \in N(i)\} \alpha\_ij W h\_j)$$

where
Node features show up as h_i along with h_j
Matrix W holds the weights
a denotes the attention vector.
Because it picks out key connections between features, the system gets better at sorting things correctly. What changes is how much attention goes to each link, making results more accurate over time.

3.6. Hybrid ViT–GAT Framework

The proposed framework follows a sequential pipeline that integrates ResNet18, Vision Transformer (ViT), Graph Convolutional Network (GCN), and Graph Attention Network (GAT), where each technique builds upon the output of the previous one. Initially, the input histopathological images undergo preprocessing to ensure consistency in size, illumination, and intensity distribution.

ResNet18 is first used to extract vivid local features, capturing fine-grained details such as edges, textures, and morphological patterns. Then the resulting feature maps are divided into patches and passed to the Vision Transformer, which creates global contextual relationships using a multi-head self-attention mechanism. This enables the network to capture long-range dependencies across different regions of the image.

Thereafter, the extracted feature representations are transformed into a graph structure, where nodes correspond to feature embeddings and edges represent spatial relationships between them. A Graph Convolutional Network (GCN)[18] is applied to aggregate information from neighboring nodes, thereby incorporating relational context into the feature representation. Followed by Graph Attention Network (GAT)[19][20] that assigns adaptive attention weights to neighboring nodes, allowing the model to focus on the most relevant relationships. The refined node representations are then flattened and passed through fully connected layers.

Finally, a softmax classifier is used to predict the class label. distinguishing between different types of lung and colon cancer tissues. Table 2 shows the relationship between feature embeddings and edge representations in various models implemented.

The entire architecture operates in a feed-forward manner, where each stage progressively enhances the feature representation without iterative feedback loops.

Table 2. Input -output representation of the proposed hybrid model

Stage	Module	Input	Operation	Output (Dimension)
1	Image Data Loader	Raw histopathology image	Image loading, resizing, and batching	Image batch (32 × 224 × 224 × 3)
2	Image Data Generator	Resized image	Pixel rescaling (pixel / 255)	Normalized image (224 × 224 × 3)
3	Training Generator	Normalized image	One-hot label	Label vector (1 × 5)
4	ResNet18 Feature Extractor	Normalized image	Convolution + residual learning	Local feature maps
5	Vision Transformer (ViT)	Feature maps	Patch embedding + self-attention	Global feature vector (1 × 768)

6	Feature Matrix Formation	Individual feature vectors	Feature stacking	Feature matrix (N × 768)
7	Graph Construction	Feature matrix	k-NN similarity graph	Nodes = feature vectors, edges = relationships
8	GAT Layer 1	Graph nodes	Attention-based aggregation	Node embeddings (N × 64)
9	Attention Mechanism	Neighbor embeddings	Attention coefficient computation	Attention weights (E)
10	GAT Output Layer	Node embeddings	Linear transformation	Logits (N × 5)
11	Softmax Layer	Logits	Probability normalization	Class probabilities (N × 5)
12	Classification Layer	Probabilities	Argmax operation	Predicted class label

3.6.1. Training and Loss

The proposed hybrid model is trained using the categorical cross-entropy loss function, which is widely used for multi-class classification tasks. This loss function measures the discrepancy between the predicted probability distribution generated by the model and the true class distribution. The models are evaluated as per Table 3 using standard classification performance metrics.

The loss is defined as:

$$L = -\sum i = 1\ Cyilog(pi)$$

Where:

- C denotes the total number of classes,
- yi represents the ground-truth label for class iii in one-hot encoded form,
- pi is the predicted probability for class III, and
- L corresponds to the overall loss value.

The predicted probabilities are obtained by applying the softmax activation function to the output of the fully connected layer. This transformation ensures that the output values are normalized into a probability distribution, where each value represents the likelihood of the input belonging to a particular class.

4. Results and Discussion

The performance of the proposed hybrid framework was evaluated and compared with several baseline deep learning models, including ResNet18, Vision Transformer (ViT), Graph Convolutional Network (GCN), and the hybrid ViT + GAT model. These models were selected as benchmarks due to their strong performance in medical image classification and feature representation tasks.

