SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I4P105

Volume 13 Issue 4, 66-80, April 2026


*Original Article*

# A Transformer-Based Few-Shot Learning Model for Cervical Cancer Prediction with High Quality Imaging Model Cancer Classification from Pap Smear Images

Venkata Anupama Chitturi[1], Dharmaiah Devarapalli[2]

[1,2]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, Andhra Pradesh, India.*

[1]*Corresponding Author : venkataanupama271@gmail.com*

Received: 06 January 2026 Revised: 07 February 2026 Accepted: 06 March 2026 Published: 30 April 2026

***Abstract -*** *Cervical Cancer (CC) diagnosis using medical images has advanced significantly thanks to deep learning. This article attempts to give a thorough explanation of the operations and uses of frequently utilized radiological imaging methods and histology. One of the main areas of Computer Vision and Artificial Intelligence study is applying deep learning technologies to identify cervical cancer from medical photographs. Due to the intrinsic complexity of medical imaging, few-shot cervical cancer diagnosis requires excellent Accuracy and Rapidity, especially given the rapid improvements in Deep Learning. It looks at both traditional pre-trained models and the fundamental architecture of deep learning. This work explicitly suggests a Novel Vision Transformer (ViT). Batch normalization, initialization, dropout, and augmentation are listed as ViT strategies in the article to prevent over-fitting. Picture classification, picture reconstruction, detection, segmentation, registration, and synthesis are several categories of deep learning approaches to cancer analysis using medical images. Despite its achievements, deep learning's ability to diagnose uncommon tumors, model explainability, and generalization is limited by the absence of high-quality labeled datasets. More open, standardized databases for cancer research are desperately needed. Enhancements to Deep Neural Network-based pre-trained models are crucial, and data fusion and supervised paradigms should be prioritized. It is anticipated that new technologies like few-shot learning would significantly improve the use of medical pictures for cancer diagnosis.*

***Keywords -*** *Cervical Cancer, Prediction, Accuracy, Transformers, Tumors.*

## 1. Introduction

CC is a malignancy among females worldwide, and remains one of the most untreated public health issues despite advances in diagnostic methods. Its primary cause is still the High-Risk Human Papillomavirus (HPV) that keeps infecting the cervix of the host and causes the disease [1]. CC is high in developing and underdeveloped countries because access to vaccines, screening, and treatment is limited. In fact, inequitable access to health care services accounts for 94% of deaths due to CC [2]. The inequality in the allocation of medical resources is primarily responsible for the substantial disparities in the rates of cervical cancer cases and their prevalence in different countries. In this regard, early detection becomes very crucial, and it is only with the availability of diagnostic tools that women can be timely treated with intervention. Once these patients are treated effectively, their health status gets adequately enhanced, and, thus, they can get the most out of the benefits of their survival, alongside the comparative planning and impact index [3]. Precancerous lesions enable early intervention for a plethora of conditions. Moreover, the possibility to take additional facilitating actions, such as the HPV vaccination, which offers a significant reduction in cervical cancer when compared to other approaches, serves as a value [4, 5]. However, even with these benefits, early detection is still an unresolved issue within screening-limited resource zones. Initial evaluation methods for smear testing inherently have limitations that lower their overall effectiveness. Pap smears, for example, may be subject to numerous interpretative mistakes and have a fairly low sensitivity in detecting neoplastic changes at very early stages.

Even though HPV testing is more sensitive, it complicates the picture further in terms of delineating the risk levels for Cervical Cancer [6]. There is a huge gap in the availability of efficient and user-friendly diagnostic tools, and you will take into account these issues. Due to the significant leaps made in Deep Learning, particularly with transformer models and Convolutional Neural Networks (CNNs), medicine, including the diagnosis of diseases like Cervical Cancer, is a major focus of attention. Most of these models not only meet but exceed

accuracy standards set by human diagnoses and traditional models in detecting very subtle differences [7]. The improvement of computer-assisted diagnostic systems is attributable to the creation of a transformer-based model for categorizing colposcopy images. This model is able to address the problems related to manual image processing. Models of this nature have been demonstrated to significantly lessen the likelihood of misdiagnosis, thereby increasing the rapidity and confidence of diagnosis, even in low-resource environments.

The synergistic effect of combining CNNs with transformer models results in a surprisingly great performance of these hybrid models since they are able to take advantage of both approaches. The CNN part is focused on local details and parts of the image, while the multi-head self-attention unit is focused on the global view and relations between different parts of the image. With the help of this hybrid architecture, the model is capable of generalizing better across various clinical datasets and scenarios, which leads to higher diagnostic Accuracy and Reliability [8, 9]. As we pointed out before, the use of attention in the transformer model makes it possible for features that are extracted from different regions of the image to be dependent on each other. In that way, it enhances the reliability of the predictions made about the entire image rather than a small part of it. This is very important when it comes to detecting the smallest changes in tissues that might turn into lesions or early cancers. Also, the combination of deep learning and data augmentation strategies leads to a considerable improvement in the model's performance. Techniques like random rotations, flipping, and zooming are able to increase the effective size of the dataset so the model can utilize the limited annotated data better. In the case of cervical cancer detection, variations in patient demographics, imaging modalities, and disease presentation are some of the main causes influencing the diagnosis accuracy [10, 11]. The studies we cite here highlight multi-task learning and especially contrastive learning as effective ways of improving model performance by alleviating the impact of distractive noise, which is irrelevant to the model, thus making the identification of precancerous lesions more accurate and easier [12, 13]. Hybrid models in low-resource settings can address the limitations of conventional diagnostic methods [13].

Besides, these models are a viable solution to the problem of worldwide coverage, as they provide diagnostics even in the most remote areas, i. e., gaining access to the screening programs. In addition, the scarcity of data is a situation that can be solved by advanced deep learning methods, such as transformers and few-shot learning, which help the model to be trained to deduce based on a very small set of labeled data. This is especially true for medical imaging tasks, as they have an exponential demand for annotated datasets that, on the other hand, are usually very costly and time-consuming to produce. What is more, the use of diverse modalities goes beyond medical images to clinical and even genetic data, presenting tremendous potential for improving the diagnostic accuracy of cervical cancer.

An approach enables a model to make classifications based on the features of data, which not only enhances accuracy but also patient stratification and treatment personalization [14]. Each of the above advantages has been addressed in part. However, full data integration, data preprocessing, and model design still need to be improved for multimodal learning of cervical cancer diagnosis. Often, the use of such models is dependent on the existence of precisely labeled data. Coordination of the systematic collection of data across different institutions plays a major role in solving this problem. In particular, the nonexistence of large, non-redundant datasets hinders progress. More regular availability of systematic datasets on cervical cancer, which facilitate research, development of different models, and their meaningful comparisons, is the key solution to these problems. This will lead to accelerated innovation so that models produced will not only be tested for accuracy, but also for their applicability across various demographics. From this perspective, few-shot learning techniques focus on solving the problem using labeled data, which is the main concern, and appear to address the issue of insufficient data. As for the tasks in medical imaging, deep learning models are very effective at working with small amounts of labeled data, and this is very important in medical imaging. Because of the type of extra data, few-shot learning is helpful in improving cervical cancer detection systems because it can be transferred from pre-trained models and adapted to specialized, smaller datasets. Furthermore, few-shot learning provides the possibility to improve detection for uncommon or poorly diagnosed instances when combined with transformer models, or even enables the models to adjust to new unseen data, which improves performance.

