Efficient Fine-Tuning of Vision Transformers for Histopathological Image Classification via Low-Rank Adaptation

Keshab Bashyal[1]

[1]Faculty Member, Divya Gyan College, Tribhuvan University

Article History

Received: 21/12/2025

Revised: 29/12/2025

Accepted: 07/01/2026

Published: 15/01/2026

Corresponding Author

Name: Keshab Bashyal

Email: keshabbashyal@divyagyan.edu.np

ORCiD: https://orcid.org/0009-0000-8461-4824

Abstract

Excessive computational and memory requirements associated with traditional full-fine tuning, despite their remarkable performance, significantly hinder the pragmatic application of modern vision transformers especially for histopathological image analysis. To alleviate this problem, modern transformers like Swin and DeiT are systematically evaluated using Low Rank Adaption (LoRA) technique, which is a parameter efficient fine-tuning technique, especially designed to shorten training time in natural language processing. When LoRA is applied to histopathological image classification, surprisingly, LoRA adapted Swin and DeiT models performs comparable performance across all evaluation metrics: accuracy, precision, specificity and F1 score, compared to their full- fine-tuned counterparts by updating less than 2% of the model's parameters. The results show that LoRA not only accelerate training speed by updating fewer than 2% of the model's parameters but also achieves superior accuracy for both Swin (99.42% vs. 99.21%) and DeiT (99.27% vs. 98.91%) compared to their fully fine-tuned counterparts on NCT-CRC-HE dataset. Consequently, efficient fine-tuning using LoRA can provide an alternative way to traditional full fine-tuning without scarifying performance while boosting training speed, opening new avenues for various medical image classification problems.

Keywords: ***Vision Transformers, Parameter-efficient, Histopathological image, Fine-tuning, Superior accuracy, Performance, Classification.***

Introduction

A fundamental component of contemporary medicine is histopathology, the microscopic analysis of tissue to identify illness. This field is especially important for the diagnosis and treatment of cancers, especially colorectal cancer (CRC). Being the third most common cancer to be diagnosed and a major cause of cancer-related death, colorectal cancer (CRC) is one of the most common and deadly cancers in the world and a major public health concern.(Mármol et al., 2017). The histopathological examination of biopsied or surgically removed tissue, in which pathologists examine cellular morphology and tissue architecture to ascertain the presence and severity of the disease, is the gold standard for diagnosing colorectal cancer. Nevertheless, this manual process has many challenges despite its fundamental role. It is a labor-intensive and time-consuming task with significantly inter-observer variability that may affect the performance of the diagnosis.(Demir & Yener., 2005).

Current developments in artificial intelligence, especially in the area of deep learning, have provided computational tools that can automate and improve digital pathology diagnostic accuracy. Despite being fundamental, Convolutional Neural Networks (CNNs) are not able to capture the long-range dependencies within the image which are essential for understanding complex tissue structures due to their high inductive biases (Romero et al., 2022). Vision Transformers (ViTs), on the other hand, can capture long-term dependencies in the images using self-attention mechanisms(Dosovitskiy et al., 2020). Since then, this concept has developed into a new class of advanced transformers, such as the Swin Transformer and DeiT. The requirements of medical imaging are perfectly met by their multi-scale designs. Nevertheless, the colossal computational cost of these large-scale models severely limits their practical application, thereby, making conventional fine-tuning a major bottleneck for many research and clinical settings.

In this paper, the multiclass classification of colorectal histopathological tissue across nine classes has been implemented and evaluated using the Parameter-Efficient Fine-Tuning of two top vision transformers, Swin and DeiT. Each of these models has been modified using the Low-Rank Adaptation (LoRA) technique using the NCT-CRC-HE dataset. The LoRA-adapted models' performance has been compared to that of their counterparts that have undergone standard fine-tuning. The comparison is thorough and includes important indicators of

computational efficiency like training and inference times in addition to common classification metrics like accuracy, recall, precision, F1-score, and specificity.

Literature Review

The usage of neural networks and deep learning in colorectal cancer detection and classification task has currently been leveraged by the integration of recent transformer architectures to mitigate the limitations of conventional convolution neural networks. Current studies have demonstrated powerful methods that facilitates these implementations to enhance medical diagnostic accuracies and precision. For instance, the standard U-Net architecture is enhanced by using skip connection. Which has used a Swin transformer for feature extraction and achieved a 95.8% accuracy on the NCT_CRC_HE_100K dataset (Qin et al., 2024). Using this, the Colorectal cancer detection network was introduced to integrate dilated convolutions with coordinate attention model with a cross-shaped window transformer in order to capture local, global and subtle tissue changes, thereby, reaching an accuracy of 98.96% on the same dataset.

The excessive computational and memory requirements of these large-scale ViTs models have demanded the development of more efficient fine-tuning strategies for fast training. Scaling and Shifting Features (SSF) has been developed as a highly efficient method for fine-tuning pretrained models by learning only to scale and shift features within frozen network blocks (Tay et al., 2023). This approach frequently outperforms full fine-tuning in various domain and problems while training significantly fewer parameters and providing the benefits of being mathematically merged back into the model to ensure no inference latency. Likewise, the application of Adapters and Low-Rank Adaptation (LoRA) to large models like SEEM and Mask DINO has a competitive performance is achievable by updating around 1–6% of model's total parameters. That can significantly reduce the costs to train the model compared to full-fine tuning (Abou Baker et al., 2024). Addressing the heuristic nature of many Parameter-Efficient Fine-Tuning (PEFT) methods. Sensitivity-aware PEFT (SPT) was designed to identify and prioritize most essential parameters for specific problems (Yin et al., 2023). This technique effectively allocates tuning budget to the most critical and sensitive weight matrices, which in turn, boost performance and achieving the state-of-the-art performance across major benchmarks (Xin et al., 2024).