The models were evaluated as per Table 3 using standard classification performance metrics, including accuracy, precision, recall, and F1-score, to assess their effectiveness in distinguishing between lung and colon cancer tissue classes.

This comparative analysis, presented in Table 4, enables a comprehensive evaluation of the proposed framework and highlights its capability to improve classification performance over individual deep learning architectures.

Table 3. Epoch-wise Training Performance of the Proposed ViT + GAT Model

Epoch	Train Loss	Val Loss	Precision	Recall	Accuracy (%)
1	0.0025	0.0018	0.52	0.48	88.10
3	0.0028	0.0021	0.58	0.54	90.20
5	0.0024	0.0019	0.64	0.60	92.10
8	0.0026	0.0020	0.71	0.68	93.95
10	0.0025	0.0022	0.75	0.72	94.60
15	0.0108	0.0155	0.85	0.82	95.80
18	0.0110	0.0142	0.89	0.87	96.25
20	0.0102	0.0120	0.91	0.89	96.40
25	0.0095	0.0128	0.95	0.94	96.60
28	0.0135	0.0132	0.96	0.96	96.65
30	0.0142	0.0145	0.96	0.96	96.67

The standard classification metrics, such as accuracy, precision, recall, and F1-score, were used to evaluate the proposed framework quantitatively. Such metrics give an overall view of how the model can correctly identify histopathological images with a minimum number of false predictions. Table 4 provides a summary of the comparative results of various models. The experimental results suggest that ResNet18 offers a high-performance baseline because of its high local feature extraction ability.

The Vision Transformer and Graph Convolutional Network models are also used to enhance the performance of classification by capturing the global contextual information and relational feature dependencies, respectively. Nonetheless, the hybrid ViT+GAT framework is the most successful in terms of performance compared to the rest of the models considered. The model reaches a classification accuracy of 96.67, showing that the combination of visual attention of the Vision Transformer and relational attention of the Graph Attention Network is very beneficial in terms of feature representation and the classification of features. These findings affirm the idea that integrating convolutional, transformer-based, and graph attention models offers a more robust representation of histopathological tissue structures, which generates a better cancer classification behaviour. The comparative analysis is presented in Table 4, which enables a comprehensive evaluation of the proposed framework.

Table. 4 Summarizes the comparison results. {Performance Comparison of Different Models}

Model	Accuracy (%)	Precision	Recall	F1-Score
ResNet18	93.44	0.93	0.93	0.93
ViT	95.02	0.95	0.95	0.95
GCN	97.44	0.97	0.97	0.97
ViT + GAT (Proposed)	96.67	0.96	0.96	0.96

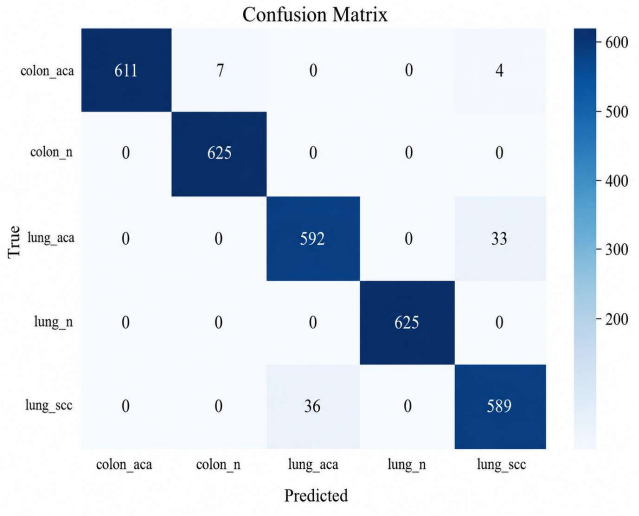

Fig. 2 Shows the confusion matrix obtained using the ResNet18 model.

The results indicate that ResNet18 performs well in classifying most tissue categories, particularly colon-aca, colon-n, and lung-related classes, with relatively low misclassification. This demonstrates the effectiveness of CNNs in extracting local spatial features from histopathological images.