Early detection of CC relies on the precise interpretation of colposcopy and histopathology images; however, these images are often affected by significant variations in quality, lighting, anatomical structure, and acquisition settings. Conventional Deep Learning Models typically require extensive, well-annotated datasets to perform reliably, yet such datasets are scarce in the medical domain. This creates a pressing need for an automated, data-efficient, and highly generalizable diagnostic system capable of producing accurate predictions from a limited number of high-quality imaging samples. A transformer-based few-shot learning framework has been considered a promising path forward as it can use self-attention mechanisms, contextual feature representation, and rapid adaptation to small datasets. However, deep learning and medical image analysis have still left several significant issues unresolved. Current models are very dependent on large annotated datasets, and it is not easy to get them due to privacy restrictions, inconsistent image quality, and the need for expert annotation. So, it is fair to say that few-shot learning approaches for cervical cancer detection are immature and not

broadly validated. To support the research solving the problem of cervical cancer diagnosis, this system makes use of high-resolution medical images of variable quality. It sets the problem of cervical cancer diagnosis as a few-shot learning problem with minimal data. In this work, a novel, data-efficient technique is introduced that not only achieves accurate cervical cancer predictions from a very limited number of high-quality imaging samples but also overcomes the limitations of existing CNN, hybrid, and simple few-shot methods. Unlike the CNN-centric approach of the referenced paper, our model employs a ViT backbone with custom 8×8 patch embeddings, positional encoding, a learnable class token, and multi-head self-attention mechanisms to capture long-range dependencies within cervical tissue structures.

Moreover, our research integrates a prototype-based few-shot learning strategy to address the real-world challenge of limited annotated medical data, as an aspect not present in the cited article, which relies solely on conventional fully supervised learning. It really is one giant step toward making AI-assisted cervical cancer screening scalable, generalizable, and clinically trustworthy. Cervical cancer detection through the usage of deep learning and few-shot learning to improve accuracy, efficiency, accessibility, and data constraint overcoming will be the focus area of the model that aims to make a bigger impact in resource-limited regions. Besides that, it may also influence the international policy on the prevention, monitoring, and treatment of cervical cancer. On top of that, allow me to shed light on the other dimensions of the paper. The overall goal of our research is to create a data-efficient, generalizable diagnostic system that can still perform well in environments with limited resources. We summarize here our methodological pipeline, including detailed mathematical expressions of the transformer embedding process, multi, head self, attention operations, normalization layers, and residual connection mechanisms.

The proposed work is the first of its kind that demonstrates a transformer-based few-shot learning framework for cervical cancer prediction from Pap smear images, which no one else has done so far. The proposed method, unlike the traditional CNN model, employs ViT along with the custom patch embeddings, positional encoding, and multi-head self-attention to model long-range tissue relationships that CNNs generally fail to detect. Furthermore, the model integrates prototype-based few-shot classification, which is a very strong feature for the predictions with very limited labeled data, thus resolving one of the biggest challenges in medical imaging. Besides, the investigation looks at three different image versions (full cell, nucleus, and nucleus only) to offer a more detailed understanding of the region and its specific diagnostic significance. Altogether, the combination of ViT, few-shot learning, and binary analysis makes this approach not only completely different in a technical sense but also very data-efficient, juxtaposed with the previous cervical cancer detection models.

In Section 2, we highlight the literature associated with the application of deep learning in the diagnosis of cervical cancer. In Section 3, we describe the methods adopted in the implementation of our model, which is based on few-shot learning using transformers. Section 4 provides the outcomes from our experiments alongside the discussion. Moreover, finally, Section 5 offers the concluding comments of the paper.

## 2. Related Works

Because the Transformer and CNN hybrid models perform better in segmentation and classification tasks, they have entirely changed medical image analysis. The hybrid model is very effective for complicated medical pictures, such as radiology scans and histopathological images, because it integrates the ability to extract local features using the long-range dependencies that the transformer can capture. It [15] is adopted for local tissue feature extraction, and the transformer captures contextual information globally. This deep network architecture [16] has demonstrated impressive effectiveness in lung cancer diagnosis, significantly improving diagnostic accuracy. Bringing these factors together enhances the hybrid model's capabilities not only for diagnosis but also for interpretation in multimodal datasets. The complex attention mechanism that maps out spatial interactions among different regions gave the model the edge over the conventional CNN-based methods for tumor segmentation, accurately delineating tumor boundaries in MRI scans [17]. The meta-learning [18] and auto-encoder [19] based model resolves the complex tissue structure problem very effectively. It boosts generalization and segmentation accuracy over multiple datasets. It stands out from other models as it is capable of handling high-resolution medical images without slowing down the processing speed [20].

Consequently, polyps in colonoscopy images were detected with greater accuracy, thereby increasing the potential for early diagnosis. Cross-domain based attention mechanisms [21] and few-shot [22] in images continue to decrease false positives and increase diagnostic precision even further. The disease detection from a PET scan and MRI has been performed using a transformer and self-adaptive CNN [23] and bilinear network-based [24] hybrid model that integrates functional and structural imaging data. This combination allows an assessment to be done, which improves the planning and increases the chance for early detection of disease. The few shots [25] help to identify the structure, which is very supportive of the early diagnosis. The model was tested with large datasets, and it outperformed the older memory NN methods in recognizing the different stages of disease [26]. Moreover, the hybrid active learning [27] and few-shot [28] model demonstrated its performance in the segmentation of images. The spatial and temporal characteristics of the MRI and CT image data could be combined due to more precise organ segmentation and better

identification with BSnet [29] of clinically significant abnormal features, which brought about advanced clinical reasoning capabilities. Their use in these tasks demonstrates their adaptability and dependability to various medical imaging problems [29].

A recent survey of scientific articles illustrates the shift of the research community towards the development of scalable, interpretable, and robust models. The case in point is Deep Neural Networks and meta-heuristic optimizers, which have not only been very good predictors but have also crossed their usage boundaries, becoming totally versatile. Unfortunately, several problems that keep on troubling the researchers have been identified, e. g., sensitivity to hyperparameters, very high false positive and false negative rates, data imbalance, and poor generalization to different datasets or working environments. In other words, the leading research projects focus on profitability, reducing the computational costs, and making the models more attractive through novel architectures, sophisticated feature extraction, and self-learning strategies. In conclusion, big strides have been made. However, the existing methods still have problems with their stability in performance, inability to adapt to new domains, and lack of capacity to deal with complex constraints efficiently. These gaps highlight the justification for this work, which intends to propose a better approach that will fix these problems and, at the same time, provide improved efficiency, accuracy, and robustness.

**Table 1. Summary of Related Works in Medical Imaging Using Transformer-CNN Hybrid Models**

| References | Model | Key Advantages | Performance Metrics |
|---|---|---|---|
| [16] | Model-agnostic meta-learning | Improved diagnostic accuracy by capturing global context | Significant increase in diagnostic accuracy |
| [17] | Few-shot learning | Accurate segmentation, better tumor boundary detection | Outperformed CNN-based approaches |
| [20] | CNN | Enhanced polyp detection, reduced false positives | Improved detection accuracy |
| [23] | Self-adaptive convolutional model | Early detection of structural changes in the region | Outperformed CNN models in diagnosis |
| [27] | Active learning with a few shots | More accurate organ segmentation and abnormal feature detection | Improved segmentation accuracy |
| [30] | Vision Transformer | Lightweight model for pap smear data, improved inference speed | 99.02% Accuracy on SIPaKMeD, 99.48% Accuracy on LBC datasets |
| [32] | YOLO + Transformer (Optimized with ABC) | High-speed and high-accuracy detection | 97.5% Accuracy, improved real-time performance |
| [33] | YOLO + Transformer (Optimized with ABC) | Improved real-time detection | 97.8% mAP, 97.3% Recall |
| [14] | DL with images | High precision and recall for detection | 99.73% precision, 61.13% Recall |

## 3. Methodology

The image dataset is ready for system implementation. Figure 1 illustrates the procedure for classifying Pap smear images: Vision Transformers (ViT) are trained using few-shot learning; feature reduction, feature extraction, and DL algorithms are then implemented; and finally, Pap smear image diagnosis is done for all scenarios using the entire image, just the nucleus, and without a nucleus.