The practical usage of these efficient fine-tuning methods are particularly vital in specialized medical imaging tasks like classification and segmentation. In cervical cancer detection task, LoRA-based models have solved data scarcity challenges and also outperformed standard CNNs by reducing trainable parameters to less than 1% of the base model (Hong et al., 2024). LoRA has been successfully utilized to adapt large vision models for lung nodule malignancy classification tasks which has achieved around 3% higher ROC AUC than previous state-of-the-art methods while using 89.9% fewer parameters and reducing training time by 36.5% (Veasey & Amini, 2025). The scalability of those models is supported by other techniques such as federated learning via Kubernetes for privacy-preserving image synthesis (Preda et al., 2025) and introduction of hierarchical cell transformers that model spatial interdependencies within whole slide images to give superior performance for survival prediction and cancer classification problems (Yang et al., 2024).

Vision Transformers (ViTs) are known for being their fundamental processing characteristics. ViTs shows a greater shape bias and closely resembles with the nature of human error patterns compared to ResNets based on evaluations on various datasets. That's why they can capture long-term dependencies within the images (Tuli et al., 2021). Research also indicates that general-purpose models like Swin Transformer V2 can outperform specialized medical encoders in various tasks like cell segmentation (Vadori et al., 2025). Those architectural advantages have been also evaluated in novel domains such as deep ultraviolet fluorescence breast cancer imaging in which patch-level transformers achieved around 98.3% accuracy and surpassed other comparable models up to nearly 13% (Afshin et al., 2025).

Despite these successes, this field continues to integrate the benefits of both transformers and specialized convolutional frameworks with hybrid designs. While transformers excel in capturing long-range dependencies, comparative analyses on datasets like BraTS 2020 show that specialized models like nnU-Net can still outperform them in certain segmentation tasks (Träff, 2023). Some research has implemented dataset shortcuts that demonstrates simpler models like EfficientNet-B0 can occasionally outperform complex transformers on clean data, which underscores the importance of combining high-quality data with innovative architectures like the CViTS-Net hybrid (Ignatov & Malivenko, 2025; Kanadath et al., 2024). Multi-modal integration has also seen progress through models like TransMed, that uses a cross-modal

transformer encoder to fuse PET and CT data for nasopharyngeal carcinoma prognosis. Those model's performance has surpassed traditional CNN-based fusion techniques (Dai et al., 2021).

The research has expanded in clinical interpretability and practical implication as those models perform exceptionally well. Hierarchical Vision Transformers have used prototype-based learning in order to provide clear explanations for their high-performance decisions, which helps to build clinical usability (Gallée et al., 2025). Hybrid models like EffNetV2_ViT and the BreaST_Net ensemble have achieved accuracies as high as 99.8% by processing images in multiple magnifications (Hayat et al., 2024; Tummala et al., 2022). To address the constraints of clinical environments, RMT-Net integrates ResNet-50 with transformers to maintain a condensed model size suitable for quick diagnosis (Ren et al., 2023). However, as noted in studies of lung disease detection and skin cancer classification, the high computational demands of the most powerful transformers still pose several challenges for resource-constrained edge devices (Aladhadh et al., 2022; Uparkar et al., 2022). This highlights the ongoing need for optimization techniques like pruning, quantization, and parameter-efficient fine-tuning to ensure that powerful diagnostic tools can be equitably deployed in diverse medical settings.

