JPPIPA 12(7) (2026)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Comparative Analysis of Convolutional Neural Network (ResNet-50) and Vision Transformer (ViT-B/16) for Histopathological Image Classification of Colorectal Cancer

Muhammad Fazly Qusyairy[1], Habibullah Akbar[1*], Wahyu Purnama Magribi[1], Khusnul Fajri Rhomadon[1]

[1] Master of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia.

Received: March 30, 2026
Revised: June 09, 2026
Accepted: July 25, 2026
Published: July 31, 2026

Corresponding Author:
Habibullah Akbar
habibullah.akbar@esaunggul.ac.id

DOI: 10.29303/jppipa.v12i7.14880

Abstract: The diagnosis of colorectal cancer (CRC) through histopathological images requires high accuracy to support appropriate clinical decisions. Although Convolutional Neural Networks (CNN) have become the gold standard in medical image analysis, the emergence of Vision Transformer (ViT) architecture offers a new paradigm based on global attention mechanisms (self-attention) that is claimed to be superior on large-scale datasets. However, the effectiveness of ViT-B/16 on medical datasets with limited sample sizes and high texture variation remains debatable. This study aims to comprehensively evaluate the performance of the ViT-B/16 architecture compared to ResNet-50 on the NCT-CRC-HE-100K histopathology dataset, which consists of 9 network classes. The performance of both models was tested using equivalent training scenarios. The evaluation was conducted multidimensionally, covering classification metrics (Accuracy, F1-Score), training stability, feature space separability (t-SNE), visual interpretability (Grad-CAM), and computational efficiency. The experimental results show that ResNet-50 significantly outperforms ViT-B/16 with a test accuracy of 93.24%, compared to ViT-B/16 which only achieves 57.11%. The t-SNE analysis revealed that ViT-B/16 failed to form well-separated feature clusters due to a lack of inductive bias to recognize local features such as cell membrane edges. Failure analysis shows that ViT-B/16 often misclassifies adipose cells as mucus and smooth muscle as tumors. In terms of efficiency, ResNet-50 is 5.8 times lighter in storage size and has lower inference latency. This study concludes that CNN-based architecture (ResNet-50) is still far superior, more stable, and more feasible for clinical implementation than ViT-B/16 in the context of medium-scale histopathological image classification.

Keywords: Colorectal cancer; Computational efficiency; Convolutional neural network; Histopathology; ResNet-50; Vision transformer

Introduction

Colorectal cancer (CRC) has become one of the most persistent global oncology challenges, ranking third as the most commonly diagnosed malignancy and the second leading cause of cancer deaths worldwide (Zhu et al., 2025). In the modern clinical management paradigm, histopathological examination of biopsy or surgical resection specimens stained with Hematoxylin and Eosin (H&E) remains the irreplaceable gold standard procedure (Dunn et al., 2024). Through this microscopic examination, pathologists conduct an in-depth evaluation to determine tumor phenotype, malignancy grade, lymphovascular invasion, and tumor microenvironment characteristics (Lan et al., 2024). This information is crucial in formulating precision treatment strategies and predicting patient survival prognosis. However, reliance on manual visual assessment by humans presents fundamental obstacles in terms of scalability and objectivity. The histopathology analysis

How to Cite:
Qusyairy, M. F., Akbar, H., Magribi, W. P., & Rhomadon, K. F. (2026). Comparative Analysis of Convolutional Neural Network (ResNet-50) and Vision Transformer (ViT-B/16) for Histopathological Image Classification of Colorectal Cancer. *Jurnal Penelitian Pendidikan IPA*, *12*(7), 566–576. https://doi.org/10.29303/jppipa.v12i7.14880

process is a highly complex, time-consuming task that demands a high cognitive load (Wen et al., 2023). Pathologists face the challenge of distinguishing highly similar morphological structures across thousands of fields of view per day, which directly contributes to visual fatigue and high levels of inter-observer variability. This diagnostic inconsistency is not merely a technical issue, but a serious medical risk that can have fatal implications for the accuracy of patient therapy (Snead et al., 2025).

In response to these challenges, the integration of artificial intelligence (AI) into computational pathology has become the most dynamic area of research in the last decade. In particular, Deep Learning algorithms have demonstrated remarkable capacity in automating complex pattern recognition tasks. Since the rise of AlexNet, Convolutional Neural Networks (CNN) architectures have cemented their hegemony as the state-of-the-art method in medical image analysis (Y.-L. Wang et al., 2024). Modern CNN architectures, such as the ResNet (Residual Networks), VGG, and DenseNet families, are designed with specific inductive biases, namely the principles of locality and translation invariance. Through mathematical convolution operations, CNNs process images by hierarchically scanning local features, starting from the detection of cell membrane edges and nuclear chromatin texture to glandular architecture patterns. This characteristic makes CNNs highly compatible with the nature of histopathological data, where the biological identity of a tissue is often determined by local texture and clear morphological boundaries, making it the de facto standard in the development of computer-aided diagnosis (CAD) systems (Balasubramanian et al., 2024; Dabass et al., 2022; De et al., 2024; Ghosh et al., 2021; Kumar et al., 2023; Martínez-Fernandez et al., 2023; C. Zhang et al., 2024).

Although CNN supremacy has been well established, the computer vision landscape has recently undergone a paradigmatic disruption with the introduction of the Vision Transformer (ViT-B/16) architecture. Adapting the Multi-Head Self-Attention (MSA) mechanism that underpinned the Transformer revolution in natural language processing (NLP), ViT-B/16 offers a radically different approach. Instead of using local convolution filters, ViT-B/16 breaks images down into a sequence of patches and calculates long-range dependencies between all of these patches. Theoretically, this capability allows ViT-B/16 to understand the global context of an image in its entirety from the early layers, overcoming the limitations of CNNs, which have a limited receptive field (T. Zhang et al., 2025). Various recent publications report that ViT-B/16 is capable of achieving, and even surpassing, the performance of CNNs on massive-scale general image datasets such as ImageNet-21k or JFT-300M, sparking a wave of academic enthusiasm for adopting this technology in medical diagnostics with the hope of capturing broader global network features.