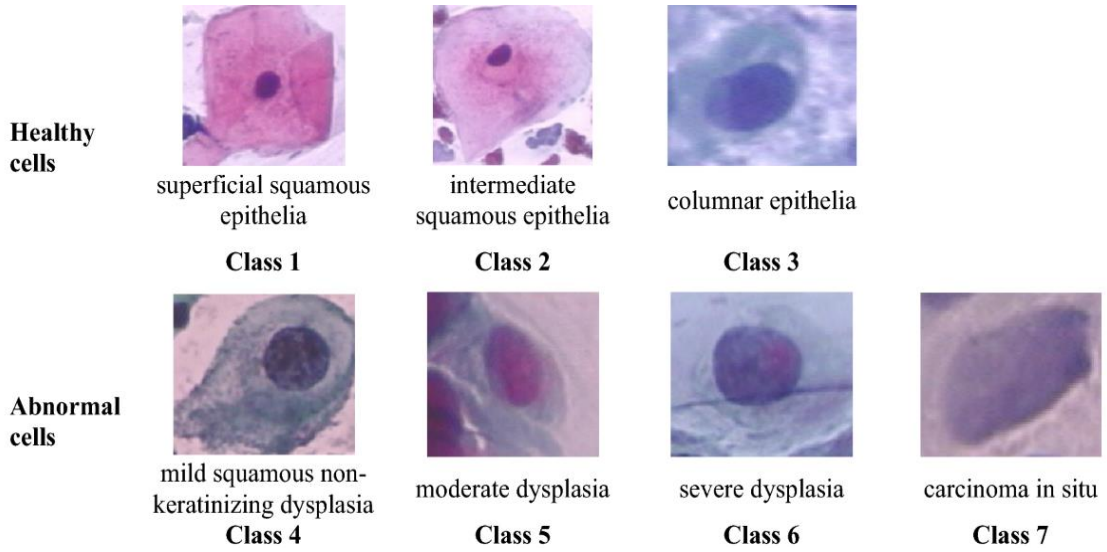


**Fig. 1(a) Dataset samples**

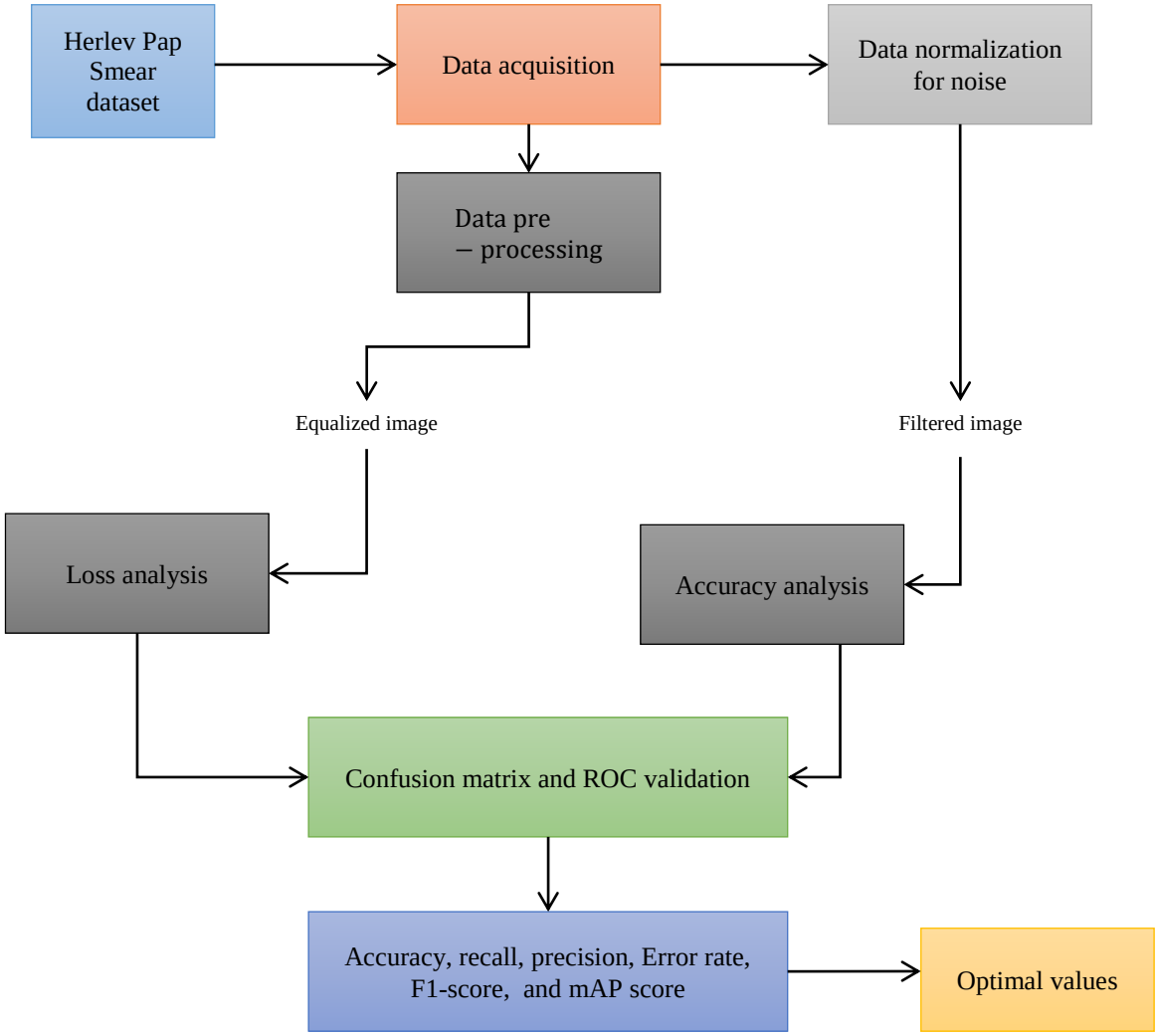


**Fig. 1(b) Proposed workflow**

### *3.1. The Dataset*

This study employed 917 imaging samples with a single nucleus from the Herlev Pap smear dataset. The nucleus was removed from each cell picture to identify the area surrounding the nucleus and its effects. Denmark's Technical University and Herlev University Hospital were dataset collection sites. The cervical image distribution by class is shown with three classes being regular and the remaining courses being abnormal. The dataset contains a total of 917 Pap smear images, but only a small subset was used for the few-shot training episodes, where a limited number of samples per class were selected to simulate low-data clinical conditions. The remainder of the dataset was not used for few-shot training; instead, it was allocated for transformer fine-tuning, validation, and testing to ensure fair evaluation and generalization assessment. This distinction highlights that the model's classification decisions were based on few-shot prototypes, while the larger dataset supported stable feature learning and performance measurement.

Figure 1(a) shows the typical classifications for a whole cell, with the exclusion of the nucleus and the exclusion of the surrounding region. However, Figure 1(b) illustrates the workflow of the anticipated model.