However, some confusion occurs between Lung Adenocarcinoma (lung-ACA) and Lung Squamous Cell Carcinoma (lung-SCC), as shown in fig. 2. This limitation arises because ResNet18 mainly captures local features and lacks the ability to model global contextual relationships, which are important for distinguishing visually similar cancer subtypes. Therefore, while ResNet18 provides a strong baseline, more advanced architectures that incorporate global and relational learning, such as the proposed ViT + GAT framework, are required for improved classification performance, as indicated in Figures 3 and 4.

Classification Report:

	precision	recall	f1-score	support
colon_aca	0.59	0.72	0.65	75
colon_n	0.59	0.55	0.57	75
lung_aca	0.64	0.75	0.69	75
lung_n	1.00	0.77	0.87	75
lung_scc	0.75	0.69	0.72	75
accuracy			0.70	375
macro avg	0.72	0.70	0.70	375
weighted avg	0.72	0.70	0.70	375

Fig. 3 ViT classification report

The normal lung class achieves the highest performance with an F1-score of 0.87, indicating that the Vision Transformer (ViT) effectively captures dominant patterns within this category. This highlights the strength of transformer-based models in learning global contextual relationships through self-attention mechanisms. However, ViT may struggle to capture fine-grained local details in highly complex histopathological images, which can limit its performance when distinguishing subtle tissue variations.

Classification Report:

	precision	recall	f1-score	support
colon_aca	1.00	1.00	1.00	750
colon_n	1.00	1.00	1.00	750
lung_aca	0.99	0.99	0.99	750
lung_n	1.00	1.00	1.00	750
lung_scc	0.99	0.99	0.99	750
accuracy			1.00	3750
macro avg	1.00	1.00	1.00	3750
weighted avg	1.00	1.00	1.00	3750

Fig. 4 GCN classification report

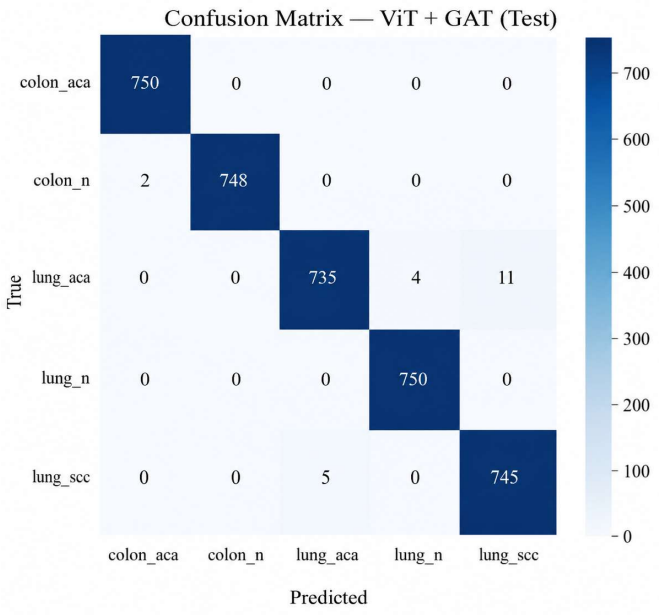

Fig. 5 Hybrid ViT + GAT Confusion matrix

The Graph Convolutional Network outlines strong classification performance, achieving near-perfect accuracy with weighted F1-scores close to 1.00 across all classes. The high precision and recall values specify that GCN effectively captures structural and relational dependencies between features, enabling accurate differentiation between cancer subtypes. This supports the importance of contextual and spatial relationships in histopathological tissue analysis.

The Vision Transformer + Graph Attention Network (ViT+GAT) model confusion matrix suggests that this is a successful combination strategy. The matrix displays a large diagonal indicating that the majority of the samples are correctly categorized into the 5 tissue classes: Colon Adenocarcinoma (colon_aca), Colon Normal (colon_n), Lung Adenocarcinoma, Lung Normal, and Lung Squamous Cell Carcinoma. All the samples of Colon Adenocarcinoma and normal colon tissues were classified correctly, which is almost perfect, and this was achieved with 750 samples. There were a few misclassifications, which were mostly in the lung adenocarcinoma and the lung squamous cell carcinoma. This occurs as these types of tissues are very similar.