However, the technological transition from CNN to ViT in the medical domain is not as smooth as it appears in the general image domain. There are fundamental problems related to the "data-hungry" characteristics of the Transformer architecture (Oh et al., 2024). Unlike CNNs, which have a built-in inductive bias for understanding 2D spatial structures, ViT-B/16 has no prior assumptions about image structures, requiring exposure to massive amounts of training data, often in the order of millions of images, to learn effective feature representations and build its own inductive bias (Zaman et al., 2025). This is where the main research gap lies, which is the focus of this study. In the medical ecosystem, acquiring high-quality annotated histopathology datasets is a very expensive and rare process. Available public datasets, such as NCT-CRC-HE-100K used in this study, are generally "medium" in scale (hundreds of thousands of images), not "massive" (Kather et al., 2021). The current literature still lacks critical evaluation of how ViT-B/16 behaves under these conditions of data scarcity. Most existing comparative studies only focus on surface metrics such as final accuracy, without conducting in-depth investigations into the quality of the feature space formed or specific model failure patterns (Le et al., 2025).

To address this research gap, this study was designed to present a comprehensive, critical, and in-depth comparative analysis between CNN (ResNet-50) and Vision Transformer (ViT-B/16) architectures in the task of multi-class colorectal cancer classification. This research aims to provide strong empirical evidence regarding the limitations and advantages of each architecture through a multidimensional evaluation approach. The problem-solving plan in this study is not limited to statistical performance evaluation (Accuracy, F1-Score), but also includes feature space topology analysis using t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the separability of inter-class clusters (Hossain et al., 2023). Additionally, interpretability validation is performed using the Gradient-weighted Class Activation Mapping (Grad-CAM) technique to ensure the medical relevance of the model's focus areas, complemented by computational efficiency analysis covering model size and inference latency (Musthafa et al., 2024). Through this systematic framework, the study is expected to provide a valid scientific basis for recommending the most robust, efficient, and rational Deep Learning architecture to be implemented in real clinical diagnosis systems, balancing technological innovation trends and the reality of medical data limitations.

Method

Research Design and Flow

This study adopts a quantitative comparative experimental design that aims to critically evaluate the effectiveness of two fundamentally different Deep Learning architecture paradigms: Convolutional Neural Networks (CNN) based on locality, and Vision Transformers (ViT) based on global attention mechanisms in the context of histopathological image classification. This approach is designed to test the hypothesis regarding the impact of architectural inductive bias on model generalization capabilities in medical datasets with limited data availability.

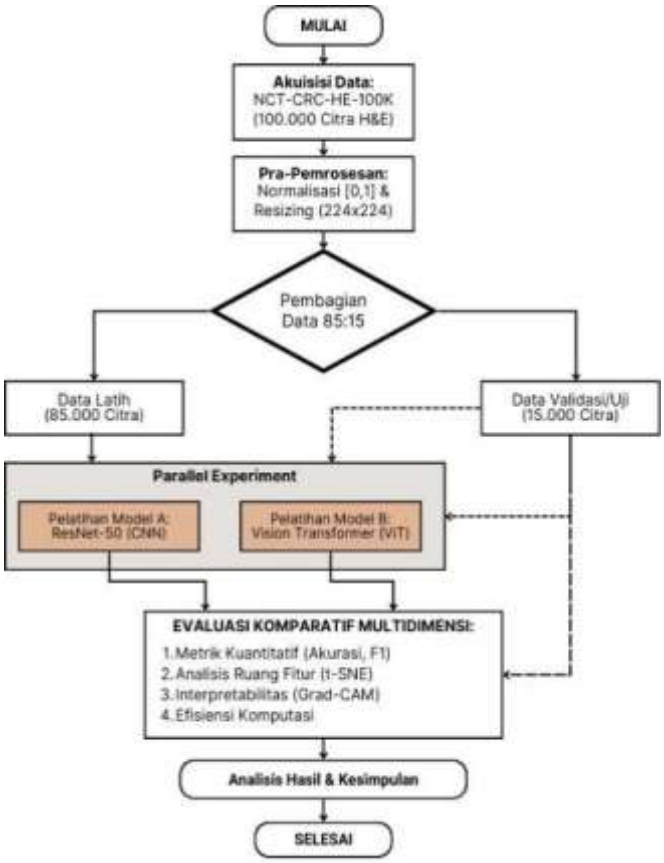


Figure 1. Research methodology flowchart

The research framework was implemented through a systematic end-to-end computational pipeline. The stages began with raw data acquisition, followed by a series of pre-processing techniques to standardize the input. Next, the data was strictly partitioned to prevent data leakage. The core of the experiment involves the parallel training of two comparative models, namely ResNet-50 and Vision Transformer, under controlled hyperparameter conditions to ensure fairness of comparison (ceteris paribus). The final stage involves multidimensional evaluation that goes beyond standard performance metrics, including feature space topology analysis, visual interpretability validation, and computational load measurement. The entire workflow of this methodology is visualized in Figure 1.

Dataset Acquisition and Characterization

The main material for this study is the public dataset NCT-CRC-HE-100K (Kather et al., 2021). This dataset is a large-scale collection of 100,000 non-overlapping histopathological images with dimensions of 224 x 224 pixels. These images were extracted from human colorectal cancer (CRC) tissue slides stained with Hematoxylin & Eosin (H&E), a standard staining technique that highlights cell nuclei (blue/purple) and cytoplasm and extracellular matrix (pink).

This dataset was selected because it represents the complex biological heterogeneity of the tumor microenvironment. The data was categorized into 9 morphologically distinct tissue classes, namely: (1) Adipose (ADI; fatty tissue), (2) Background (BACK; empty area without tissue), (3) Debris (DEB; necrosis or artifacts), (4) Lymphocytes (LYM; immune cells), (5) Mucus (MUC; mucus), (6) Smooth Muscle (MUS; smooth muscle), (7) Normal Colon Mucosa (NORM; healthy intestinal mucosa), (8) Cancer-Associated Stroma (STR; tumor-supporting connective tissue), and (9) Colorectal Adenocarcinoma Epithelium (TUM; epithelial cancer cells).

Pre-processing and Data Partitioning

Before being fed into the neural network model, raw images undergo crucial pre-processing steps to ensure numerical stability and training convergence:

a) Intensity Normalization

The original pixel intensity values, which are in the range [0, 255] for each RGB color channel, are normalized to the range [0, 1] through linear scaling. This step aims to reduce internal covariate *shift* and accelerate the gradient optimization process.

b) Resolution and Batch Management

Images are maintained at their original resolution of 224 x 224, which is compatible with the input requirements of both model architectures. The data is then organized into *mini-batches* of 32 samples to balance GPU memory usage efficiency and stable gradient estimation.