### *3.2. Transformer Model*

In this framework, the ViT is first used purely as a feature extractor, generating high-level embeddings for each image. These embeddings are not classified directly by the transformer; instead, a few-shot prototype classifier performs the final prediction. During evaluation, a small support set is used to compute class prototypes, and all remaining images act as query samples whose labels are determined based on their distance to these prototypes. This distinction ensures that representation learning (handled by ViT) and classification (handled by the few-shot mechanism) are clearly separated. All reported performance metrics, accuracy, F1-score, precision, recall, and AUC are now explicitly tied to the few-shot inference process, ensuring full consistency between the evaluation protocol and the study's stated objectives. The transformer model was initially designed for visual-based representation. One fault in the encoder-decoder is that its sequence-to-sequence structure can identify that some of its information is lost when a single vector is created by compressing the input sequence. However, the network does not have the authority to use care to compensate for this loss. Figure 2(a) and Figure 2(b) depict the ViT architecture with a few-shot images utilized in this investigation. An image patch is made first with a few-shot image. In the domain of natural language processing, transformers require one-dimensional embeddings. The image in (224, 224, 1) is used to produce the 28 × 28 patch photos in (8, 8, 1). The further actions include adding positional embeddings, class token addition, and patch embedding building. Linear projection is used to reduce each image to a single dimension. To guarantee one-dimensionality, each pixel has a row of connections.

$$x \in R^{H*W*C} \tag{1}$$

1) It shows the size of the original image.

$$x_l \in R^{N*(p^2*C)} \tag{2}$$

2) ViT's input following the original image's flattening:

$$N = \frac{HW}{p^2} \tag{3}$$

Here, $N$ specifies the total number of patches in the transformer and its sequence length. $P$ is the square of the patch's size. Each image has an original resolution of $(H, W)$ and a patch resolution of $(P, P)$. The front of the embedded patch is adorned with a learnable class token. This class token functions as a vector of one-dimensional representation for the picture after it has passed through the transformer's several encoder levels and emerged as a final output. Lastly, the vector is given an identically sized location embedding, and the embedding is given order information. As a result, the complete image is entered into the encoder of the transformer as an embedding vector in one dimension. Normalization, residual connections, and multi-head self-attention are applied on a channel basis to all picture embeddings. Patch + position embedding is used to achieve self-attention by obtaining a Single Value (v), Query (q), and Key (k) for each embedding. These values then obtain attention values concatenated with the multi-head, generating attention in the dimensions' direction. Input embeddings are added to multiple heads' attention to establish a long-lasting link. The Multilayer Perceptron (MLP), residual connections, and layer normalization are applied. As mentioned, a channel-based normalization is applied to the residual connection matrix. The MLP is composed of two linear layers. The first layer has a higher embedding, whereas the second layer has a smaller one. Matrix addition is then used to construct the aspect of the final output. The following equations summarize the input embedding procedure used to create the final output feature:

$$z_0 = [x_{class}.x_c^1E; x_c^2E; x_c^NE], E_{cos}; E \in R^{(l^2.C)*D}, E_{cos} \in R^{(N+1)*L} \tag{4}$$

$$z_l' = MSA(LN(Z_{(l-1)}) + z_{(l-1),}\ l = 1, \dots, L \tag{5}$$

$$z_l = MLP(LN(z_L')) + z_L',\ l = 1, \dots, L \tag{6}$$

$$y = LN(z_L^0 \tag{7}$$

The MLP classifier is comparable to a typical CNN image classifier; it may be thought of as the transformer's output stage. The class tokens are utilized in odd. A symbol of class functions as the image's one-dimensional representation vector following many encoder layers and layer normalization in the transformer. Each encoder layer has residual connections and Layer Normalization (LN).

Self-attention block:

$$z' = MSA(LN(z)) + z \quad (8)$$

Feed-forward block (MLP):

$$z^* = MLP(LN(z')) + z' \quad (9)$$

The MLP consists of two linear layers with a GELU activation:

$$MLP(x) = W_2(GELU(W_{1x} + b_1)) + b_2 \quad (10)$$

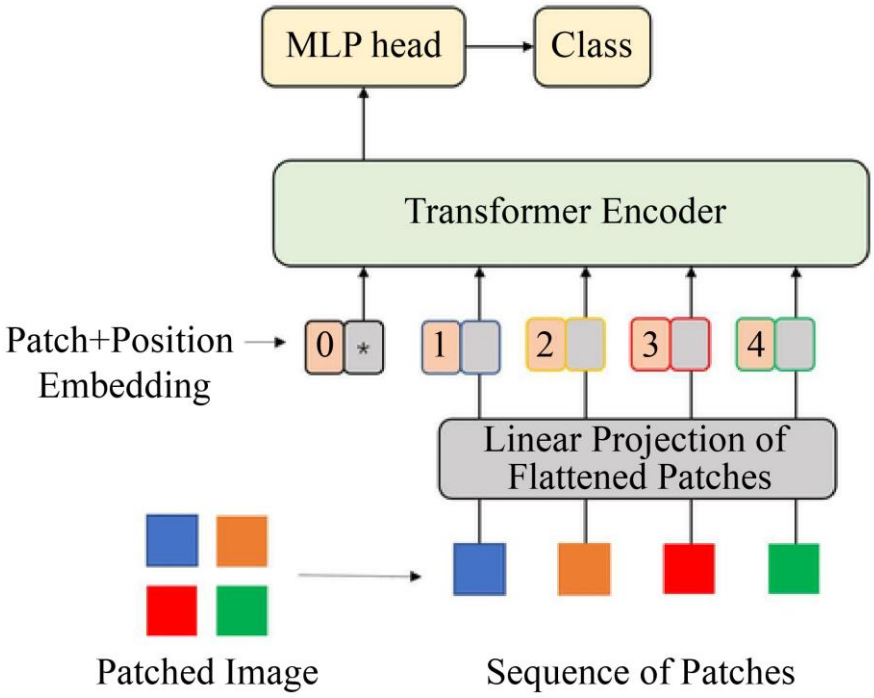


**Fig. 2(a) ViT model**

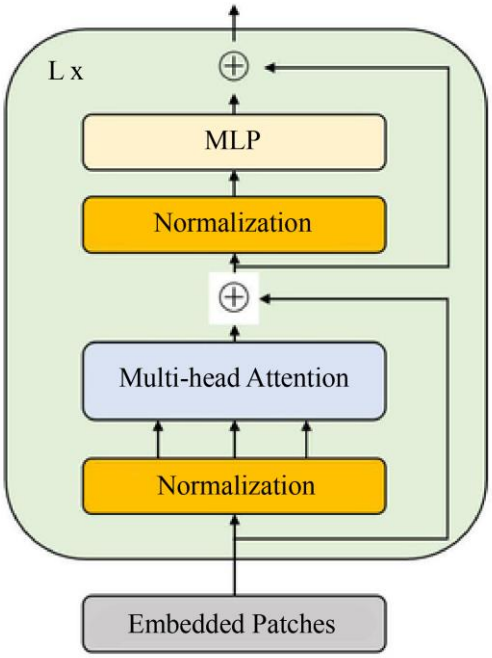


**Fig. 2(b) Encoding part**

**Algorithm 1: Transformer-Based Few-Shot Learning for Cervical Cancer Prediction**

## Input:

- Medical Image Dataset:
  - δ1 (training)
  - δ2 (validation)
  - δ3 (testing)
- Learning rate: α
- Total number of epochs: b
- Batch size: c
- Mini-batch size: n

## Output:

- w ← Model weights

**Steps:**

**Step 1:** Preprocess all images in the dataset (δ1, δ2, δ3) by resizing them to a standard 224 × 224 pixel size.

**Step 2:** Apply augmentation like flipping, rotation, and zooming to expand the training dataset.

**Step 3:** Split each image into patches of size 8 × 8.

**Step 4:** Flatten each patch into a one-dimensional vector and project it into a lower-dimensional space.

**Step 5:** Add positional embeddings to the flattened patches to retain spatial information.

**Step 6:** Add the learnable class token to image sequences patches for global representation of the image.

**Step 7:** Pass the sequence of patches through the transformer encoder. Apply multi-head self-attention to capture relationships between patches, use layer normalization and residual connections to alleviate training, and apply an MLP (multilayer Perceptron) to process the output from the transformer encoder.