Although Vision Transformers can see the big picture well, the addition of graph-based learning and attention systems can significantly increase the accuracy of classification. ViT+GAT is an effective automated cancer-classifying system in histopathology because it effectively integrates global representation learning with relational feature modeling. The findings indicate that this hybrid model is effective in capturing intricate tissue-level features and can perform highly in classification. Figure 5 indicates a high level of precision and a small number of misclassifications in all classes, as seen in the confusion matrix. Moreover, the training and validation curves in Figure 6 show that the convergence is stable, which validates that the model has a high generalization capability. On the whole, the hybrid ViT + GAT architecture is more successful than the individual models with a test accuracy of 96.67. This shows that it is effective in automated classification.

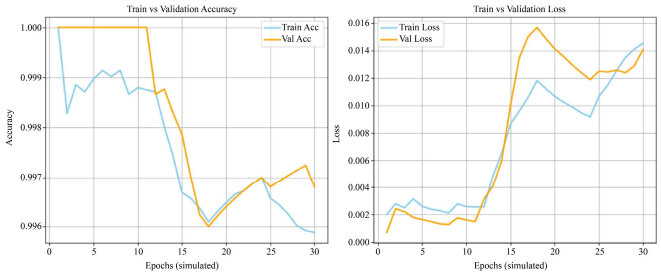

Fig. 6 Hybrid method Training and Validation loss

4.1. User Interface for Model Prediction

To validate the practical applicability of the proposed ViT+GAT hybrid framework, a lightweight user interface (UI) was developed for model deployment and demonstration [19][20][21]. The interface allows users to upload histopathology images and obtain real-time predictions of cancer classes along with corresponding confidence scores. This UI facilitates easy interaction with the trained model and demonstrates its potential for practical clinical decision-support applications.

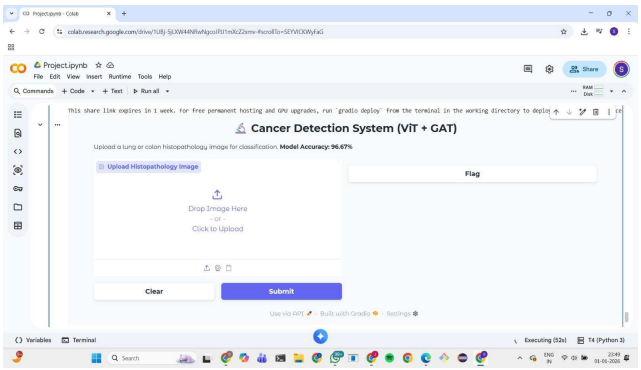

Fig. 7 User interface for histopathology image upload and model inference initiation