For unbiased performance evaluation, the dataset was partitioned using a strict hold-out validation scheme with a ratio of 85:15. A total of 85,000 images were allocated as the training set for model parameter learning, while 15,000 completely separate images were used as the validation/test set to measure the model's

generalization ability on previously unseen data (Ma et al., 2025).

Comparative Model Architecture

This study compares two state-of-the-art architectures that represent fundamental paradigm shifts in computer vision, as illustrated in Figure 2.

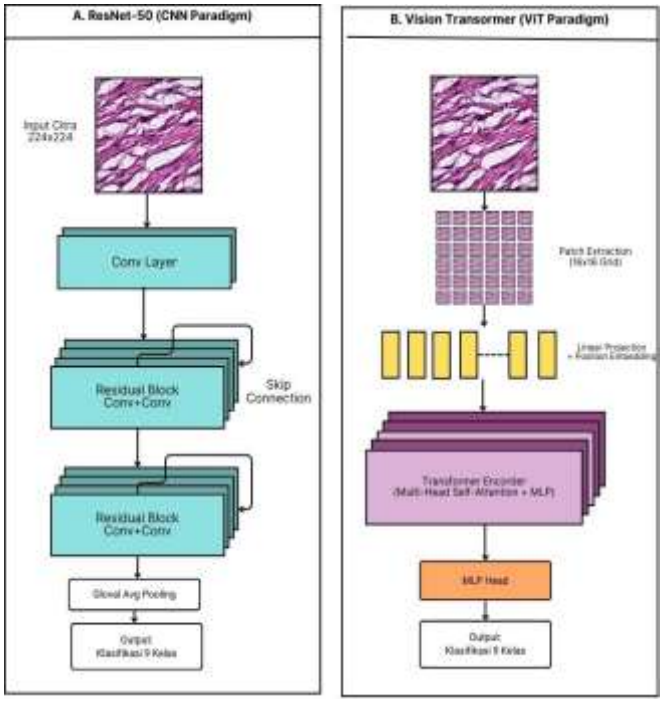


Figure 2. Architecture comparison scheme: (a) ResNet-50 (CNN); and (b) Vision Transformer (ViT-B/16)

ResNet-50 (Convolutional Neural Network Paradigm)

As a CNN representation, this study uses ResNet-50. This architecture is built on the foundation of convolution operations that have strong inductive biases toward spatial locality and translation invariance, making it highly effective in extracting hierarchical textural features from histopathological images. ResNet-50 overcomes the vanishing gradient problem in deep networks through the introduction of Residual Blocks that utilize skip connections, allowing direct gradient flow to the early layers and facilitating the training of very deep networks (Desai et al., 2025).

Vision Transformer (Self-Attention Paradigm)

As the main comparison, this study implements Vision Transformer (ViT). Unlike CNN, ViT-B/16 does not process images through local convolution filters. Instead, the input image is divided into a sequence of patches of fixed size (e.g., 16 x 16 pixels), which are then linearly projected into embedding vectors. ViT-B/16 utilizes the Multi-Head Self-Attention (MSA) mechanism in the Transformer Encoder block to capture long-range dependencies between all patches globally. The absence of built-in spatial inductive bias in ViT-B/16 makes it highly dependent on data scale to learn image structure (Y. Wang et al., 2025).

Configuration and Training Environment

To ensure a fair comparison (apple-to-apple comparison), both models were trained under a uniform protocol. The loss function used was Categorical Cross-entropy, which is suitable for multi-class classification. Parameter optimization was performed using the Adam (Adaptive Moment Estimation) algorithm with an initial learning rate set at 1 x 10-4 (Jeddah et al., 2025).

The training strategy is complemented by callback mechanisms to prevent overfitting and ensure optimal convergence:

a) Early Stopping

Training is automatically stopped if the accuracy metric on the validation data does not show improvement for 5 consecutive epochs.

b) ReduceLROnPlateau

The learning rate is reduced by a factor of 0.1 if the validation loss stagnates for 3 epochs, allowing the model to explore local minima more smoothly.

All computational experiments were performed in a cloud environment powered by NVIDIA Tesla T4 GPU hardware accelerators with 16 GB of video memory, using the TensorFlow and Keras deep learning frameworks.

Multidimensional Evaluation Framework

Model performance is not only assessed based on single accuracy, but through a comprehensive evaluation framework that covers four dimensions:

a) Quantitative Performance Evaluation:

Standard metrics are calculated based on the confusion matrix in the test data, including Accuracy (global accuracy), Precision (positive predictive value), Recall (sensitivity), and F1-Score (harmonic mean of precision and recall). The Receiver Operating Characteristic (ROC) curve is also analyzed to assess the trade-off between true positive and false positive rates (Araújo et al., 2025).

b) Feature Space Topology Analysis (Qualitative)

To understand the quality of the internal representation learned by the model, a manifold learning technique was applied using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm. This technique projects high-dimensional features from the last layer before classification into a 2D space, enabling visualization of cluster separability between network classes (Shi et al., 2023).

c) Medical Interpretability Validation

Gradient-weighted Class Activation Mapping (Grad-CAM) technique is used to generate visual attention heatmaps. This analysis aims to verify whether the areas considered important by the model in decision-making correlate with medically relevant pathological morphological features (e.g., focusing on tumor cell nuclei rather than background artifacts) (Gamage et al., 2024).

d) Computational Efficiency and Implementation Feasibility

The model's physical parameters were measured to assess its clinical feasibility, including model *storage size* (in MB), average inference latency per image (in milliseconds), and processing *throughput* (frames per second/FPS) (Howell et al., 2024).

