SegFormer : Semantic Segmentation for Cervical Cancer detection based on Mix Transformer (MiT)

Inda Rusdia Sofiani1,a), Hadi Suyono2,b), Erni Yudaningtyas2,c), and Fitri Utaminingrum3,d)

1Department of Vocational, Muhammadiyah Malang University, Indonesia

2,3Department of Engineering, Universitas Brawijaya, Malang, Malang, Indonesia

3Department of Computer Science, Universitas Brawijaya, Malang, Indonesia

a) Corresponding author: [indarusdias05@umm.ac.id](mailto:indarusdias05@umm.ac.id)

b) [hadis@ub.ac.id](mailto:hadis@ub.ac.id)

c) [erni@ub.ac.id](mailto:erni@ub.ac.id)

c) [f3\_ningrum@ub.ac.id](mailto:f3_ningrum@ub.ac.id)

**Abstract.**  Cervical cancer is a leading cause of cancer-related deaths in women, with the highest prevalence in developing countries. According to WHO, there were 528,000 new cases in 2012, and 87% of cervical cancer deaths occurred in low-resource regions. Despite its high cure rate when detected early, many developing countries lack effective early detection systems. Medical image segmentation using artificial intelligence offers a potential solution to improve the accuracy of cervical cancer diagnosis and treatment. This study explores the use of image segmentation techniques through SegFormer to address challenges in colposcopy image segmentation. The research aims to enhance diagnostic accuracy through image classification and segmentation, focusing on overcoming challenges such as medical data variability, anatomical structural complexity, and limited labeled data. The methods are evaluated using the Dice Coefficient (dc) and Intersection Over Union (IoU) metrics to determine model performance. This research is expected to significantly contribute to the development of cervical cancer diagnostic technologies and improve access to accurate early diagnosis in resource-limited areas.

**Keywords:** cervical cancer, image segmentation, artificial intelligence, SegFormer

# INTRODUCTION

Cervical cancer is one of the most common types of cancer affecting women and ranks seventh as the leading cause of cancer-related deaths worldwide. In 2012, according to data from the International Agency for Research on Cancer (IARC) of the WHO, 528,000 new cases of cervical cancer were recorded. The majority of these cases, around 85%, occurred in developing countries, where cervical cancer accounts for approximately 12% of all cancer cases. [1]. Furthermore, the data also revealed that 87% of deaths due to cervical cancer occurred among women living in less developed regions, with a patient mortality rate of 7.3%.

Although cervical cancer is highly treatable if detected early, unfortunately, in many developing countries, effective early detection and prevention programs are still lacking. [2]. The significant geographic variation in the incidence of cervical cancer reflects differences in access to screening and the prevalence of Human Papillomavirus (HPV) infection. In countries with low cervical cancer incidence, the prevalence of chronic HPV infection is nearly half that of countries with high cancer incidence. [3].

HPV has been identified as the primary cause of cervical cancer, and it takes years to develop into invasive cancer[4]. Therefore, early detection and proper treatment are crucial to prevent further progression of cervical cancer. The development of better diagnostic tools can assist medical professionals in detecting and treating cervical cancer more quickly and accurately.

AI-based image classification techniques offer a promising solution in assisting with the diagnosis of cervical cancer through medical imaging. In many studies, the use of image classification methods enables the recognition of patterns and features that are difficult for the human eye to detect, such as tumors, lesions, or structural anomalies. Moreover, with image segmentation techniques, critical areas of medical images can be isolated and analyzed more thoroughly, enhancing the accuracy of diagnosis and personalizing patient care. [5].

However, medical image segmentation still faces various challenges, particularly related to data inconsistency, small details in anatomical structures, and the limited availability of labeled data. Cervical cancer, especially in its early stages, is often difficult to diagnose accurately due to these factors. The development of more advanced segmentation techniques, such as SegFormer, is expected to help overcome these challenges, thereby improving the accuracy of cervical cancer detection and diagnosis [6].

Thus, this study aims to explore and develop AI-based segmentation techniques for the early detection of cervical cancer, with the goal of improving diagnostic effectiveness and enabling more timely medical intervention.

