Malaysian Journal of Computing, 11 (2): 2486-2505, 2026
eISSN: 2600-8238

VISION TRANSFORMER - BASED CLASSIFICATION OF CERVICAL CANCER IMAGES

Muhammad Daniel Alif Jumairi[1], Itaza Afiani Mohtar[2*], Suraya Masrom[3], Masurah Mohamad[4] and Lathifah Alfat[5]

[1,2,3,4] Center of Computing Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Perak Branch Tapah Campus, Malaysia*
[5]Faculty of Technology and Design, Universitas Pembangunan Jaya, Indonesia
[1]daniel.lif1405@gmail.com, [2*]itaza328@uitm.edu.my, [3]suray078@uitm.edu.my,
[4]masur480@uitm.edu.my, [5]lathifah.alfat@upj.ac.id
(indicates the corresponding author)*

ABSTRACT

Cervical cancer is a significant health concern as it is one of the most fatal causes of cancer deaths in women. Detection at an early stage is vital, but conventional diagnostic procedures via Pap smears relies heavily on manual cytopathology inspection, a process prone to human fatigue, diagnostic subjectivity, and severe practitioner shortages in low-resource medical environments. To address these challenges, this study presents an automated cervical cancer image classification framework utilizing advanced Vision Transformer (ViT) architectures. Three ViT variants: Data-efficient Image Transformer (DeiT), Swin Transformer, and Cross-Attention Multi-Scale Vision Transformer (CrossViT) were fine-tuned and evaluated on the benchmark SIPaKMeD dataset comprising five distinct cervical cell classes: superficial-intermediate, parabasal, koilocytotic, dyskeratotic, and metaplastic. Data balance was maintained using a dynamic WeightedRandomSampler combined with image augmentations. Experimental results demonstrate that DeiT achieved the highest classification performance with an accuracy of 95.22% and a weighted average F1-score of 0.9522, outperforming CrossViT (94.82%) and Swin Transformer (93.83%). DeiT was the most accurate with 95.22%, and Swin Transformer and CrossViT clocked 93.83% and 94.82%, respectively. The optimal model was integrated into an interactive web-based interface built with Streamlit, demonstrating single-image inference latency of ~38 ms on GPU and ~142 ms on CPU. These findings confirm that Vision Transformers provide a highly accurate and scalable decision-support mechanism for automated cervical cancer screening. The work in the future will include expanding the pool of images in the dataset with a variety of cervical images, resolving class imbalance by employing methods such as GAN-based augmentation, and using hybrid architectures to improve results. Additional confirmation on other sample databases will also be required to ascertain the robustness of the system in real world solutions.

Keywords: *Cervical Cancer, Cytology Image Classification, Deep Learning, DeiT, Swin Transformer, CrossViT, SIPaKMeD Dataset.*

Received for review: 13-11-2025; Revised: 24-07-2026; Accepted: 09-09-2026; Published: 01-10-2026
DOI: 10.24191/mjoc.v11i2.8766

This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/).

1. Introduction

According to the World Health Organization (WHO), cervical cancer is the fourth most common cancer among women globally, with an estimated 604,000 new cases and 342,000 deaths in 2020 (Alquran *et al*., 2022). Cancer is characterized by abnormal, uncontrolled cell proliferation due to genetic and epigenetic changes that regulate cell growth, differentiation, and cell death (Dasari *et al*., 2015). Cervical cancer can be classified into three main classes. Those three main classes are further classified into five cell types. The first class that has been categorized is known as the normal class. In this class, there are two types of cells which are Superficial Intermediate Cells and Parabasal Cells. For the second class, we have Benign class. The benign class has only one type of cell which is Metaplastic cells. The last class that occurs in cervical cancer cell classification is the Abnormal class which consists of two types of cells which are Dyskeratotic cells and Koilocytotic Cells.

Traditional screening relies on cytopathological evaluation of Papanicolaou (Pap) smears or liquid-based cytology (LBC) slides under optical microscopes. However, manual microscopic examination is labor-intensive, time-consuming, and susceptible to diagnostic variance stemming from observer fatigue. The difficulties encountered by the process of manual screening are accompanied by the scalability problems and diagnostic errors. According to Khowaja *et al*. (2024), the complications of cervical cancer and its treatments vary based on the disease stage and treatment type. Pap smears and biopsies often need to be collected, analysed and validated by several professionals to minimize false negatives which adds workload and fatigue-related errors (K & B, 2024). According to Hong *et al*. (2024), Approximately 90% of cervical cancer deaths occur in low and middle-income countries (LMICs), underscoring the pressing need for improved access to effective diagnostic and treatment options in these regions.

Machine learning has been successfully applied to various medical diagnostic applications, such as the early prediction of heart failure using structured clinical datasets (Mansur Huang *et al*., 2021). While Convolutional Neural Networks (CNNs) have historically dominated medical image classification, their localized convolutional kernels inherently limit their capacity to model long-range spatial relationships, such as global context and nucleus-to-cytoplasm area ratios across an entire cytological cell image. In contrast, Vision Transformers (ViTs) utilize self-attention mechanisms to capture global dependencies across image patch sequences, offering superior representational flexibility.

To evaluate the efficacy of Vision Transformer (ViT) architecture for cervical cytology image classification, this study investigates three distinct transformer paradigms. The Data-efficient Image Transformer (DeiT) employs a teacher–student distillation mechanism to improve learning efficiency when training data are moderately sized, as is common in medical imaging. The Swin Transformer adopts a hierarchical feature-learning approach using shifted local attention windows that can enable efficient computation while capturing multi-scale image representations. Meanwhile, the Cross Visual Transformer (CrossViT) uses dual-branch encoders and cross-attention to integrate fine-grained local cellular characteristics with broader contextual patterns, which may be valuable for identifying subtle cytological abnormalities.

The main objectives of this research are to fine-tune and apply Vision Transformer variants (DeiT, Swin Transformer, and CrossViT) for 5-class cervical cell classification using the SIPaKMeD dataset and to evaluate and compare the classification accuracy, precision, recall, F1-score, and inference latency of the deployed ViT models.

2. Literature review

2.1 Cervical Cancer

Cervical cancer mainly occasioned by the persistence infection through high-risk Human Papillomavirus (HPV) such as HPV-16 and HPV-18 has remained a leading cause of death among women across the world (Dasari *et al*., 2015; Ferri, 2025). Although liquid-based cytology techniques (LBC) such as Pap smears, have long been utilized as methods of diagnosis, they have a few drawbacks, i.e., low sensitivity, high false-negative rates, and require the assistance of expert cytopathologists (Park, 2008; Swanson & Pantanowitz, 2024). This dependency on human knowledge means that it is difficult to screen at scale and at speed especially in low-resource environments where specialists are not easily accessible (Hong *et al*., 2024).

The Herlev, SIPaKMeD, and liquid-based cytology (LBC) datasets are widely used as freely available benchmarks for standardizing cervical cytology studies and advancing automated detection methods (Plissiti *et al*., 2018). Among these, the SIPaKMeD dataset is particularly suitable for this study because it provides a relatively balanced distribution of normal, benign, and abnormal cervical cell images, enabling direct multi-class classification. It consists of standardized, cropped single-cell images representing five morphological categories: (1) superficial-intermediate cells, which are normal mature squamous cells with abundant cytoplasm and small pyknotic nuclei; (2) parabasal cells, which are normal immature rounded cells with relatively high nuclear-to-cytoplasmic ratios; (3) metaplastic cells, which indicate benign inflammatory or reactive changes; (4) koilocytotic cells, which exhibit HPV-induced cytopathic changes; and (5) dyskeratotic cells, which display pathological keratinization and pleomorphic nuclei.