Methodology

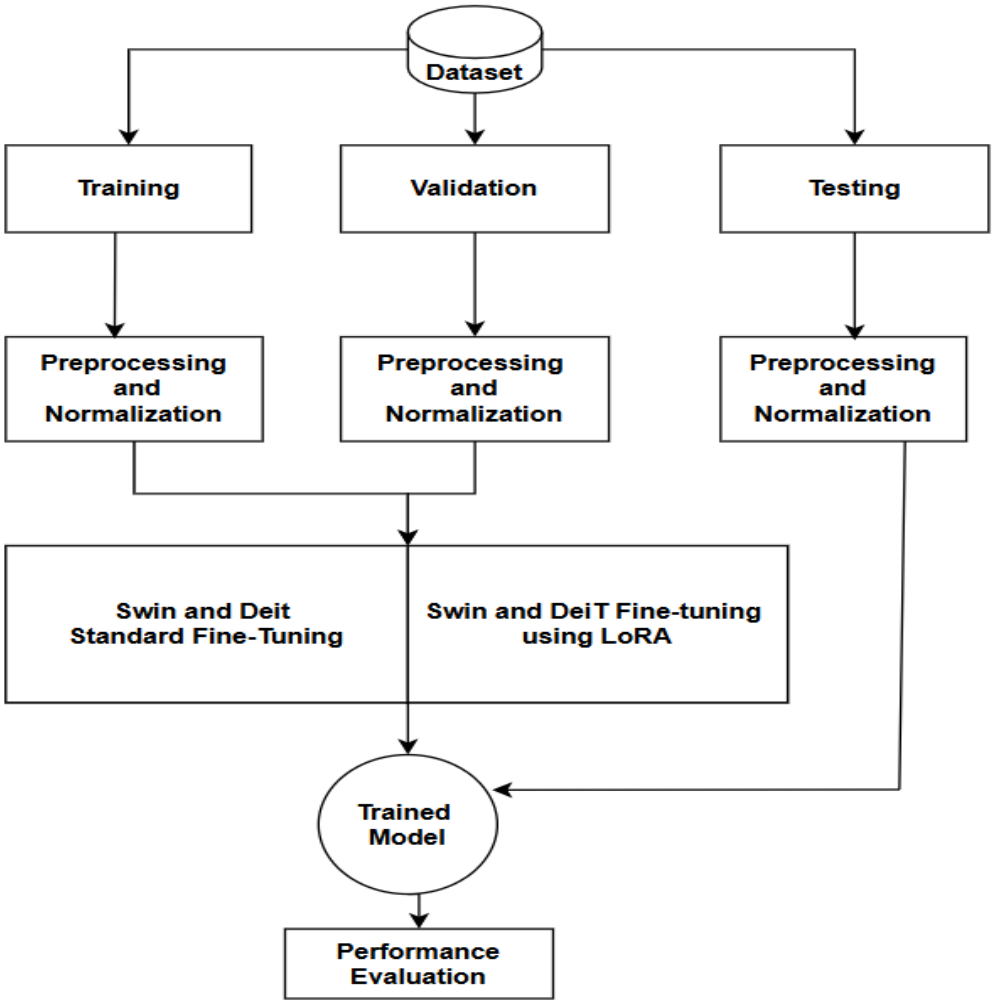


Figure 1: Research Methodology

Dataset Description

In this research the NCT-CRC-HE dataset collected from the National Center for Tumor Disease in Heidelberg is used, which consists of 100000 H&E stained images of colorectal cancer tissue parts. And the images are collected from 86 patients. All the images are standardized into 224*224 pixels for uniformity in the models. This dataset consists of nine different classes of different tissues. For better performance and evaluation, the dataset is partitioned into 80/10/10 split, leaving 80,000 images for training, 10,000 for validation tasks such as hyperparameter tuning and early stopping, and 10,000 as a strictly held-out test set for fair model evaluation.

Explanation of each 9 classes

- **Adipose (ADI)** -- A fat tissues, in which a large and empty appearing adipocytes that store lipids and provides structural cushioning.
- **Background (BACK)** -- It is a non-tissue area on the slide such as empty glass or regions without cellular material.
- **Debris (DEB)** -- It accumulates dead cells necrotic fragments and cellular breakdown products that is found in damaged tissue regions.
- **Lymphocytes (LYM)** -- It is a cluster of small and round immune cells that represent the body's immune response within or around the tumor.
- **Mucus (MUC)** -- Pools of mucin secreted by glands, which is often found in normal tissue or associated with mucinous tumors. It is a healthy tissue.
- **Smooth Muscle (MUS)** --- A long and spindle-shaped muscle fibers that form a part of the bowel wall and provide contractile function.
- **Normal Colon Mucosa (NORM)** -- Healthy epithelial glands and supporting tissue of the colon, which shows typical glandular architecture.
- **Cancer-associated Stroma (STR)** -- Fibrous and dense connective tissue that surrounds and supports tumor cells and often remodeled by cancer.

Table 1: Number of Training, Validation and Testing Images per class

Class	Training Images(80k)	Validation Images(10K)	Testing Images(10K)
ADI	8,295	1,080	1,032
BACK	8,470	1,020	1,076
DEB	9,295	1,130	1,087
LYM	9,322	1,090	1,145
MUC	7,079	930	887
MUS	10,891	1,290	1,355
NORM	6,927	940	896
STR	8,261	1,060	1,125
TUM	11,460	1,460	1,397

Data Preprocessing

For better model evaluation, a uniform preprocessing protocol without data augmentation is applied across all data partitions. To provide stable and numerical optimization, the pipeline first converts 224x224 images to 32-bit floating-point PyTorch tensors. Furthermore, for efficient transfer learning, the tensors have been normalized using standard ImageNet mean and standard deviation statistics.

Low Rank Adaptation (LoRA)

Rather than fine-tuning all the model's parameters, the central idea of the LoRA technique is in training time in which the original pre-trained weights remain frozen. Trainable low-rank decomposition matrices, the LoRA adapters-are injected into the self-attention blocks of the models to adapt the model for the problem specific task. These LoRA modules are applied to the query (Wq) and value (Wv) projection matrices within all multi-head self-attention block across all of the model's stages. To reduce training time, gradient updates are only computed and applied to these newly introduced LoRA adapters. However, the rest of the model's parameters including the MLP layers and other architectural components remain unchanged.

The final classification process follows a fully connected layer. After the last stage, the output vectors are aggregated by global average pooling layer to produce a single feature vector of the entire image for stable and compact output representation. The vector is then passed to a final fully connected classification head that projects it to a logit vector with a dimension equal to the number of target classes. This classification head and the LoRA adapters are the only trainable components of the model which drastically reduce the training time. Lastly, a softmax function is applied to the logits to generate a normalized probability distribution over the classes, from which the final prediction is obtained.

The core equation of LoRA is,

$$h = W_0 x + \frac{a}{r}(B A) x, \quad \text{where}$$

- "h" is the output vector of the layer
- "x" is input vector of the layer
- "a" is hyperparameter that modules the magnitude of the adaptation
- "r" is rank of the adaptation, a key hyperparameter

- "W_0" is the original, pre-trained weight matrix, which is frozen and does not get updated during training time.
- "A" and "B" are the two new small, low-rank adaptation matrices. These are the only matrices that are trained.