# METHODS

The administration of datasheets for image processing in AI involves several key steps. First, collect images from various sources while ensuring their variation and quality align with the research objectives. Second, annotate each image with appropriate labels, using annotation software to mark important objects or features. Third, store the images and annotations in an organized format, such as separate folders or a database, and ensure each image contains relevant metadata such as resolution, file format, and description. Fourth, ensure the datasheet is well-structured and documented to facilitate the processing and analysis by the AI model. The research design to be conducted by the researcher is illustrated in **FIGURE 1**.



**FIGURE 1**. Research Design

## Research Stages

The research stages begin with a literature review and conclude with the final results in processing image datasheets using Segformer, involving several systematic steps.

## Dataset Collection

The AnnoCerv dataset, created by a group of researchers including Dorina Adelina Minciună and Demetra Gabriela Socolov, is a valuable resource for automated colposcopy analysis [7]. It contains 527 images taken from 100 medical records, all of which have been carefully annotated by specialists to distinguish between healthy and abnormal cervical tissues. This dataset helps advance machine learning applications for detecting and classifying cervical lesions, aiding in early diagnosis and treatment planning for cervical cancer.

Additionally, the dataset includes Swede scores that quantify the severity of cervical abnormalities. Available under a CC-BY-4.0 license on platforms like GitHub, AnnoCerv supports researchers developing tools for medical image analysis, providing a useful foundation for advancements in automated cervical cancer screening technology.

## Data Preprocessing

Contrast Limited Adaptive Histogram Equalization (CLAHE) works by dividing an image into small regions (tiles) and applying histogram equalization to each region, allowing for local contrast enhancement without excessively increasing noise. CLAHE can address the over-contrast issue in Adaptive Histogram Equalization (AHE) by limiting the use of boundary values in the histogram. CLAHE offers the best sensitivity compared to other enhancement methods. The most important factor in CLAHE is the clip limit. The clip limit is a system designed to reduce image blurring, ensuring that image quality enhancement does not exceed a predefined limit or clip. This technique is commonly used in digital images, particularly in the medical field, to improve contrast levels. In this study, a clip limit value of 2.0 and a grid size of 8x8 were used. After applying the CLAHE technique, the resulting data is stored in a new folder, which will be used in the next steps of the process.

The next step is typically to preprocess the data to ensure it is in a suitable format for machine learning models. This involves tasks such as normalizing or resizing images, augmenting the data to create more training examples, and converting annotation files (such as JSON) into the appropriate input format for deep learning frameworks. Once the data is prepared, it is then split into training, validation, and test sets to ensure that the model can generalize to unseen data. The annotated dataset is then used to train machine learning models, such as convolutional neural networks (CNNs) for image segmentation, object detection, or classification tasks, depending on the project requirements.

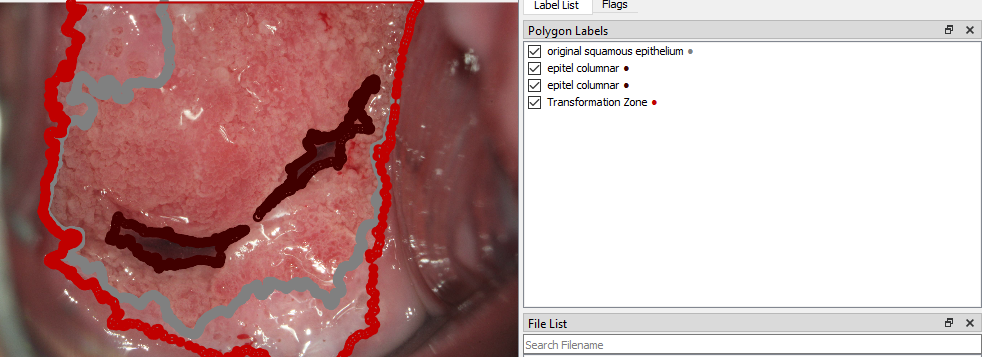
After the annotation process, expressing an image typically involves visualizing both the original image and the corresponding annotations to assess the quality and accuracy of the labeled data. This can be done by overlaying the annotated regions or segmentation masks onto the image, which highlights the labeled areas with different colors representing various classes. This visual expression helps in identifying how well the annotations align with the objects in the image and allows for quick detection of any discrepancies or errors in the labeling, as show in **FIGURE 2**.