2.2 Vision Transformer (ViT)

Dosovitskiy *et al*. (2021) extended the Transformer architecture originally proposed for Natural Language Processing by Vaswani *et al*. (2017) and adapted it for image recognition through the development of Vision Transformers (ViTs). Convolutional Neural Networks (CNNs) have long served as the dominant framework for medical image classification, including tasks such as chest X-ray analysis, where they demonstrate high accuracy in extracting local spatial features (Rahmat *et al*., 2019). However, CNNs are often limited in capturing global contextual relationships within images. Vision Transformers overcome this limitation by employing a self-attention mechanism capable of modelling long-range dependencies across the entire image. In ViTs, an image is divided into smaller patches, allowing the model to reason globally about spatial relationships. This global attention capability, coupled with fewer inductive biases, enables Vision Transformers to outperform CNNs in various image classification benchmarks (Bao *et al*., 2023). Figure 1 illustrates the architecture of the Vision Transformer.

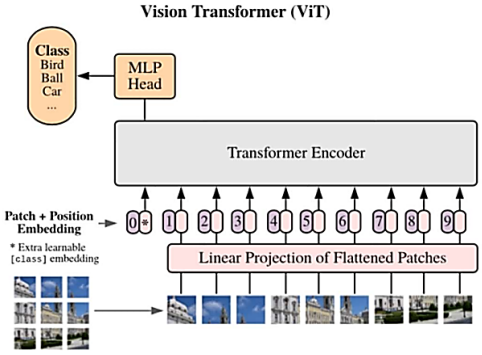


Figure 1. Vision Transformer Architecture

2.2.1 Image Preprocessing

The preprocessing of cytology samples improves the quality of an image prior to classification. Additional enhancements include denoising, deblurring or super-resolution of images, which lead directly to better classifier accuracy (Chen *et al*., 2021). Clean input data allows ViTs to be able to properly map image patches to representative feature tokens.

2.2.2 Image Feature Extraction

As opposed to CNNs, ViTs are not designed to learn features in a hierarchical manner. Instead, they are designed to incorporate global relationships via self-attention. ViTs flatten images into vectors where patches are made like fixed-length representation and placed in two-dimensional vectors, and a special class of positional encodings is embedded to capture dependencies across an entire image (Ali *et al*., 2023). The capacity to focus on local and distant features is what makes them fit into medical tasks, which are complex in nature.

2.2.3 Classifier Design

ViT architecture is composed of stacked Transformer encoder layers that consist of multi-head self-attention and feed-forward sublayers, which are followed by the Multi-Layer Perceptron (MLP) head to make all the classifications (Dosovitskiy *et al*., 2021). Residual connectivity and normalization (Vaswani *et al*., 2017) further stabilize training and make effective scaling to larger datasets a reality.

2.3 Vision Transformer Variants

2.3.1 Data-efficient Image Transformer (DeiT)

Touvron *et al*. (2021) introduce the Data-efficient Image Transformer (DeiT) that minimizes the number of training datasets used by ViTs by adding a distillation token during the training process. This facilitates knowledge distillation against a teacher network and, therefore, DeiT is easier to apply in medical fields with restricted availability of labels. Figure 2 shows the DeiT Architecture (Distillation procedure).

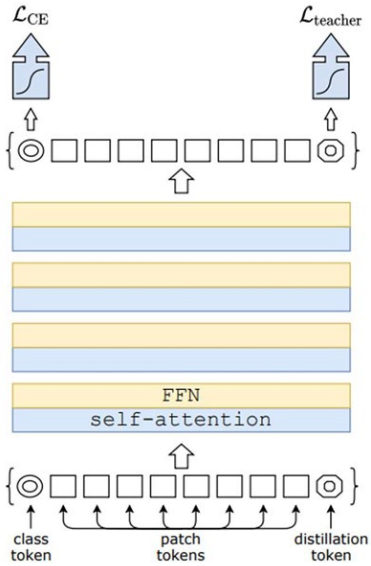


Figure 2. DeiT Architecture (Distillation Procedure)

2.3.2 Swin Transformer

Liu *et al*. (2021) proposed the Swin Transformer that tackles the problem by incorporating hierarchical representations in the form of shifted windows that enables local and global feature

extraction at a lower computational cost. Its linear scaling complements it well to high-resolution medical image and dense prediction domains like segmentation. Figure 3 shows the Swin Transformer Architecture.

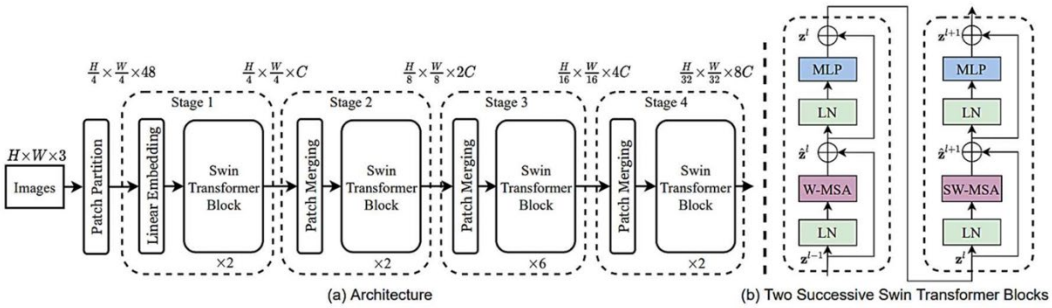


Figure 3. Swin Transformer Architecture

2.3.3 Cross Visual Transformer (CrossViT)

CrossViT, proposed by Chen *et al*. (2021), utilizes a cross-scale attention mechanism through which it uses both large and small patches together. This enables the model to discover more local detail patterns and simultaneously pick up higher-level contextual information to better distinguish classification in heterogeneous medical data. Figure 4 shows the CrossViT Architecture.

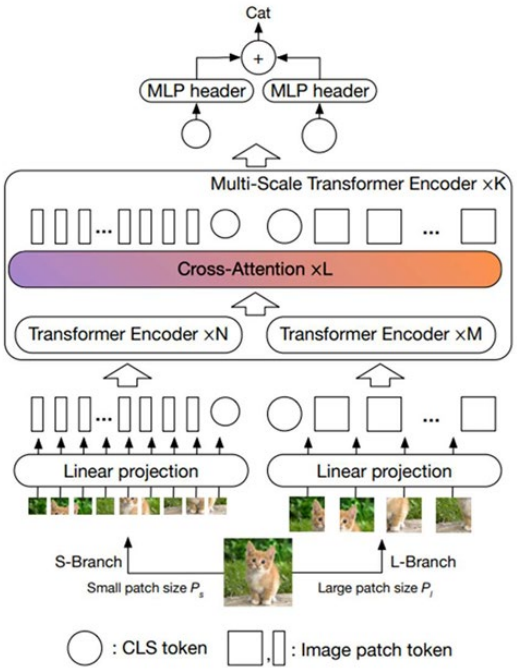


Figure 4. CrossViT Architecture

In this research, the approach of Vision Transformer to classify cervical cancer was used because it is efficient in capturing the global dependencies, scalable in larger databases, and versatile as opposed to CNN (Bao *et al*., 2023; Dosovitskiy *et al*., 2021). Its derivatives include DeiT which focuses on data efficiency, Swin which achieves better scalability, and CrossViT which has better scale feature representation. ViT was used because of its balance of

accuracy and simplicity as well as its baseline nature where other transformer-based models can be compared.