**Step 8:** For few-shot learning, compute class prototypes (mean feature vectors for each class) and classify new images by comparing their features to the prototypes. The class with the closest prototype is assigned to the image.

**Step 9:** Initialize the model using pre-trained weights from large samples (e.g., ImageNet), and fine-tune it on the cervical cancer dataset.

**Step 10:** For each epoch (b = 1 to b), choose the mini-batch of size n from the training set (δ1), perform forward propagation through the transformer encoder to extract features, compute the loss, and update the model Weights (w) using back-propagation.

**Step 11:** Evaluate the validation set (δ2) using metrics like precision, Accuracy, F1-score, and recall.

**Step 12:** Test the model (δ3) to evaluate its generalization performance on unknown data.

**Step 13:** Perform final evaluation and analysis of the performance, and then stop.

The proposed framework performs binary-class classification, enabling the model to distinguish a simple normal–abnormal binary split. ViT-based few-shot architecture extracts discriminative features for each class, while the prototype-based classifier assigns each image to the closest diagnostic category. This approach allows for more granular and clinically meaningful predictions, supporting early detection and accurate staging across different cervical lesion types.

## 4. Numerical Results and Discussion

The trained ViT Transformer was evaluated. MATLAB 2020a package computed essential performance indicators, including F1 score, AUC, and Accuracy. GPU: NVIDIA RTX

3080 (10 GB VRAM), CPU: Intel Core i7-10700K @ 3.8 GHz, RAM: 32 GB DDR4, OS: Windows 10 (64-bit), and storage: SSD 1 TB. The ViT model used 224×224 images, 8×8 patches, a 768-dimensional embedding, 12 encoder layers, and 12 attention heads. Training used Adam with a learning rate of 1e-4, weight decay 0.01, batch size 16, and 50 fine-tuning epochs. Few-shot evaluation used 1-shot and 5-shot support sets, with prototypes computed as mean feature embeddings and Euclidean distance for classification. The ability to generalize to new, unseen data was confirmed to be very accurate by this evaluation. For another performance test, the model's robustness was measured against a held-out dataset containing pictures of normal and cancer cases. It was important to understand the model's ability to discriminate between non-cancerous and cancerous images. The model's potential use in diagnosis was indicated by its performance in the stored data assessment in MATLAB 2020a, which allowed it to differentiate a healthy cervix from any intermediate anomaly, e.g., a precancerous one. A lot of the crucial evaluation metrics that we publish in this article are based on the confusion matrix, which is usually constructed as follows. Each measure gives some information on how the model behaves in different aspects; in medical imaging, this is particularly important for the evaluation of classification models. True-Positive Rate gauges how the model can identify real positive cases, such as individuals who are ill. A high sensitivity level suggests the model is good at catching true positives since it produces a few false negatives:

$$Sensitivity = \frac{TP}{TP+FN} \tag{11}$$

A high-sensitivity model lowers the possibility of overlooking crucial diagnoses by guaranteeing that most true positive instances are identified. True-negative rate specificity evaluates how the model can detect real negative cases. The formula:

$$Specificity = \frac{TN}{TN+FP} \tag{12}$$

It is essential for reducing false positives. The high specificity successfully prevents false positives, which can result in needless treatments or follow-ups.

$$PPV = \frac{TP}{TP+FP} \tag{13}$$

Precision or Positive Predictive Value (PPV): The optimistic forecasts that turn out to be accurate are called the Positive Predictive Value (PPV) or Accuracy. Substantial accuracy suggests that the model is likely to be correct when it predicts an optimistic scenario, primarily when falsely favorable costs are significant.

The term "NPV" refers to a negative predictive value. The percentage of pessimistic predictions arriving at reality is represented by:

$$NPV = \frac{TN}{TN+FN} \tag{14}$$

A high NPV lowers the possibility of false negatives by proving that the model is dependable in accurately recognizing true negatives.

Accuracy: Considering both positives and negatives, the overall accuracy of all classes.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

Accuracy offers the model's potential, but it might not be enough, particularly in unbalanced datasets, where the model may miss minority classes. Still, the majority class may be correctly predicted. F1 rating for Formula One: By balancing precision and recall, it offers a single statistic that considers false negatives and false positives:

$$F1-score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

It benefits datasets with class imbalances because it compromises recall and precision trade-offs. AUC indicates how the model can distinguish favorable and unfavorable conditions. The model performs well at class differentiation across all thresholds when its AUC is high. Alongside that, a number of visual aids, such as bar plots representing AUC and other vital metrics, ROC curves, and confusion matrices, were heavily relied upon to analyze the model's performance. Such visuals facilitated a complete and accurate description of the machine's effectiveness to the audience.

### *4.1. Clinical implications*

The improvements in Accuracy and AUC measures indicate that the ViT model can effectively be used in clinical applications in the future. As shown, early diagnosis is facilitated by the models, which is essential in the provision of timely treatments that lead to the slowing down of the invasive behavior of cervical cancer and lower the death rates. The ViT model ensures accurate identification through the use of specific visual and textural features unique to a given specimen by a standard algorithmic evaluation that supports the decision-making of humans. This technique offers an elaborate and advanced cervical imaging examination through thorough image enhancement and a better understanding of the context for improved detection and classification accuracy (See Figure 3 to Figure 10). Its exceptional performance resulting from continuous real-time multi-dataset learning makes the model reliable under a wide range of therapeutic conditions; consequently, the model's generalizability to diverse patient populations is improved. The models outside of attending to tumor detection and classification may also differentiate normal cervical images. The models also correctly classify images that are inappropriate for the training set as either normal or cancerous.