The Equation for Merging (Inference): Matrices A and B are fixed during inference after the training. To avoid any inference latency, one time merge is performed.

$W_{Final} = W_0 + \frac{a}{r}$(B A)

Here, the output matrix W_{Final} has the same dimension as the original W_0. and A and B are discarded and deployed a model with this single, unified weight matrix. The inference pass then becomes a simple h = W_{Final} * x , which is identical in speed to a standard fine-tuned model.

Swin Model

The Swin base model classifies 224×224 colorectal images using a hierarchical design which divides the input image into 4×4 non-overlapping image patches. This process first generates 3136 tokens, to effectively capture both the local details and global context, layers are projected into 128-dimensional embedding vectors. The architecture is structured into four progressive stages in which the number of Swin Transformer blocks per stage is set to [2, 2, 18, 2]. Each block contains two key components.

1. In Swin transformer multi-head self-attention works within local 7×7 windows which employs a shifted window mechanism so as to provide cross-window communication to detect long-range dependencies. The number of self-attention heads is increased hierarchically [4, 8, 16, 32] across stages to progressively enrich feature representation of the model.
2. A two multi-layer perceptron (MLP) is used with the GELU activation function, which introduces non-linearity and helps in learning involved hierarchical transformations. Each MLP sub-layer is preceded by layer normalization and followed by residual connections, offering stable training and gradient flow.

By doubling feature dimensionality to 1024, patch merging layers down sample spatial resolution that balance computational efficiency with semantic depth. Eventually, global

average pooling aggregates these features for a fully connected classification head, which applies softmax to predict the final nine tissue subtypes.

Multi-head Self-attention within 7×7 Shifted Windows: In multi-head self-attention the 7*7 shifted windows divide the image tokens into fixed sized local windows, which computes self-attention separately in each window in order to reduce computational complexity. To enable cross-window connections and better capture global context, those windows are shifted between consecutive layers. This process balances efficiency and expressive power in formulating spatial relationships.

DeiT Model

By dividing 224×224 RGB images to 196 non-overlapping 16×16 patches, the DeiT base_16 model classifies colorectal histopathology tissue. These patches are linearly projected into 768_ dimensional embedding vectors. That generates a token sequence which optimizes the transformer architecture for making training more efficient on limited medical datasets. In this model two different special learnable tokens are prepended.

1. A CLS token whose output embedding is used during inference to perform final class prediction.
2. A DIST token which is introduced uniquely in the DeiT framework that is designed to absorb signals from a teacher model through knowledge distillation technique during training. This enables the model to mimic the behavior of a stronger and typically pre-trained CNN-based teacher.

Learnable positional embeddings are appended to the sequence of 198 tokens to ensure spatial awareness during training so that the model can track the position of each pixel. Which includes the special “CLS” and “DIST” tokens. And then after, input is processed through 12 Transformer encoder blocks in a non-hierarchical design. This maintains constant feature dimensionality and resolution throughout the network, which is different from architectures like Swin. Each encoder block is composed of two main sub-layers.

1. A global multi-head self-attention module with 12 attention heads which allows each token to interact with all others globally, which in turn captures better long-range dependencies.

2. A two-layer multilayer perceptron “MLP” that introduces non-linearity via the GELU activation function, enhancing the network's capacity to learn intricate feature transformations.

The architecture uses residual connections and layer normalization, while maintaining a constant 768-hidden-dimensional state. For instance, the “CLS” token drives a linear classifier to identify the nine tissue subtypes. The “DIST” token facilitates knowledge distillation from a teacher CNN that enables the model to achieve high performance and data efficiency despite limited annotated training samples.

Experimental Setup & Environment

The models are implemented using PyTorch, which is the most popular deep learning library in the ML research community. NumPy is used for numerical computation. For visualization and preprocessing Scikit- learn and matplotlib are used. The experiments have been conducted on Google colab by utilizing NVDIA T4 GPU of around 16 GB RAM to accelerate training and testing through CUDA optimized kernels.

Configuration for Standard Fine-tuned models

For both Swin base and DeiT base architectures, a conventional fine-tuning strategy has been adopted to update their pretrained weights for the 9_class classification task. The original classification heads are replaced with a new custom head consisting of a dropout layer with a probability of 0.20. This is followed by a fully connected linear layer. To provide stable convergence while preserving discriminative representations learned by the pretrained backbones, the AdamW optimizer has been used with discriminative learning rates: $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 1e^{-5}$. The base parameters are optimized using a small learning rate of 1×10^{-6} whereas the newly initialized classification head is trained with higher learning rate of 1×10^{-5} along with a weight decay of 0.05 for regularization.

Both models have been trained over 10 epochs. Which use a OneCycleLR learning rate scheduler with cosine annealing. The scheduler incorporates a warm up phase occupying 30% of the total training steps “pct_start = 0.3” and utilizes peak learning rates of $[1 \times 10^{-6}, 1 \times 10^{-5}]$ for backbone and head respectively. The learning rate is initialized at one-tenth of the maximum value “div_factor =10” and gradually decays to one-hundredth of the initial rate

"final_div_factor = 100". Which ensures a smooth transition from warm up to cooldown. This scheduling strategy enables precise weight updates while preventing disruption of the pretrained spatial and semantic hierarchies during fine-tuning.