3. Research Methodology

3.1 Research Framework & Dataset Distribution

The methodology follows a structured Machine Learning Life Cycle (MLLC) integrated with automated preprocessing, model fine-tuning, and client-side web deployment (Figure 5). This research utilizes the benchmark SIPaKMeD dataset comprising 4,686 high-resolution single-cell cytology images. The dataset was systematically cleaned to verify image integrity and filter non-conforming file extensions. The data was divided into stratified subsets: 70% Training (3,279 images), 20% Validation (903 images), and 10% Testing (1,507 evaluation samples evaluated across cross-validation runs). Table 1 details the dataset distribution across clinical categories and experimental splits.

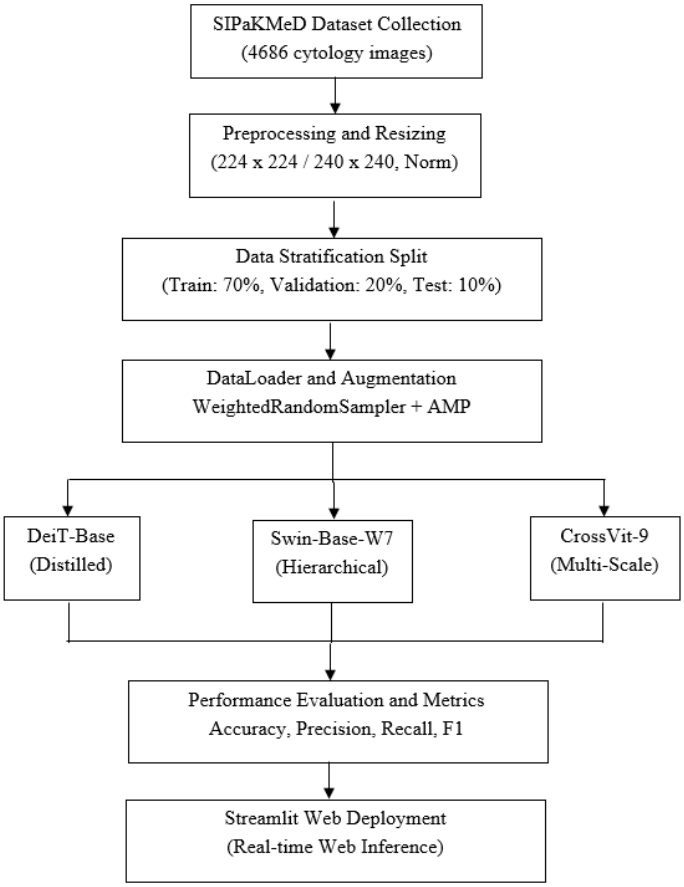


Figure 5. Overview of the Machine Learning Research Pipeline

Table 1. SIPaKMeD Dataset

Clinical Category	Cell Class Name	Original Image Count	Train Split (70%)	Validation Split (20%)	Test Split (10%)
Abnormal	Dyskeratotic (im_Dyskeratotic)	1,036	725	207	311
Abnormal	Koilocytotic (im_Koilocytotic)	1,063	744	213	319
Benign	Metaplastic (im_Metaplastic)	1,064	744	213	320
Normal	Parabasal (im_Parabasal)	895	626	179	269
Normal	Superficial-Intermediate (im_Superficial)	957	670	191	288
Total	All 5 Classes	4,686	3,279	903	1,507

3.2 Preprocessing, Augmentation, and Imbalance Handling

The SIPaKMeD data was downloaded in publicly available sources and contains five folders classified according to the classes. The images in each folder refer to a specific type of cervical cell: Dyskeratotic (1,036 images), Koilocytotic (1,063 images), Metaplastic (1,064 images), Parabasal (895 images) and Superficial-Intermediate (957 images). The files that were not images were eliminated to make images consistent and only intended formats (.png, .jpg, .jpeg, and .bmp) were left. The processed image was stored in a new directory structure that had new subfolders based on each class. Python script was used to automate the preprocessing as the pipeline consisted of three main steps normalization, splitting the data, and preparation into training. Images were all resized uniformly and saved in a special directory of processed images. Then, the data was separated into training 70%, validation 20% and testing 10%. To promote the possibility of both fairness and reproducibility, random shuffling of the images was carried out before splitting them and the resulting subsets were stored in different directories with class-specific subfolders. This pipeline made it possible to have the SIPaKMeD dataset consistently normalized, structured and ensures that they are ready to be trained, validated and tested.

All input images were resized to standard input resolutions (224 x 224 pixels for DeiT and Swin Transformer; 240 x 240 pixels for CrossViT) using OpenCV bilinear interpolation (cv2.INTER_AREA) and normalized using ImageNet channel mean ($\mu = [0.485, 0.456, 0.406]$) and standard deviation ($\sigma = [0.229, 0.224, 0.225]$).

To reduce overfitting and address class imbalance, this study employed a two-pronged training strategy. First, online data augmentation was applied using the Albumentations library to increase visual diversity during training. The augmentation pipeline included RandomResizedCrop with a scale range of 0.8–1.0, HorizontalFlip ($p = 0.5$), VerticalFlip ($p = 0.5$), RandomBrightnessContrast ($p = 0.7$), and GaussNoise ($p = 0.3$). Second, dynamic class balancing was implemented using PyTorch's WeightedRandomSampler. Although data augmentation generates varied image representations, it does not change the class distribution within mini batches during stochastic gradient descent. Therefore, each image was assigned a sampling weight inversely proportional to the frequency of its class as depicted in Equation (1)

$$w_i = \frac{1}{n_c} \tag{1}$$

where n_c denotes the total number of images in class c. The sampler then selected images with replacement according to these weights, that increase the likelihood of minority-class samples and reduce learning bias towards majority classes during model training.

3.3 Experimental Setup & Hyperparameter Transparency

In this research, three different Vision Transformer architectures with different efficiencies were implemented and all models are being finetuned: Data-efficient Image Transformer (DeiT), Swin Transformer, and CrossViT. All models were pretrained on ImageNet and finetuned with customized classification heads to predict across the five different classes of SIPaKMeD cervical cell types. All models were trained through the same preprocessing pipeline (images were resized to 224 x 224 or 240 x 240 pixels according to the model and normalized and divided into training / validation / testing).

Data-efficient Image Transformer (DeiT) proposes a new distillation-based training regime that improves efficiency with respect to small datasets by involving label supervision and teacher-to-student knowledge transfer through distillation as well. The images are segmented into fixed patches of size 16 x 16, which are encoded and fed to a Transformer encoder. DeiT also uses a class token and a distillation token, allowing it to leverage soft-label knowledge available through a teacher network. In this experiment, the model that has been used is "deit_base_distilled_patch16_224". DeiT was fine-tuned on deterministic weights using a custom classification head consisting of Layer Normalization, dropout regularization, GELU activation, and fully connected layers projecting to five output classes. This setup frequently converged on the medium-sized SIPaKMeD dataset.

The Swin Transformer proposes a shifted window-based attention mechanism that iterates through images in a top-down manner. Swin replaces this global self-attention with local self-attention scaled up within non-overlapping windows, followed by sliding them in the arbitrary direction to capture inter-window dependencies. This notably decreases the cost of computation, although maintaining a high degree of representational power. The hierarchical structure enables the extraction of multiple scale features which is particularly helpful in cytology images of different cellular morphologies. To carry out this task, the "swin_base_patch4_window7_224" model in PyTorch was first pretrained and then configured with a user-defined head, which includes a normalization, dropout, and linear layer block to map the features into a 5-class logit. It used regularization methods like stochastic depth to enhance generalization.