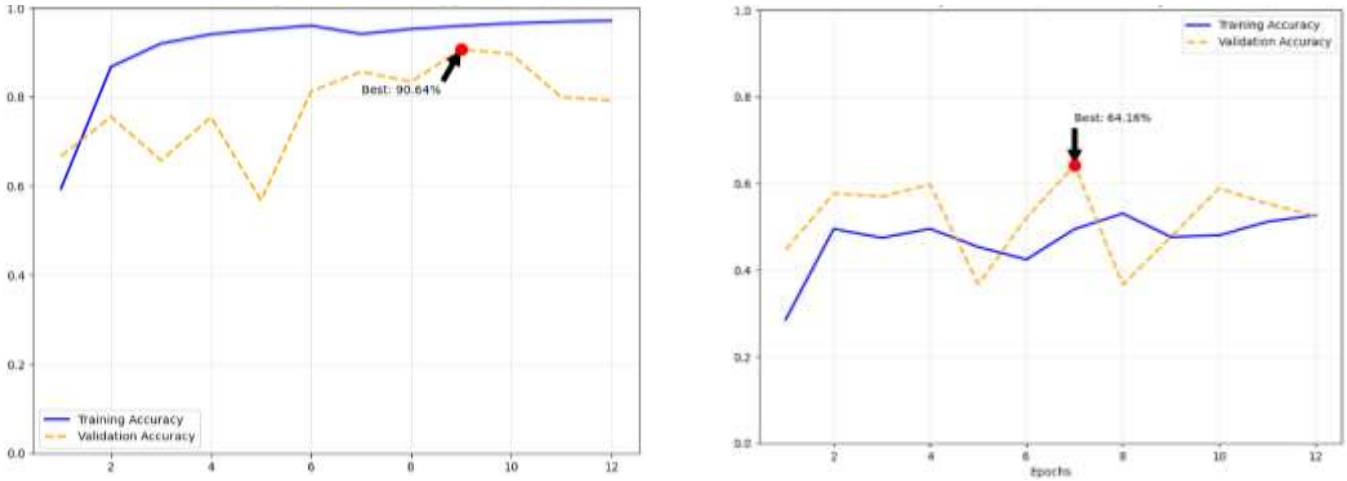
Result and Discussion

This chapter presents a comprehensive analysis of experimental findings obtained from a head-to-head comparison between Convolutional Neural Network (ResNet-50) and Vision Transformer (ViT-B/16) architectures in colorectal cancer histopathology classification. The discussion is organized systematically, starting from the dynamics of the training process, quantitative performance evaluation, qualitative analysis of feature space and interpretability, specific failure case studies, to the evaluation of technical implementation feasibility.

Training Dynamics and Training Stability

The evaluation began by analyzing the behavior of both models during the learning phase to understand convergence stability and initial generalization capabilities. The learning curve, which shows the movement of training and validation accuracy throughout the epoch, is presented in Figure 3.

The observation results show a sharp contrast in training stability. ResNet-50 shows a consistent and smooth increase in validation accuracy, peaking at around 90% at epoch 9 before the early stopping mechanism is activated. This stable convergence pattern indicates that the spatial inductive bias of CNNs is highly effective in guiding the model to quickly learn relevant histopathological features, even with a moderate amount of data.



(a) (b)

Figure 3. Training and validation accuracy dynamics: (a) Convergence stability (ResNet-50) and (b) Vision transformer performance fluctuations (ViT-B/16)

In contrast, Vision Transformer exhibits significant instability, characterized by a "jagged" or fluctuating validation graph. The validation accuracy of ViT-B/16 often experiences a drastic decline between epochs (for example, falling from 64% to 36%). This phenomenon confirms the hypothesis that without inherent inductive bias, ViT-B/16 struggles to find stable global minima in the loss landscape when trained on limited-scale datasets. ViT-B/16 shows signs of early overfitting, where the model struggles to generalize patterns from training data to previously unseen validation data.

Quantitative Performance Evaluation

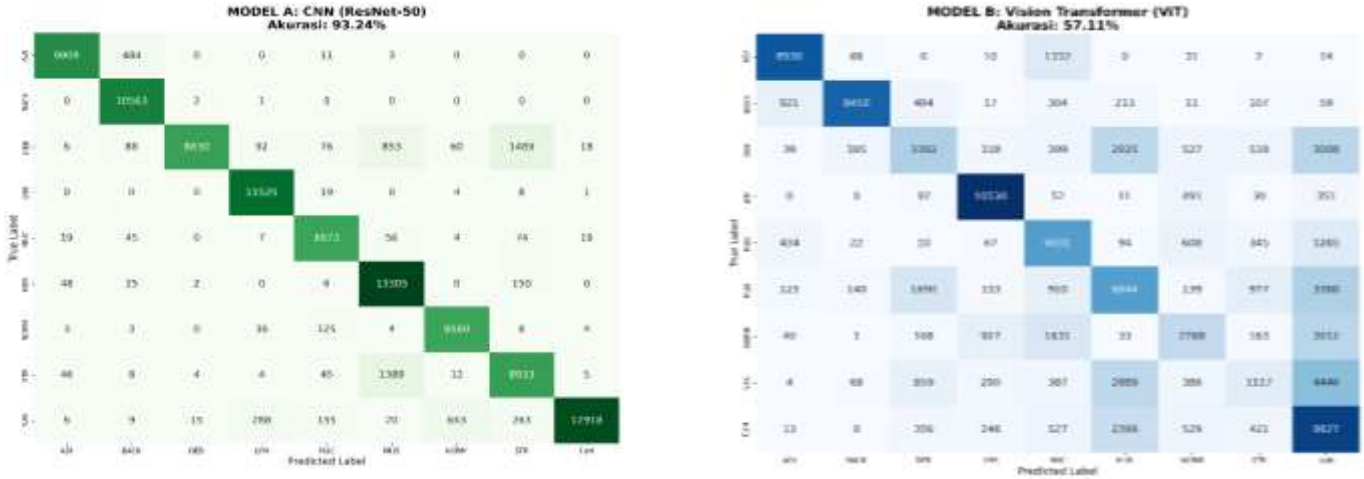
To measure the final performance of the model on independent test data (15,000 images), an evaluation was conducted using standard classification metrics. A comprehensive summary of the performance is presented in Table 1, while details of the inter-class error distribution are visualized through the confusion matrix in Figure 4 and the ROC curve in Figure 5.