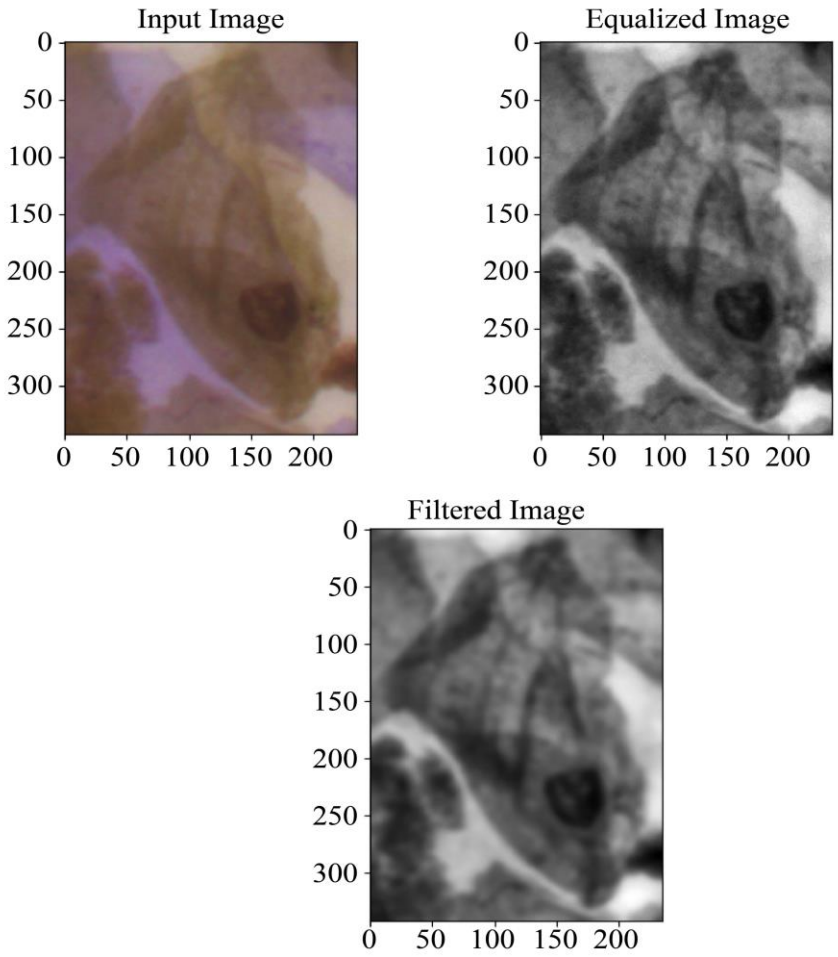


**Fig. 3 Preprocessed image**

This study demonstrates the integration of ViT transformers of CC from the provided images. The AI-based solution has the potential to automate and streamline the analysis of colposcopy images, thus reducing the inconsistencies commonly associated with traditional methods such as Pap smears and HPV testing. Additionally, it stands to enhance diagnostic accuracy, which may reduce the financial and psychological burdens placed on patients and healthcare systems by minimizing false positives.