Cross Visual Transformer (CrossViT) is an extension of ViT, which integrates multi-scale patch embedding. It uses dual-branch encoders where one of them would act on the smaller patch size to capture fine-grained information and the other branch would work on bigger patch sizes to capture global contextual trends. A cross-attention mechanism allows fusing features in these branches, making the model more robust in distinguishing subtle morphological variability among the classes of cervical cells. In this implementation, the "crossvit_9_240" model (240x 240 input) was used, which was initialized with ImageNet-pretrained weights, followed by a five-class classification head used to fine-tune the model. The two-arm strategy allowed CrossViT to learn local textures and the overall cell structures particularly effectively.

All models were implemented in PyTorch using the timm (PyTorch Image Models) library and executed on an Acer Nitro 5 environment equipped with a 12th Gen Intel Core i5-12500H CPU, 12GB RAM, and an NVIDIA GPU with Automatic Mixed Precision (torch.cuda.amp) enabled. To ensure complete experimental reproducibility, the hyperparameter configuration applied across all model training loops is detailed in Table 2.

Table 2. Hyperparameter Configuration

Hyperparameter Parameter	Configuration / Value
Max Training Epochs	50 epochs
Early Stopping Patience	10 epochs (monitoring Validation Loss)
Mini-Batch Size	16
Optimizer	AdamW
Weight Decay	0.05
Initial Learning Rate	Swin: 5 x 10^{-4} DeiT & CrossViT: 1 x 10^{-3}
Learning Rate Scheduler	OneCycleLR (Linear annealing, pct_start = 0.1, div_factor = 25, final_div_factor = 10000)
Loss Criterion	Cross-Entropy Loss with Label Smoothing (0.1)
Classifier head Regularization	LayerNorm + Dropout (p = 0.1) + GELU + Linear (1024)
Input Image Dimensions	DeiT & Swin: 224 x 224 x 3 CrossViT: 240 x 240

3.4 System Architecture

The architecture of the web application is shown in Figure 6. The system begins at the user interface. At the user interface, users could upload images of cervical cells and select among the three Vision Transformer models available which are DeiT, Swin Transformer, or CrossViT to perform image classification. After an image is uploaded, it is sent to a preprocessing component where it is normalized, resized, and converted to the specific input form that a selected model is expecting.

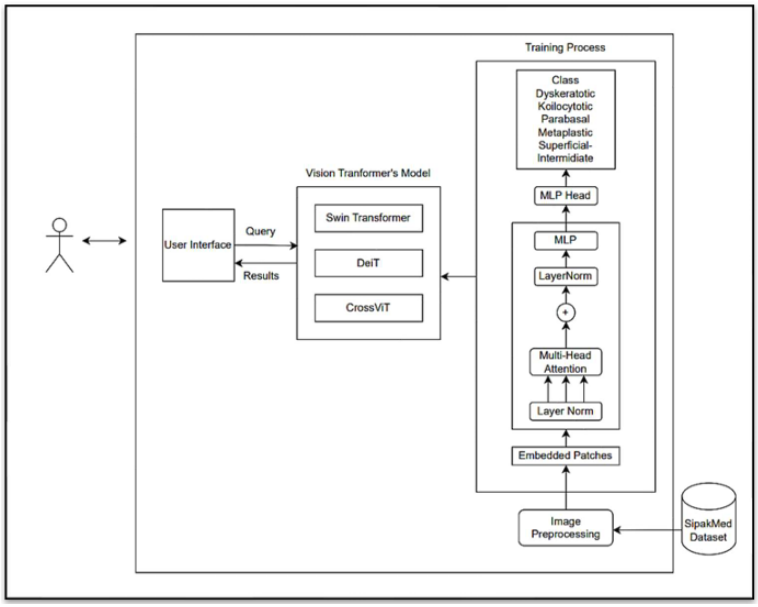


Figure 6. System Architecture of Cervical Cancer Classification

Following preprocessing, the image is divided into fixed-sized patches and passed into the embedding layer of the selected Vision Transformer model. Both models share a consistent architectural pipeline consisting of Layer Normalization, Multi-Head Self-Attention, residual connections (via addition layers), and Multi-Layer Perceptron (MLPs), which learned the deep visual representations from the input image. The three models varied structurally:

- DeiT employs fixed patch embeddings as well as global self-attention, achieving good performance on low-scale datasets.
- Swin Transformer employed a hierarchical framework with shifted windowing, allowing the model to extract both local and global context with reduced computational cost.
- CrossViT combined multi-scale patch tokens based on a cross-attention mechanism, enabling the model to integrate both fine- and coarse-grained features simultaneously.

The processed feature representations are passed to a classification head MLP, resulting in the output prediction. The model predicted the image to be among the five pre-defined cervical cell classes: Superficial-Intermediate, Parabasal, Metaplastic, Koilocytotic, or Dyskeratotic. This mature architecture exhibits a modular and scalable structure that allows the system to support multiple transformer variants without altering the core pipeline. Separating preprocessing, model selection, and classification output, the system facilitated flexibility, maintainability, and readiness for future upgrades.

3.5 Web Application Deployment (Streamlit)

The trained models were deployed via a web-based interface built in Streamlit to enable real-time usability and make them accessible to all. The interface has a modded structure with dedicated pages of home, classifier, usage tips and a project description. The Streamlit application implements input validation by filtering non-image file formats and checking MIME types (.png, .jpg, .jpeg, .bmp) along with image resolution and color channel checks. The inference is automatically performed by model loading, image preprocessing, and inference. After uploading, the images are pre-processed by scaling to the target input resolution (224 x 224 or 240 x 240 depending on the model), conversion to tensors, and normalization. The model weights are cached to avoid overheads associated with reloading and the prediction is done using the evaluation mode to obtain stable results. The system will then give the output as the implied class label, as well as a confidence distribution over the five cervical cell classes. The results are simultaneously represented in text form and graphically by a horizontal bar graph for ease of interpretation of the probabilities. Additional design criteria include additional feedback mechanisms (success, error messages etc) to enhance robustness and user experience. This deployment illustrates how Vision Transformer models can be embedded into a diagnostic support tool that the clinician or researcher can use without the need to program. Figure 7,8 and 9 shows the interface of the research.