Table 1. ResNet-50

Class	Precision	Recall	F1-Score
ADI	0.98	0.99	0.99
BACK	1	1	1
DEB	0.95	0.85	0.9
LYM	0.93	0.96	0.94
MUC	0.94	0.95	0.94
MUS	0.88	0.95	0.91
NORM	0.96	0.98	0.97
STR	0.91	0.95	0.93
TUM	0.98	0.95	0.96
accuracy			0.93
macro average	0.95	0.95	0.95
weighted average	0.93	0.93	0.93

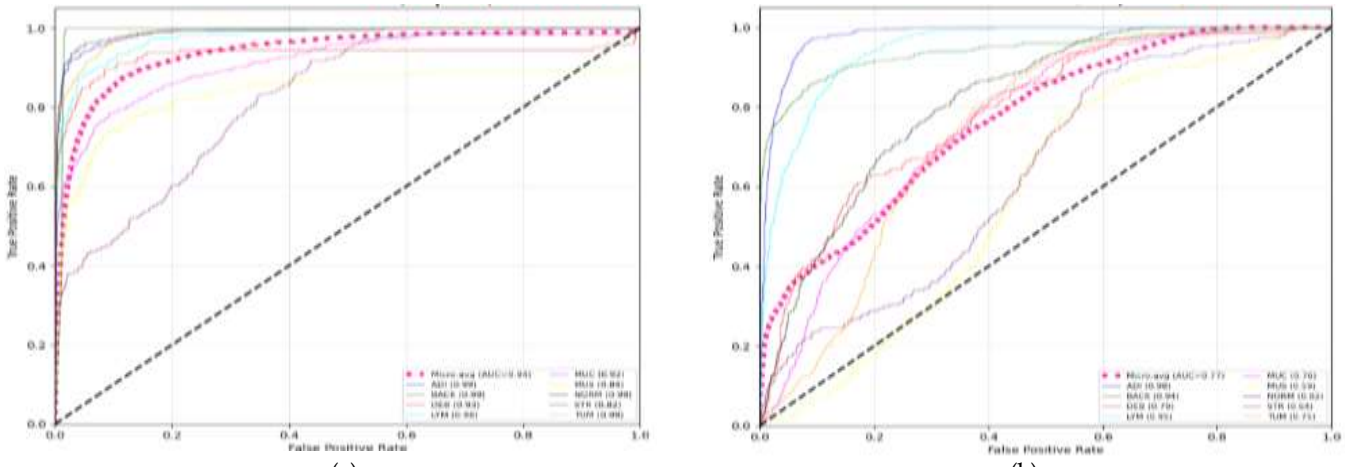
Table 2. ViT-B/16

Class	Precision	Recall	F1-Score
ADI	0.67	0.65	0.66
BACK	0.85	0.94	0.89
DEB	0.57	0.43	0.49
LYM	0.58	0.69	0.63
MUC	0.53	0.54	0.53
MUS	0.36	0.24	0.29
NORM	0.57	0.61	0.59
STR	0.18	0.14	0.16
TUM	0.52	0.72	0.61
accuracy			0.57
macro average	0.54	0.55	0.54
weighted average	0.55	0.57	0.55



(a) (b)

Figure 4. Confusion matrix (a) ResNet-50 shows a dominant diagonal; (b) Vision transformer shows a high misclassification rate



(a) (b)

Figure 5. Multiclass ROC Curve: (a) ResNet-50 (AUC Approaches 1.0); and (b) ViT-B/16 (Low AUC)

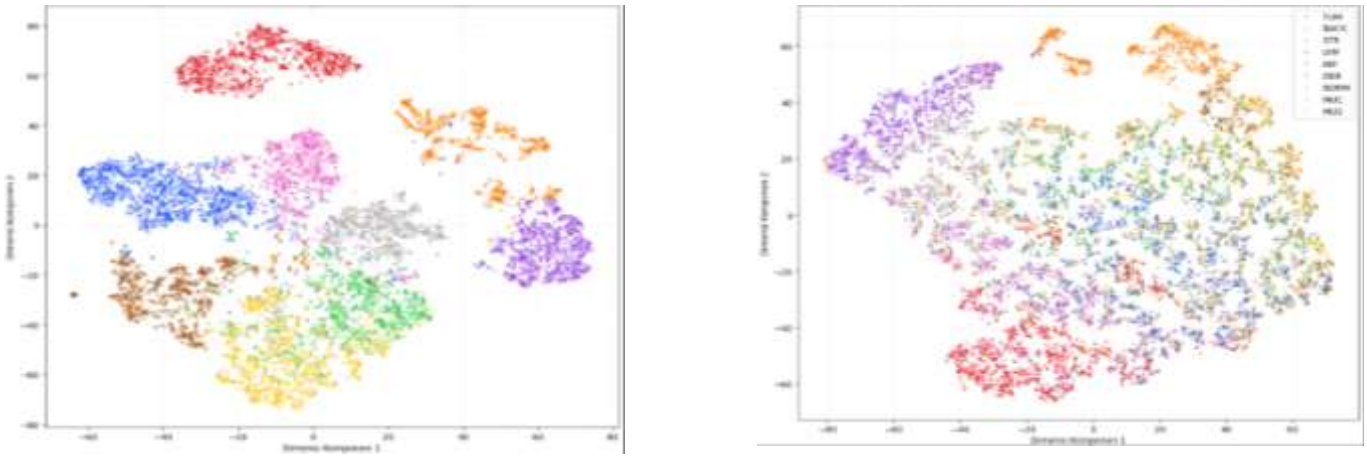
Statistically, ResNet-50 shows absolute superiority with an overall testing accuracy of 93.24%, far surpassing ViT-B/16, which only reached 57.11%. Analysis of the Classification Report (1) reveals that ResNet-50 maintains an F1-Score above 0.90 for almost all majority classes such as ADI, MUC, and TUM. In contrast, ViT-B/16 experienced catastrophic failure in certain classes, particularly in the Cancer-Associated Stroma (STR) class, which only achieved an F1-Score of 0.16.

The confusion matrix (Figure 4) clarifies the source of ViT-B/16's failure. This model shows severe confusion in distinguishing classes with visually similar but biologically different textures. As a significant

example, a large number of stromal tissue (STR) and smooth muscle (MUS) samples were misclassified as tumors (TUM) by ViT-B/16, which in a clinical context is a dangerous false positive error. The ROC curve (Figure 5) further validates these findings, where the ResNet-50 curve approaches the upper left corner (ideal performance) for all classes, while the ViT-B/16 curve tends to slope toward the random line, indicating weak discriminatory ability.

Qualitative Feature Space Analysis (t-SNE)

To understand why ViT-B/16 failed to achieve performance equivalent to CNN, an investigation was conducted into the internal feature space structure learned by both models using t-SNE visualization. The 2D projection of high-dimensional features in the last layer before classification is shown in Figure 6.