Configuration for LoRA adapted models

Both Swin and DeiT models have been optimized using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 1e^{-5}$ and a weight decay of 0.05. Separate learning rates of 3×10^{-5} and 1×10^{-4} is assigned to the LoRA parameters and the classification head for both models. A cosine annealing learning rate scheduler with warm restarts has been employed with an initial restart period $T_0 = 15$ and T_mult = 1. At first, the minimum learning rate is set to 5×10^{-6} to prevent premature convergence. To reduce memory requirement the validation phase is conducted in evaluation mode without gradient computation. Mixed precision inference has been applied using automatic casting to improve computational efficiency and the validation loss is accumulated over all batches. While accuracy is computed by comparing predicted class labels with ground truth labels, final validation loss and accuracy is obtained by normalizing over the total number of batches and samples respectively.

A Swin base Transformer model pretrained on ImageNet-22K is used as the backbone model. Low-Rank Adaptation (LoRA) blocks with rank r = 8 and scaling factor $\alpha = 16$ have been integrated to the attention (QKV) and MLP layers of each transformer block. All original backbone parameters are kept frozen to preserve pretrained pattern representations. Only the LoRA parameters and the modified classification head with dropout of 0.20 are trainable. This design has significantly reduced the number of trainable parameters while maintaining strong fine-tuning capability. The Swin model is trained over 10 epochs.

A pretrained DeiT base Vision Transformer has been employed as the backbone model. Low-Rank Adaptation (LoRA) modules with rank r = 8 and scaling factor $\alpha = 16$ have been injected into the attention (QKV) and MLP layers of each transformer block. All original DeiT parameters are kept frozen to preserve pretrained knowledge. Only the LoRA parameters and a modified classification head with dropout of 0.20 is updated during training to reduce number of trainable parameters by offering efficient fine-tuning.

Result Analysis

This data depicts the performance of LoRA adapted fine-tuning and full fine- tuning, LoRA significantly reduces training time by around 16% and 12% for both Swin and DeiT models respectively by reducing the number of trainable parameters from approximately 86-85 million to just 1.3-1.2 million. This result is achieved with no impact on inference speed and even leads to a slight better performance than full-fine tuning. As compared to the base models, the results show the Swin model with LoRA technique achieves a highest accuracy with slight margin.

The macro averaged results in the charts are calculated, the practical application of LoRA offers a dramatic performance improvement in both Swin and DeiT models. The LoRA adapted Swin model has achieved the best results among 4 models in all evaluation metrices: accuracy, precision, recall and F1 score. All of which are around 99.4%. This points a slight improvement than baseline Swin model. Likewise, similar results is obtained in LoRA adapted DeiT model in which LoRA boosts its key metrics from around98.9% to 99.27%, making LoRA an alternative method to full-fine tuning.