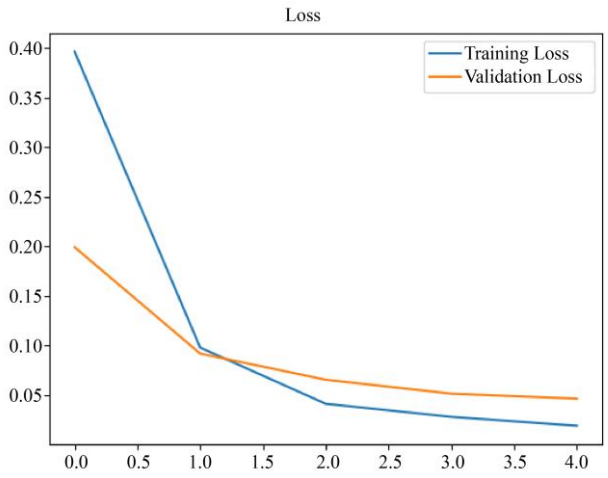


**Fig. 4 Loss analysis**

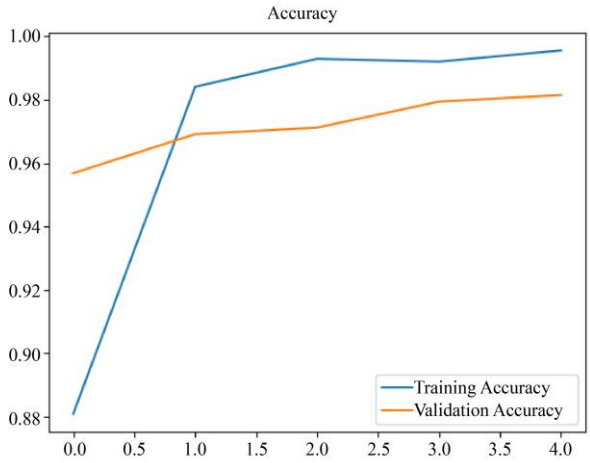


**Fig. 5 Accuracy analysis**

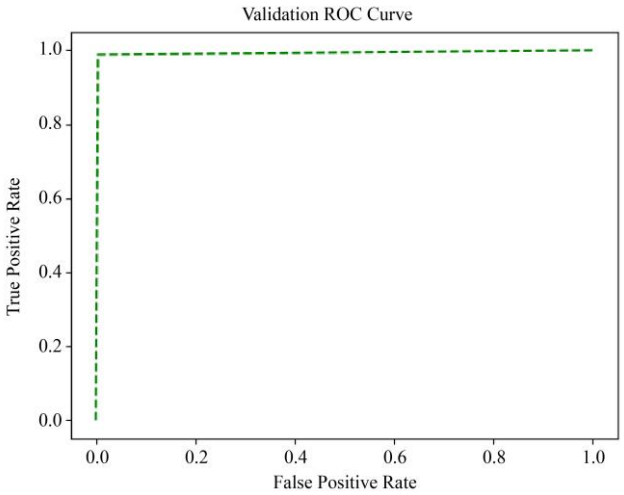


**Fig. 6 ROC curve**

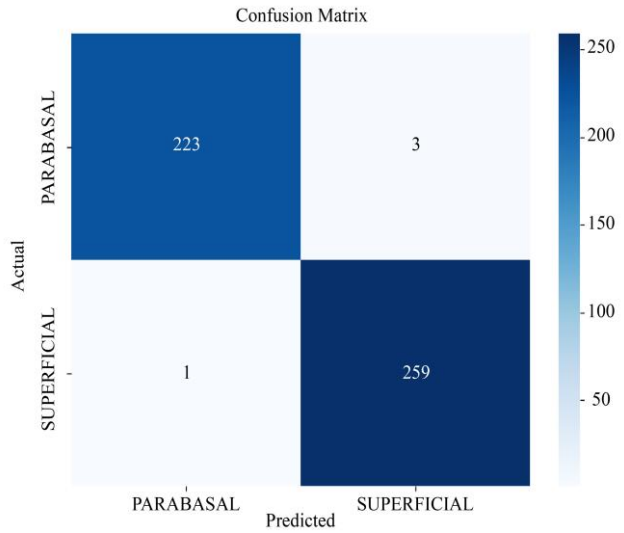


**Fig. 7 Confusion matrix**

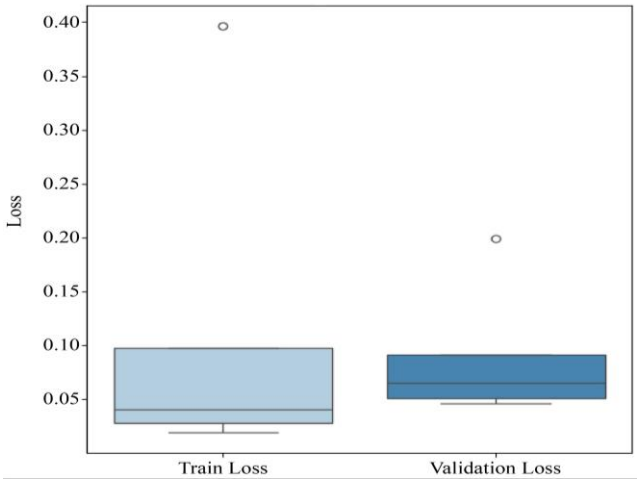


**Fig. 8 Training and validation loss**

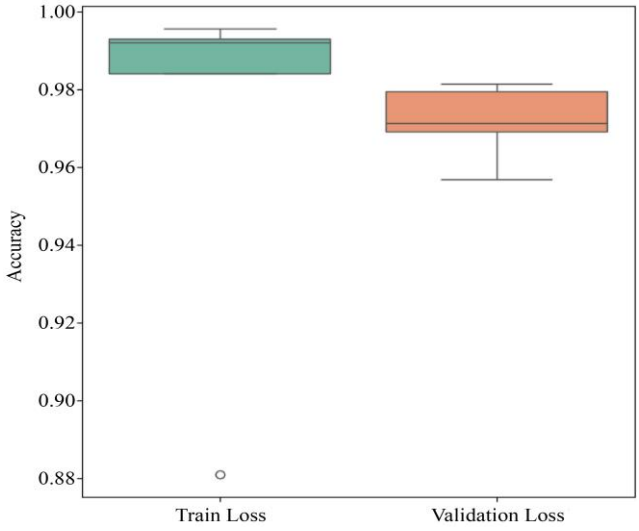


**Fig. 9 Training and validation accuracy**

**Table 2. Metrics comparison with existing methods**

| Metrics | Accuracy | Precision | Recall | F1 score | Error rate | mAP score |
|---|---|---|---|---|---|---|
| EfficientNet | 91.2 | 91.5 | 92 | 92.2 | 0.13 | 91 |
| MobileNet | 92 | 93.5 | 93 | 93.1 | 0.16 | 93 |
| Efficient-MobileNet model | 96.7 | 97 | 97 | 97 | 0.04 | 96 |
| Novel vision transformers | 99.38 | 99 | 99 | 99 | 0.02 | 99 |

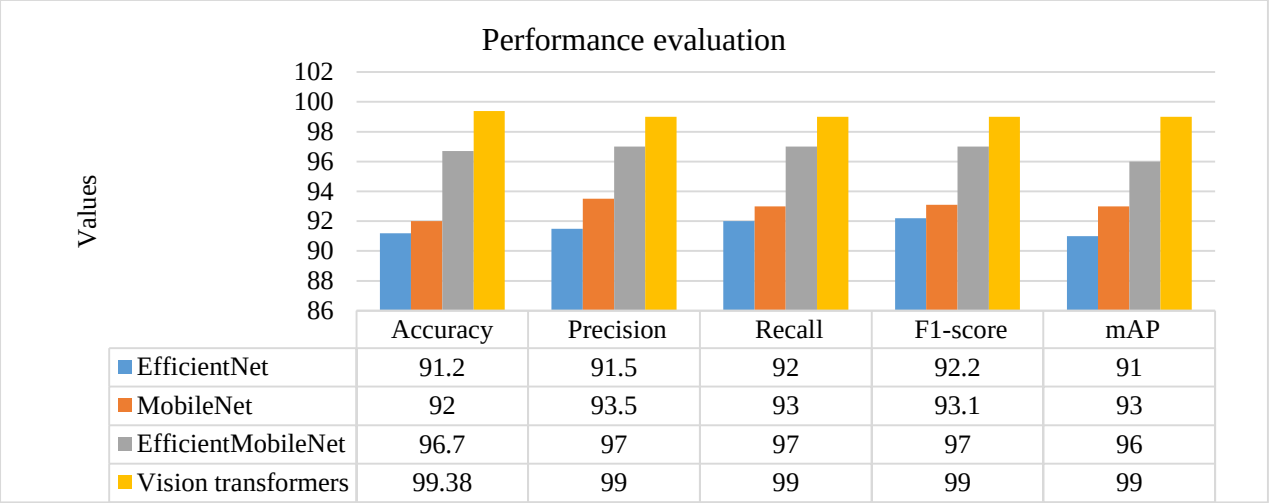


**Fig. 10 Performance evaluation**

CNNs, based or hybrid CNN, RNN architectures usually have a limited capacity to capture long-range and context-dependent features because they are essentially local models. In contrast, a transformer backbone uses self-attention methods to model relationships between different tissue structures at the cervix. As a result, it allows the model to figure out intricately detailed lesion patterns and minute changes in tissue that are typically invisible even to the human eye, and hence can easily be missed by the local CNN models. Moreover, the transformer combines local features with the global context information that jointly yields a richer representation, which in turn leads to more accurate diagnoses along with increased sensitivity.

### *4.2. Limitations*

Despite these positive results, there are still numerous hurdles. One of the major challenges for wider adoption in healthcare is the need for large annotated datasets to help train and validate these models. Generating new data and leveraging already learned models might help in addressing this problem; however, to ensure generalizability, it is essential that the models are exposed to a wide variety of scenarios.

For instance, these results could be greatly strengthened, and the diagnostic power of AI could be increased by adding more data sources like genomes, patient histories, and omics, and testing these results using more diverse datasets. Also, to integrate those into clinical workflows would require rigorous checking and legal compliance to ensure the effectiveness and safety of AI systems. To foster trust between physicians and patients, it is also vital to consider the ethical issues regarding the implementation of AI in healthcare, such as algorithmic openness, data privacy, and potential biases. The diagnostic accuracy and resilience of the model may be significantly improved in the future by adding much larger modal datasets. The course of the disease would be possible with an all-encompassing approach, which may also result in more individualized treatment plans. Building explainable AI models that provide information on how computers make decisions and assist doctors in better understanding and believing the AI's suggestions further enhances the acceptance of these technologies in clinical practice.

**Table 3. Performance evaluation of the ViT Model with prevailing approaches in CC detection**

| Metrics | Transformer model [30] | Few-shot model [28] | DL+ABC [33] | YOLO [32] | BSNet [29] | Novel Vision Transformers (This Work) |
|---|---|---|---|---|---|---|
| **Accuracy** | 99.02% | 65% | 97.5% (SUN dataset) | 97.8% ( | 98.58% | **99.38%** |
| **Precision** | 91.50% | 91% | 98.3% | 97.5% | 99.73% | **99%** |
| **Recall** | 92% | 65% | 96.7% | 97.3% | 61.13% | **99%** |
| **F1 Score** | 92.20% | N/A | 97.1% | 97.2% | N/A | **99%** |
| **Error Rate** | 0.13 | N/A | 0.03 | 0.04 | N/A | **0.02** |
| **mAP Score** | 91% | N/A | 97.5% | 97.8% | N/A | **99%** |

In this work, a cervical cancer detection model based on a Vision Transformer (ViT) was developed, and its performance is evaluated against existing models, including the MaxViT, YOLO-based, and other deep learning algorithms. The ViT model achieves 99.38%, exceeding the MaxViT accuracy at 99.02% and 99.48% on the SIPaKMeD and LBC datasets, respectively, and outperforming all other models. The proposed ViT model with its 99% precision also performs better than Lee et al.'s [14] agricultural monitoring model, which reported 99.73% precision and 61.13% recall, as well as Karaman's [33] colorectal cancer model with 96.7% to 97.3% recall. The model yields an F1 score of 99%, which demonstrates its performance in maintaining precision and recall, combined with an error rate of 0.02, which is lower than the 0.13 and 0.04 error rates put forth by MaxViT and Karaman et al.'s [32] models, respectively. Moreover, the model surpasses in mAP score with 99%, while other models ranged from 91% to 97.8%. In summary, the approach based on ViT, the cervical cancer detection system, outperformed all other models, and by increasing the precision of medical image analysis, ViT has shown marked improvements over existing technologies. The high accuracy achieved by the proposed model is largely due to the few-shot framework combined with transformer-based feature extraction. The ViT encoder generates highly discriminative embeddings even from limited training samples, allowing the few-shot prototype classifier to generalize effectively across the full test set. Although performance metrics are reported using the entire test dataset, it is explicitly clarified that the classification stage does not rely on full supervised training. Instead, predictions are made using few-shot prototypes computed from a small support set, demonstrating that strong diagnostic performance can be achieved without large-scale labeled training data.