(a) (b)

Figure 6. t-SNE feature space visualization: (a) Optimal cluster separability in ResNet-50; and (b) Significant overlapping in vision transformer

This visualization provides strong empirical evidence regarding the quality of feature representation. In ResNet-50 (Figure 6a), nine well-separated clusters are clearly visible, where images from the same class are grouped closely together and have significant inter-class distances. This shows that ResNet-50 successfully learned unique discriminative features for each type of network.

In contrast, the ViT-B/16 feature space (Figure 6b) shows significant disorder. Clusters from different classes (such as MUS, STR, and TUM) overlap each other, forming a large "cloud" with no clear boundaries. The inability of ViT-B/16 to separate data in this latent feature space directly explains the low classification accuracy. This confirms that the global attention mechanism of ViT-B/16 fails to capture the local textural details that are crucial for distinguishing histopathological tissue types in datasets with a limited number of samples.

Visual Interpretability Validation (Grad-CAM)

To ensure that the high performance of ResNet-50 is based on medically relevant pathological features, rather than background artifacts, an interpretability analysis was performed using Grad-CAM. Examples of the model's attention heatmaps for several key classes are presented in Figure 7.

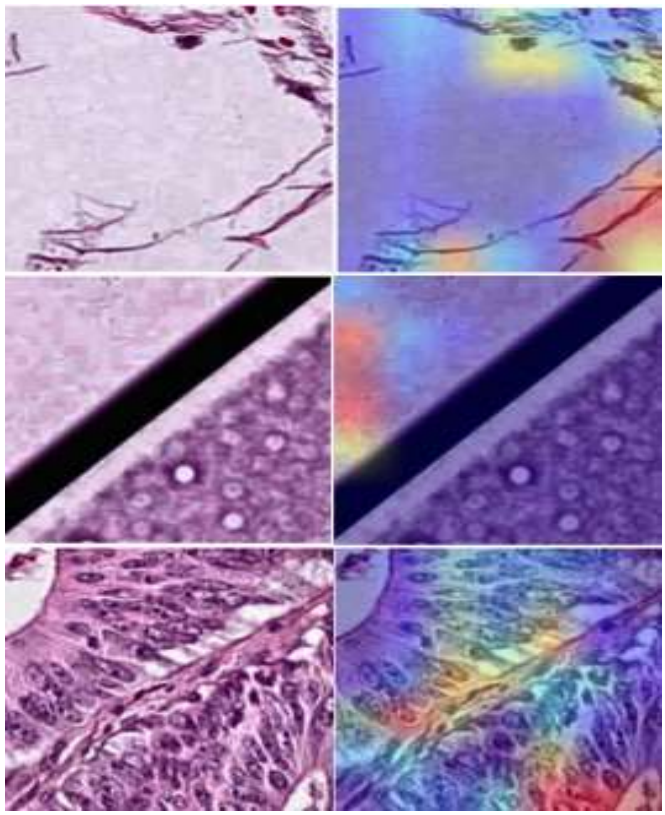


Figure 7. Grad-CAM interpretability analysis of ResNet-50 on the Adipose (ADI), Background (BACK), and Tumor (TUM) classes

The Grad-CAM results confirm the medical validity of the ResNet-50 predictions. In adipose tissue images (Adipose/ADI), the heat map accurately highlights the edges of adipocyte cell membranes, which are key diagnostic features. In background images (Background/BACK), the model does not show significant focal activation, indicating that the model does not "hallucinate" in empty areas. This validation is crucial for building clinical confidence in the decisions made by AI models.

Specific Failure Analysis

This subsection qualitatively examines specific examples where the ViT-B/16 model fails, providing deep insights into the limitations of its architecture. Figure 8 shows a comparison of predictions on challenging test samples.

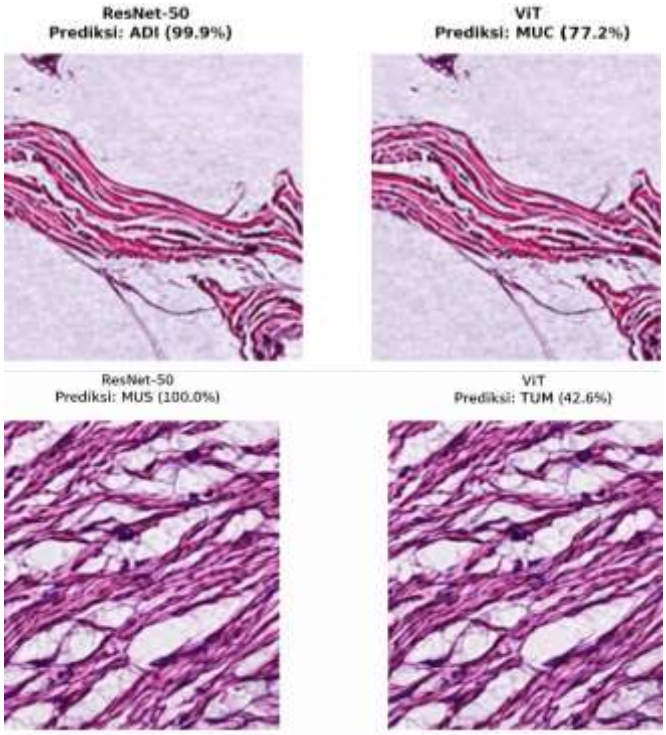


Figure 8. Vision transformer prediction failure analysis: classification errors on empty textures and fibrous networks

Two main failure patterns were identified:

Empty Texture-Based Failures (ADI vs. MUC)

ViT often misclassifies fat cells (ADI) as mucus (MUC). Visually, both tissues are dominated by empty white areas. ResNet-50 successfully distinguishes between them because it can detect local features in the form of thin cell membrane lines in ADI. ViT-B/16, which processes images as global *patches*, tends to only see the dominance of white areas without paying attention to these fine edge details, causing classification errors.

Fiber-Based Failure (MUS vs. TUM)

ViT-B/16 often misinterprets dense fibrous tissue such as smooth muscle (MUS) as tumor tissue (TUM). This indicates the inability of the ViT-B/16 attention mechanism to capture the subtle tissue architecture patterns that distinguish benign supporting tissue from epithelial malignancy.

Evaluation of Computational Efficiency and Implementation Feasibility

In addition to diagnostic performance, practical operational aspects are crucial for the feasibility of implementing the model in clinical environments, which often have limited hardware resources (**).** Table 3.2 presents a comparison of computational efficiency between the two models.