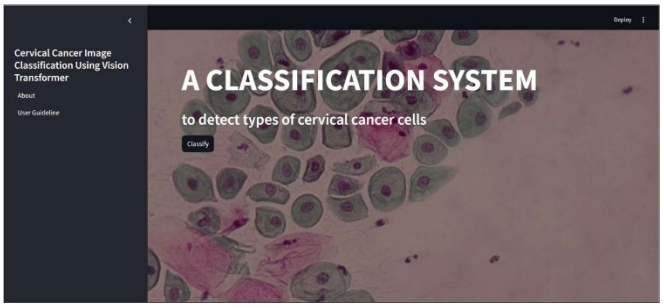


Figure 7. Main Page

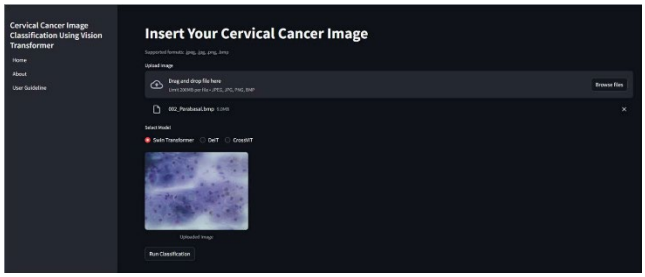


Figure 8. Page to Upload Image for Classification

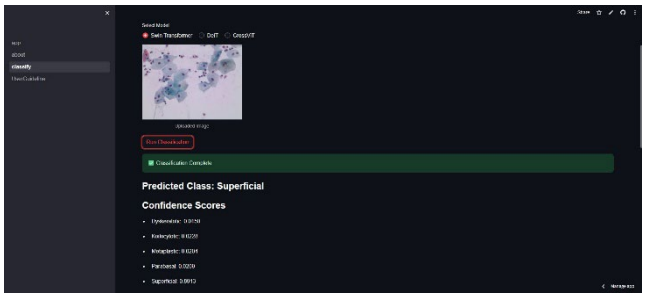


Figure 9. Result Page

4. Results and Discussion

4.1 Swin Transformer

The Swin Transformer (swin_base_patch4_window7_224) achieved an overall accuracy of 93.83% and a weighted average F1-score of 0.9375 on the test dataset. Table 3 displays the detailed class-wise metrics.

Table 3: Classification Performance of Swin Transformer Model

Class Index	Cell Class Name	Precision	Recall	F1-Score	Support
0	Dyskeratotic	0.9074	0.9453	0.9260	311
1	Koilocytotic	0.9388	0.8182	0.8744	319
2	Metaplastic	0.8940	0.9750	0.9327	320
3	Parabasal	1.0000	0.9703	0.9849	269
4	Superficial-Intermediate	0.9695	0.9931	0.9811	288
Macro Avg	*Unweighted Mean*	0.9419	0.9404	0.9398	1,507
Weighted Avg	*Sample Weighted Mean*	0.9396	0.9383	0.9375	1,507
Overall Accuracy	93.83%				1,507

As shown in Table 3, Swin Transformer achieved perfect precision (1.0000) for Parabasal cells and high recall (0.9931) for Superficial-Intermediate cells. However, recall for Koilocytotic cells was lower (0.8182), as 30 koilocytotic samples were misclassified as Metaplastic and 23 as Dyskeratotic due to overlapping nuclear halo boundaries. This can be further backed by the training and validation graphs as indicated in Figure 10.

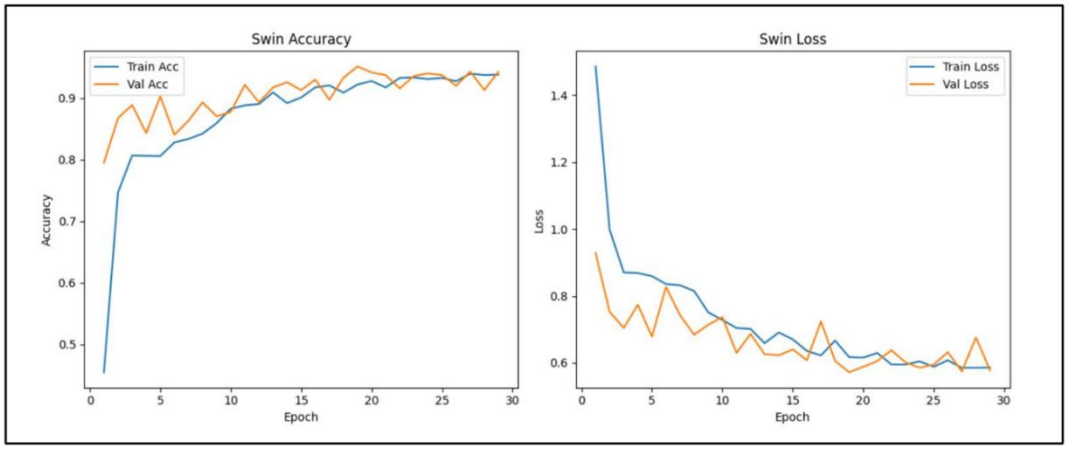


Figure 10. The Graphs of Training and Validation Accuracy (Left) and Losses (Right)

The graph on left is the graph of training accuracy and validation accuracy. It displays a general trend of the increase with 30 epochs due to its early stop of 50 epochs. First, the validation accuracy is getting increased at a faster rate than the training accuracy thus there is good generalization. An epoch of 10 is where the two curves start coming toward each other, and now the validation accuracy becomes constant at approximately between 0.93 and 0.95. The model does not demonstrate major declines and variability, which indicates that it is learning successfully and generalizing well on the upcoming information. The curves increasing steadily and coming closer virtually prove that the model does not overfit and is strong in its classification performance. As for the loss graph for both train loss and validation loss, indicate a significant reduction during the early epochs, particularly in the first five, where the model quickly learns core features. Training loss continues to decline gradually and stabilizes around 0.6, while the validation loss follows a similar pattern with slight fluctuations. These small oscillations in validation loss are normal and expected in deep learning training. The lack of sudden spikes or divergence between the two curves shows that the model maintains a stable learning rate and avoids overfitting. The overall decline in both losses confirms the model's effective optimization throughout the training process.

The model’s performance in classifying each class can also being shown by using confusion matrix in Figure 11.

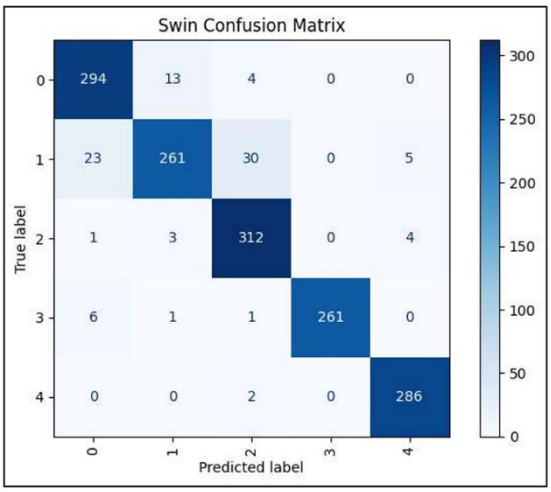


Figure 11. Confusion Matrix of Swin Transformer Model

In general, the Swin Transformer model demonstrates good classification stability, the most predictions are in the right place in the confusion matrix along the diagonal. This justifies its very high accuracy of 93.83%.

In case of Dyskeratotic cells (Class 0), 294 samples out of 311 were accurately predicted using the model. The under-represented categories were 13 cells of Koilocytotic, 4 cells of Metaplastic, and none in any other category. This implies a high probability of true detection with minimal confusion.

In the case of the Koilocytotic cells (Class 1), 261 samples were accurately identified by the model out of 319 samples. The total number of misclassified samples was 30 cells as Metaplastic and 23 cells as Dyskeratotic with a significant overlap with visually similar abnormal classes.

In the case of Metaplastic cells (Class 2), 312 of 320 were correctly forecasted. Some were incorrectly dictated as Dyskeratotic (1 cells), Koilocytotic (3 cells), and Superficial-Intermediate (4 cells), pointing to a relatively high degree of success in detection.

In the case of Parabasal cells (Class 3) the model predicted 261 correct out of 269. Misclassifications were very minimal with 6 cells being misclassified into Dyskeratotic, and one each into Koilocytotic and Metaplastic.

In case of Superficial-Intermediate cells (Class 4), 286 of 288 samples were correctly classified. 2 cells were misclassified as Metaplastic, which indicates a very accurate classification of normal cells.