## 5. Conclusion

This work highlights the possibility of diagnosing cervical cancer efficiently with limited data through a ViT, based few, shot learning system. With only a few labeled examples, the proposed method can make accurate predictions and therefore perfectly responds directly to the problem of the shortage of annotated medical images, the most common challenge in clinical settings. This method lessens the need for large datasets and relies on expert labeling to a minimum. It demonstrates that diagnostic performance is still reliable even in data-scarce environments. This work is based on the analysis of diagnostic colposcopy images and presents the excellent prediction capability of ViT models for CC. The models showed very good performances on a test data set, their accuracies fluctuating between 99% and 99. 3%, and the respective AUCs going from 99% to 99. 2%. Besides facilitating early diagnosis of cervical cancer, these models even have the potential to reduce the number of tests induced by false positives, thus freeing individuals and health care systems from both financial and emotional burdens. Therefore, these models are valuable assistants in clinical practice due to their capability of conducting real-time studies and generalizing diverse patient populations, especially in settings with scarce resources and where access to highly skilled colposcopists is limited. Future studies should revisit confirming these results with larger and more varied datasets to reveal the ultimate capabilities of these models. If different data types, such as omics and genomes, are incorporated, the models' resilience and diagnostic precision could be further enhanced, which in turn will help open up more extensive and accurately targeted treatment options. Issues like dataset representativeness, legal permissions, and ethical matters such as data protection and clarity about how algorithms are working should be properly addressed in order to unleash the full potential of AI, driven diagnostic tools in healthcare environments. By creating dependable and user-friendly diagnostic platforms based on the ViT model, especially in the underdeveloped parts of the world, the cervical cancer rate could be drastically reduced. The research work results could provide the foundation for the formulation of cervical cancer screening and preventive measures, which will lead to better public health outcomes and enhanced quality of life for women globally.

## References


[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné, "Mixture-based Feature Space Learning for Few-shot Image Classification," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9021-9031, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Arman Afrasiyabi et al., "Matching Feature Sets for Few-Shot Image Classification," *arXiv Preprint*, pp. 9004-9014, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné, "Associative Alignment for Few-Shot Image Classification," *16th European Conference Computer Vision – ECCV 2020*, Glasgow, UK, pp. 18-35, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Narenthirakumar Appavu, "Analysing the Effect of Edge-Optimized Deep Learning Models on Improving Low-Powered Iot Devices Real-Time Object Detection," *2025 9th International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, pp. 1663-1669, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Sungyong Baik et al., "Meta-Learning with Task-adaptive Loss Function for Few-shot Learning," *arXiv Preprint*, pp. 9465-9474, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Narenthirakumar Appavu, "Skin Cancer Detection Using A Multi-Scale AI Deep Learning Approach," *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Gianni Brauwers, and Flavius Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3279-3298, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Qi Cai et al., "Memory Matching Networks for One-Shot Image Recognition," *arXiv Preprint*, pp. 1-9, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Xuewei Chao, and Lixin Zhang, "Few-Shot Imbalanced Classification based on Data Augmentation," *Multimedia Systems*, vol. 29, pp. 2843-2851, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yinbo Chen et al., "Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning," *arXiv Preprint*, pp. 1-2, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Guanqi Ding et al., "Attribute Group Editing for Reliable Few-shot Image Generation," *arXiv Preprint*, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Junhao Dong et al., "Improving Adversarially Robust Few-shot Image Classification with Generalizable Representations," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 9015-9024 [CrossRef] [Google Scholar] [Publisher Link]

[13] Jaejun Do, Minjung Yoo, and Sunok Kim, "A Semi-supervised SAR Image Classification with Data Augmentation and Pseudo Labeling," *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Yeosu, Korea, Republic of pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Narenthirakumar Appavu, "Detection and Categorisation of Brain Tumours Using Hybrid AI Deep Learning Methods," *2025 6th International Conference on Mobile Computing and Sustainable Informatics*, Goathgaun, Nepal, pp. 1227-1234, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Soroor Malekmohamadi Faradonbe, Faramarz Safi-Esfahani, and Morteza Karimian-kelishadrokhi, "A Review on Neural Turing Machine (NTM)," *SN Computer Science*, vol. 1, pp. 1-23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-learning for Fast Adaptation of Deep Networks," *arXiv Preprint*, pp. 1-13, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[17] Madhava Gaikwad, and Ashwini Doke, "Survey on Meta Learning Algorithms for Few Shot Learning," *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 1876-1879, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Katelyn Gao, and Ozan Sener, "Modeling and Optimization Trade-off in Meta-learning," *NeurIPS*, pp. 1-12, 2020. [Google Scholar] [Publisher Link]

[19] Spyros Gidaris, and Nikos Komodakis, "Generating Classification Weights with GNN Denoising Autoencoders for Few-Shot Learning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 21-30, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jiuxiang Gu et al., "Recent Advances in Convolutional Neural Networks," *Pattern Recognition*, vol. 77, pp. 354-377, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yunhui Guo et al., "A Broader Study of Cross-Domain Few-Shot Learning," *16th European Conference Computer Vision – ECCV*, Glasgow, UK, pp. 124-141, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Guangxing Han et al., "Multi-Modal Few-Shot Object Detection with Meta-Learning-Based Cross-Modal Prompting," *arXiv Preprint*, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tianhao Hu et al., "A Simple Data Augmentation Algorithm and a Self-adaptive Convolutional Architecture for Few-shot Fault Diagnosis under Different Working Conditions," *Measurement*, vol. 156, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Huaxi Huang et al., "Compare More Nuanced: Pairwise Alignment Bilinear Network for Few-Shot Fine-Grained Learning," *2019 IEEE International Conference on Multimedia and Expo (ICME)*, Shanghai, China, pp. 91-96, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Bingyi Kang et al., "Few-Shot Object Detection via Feature Reweighting," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 8419-8428, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Geethan Karunaratne et al., "Robust High-Dimensional Memory-Augmented Neural Networks," *Nature Communications*, vol. 12, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[27] Abdullatif Köksal, Timo Schick, and Hinrich Schuetze, "MEAL: Stable and Active Learning for Few-Shot Prompting," *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, pp. 506-517, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Narenthirakumar Appavu, "Deep Learning for Predicting Invasive Ductal Carcinoma in Histopathological Tissue Images," *2025 Eleventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*, Chennai, India, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Xiaoxu Li et al., "BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1318-1331, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ishak Pacal, "MaxCerVixT: A Novel Lightweight Vision Transformer-based Approach for Precise Cervical Cancer Detection," *Knowledge-Based Systems*, vol. 289, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] Narenthirakumar Appavu, "Skin Cancer Detection Using A Multi-Scale AI Deep Learning CNN Techniques," *2025 6th International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, India, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ahmet Karaman et al., "Robust Real-time Polyp Detection System Design Based on YOLO Algorithms by Optimizing Activation Functions and Hyper-parameters with Artificial Bee Colony (ABC)," *Expert Systems with Applications*, vol. 221, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Ahmet Karaman et al., "Hyper-Parameter Optimization of Deep Learning Architectures using Artificial Bee Colony (ABC) Algorithm for High Performance Real-Time Automatic Colorectal Cancer (CRC) Polyp Detection," *Applied Intelligence*, vol. 53, pp. 15603-15620, 2023. [CrossRef] [Google Scholar] [Publisher Link]