Table 3. Computational Efficiency Evaluation: Model Size, Latency, and Throughput

Evaluation Parameters	ResNet-50 (Proposed)	ViT-B/16 (Comparator)
Model Size (Storage Size)	282.36 MB	1635.21 MB
Average Latency (Inference Time)	92.81 ms/image	99.23 ms/image
Throughput (Processing Speed)	10.8 FPS	10.1 FPS

In addition to diagnostic performance, evaluation of computational load is a crucial parameter for assessing the feasibility of model implementation in real clinical ecosystems. As summarized in Table 3.2, the measurement results show a significant disparity in efficiency between the two architectures. The proposed ResNet-50-based model proved to be much more efficient in terms of storage space usage, with a model size of only 282.36 MB. This figure is in stark contrast to the Vision Transformer (ViT-B/16) comparison model, which takes up 1635.21 MB (approximately 1.6 GB) of storage, making ResNet-50 5.8 times more compact and easier to distribute on medical devices with limited memory capacity (resource-constrained edge devices).

In terms of processing speed, ResNet-50 also demonstrates superiority with a lower average inference latency of 92.81 milliseconds per image, compared to ViT-B/16 which requires 99.23 milliseconds. This correlates directly with processing throughput, where ResNet-50 is capable of handling 10.8 frames per second (FPS), slightly outperforming ViT-B/16 at 10.1 FPS. Although the difference in inference speed appears slight, the combination of a much smaller model size and lower latency confirms that the ResNet-50 architecture offers the most optimal trade-off between accuracy and efficiency, making it a much more feasible candidate for integration into real-time digital pathology systems than

the computationally resource-intensive Transformer architecture.

Conclusion

Based on a comprehensive experimental investigation comparing the Convolutional Neural Network (CNN) and Vision Transformer (ViT) paradigms, this study empirically confirms that the ResNet-50 architecture demonstrates absolute superiority and significantly higher implementation feasibility compared to ViT-B/16 for the classification of colorectal cancer histopathology images on a medium-scale dataset. The significant performance gap—marked by ResNet-50's testing accuracy reaching 93.24% compared to 57.11% for ViT-B/16, as well as the highly fluctuating training dynamics of ViT-B/16—validates the fundamental hypothesis that the absence of spatial inductive bias in the Transformer architecture is a critical obstacle in learning effective feature representations without the availability of massive-scale data. This quantitative finding is reinforced by qualitative evidence through t-SNE visualization, which reveals ViT-B/16's failure to form separable inter-class clusters in the network, as well as failure analysis highlighting the inability of global attention mechanisms to capture crucial local textural details, such as distinguishing the morphology of fat cells from mucus artifacts. Considering the additional fact that ResNet-50 proved to be 5.8 times more efficient in storage size and offered lower inference latency, this study concludes that despite the surge in popularity of attention mechanisms in computer vision, CNN-based architectures with their inductive bias of locality remain the most robust, accurate, and rational gold standard for the development of computer-aided diagnosis systems in the current digital pathology ecosystem. As a future development direction, further research is recommended to explore domain-adaptive pre-training strategies on massive medical datasets or the application of hybrid architectures (CNN-ViT) to overcome data limitations in Transformer-based models.

Acknowledgments
The authors would like to thank Universitas Esa Unggul for its support in conducting this research.

Author Contributions
Conceptualization, H.A.; methodology, M.F.Q.; software, M.F.Q.; validation, M.F.Q., W.P.M. and K.F.R.; formal analysis, M.F.Q.; investigation, M.F.Q.; resources, H.A.; data curation, W.P.M. and K.F.R.; writing—original draft preparation, M.F.Q.; writing—review and editing, H.A. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Conflicts of Interest
The authors declare no conflict of interest.

References

Araújo, A. L. D., Sperandio, M., Calabrese, G., Faria, S. S., Cardenas, D. A. C., Martins, M. D., Vargas, P. A., Lopes, M. A., Santos-Silva, A. R., Kowalski, L. P., & Moraes, M. C. (2025). Artificial intelligence in healthcare applications targeting cancer diagnosis—part II: interpreting the model outputs and spotlighting the performance metrics. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *140*(1), 89–99. https://doi.org/10.1016/j.oooo.2025.01.002

Balasubramanian, A. A., Al-Heejawi, S. M. A., Singh, A., Breggia, A., Ahmad, B., Christman, R., Ryan, S. T., & Amal, S. (2024). Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology. *Cancers*, *16*(12), 2222. https://doi.org/10.3390/cancers16122222

Dabass, M., Vashisth, S., & Vig, R. (2022). A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images. *Computers in Biology and Medicine*, *147*, 105680. https://doi.org/10.1016/j.compbiomed.2022.105680

De, A., Mishra, N., & Chang, H.-T. (2024). An approach to the dermatological classification of histopathological skin images using a hybridized CNN-DenseNet model. *PeerJ Computer Science*, *10*, e1884. https://doi.org/10.7717/peerj-cs.1884

Desai, A., & Mahto, R. (2025). Multi-Class Classification of Breast Cancer Subtypes Using ResNet Architectures on Histopathological Images. *Journal of Imaging*, *11*(8), 284. https://doi.org/10.3390/jimaging11080284

Dunn, C., Brettle, D., Cockroft, M., Keating, E., Revie, C., & Treanor, D. (2024). Quantitative assessment of H&E staining for pathology: development and clinical evaluation of a novel system. *Diagnostic Pathology*, *19*(1), 42. https://doi.org/10.1186/s13000-024-01461-w

Gamage, L., Isuranga, U., Meedeniya, D., De Silva, S., & Yogarajah, P. (2024). Melanoma Skin Cancer Identification with Explainability Utilizing Mask Guided Technique. *Electronics*, *13*(4), 680. https://doi.org/10.3390/electronics13040680

Ghosh, S., Bandyopadhyay, A., Sahay, S., Ghosh, R.,

Kundu, I., & Santosh, K. C. (2021). Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network. *Engineering Applications of Artificial Intelligence*, *100*, 104202. https://doi.org/10.1016/j.engappai.2021.104202