Swin Model Evaluation

Swin Standard Fine-Tuning Results

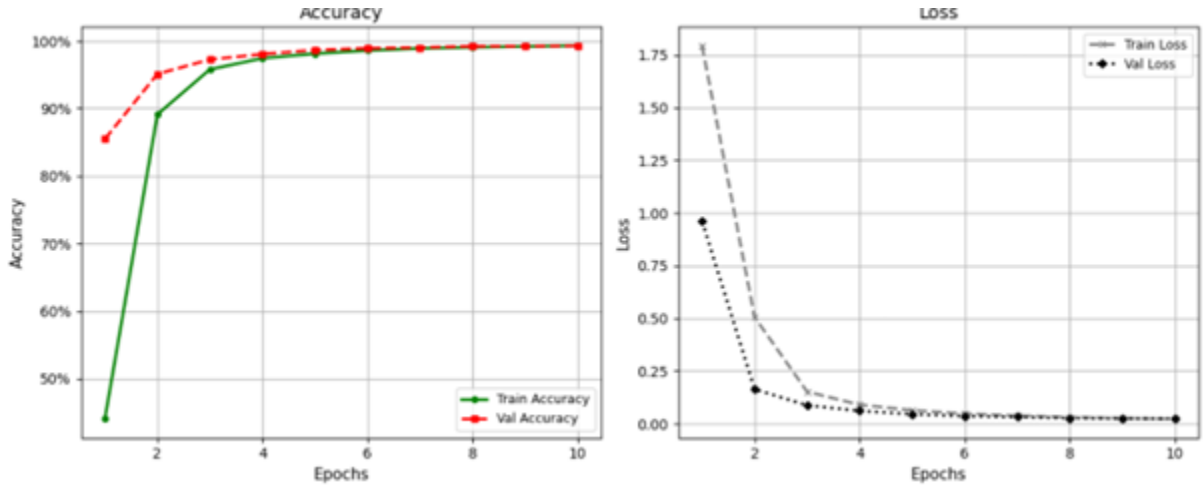


Figure 2: Swin Fine-tuned Model Accuracy and Loss curves

Swin Model with LoRA-Adapted Results

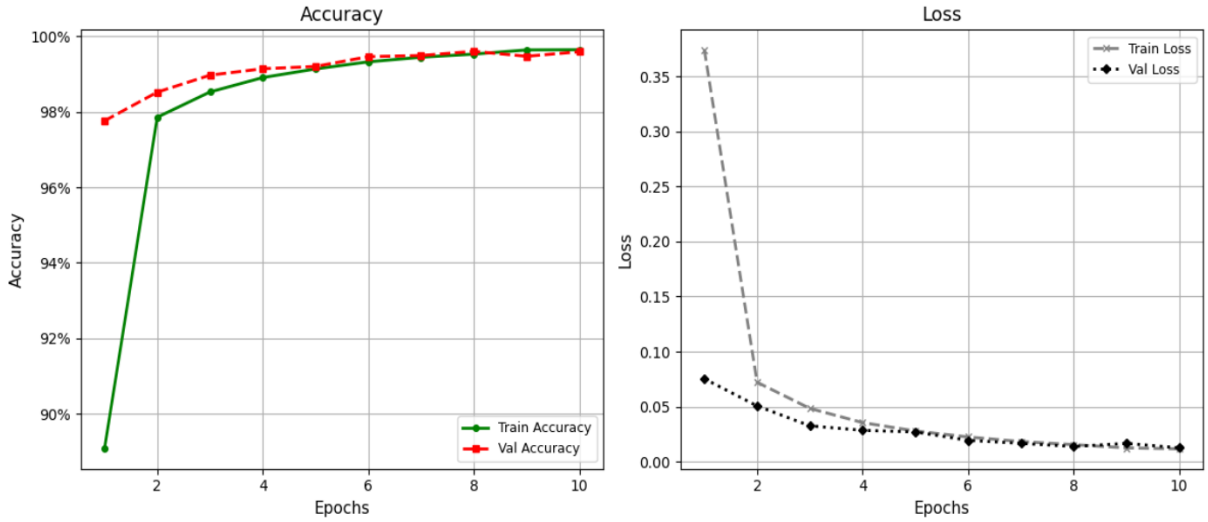


Figure 3: Swin Model with LoRA Accuracy and Loss Curves

DeiT Model Evaluation

DeiT Standard Fine-Tuning

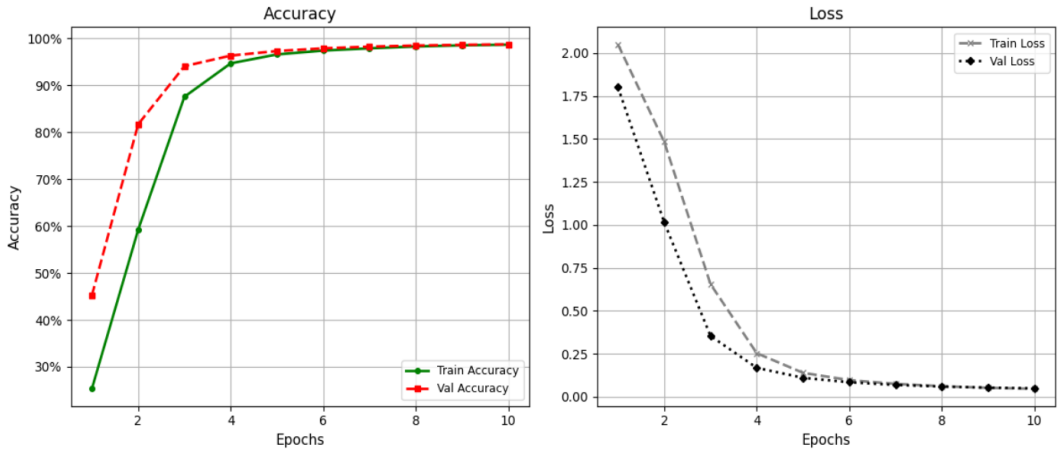


Figure 4: DeiT Fine-tuned Model Accuracy and Loss Curves.

DeiT with LoRA-Adapted Results

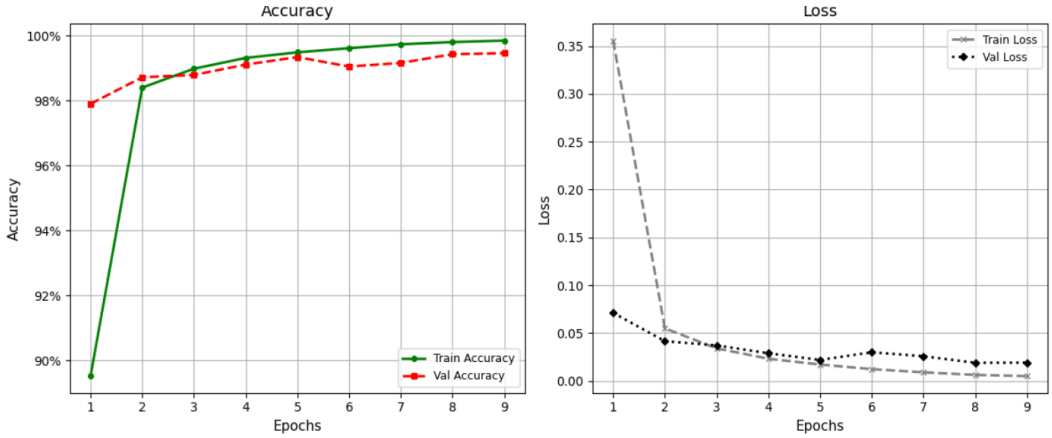
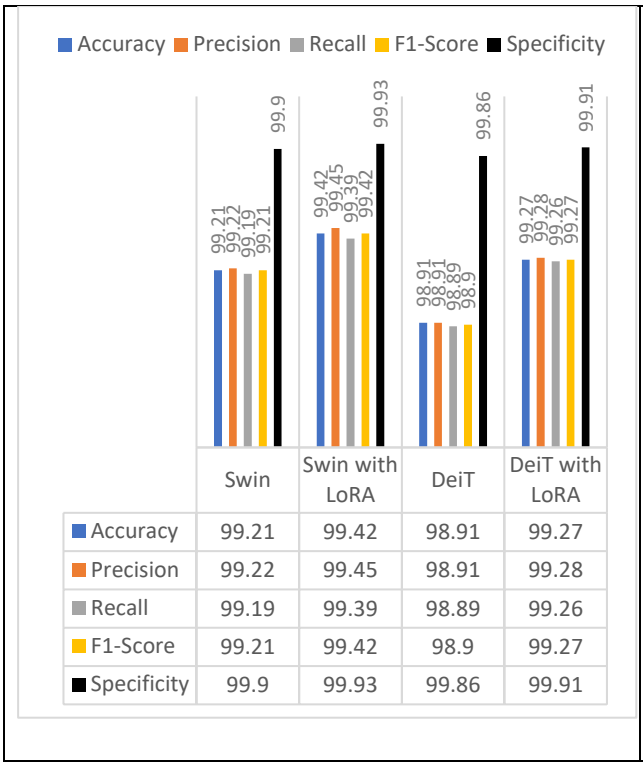