4.2 Data-efficient Image Transformer (DeiT)

The DeiT model (deit_base_distilled_patch16_224) achieved the highest overall performance, reaching an accuracy of 95.22% and a weighted average F1-score of 0.9522. Table 4 details the classification report for DeiT.

Table 4: Classification Performance of DeiT Model

Class Index	Cell Class Name	Precision	Recall	F1-Score	Support
0	Dyskeratotic	0.9304	0.9453	0.9378	311
1	Koilocytotic	0.9206	0.9091	0.9148	319
2	Metaplastic	0.9528	0.9469	0.9498	320
3	Parabasal	0.9851	0.9851	0.9851	269
4	Superficial-Intermediate	0.9792	0.9826	0.9809	288
Macro Avg	*Unweighted Mean*	0.9536	0.9538	0.9537	1,507
Weighted Avg	*Sample Weighted Mean*	0.9522	0.9522	0.9522	1,507
Overall Accuracy	95.22%				1,507

DeiT demonstrated improved balance across all classes. Notably, Koilocytotic recall increased to 0.9091 (F1-score 0.9148), representing a 9.09% improvement over Swin Transformer. Precision for Metaplastic cells reached 0.9528, demonstrating effective discrimination between normal and benign cellular variations. This can be further backed by the training and validation graphs as indicated in Figure 12.

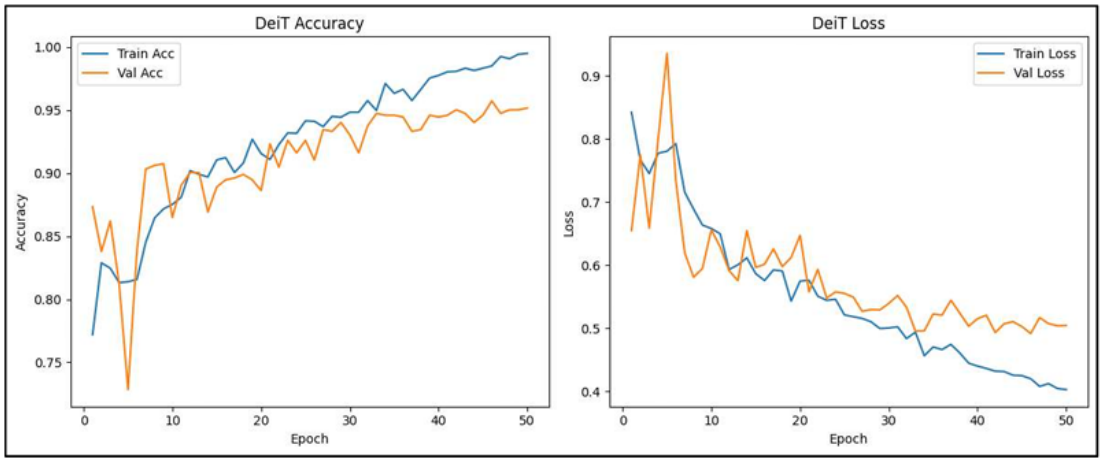


Figure 12. The Graphs of Training and Validation Accuracy (Left) and Losses (Right)

In the DeiT model, both training accuracy and validation accuracy graphs are observed to be rising steadily over 50 iterations. There is an initial low performance, but the accuracy increases continuously, particularly beyond 10 epochs in the model. After epoch 30, the accuracy of validation remains around 95 percent, whereas the training accuracy keeps increasing, ultimately reaching 77 or nearly reaching 99 percent. It means that the model learns effectively and generalizes well, although the fact that the accuracy on training and validation data is slightly different means that some overfitting occurs in later stages. Meanwhile, both the train loss and validation loss curves exhibit an obvious negative trend where the training loss continues to drop lower after the first epoch at 0.85 to less than 0.4 by the 50 epoch. The trend of validation loss is slightly more volatile, but again, will tend to converge towards 0.50 near the end. Stable learning is depicted by the meeting of both curves and the decrease of the curves. The greater error of validation loss in the initial stages is expected and represents sensitivity to the variance of batches that smooths out as the training proceeds. The model's performance in classifying each class can also being shown by using confusion matrix in Figure 13.

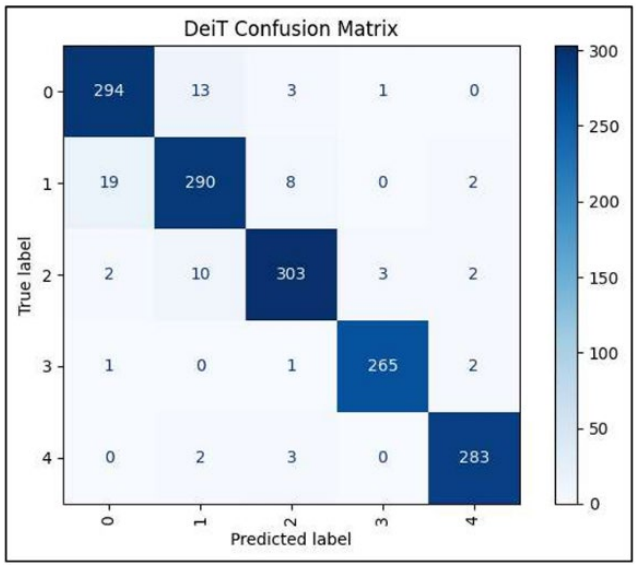


Figure 13. Confusion Matrix of DeiT Model

The confusion matrix of DeiT model in general demonstrates the high level of classification accuracy regarding all five classes of cervical cells where most of the predictions lie on the diagonal. This judges the overall accuracy of the model to be high (95.22%).

In Dyskeratotic cells (Class 0) the model could correctly predict 294 out of 311 cases. There were only 13 cells misclassified as Koilocytotic and some others as Metaplastic and Parabasal. This is evidence of great precision and recall, where results are regularly comparative and similar identifications are recurrent. In relation to Koilocytotic cells (Class 1), using the model gave 290 correct predictions against 319. Misclassifications had 19 cells being Dyskeratotic, and 8 cells being Metaplastic. This implies that the Koilocytotic cells might possess visual similarities with the other abnormal cells with a mild illustration on the sensitivity.

In the case of Metaplastic cells (Class 2), there was a better classification of 303 samples out of the total 320 samples. Few of them were mixed in Koilocytotic, Parabasal, and Superficial-Intermediate classes, which means an effective management of slight variances in benign cells. In Parabasal cells (Class 3) the number of the correctly classified sample was 265 out of 269. A few were misclassified, which portrays a great performance as well as reliability of these classes. In Superficial-Intermediate cells (Class 4), 283 of 288 cases were accurately modelled. This is a very high figure of generalization and almost complete identification of normal cells.

4.3 Cross-Attention Multi-Scale Vision Transformer (CrossViT)

CrossViT (crossvit_9_240) achieved the second-highest accuracy at 94.82%, with a macro average F1-score of 0.9499 and the lowest overall validation loss (0.5022). Table 5 presents the classification performance for CrossViT.

Table 5. Classification Performance of CrossViT Model

Class Index	Cell Class Name	Precision	Recall	F1-Score	Support
0	Dyskeratotic	0.9379	0.9228	0.9303	311
1	Koilocytotic	0.9057	0.9028	0.9042	319
2	Metaplastic	0.9439	0.9469	0.9454	320
3	Parabasal	0.9851	0.9851	0.9851	269
4	Superficial-Intermediate	0.9761	0.9931	0.9845	288
Macro Avg	*Unweighted Mean*	0.9497	0.9501	0.9499	1,507
Weighted Avg	*Sample Weighted Mean*	0.9481	0.9482	0.9481	1,507
Overall Accuracy	94.82%				1,507

CrossViT achieved high performance across both normal classes (Parabasal F1-score 0.9851; Superficial-Intermediate F1-score 0.9845) while maintaining high precision for Dyskeratotic cells (0.9379). This can be further backed by the training and validation graphs as indicated in Figure 14.