Hossain, M. M., Hossain, M. A., Musa Miah, A. S., Okuyama, Y., Tomioka, Y., & Shin, J. (2023). Stochastic Neighbor Embedding Feature-Based Hyperspectral Image Classification Using 3D Convolutional Neural Network. *Electronics*, *12*(9), 2082. https://doi.org/10.3390/electronics12092082

Howell, L., Ingram, N., Lapham, R., Morrell, A., & McLaughlan, J. R. (2024). Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound. *Ultrasonics*, *140*, 107251. https://doi.org/10.1016/j.ultras.2024.107251

Jeddah, Y. M., Hassan Abdalla Hashim, A., Omran Khalifa, O., & Ouhada, K. (2025). Enhancing Anomaly Detection Performance: Deep Learning Models Evaluation. *IIUM Engineering Journal*, *26*(2), 96–108. https://doi.org/10.31436/iiumej.v26i2.3287

Kather, J. N., Halama, N., & Marx, A. (2021). *100,000 Histological Images of Human Colorectal Cancer and Healthy Tissue (V0.1)*. Zenodo. Retrieved from https://zenodo.org/records/1214456

Kumar, A., Vishwakarma, A., & Bajaj, V. (2023). CRCCN-Net: Automated framework for classification of colorectal tissue using histopathological images. *Biomedical Signal Processing and Control*, *79*, 104172. https://doi.org/10.1016/j.bspc.2022.104172

Lan, X., Guo, G., Wang, X., Yan, Q., Xue, R., Li, Y., Zhu, J., Dong, Z., Wang, F., Li, G., Wang, X., Xu, J., & Jiang, Y. (2024). Differentiation and risk stratification of basal cell carcinoma with deep learning on histopathologic images and measuring nuclei and tumor microenvironment features. *Skin Research and Technology*, *30*(1). https://doi.org/10.1111/srt.13571

Le, T. T., Nguyen-Truong, V.-T., Van Nhat Duong, Q., Le Phan, N. T., Thien Dao, P. N., Mavuso, M. F., Nguyen, H. N. A., Mai, T. T., & Quang, K. T. (2025). Deep learning-based classification of colorectal cancer in histopathology images for category detection. *Biology Methods and Protocols*, *10*(1). https://doi.org/10.1093/biomethods/bpaf077

Ma, J., Yang, H., Chou, Y., Yoon, J., Allison, T., Komandur, R., McDunn, J., Tasneem, A., Do, R. K., Schwartz, L. H., & Zhao, B. (2025). Generalizability of lesion detection and segmentation when ScaleNAS is trained on a large multi-organ dataset and validated in the liver. *Medical Physics*, *52*(2), 1005–1018. https://doi.org/10.1002/mp.17504

Martínez-Fernandez, E., Rojas-Valenzuela, I., Valenzuela, O., & Rojas, I. (2023). Computer Aided Classifier of Colorectal Cancer on Histopatological Whole Slide Images Analyzing Deep Learning Architecture Parameters. *Applied Sciences*, *13*(7), 4594. https://doi.org/10.3390/app13074594

Musthafa, M. M., Mahesh, T. R., Kumar, V., & Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. *BMC Medical Imaging*, *24*(1), 107. https://doi.org/10.1186/s12880-024-01292-7

Oh, S., Kim, N., & Ryu, J. (2024). Analyzing to discover origins of CNNs and ViT architectures in medical images. *Scientific Reports*, *14*(1), 8755. https://doi.org/10.1038/s41598-024-58382-3

Shi, S., Xu, Y., Xu, X., Mo, X., & Ding, J. (2023). A Preprocessing Manifold Learning Strategy Based on t-Distributed Stochastic Neighbor Embedding. *Entropy*, *25*(7), 1065. https://doi.org/10.3390/e25071065

Snead, D. R., Azam, A. S., Thirlwall, J., Kimani, P., Hiller, L., Bickers, A., Boyd, C., Boyle, D., Clark, D., Ellis, I., Gopalakrishnan, K., Ilyas, M., Kelly, P., Loughrey, M., Neil, D., Rakha, E., Roberts, I. S., Sah, S., Soares, M., … Dunn, J. (2025). Variation within and between digital pathology and light microscopy for the diagnosis of histopathology slides: blinded crossover comparison study. *Health Technology Assessment*, 1–75. https://doi.org/10.3310/SPLK4325

Wang, Y.-L., Gao, S., Xiao, Q., Li, C., Grzegorzek, M., Zhang, Y.-Y., Li, X.-H., Kang, Y., Liu, F.-H., Huang, D.-H., Gong, T.-T., & Wu, Q.-J. (2024). Role of artificial intelligence in digital pathology for gynecological cancers. *Computational and Structural Biotechnology Journal*, *24*, 205–212. https://doi.org/10.1016/j.csbj.2024.03.007

Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2025). Vision Transformers for Image Classification: A Comparative Survey. *Technologies*, *13*(1), 32. https://doi.org/10.3390/technologies13010032

Wen, Z., Wang, S., Yang, D. M., Xie, Y., Chen, M., Bishop, J., & Xiao, G. (2023). Deep learning in digital pathology for personalized treatment plans of cancer patients. *Seminars in Diagnostic Pathology*, *40*(2), 109–119. https://doi.org/10.1053/j.semdp.2023.02.003

Zaman, F. H., Ng, K. M., & Abdullah, S. A. C. (2025). Comparative Analysis of Vision Transformers and CNN Models for Driver Fatigue Classification. *IIUM Engineering Journal*, *26*(2), 169–186. https://doi.org/10.31436/iiumej.v26i2.3488

Zhang, C., Aamir, M., Guan, Y., Al-Razgan, M., Awwad, E. M., Ullah, R., Bhatti, U. A., & Ghadi, Y. Y. (2024).

Enhancing lung cancer diagnosis with data fusion and mobile edge computing using DenseNet and CNN. *Journal of Cloud Computing*, *13*(1), 91. https://doi.org/10.1186/s13677-024-00597-w

Zhang, T., Xu, W., Luo, B., & Wang, G. (2025). Depth-Wise Convolutions in Vision Transformers for efficient training on small datasets. *Neurocomputing*, *617*, 128998. https://doi.org/10.1016/j.neucom.2024.128998

Zhu, M., Zhai, Z., Wang, Y., Chen, F., Liu, R., Yang, X., & Zhao, G. (2025). Advancements in the application of artificial intelligence in the field of colorectal cancer. *Frontiers in Oncology*, *15*. https://doi.org/10.3389/fonc.2025.1499223