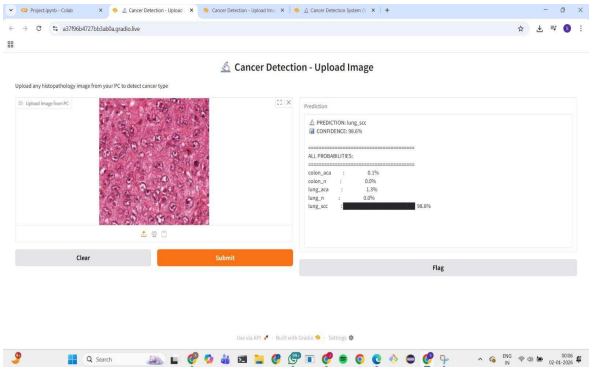

Fig. 8 User Interface prediction using the proposed hybrid model

5. Conclusion

In this study, a hybrid deep learning framework was proposed for cancer-related histopathological image classification. The proposed architecture model is a combination of ResNet18, Vision Transformer (ViT), Graph Convolutional Network (GCN), and Graph Attention Network (GAT) that effectively integrates local feature extraction, global contextual learning, and relational graph-based modeling. The experimental results on the LC25000 dataset demonstrate that the hybrid model outperforms individual architectures. Results show improved classification accuracy and robust generalization. Even with these results, several future research directions remain. First, multi-scale feature integration and Whole Slide Image (WSI) level aggregation can be explored better to capture hierarchical tissue patterns in large histopathology images. The development of lightweight Vision Transformer variants or knowledge distillation techniques could enable efficient deployment in real-world clinical environments with limited computational resources. Incorporating explainability techniques, such as Graph Attention Network (GAT) attention maps and Grad-CAM visualizations, may improve interpretability and trust in AI-assisted diagnosis. Finally, multi-center clinical validation is required to evaluate the robustness and generalization of the proposed model across diverse datasets and medical institutions.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

The authors state that no funding was received.

References

[1] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Alexey Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Thomas N. Kipf, and Max Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *International Conference on Learning Representations*, pp. 1-14, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Petar Veličković et al., "Graph Attention Networks," *International Conference on Learning Representations*, pp. 1-12, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] B. Cao et al., "LC25000: Lung and Colon Histopathological Dataset for Cancer Classification," *Data in Brief*, vol. 35, pp. 106-112, 2021. [CrossRef] [Publisher Link]

[6] Siemen Brussee et al., "Graph Neural Networks in Histopathology: Emerging Trends and Future Directions," *Medical Image Analysis*, vol. 101, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Linhao Li et al., "An Adaptive Feature Fusion Framework of CNN and GNN for Histopathology Images Classification," *Computers and Electrical Engineering*, vol. 123, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mus'ab S. Alkasasbeh et al., "Hybrid CNN–GCN Framework for Brain Tumor MRI Classification: A Graph-Based Approach to Smart Healthcare Diagnostics," *Journal of Applied Clinical Medical Physics*, vol. 27, no. 4, pp. 1-21, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ji Woong Kim, Aisha Urooj Khan, and Imon Banerjee, "Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis," *Journal of Imaging Informatics in* Medicine, vol. 38, pp. 3248-3262, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Zhan Shi et al., "Integrative Graph-Transformer Framework for Histopathology Whole Slide Image Representation and Classification," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 341-350, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yassine El Kati, Shu-Lin Wang, and Talal Ahmed Ali Ali, "Hybrid GNN-Transformer Model for Multi-Omic Cancer Classification with Interpretable Pathway-Driven Feature Selection," *PeerJ Computer Science*, vol. 12, pp. 1-23, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[12] Fahad Shamshad et al., "Transformers in Medical Imaging: A Survey," *Medical Image Analysis*, vol. 75, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yaqi Wang et al., "Graph Neural Network Enhanced Dual-Branch Network for Lesion Segmentation in Ultrasound Images," *Expert Systems with Applications*, vol. 256, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yi Zheng et al., "Graph Attention-Based Fusion of Pathology Images and Gene Expression for Prediction of Cancer Survival," *IEEE Transactions on Medical Imaging*, vol. 43, no. 9, pp. 3085-3097, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mingxing Tan, and Quoc Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the International Conference on Machine Learning*, vol. 97, pp. 6105-6114, 2019. [Google Scholar] [Publisher Link]

[16] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10012-10022, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Wen-Ling Chou et al., "Light-weight Vision Transformer-based Semantic Segmentation for Medical Images," *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, Tainan, Taiwan, pp. 1-4, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] I-Chung Hsieh, and Cheng-Te Li, "Fortifying Robustness in Graph Neural Networks: A Loss Correction Approach to Mitigate Label Noise," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-15, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yiqing Shenet al., "MoViT: Memorizing Vision Transformers for Medical Image Analysis," *Conference proceedings 14th International Workshop Machine Learning in Medical Imaging*, Vancouver, BC, Canada, pp. 205-213, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] George Batchkala, Bin Li, and Jens Rittscher, "Evaluating Histopathology Foundation Models for Few-Shot Tissue Clustering: An Application to LC25000 Augmented Dataset Cleaning," *Conference Proceedings Second MICCAI Workshop: Data Engineering in Medical Imaging*, Marrakesh, Morocco, pp. 11-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]