Figure 5: DeiT with Lora Accuracy and Losses Curves.

Models comparison

Table 2: Accuracy, Precision, Recall and Specificity using Test data

Model	Accuracy	ADI Precision, Recall, Specificity	BACK Precision, Recall, Specificity	DEB Precision, Recall, Specificity	LYM Precision, Recall, Specificity	MUC Precision, Recall, Specificity	MUS Precision, Recall, Specificity	NORM Precision, Recall, Specificity	STR Precision, Recall, Specificity	TUM Precision, Recall, Specificity
Swin Base	0.9921	0.9971, 0.9990, 0.9997	0.9991, 1, 0.9999	0.9940, 0.9898, 0.9992	0.9982, 0.9964, 0.9998	0.9921, 0.9843, 0.9992	0.9949, 0.9927, 0.9992	0.9850, 0.9895, 0.9986	0.9833, 0.9888, 0.9980	0.9853, 0.9880, 0.9976
Swin Base with LORA	0.9942	0.9990, 1.0000, 0.9999	1.0000, 1.0000, 1.0000	0.9907, 0.9941, 0.9988	1.0000, 0.9964, 1.0000	0.9955, 0.9854, 0.9996	0.9956, 0.9978, 0.9993	0.9919, 0.9907, 0.9992	0.9916, 0.9879, 0.9990	0.9860, 0.9930, 0.9977
DeiT Base	0.9891	0.9971, 0.9971, 0.9997	0.9972, 0.9991, 0.9997	0.9932, 0.9890, 0.9991	0.9982, 0.9964, 0.9998	0.9842, 0.9809, 0.9985	0.9868, 0.9897, 0.9979	0.9791, 0.9814, 0.9980	0.9831, 0.9795, 0.9980	0.9825, 0.9866, 0.9971
DeiT Base with LoRA	0.9927	0.9981, 0.9990, 0.9998	0.9991, 1.0000, 0.9999	0.9983, 0.9847, 0.9998	1.0000, 0.9991, 1.0000	0.9899, 0.9876, 0.9990	0.9941, 0.9941, 0.9991	0.9930, 0.9895, 0.9993	0.9760, 0.9879, 0.9971	0.9867, 0.9916, 0.9978



Model Name	Total Parameters	Trainable Parameters	Training time per epoch	Inference time	Accuracy (%)
Swin	86 M	86 M	17.55 mins	55 sec	99.21
Swin with LoRA	87 M	1.3 M	14.45 mins	55 sec	99.42
DeIT	85 M	85 M	10.9 mins	29 sec	98.91
DeIT with LoRA	86 M	1.2 M	9.5 mins	29 sec	99.27

Figure 6: Performance Evaluation

Conclusion and Future Works

Conclusion

It is found that the LoRA technique achieves even superior performance as compared to their full-fined tuned models for both Swin and DeiT transformers by reducing trainable parameters by more than 98%, and reducing training time significantly. For the Swin and DeiT models accuracy improved from 99.21% to 99.42% and from 98.91% to 99.27% respectively. The tiny added LoRA metrices during training can be merged back during testing time, having no impact on inference time as compared to conventional fine tuning. By using LoRA techniques rather than managing various large models for every problem domain, hospitals and labs can utilize a unique base model with its several smaller and lightweight LoRA adapters, which can considerably reduce computational costs, thereby, making it practically feasible for numerous medical tasks.

Future Works

Based on the success of LoRA for histopathological images, it is crucial to validate its efficiency and efficacy across other imaging domains such as radiological scans or satellite

images. This helps find out its generalizability. Hybrid methodological techniques are recommended to strengthen its accuracy in all imaging domains along with the standard LoRA, in which LoRA modifies the internal patterns of existing weight matrices where as other method like Adapters reshape the information flow between two consecutive layers, allowing orthogonal improvements. Furthermore, a dynamic rank allocation technique where some layers perform better in different rank while others excel at different values of rank because pretrained weights of some layers are more important than others for the specific problem domains.