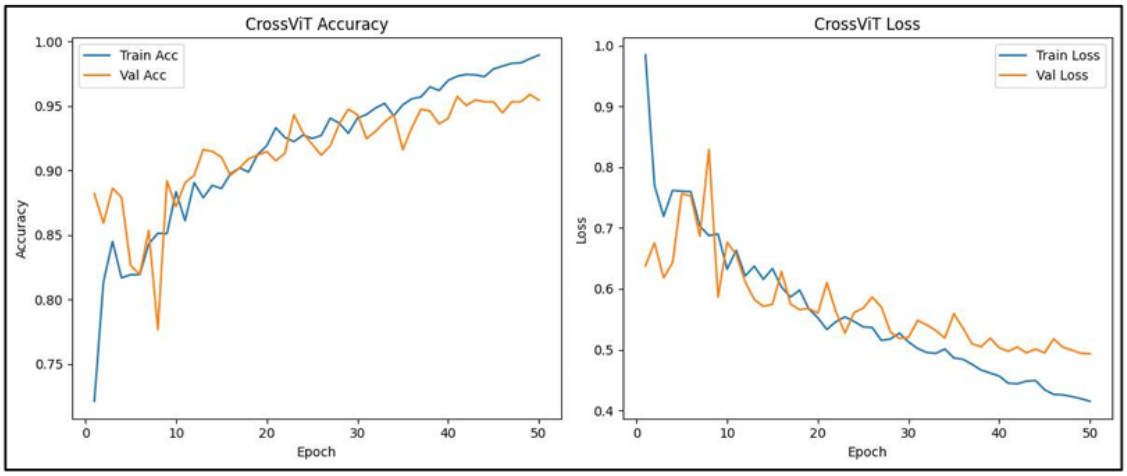


Figure 14. The Graphs of Training and Validation Accuracy (Left) and Losses (Right)

The CrossViT accuracy graph proves that there was no learning degradation during the 50 epochs of learning. The accuracy of training is roughly 73 percent and rises with time to above 98 percent in the last epoch. The validation accuracy also rises at a similar rate but reaches a value of about 95, after the epoch 30. Training vs. validation accuracy has a small spread, which means that there are strong generalizations and little overfitting. The fact that the model can hold the two composite curves very close to each other means that the model is not just memorizing the training data but is learning some meaningful patterns. Meanwhile, loss curves present a decisive downward tendency of both training set and validation set. The training loss gradually drops, starting with about 0.95 down to less than 0.42 after the 50 epoch. The loss in validation is again on the same route with some initial oscillations probably because of change in the performance of the batches but then gradually stabilizing around 0.5. The well converged training and validation loss further demonstrates that optimization happened stably and successfully, and there is no evidence of underfitting and overfitting. Such loss dynamics strengthen the fact of a significant level of accuracy and proves that there is an optimal adjustment of CrossViT to the problem of cervical cell classification. The model's performance in classifying each class can also being shown by using confusion matrix in Figure 15.

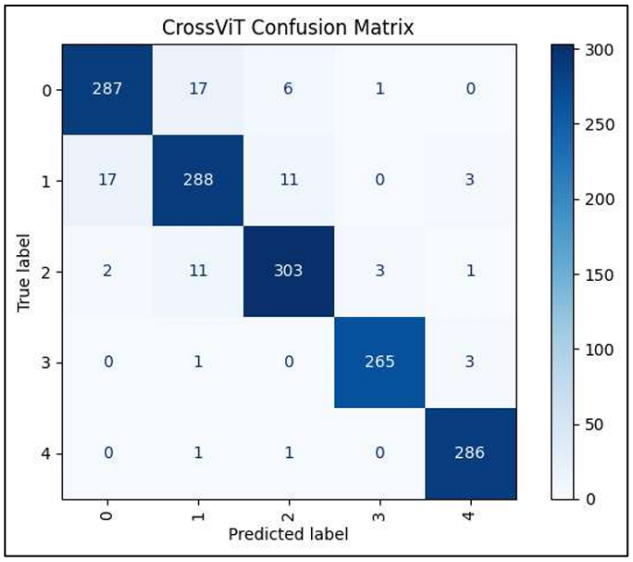


Figure 15. Confusion Matrix of CrossViT Model

In Dyskeratotic cells (Class 0) the model was able to correctly predict 287 and 311 cells in total. Very few were misdiagnosed as Koilocytotic (17 cells), Metaplastic (6 cells) and Parabasal (1 cells) and there was clear performance with slight confusions.

In Koilocytotic cells (Class 1), there were correctly identified 288 of 319 samples. Others were either wrongly classified as Dyskeratotic (17 class), Metaplastic (11 cells) or Superficial-Intermediate (3 cells) and it means that the class is still to some extent difficult because of the visual similarities. In the case of the Metaplastic cells (Class 2), the model got 303 correct predictions against 320. None of the misclassifications of this benign, high precision and recall were associated with Koilocytotic (11 cells), Parabasal (3 cells), and Superficial-Intermediate (1 cells).

In the case of the Parabasal cells (Class 3), 265 correct classes were predicted in the 269 samples. This class has a very good model performance with only a few being misclassified as Koilocytotic (1 cells) or Superficial-Intermediate (3 cells). The most populous cell, the Superficial-Intermediate (Class 4) had 286 of the 288 samples placed correctly. The other two went erroneously as Koilocytotic (1 cells) and Metaplastic (1 cells), with an extremely high sensitivity of detecting the remaining normal cells.

4.4 Discussion

Table 6 provides a comparative summary of all three Vision Transformer models across validation loss, test accuracy, precision, recall, and F1-score metrics.

Table 6. Overall Performance Comparison Across Vision Transformer Architectures

Model Architecture	Overall Accuracy (%)	Validation Loss	Macro Precision	Macro Recall	Macro F1-Score	Weighted F1-Score
Swin Transformer	93.83%	0.5903	0.9419	0.9404	0.9398	0.9375
DeiT (Distilled)	95.22%	0.5130	0.9536	0.9538	0.9537	0.9522
CrossViT (Multi-Scale)	94.82%	0.5022	0.9497	0.9501	0.9499	0.9481

DeiT achieved the highest accuracy (95.22%). Its distillation effectively incorporates pre-trained knowledge from a teacher network, compensating for the moderate sample size of the SIPaKMeD dataset. Global self-attention across standard 16 x 16 patches allows DeiT to evaluate relationships between nuclear texture and cytoplasmic region across the entire cell.

CrossViT achieved the lowest validation loss (0.5022) and a competitive accuracy of 94.82%. Its dual-branch architecture extracts fine-grained nuclear details via small patches while capturing cell contours via large patches. The cross-attention module integrates these representations, improving classification consistency.

Swin Transformer achieved 93.83% accuracy. While shifted local window attention reduces computational cost, restricting early self-attention to isolated 7 x 7 windows slightly reduced sensitivity for subtle koilocytotic features.

Training and validation curves indicated that while all models converged stably, mild overfitting emerged beyond 25–30 epochs, where training accuracy approached ~98–99% while validation accuracy stabilized around 95%. Early stopping with a patience of 10 epochs successfully prevented overtraining, saving the optimal model checkpoint based on minimum validation loss.

When deployed on the Streamlit platform, the end-to-end single-image processing pipeline includes image preprocessing, model forward propagation, and softmax-based confidence scoring demonstrated near real-time performance. Using GPU acceleration with an NVIDIA RTX 3060 and CUDA, all three ViT models processed an image in approximately 35–45 ms, equivalent to around 22–28 frames per second (FPS). In CPU mode using an Intel Core i5-12500H processor, processing required approximately 120–180 ms per image, corresponding to 5.5–8.3 FPS. These empirical latency results indicate that the proposed system can provide responsive interactive decision support for web-based clinical use.

5. Conclusion and Future Work

This research demonstrated the application of Vision Transformers for automated cervical cancer cytology image classification on the SIPaKMeD dataset. By fine-tuning three ViT paradigms, the models achieved classification accuracies exceeding 93.8%. The Data-efficient Image Transformer (DeiT) delivered the highest overall accuracy (95.22%) and weighted F1-score (0.9522), successfully resolving visual ambiguities between dyskeratotic, koilocytotic, and metaplastic cells. Deployment via a Streamlit web application demonstrated rapid single-cell classification with an inference latency of ~38 ms on GPU, supporting real-time interactive decision support.

The comparative evaluation of the three Vision Transformer variants in this study resonates with numerical modeling efforts in oncology, where different computational

approaches are analyzed for stability and accuracy in modeling cancer progression (Iyanda *et al*., 2021). These findings confirm the promise of Transformer-based methods to outperform more traditional CNN models when applied to medical image classification, notably in learning global and local dependencies in cytology samples.

Although encouraging findings were obtained, certain weaknesses are noted in this work. The SIPaKMeD dataset, although widely used, is rather small and homogeneous and may thus limit the applicability of the models to real-life clinical settings. Besides, the class imbalance, especially in minority groups like koilocytotic cells, is still a problem which is likely to affect sensitivity and leads to more likely misclassification risk. Vision Transformers also have high computation demands, which constrains their scalability in resource-limited environments.

To extend these findings, future work should focus on rigorous benchmarking, model optimization, and deployment safety. First, incorporating standard CNN baselines (e.g., ResNet-50, EfficientNet-B0) on the identical SIPaKMeD split alongside 5- or 10-fold cross-validation will establish robust empirical comparisons and statistical confidence intervals. Second, to mitigate the training-validation gap observed past 25 epochs, future iterations should explore advanced data regularizations (e.g., Mixup, CutMix) paired with Cosine Annealing learning rate schedules. Finally, to enhance real-world generalization, future studies should evaluate performance on multi-center clinical datasets (e.g., Herlev, LBC), leverage GAN-based synthetic data generation for rare cell pathologies and integrate energy-based out-of-distribution detection modules to flag non-cervical images prior to inference.

Acknowledgement

The authors acknowledge Universiti Teknologi MARA for the support of academic research activities.

Funding

This research received funding from Universiti Teknologi MARA under grant number 600-RMC 5/3/GPM (021/2023).

Author Contribution

Muhammad Daniel Alif Jumairi was responsible for the data curation, experimentation, visualization, and writing the draft of the manuscript. Itaza Afiani Mohtar was responsible for the conceptualization, methodology, and supervision of the project. Suraya Masrom was responsible for validation, formal analysis, and quality assurance. Masurah Mohamad was responsible for manuscript formatting and editing. Lathifah Alfat was responsible for proofreading.

Declaration of generative AI in the writing process

Generative AI was used solely for language refinement and grammar checking to improve the clarity, readability, and overall quality of the manuscript. All research content, analysis, interpretations, and conclusions were developed and verified by the authors.

Data Availability/Supplementary Materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Statement

The authors declare that this research did not involve human or animal subjects. All experimental procedures were performed following the institutional Safety, Health, and Environmental (HSE) protocols of Universiti Teknologi MARA.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Alquran, H., Alsalatie, M., Mustafa, W. A., Abdi, R. A., & Ismail, A. R. (2022). Cervical Net: A novel cervical cancer classification using feature fusion. *Bioengineering*, *9*(10), 578. https://doi.org/10.3390/bioengineering9100578

Ali, A. M., Benjdira, B., Koubaa, A., El-Shafai, W., Khan, Z., & Boulila, W. (2023). Vision transformers in image restoration: A survey. *Sensors*, *23*(5), 2385. https://doi.org/10.3390/s23052385

Bao, S., Ding, H., & Chen, H. (2023). A review of the development of Vision Transformer. In *2023 International Conference on Artificial Intelligence and Automation Control (AIAC)* (pp. 352-356). IEEE. https://doi.org/10.1109/AIAC61660.2023.00008

Chen, C.-F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale Vision Transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 357-366). https://openaccess.thecvf.com/content/ICCV2021/html/Chen_CrossViT_Cross-Attention_Multi-Scale_Vision_Transformer_for_Image_Classification_ICCV_2021_paper.html

Dasari, S., Wudayagiri, R., & Valluru, L. (2015). Cervical cancer: Biomarkers for diagnosis and treatment. *Clinica Chimica Acta*, *445*, 7-11. https://doi.org/10.1016/j.cca.2015.03.005

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/

Ferri, F. F. (2025). *Ferri's clinical advisor 2025*. Elsevier.

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), 468. https://doi.org/10.3390/bioengineering11050468

Iyanda, F. K., Tunde, T. A., & Isa, U. (2021). Numerical comparison of Runge-Kutta (RK5) and New Iterative Method (NIM) for solving metastatic cancer model. *Malaysian Journal of Computing*, *6*(1), 758-771. https://doi.org/10.24191/mjoc.v6i1.10359

Khowaja, A., Zou, B., & Kui, X. (2024). Enhancing cervical cancer diagnosis: Integrated attention-transformer system with weakly supervised learning. *Image and Vision Computing*, *149*, 105193. https://doi.org/10.1016/j.imavis.2024.105193

K, A., & B, S. (2024). A deep learning-based approach for cervical cancer classification using 3D CNN and Vision Transformer. *Journal of Imaging Informatics in Medicine*, *37*(1), 280-296. https://doi.org/10.1007/s10278-023-00911-z

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012-10022). IEEE. https://doi.org/10.1109/ICCV48922.2021.00986

Mansur Huang, N. S., Ibrahim, Z., & Mat Diah, N. (2021). Machine learning techniques for early heart failure prediction. *Malaysian Journal of Computing*, *6*(2), 872-884. https://doi.org/10.24191/mjoc.v6i2.13708

Park, C.-H. (2008). False-negative results of conventional Papanicolaou cervical cytology in women with cervical conization. *Korean Journal of Gynecologic Oncology*, *19*(1), 40-46. https://doi.org/10.3802/kjgo.2008.19.1.40

Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., & Charchanti, A. (2018). SIPaKMeD: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. In 2*018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3144-3148). IEEE. https://doi.org/10.1109/ICIP.2018.8451588

Rahmat, T., Ismail, A., & Aliman, S. (2019). Chest X-ray image classification using Faster R-CNN. *Malaysian Journal of Computing*, *4*(1), 225-236. https://doi.org/10.24191/mjoc.v4i1.6095

Swanson, A. A., & Pantanowitz, L. (2024). The evolution of cervical cancer screening. J*ournal of the American Society of Cytopathology*, *13*(1), 10-15. https://doi.org/10.1016/j.jasc.2023.09.007

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jegou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the 38th International Conference on Machine Learning*, 139, 10347-10357. https://proceedings.mlr.press/v139/touvron21a.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998-6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html