References

Abou Baker, N., Rohrschneider, D., & Handmann, U. (2024). Parameter-Efficient Fine-Tuning of Large Pretrained Models for Instance Segmentation Tasks. *Machine Learning and Knowledge Extraction*, *6*(4), 2783–2807. https://doi.org/10.3390/make6040133

Afshin, P., Helminiak, D., Lu, T., Yen, T., Jorns, J. M., Patton, M., Yu, B., & Ye, D. H. (2025). *Breast Cancer Classification in Deep Ultraviolet Fluorescence Images Using a Patch-Level Vision Transformer Framework*. http://arxiv.org/abs/2505.07654

Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S., & Islam, M. (2022). An Effective Skin Cancer Classification Mechanism via Medical Vision Transformer. *Sensors*, *22*(11). https://doi.org/10.3390/s22114008

Dai, Y., Gao, Y., & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, *11*(8). https://doi.org/10.3390/diagnostics11081384

Demir, C., & Yener, B. (2005). *Automated cancer diagnosis based on histopathological images: a systematic survey*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. http://arxiv.org/abs/2010.11929

Gallée, L., Lisson, C. S., Beer, M., & Götz, M. (2025). *Hierarchical Vision Transformer with Prototypes for Interpretable Medical Image Classification*. http://arxiv.org/abs/2502.08997

Hayat, M., Ahmad, N., Nasir, A., & Tariq, Z. A. (2024). Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3503413

Hong, Z., Xiong, J., Yang, H., & Mo, Y. K. (2024). Lightweight Low-Rank Adaptation Vision Transformer Framework for Cervical Cancer Detection and Cervix Type Classification. *Bioengineering*, *11*(5). https://doi.org/10.3390/bioengineering11050468

Ignatov, A., & Malivenko, G. (2025). *NCT-CRC-HE: Not All Histopathological Datasets Are Equally Useful*. https://github.com/gmalivenko/NCT-CRC-HE-experiments.

Kanadath, A., Angel Arul Jothi, J., & Urolagin, S. (2024). CViTS-Net: A CNN-ViT Network With Skip Connections for Histopathology Image Classification. *IEEE Access*, *12*, 117627–117649. https://doi.org/10.1109/ACCESS.2024.3448302

Khalid, M., Deivasigamani, S., V, S., & Rajendran, S. (2024). An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-70117-y

Li, H., Zhang, Y., Chen, P., Shui, Z., Zhu, C., & Yang, L. (2024). *Rethinking Transformer for Long Contextual Histopathology Whole Slide Image Analysis*. http://arxiv.org/abs/2410.14195

Malekmohammadi, A., Badiezadeh, A., Mirhassani, S. M., Gifani, P., & Vafaeezadeh, M. (2024). *Classification of Gleason Grading in Prostate Cancer Histopathology Images Using Deep Learning Techniques: YOLO, Vision Transformers, and Vision Mamba*. http://arxiv.org/abs/2409.17122

Mármol, I., Sánchez-de-Diego, C., Dieste, A. P., Cerrada, E., & Yoldi, M. J. R. (2017). Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. In *International Journal of Molecular Sciences* (Vol. 18, Issue 1). MDPI AG. https://doi.org/10.3390/ijms18010197

Preda, A.-A., Tăiatu, I.-M., & Cercel, D.-C. (2025). *Scaling Federated Learning Solutions with Kubernetes for Synthesizing Histopathology Images*. http://arxiv.org/abs/2504.04130

Qin, Z., Sun, W., Guo, T., & Lu, G. (2024). Colorectal cancer image recognition algorithm based on improved transformer. *Discover Applied Sciences*, *6*(8). https://doi.org/10.1007/s42452-024-06127-2

Ren, K., Hong, G., Chen, X., & Wang, Z. (2023). A COVID-19 medical image classification algorithm based on Transformer. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-32462-2

Romero, D. W., Knigge, D. M., Gu, A., Bekkers, E. J., Gavves, E., Tomczak, J. M., & Hoogendoorn, M. (2022). *Towards a General Purpose CNN for Long Range Dependencies in ND*. http://arxiv.org/abs/2206.03398

Träff, H. (2023). *Comparative Analysis of Trans-former and CNN Based Models for 2D Brain Tumor Segmentation*. www.liu.se

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). *Are Convolutional Neural Networks or Transformers more like human vision?* http://arxiv.org/abs/2105.07197

Tummala, S., Kim, J., & Kadry, S. (2022). BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. *Mathematics*, *10*(21). https://doi.org/10.3390/math10214109

Uparkar, O., Bharti, J., Pateriya, R. K., Gupta, R. K., & Sharma, A. (2022). Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. *Procedia Computer Science*, *218*, 2338–2349. https://doi.org/10.1016/j.procs.2023.01.209

Vadori, V., Peruffo, A., Graïc, J.-M., Finos, L., & Grisan, E. (2025). *Mind the Gap: Evaluating Patch Embeddings from General-Purpose and Histopathology Foundation Models for Cell Segmentation and Classification*. http://arxiv.org/abs/2502.02471

Veasey, B. P., & Amini, A. A. (2025). Low-Rank Adaptation of Pre-trained Large Vision Models for Improved Lung Nodule Malignancy Classification. *IEEE Open Journal of Engineering in Medicine and Biology*. https://doi.org/10.1109/OJEMB.2025.3530841

Xin, Y., Yang, J., Luo, S., Zhou, H., Du, J., Liu, X., Fan, Y., Li, Q., & Du, Y. (2024). *Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey*. http://arxiv.org/abs/2402.02242

Yang, Z., Qiu, Z., Lin, T., Chao, H., Chang, W., Yang, Y., Zhang, Y., Jiao, W., Shen, Y., Liu, W., Fu, D., Jin, D., Yan, K., Lu, L., Jiang, H., & Bian, Y. (2024). *From Histopathology Images to Cell Clouds: Learning Slide Representations with Hierarchical Cell Transformer*. http://arxiv.org/abs/2412.16715

Yin, D., Han, X., Li, B., Feng, H., & Bai, J. (2023). *Parameter-efficient is not sufficient: Exploring Parameter, Memory, and Time Efficient Adapter Tuning for Dense Predictions*. https://doi.org/10.1145/3664647.3680940