## SegFormer Architecture

The diagram of the SegFormer architecture consists of several key components that interact with each other[8]. First, the input image is divided into small patches, which form the foundation of the initial process. These patches are then processed by a series of Transformer blocks, each consisting of two stages: multi-head self-attention to capture global relationships between patches, and feed-forward networks to process the information more deeply [9][10]. These Transformer blocks are arranged hierarchically, allowing feature capture at various resolution scales. Each block generates feature representations that are then collected by a simple decoder. This decoder combines information from all the Transformer blocks, aligning features from different resolutions to form a detailed and accurate segmentation map. The final output is a segmentation map ready for various image processing applications.

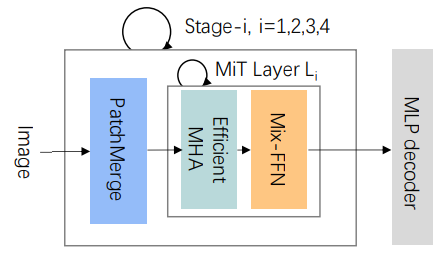
**FIGURE 3** shows the conventional SegFormer architecture using MHA, which consists of four stages. Each stage has several MiT layers, where each layer includes an MHA module followed by a mix FFN module. The decoder adopts an MLP architecture for pixel prediction.



(a)

(b)

**FIGURE 2** a) cervical image before annotation, b) after annotation with squamous epithelium, columnar epithelium and transformation zone



**FIGURE 3**. SegFormer Architecture with Multi-Head Attention (MHA)

## Implementation of SegFormer

The raw images that have undergone CLAHE preprocessing are trained using the SegFormer architecture model. During training, various attention mechanisms are applied to maximize the understanding of spatial and temporal context within the images.

The implementation of Segformer, particularly in medical imaging, involves several training models developed by Hugging Face's `transformers` and PyTorch, which provide weights for mit\_b1, mit\_b2, mit\_b3, mit\_b4, and mit\_b5, variants of MiT (Mix Transformer). MiT is used as the backbone in Segformer's architecture. These variants differ in size and complexity, with mit\_b1 being the lightest and fastest in computation due to having the fewest parameters. As the variant level increases, more parameters are used, leading to better accuracy and efficiency, making them ideal for segmentation tasks requiring high precision. The main differences between these variants lie in the number of transformer blocks and embedding dimensions, which affect accuracy and inference time.

First, medical images are preprocessed (resized, normalized) using a feature extractor, and the model is trained with pixel-specific annotation masks for the particular medical task. After training, the model predicts a segmentation map where each pixel represents a specific class (e.g., tumor or normal tissue). Post-processing steps, such as overlaying the segmentation map on the original image, help visualize the results for clinical interpretation.

## Evaluation

The evaluation was conducted using metrics such as Intersection over Union (IoU) to assess the quality of the generated segmentation. In this implementation, several MiT varians, were utilized.

# RESULTS AND DISCUSSION

In this experiment, multiple variants of the MiT (Multi-scale Vision Transformer) model were integrated into the network to assess the program’s efficiency in detecting objects. Specifically, we employed the mit\_b1 and mit\_b2 versions to determine how well the program performed. The key metric used to evaluate the program's effectiveness was the Intersection over Union (IoU) value, which is a common measure in image segmentation tasks, representing the overlap between the predicted and ground truth areas. The IoU scores provide a clear indication of how accurate the models are in identifying and segmenting the target areas.

The results revealed that both MiT variants demonstrated high levels of accuracy, with the mit\_b1 variant achieving an IoU value of 99.83%, and mit\_b2 slightly outperforming it with an IoU value of 99.85%. These extremely high IoU values indicate that the program is highly efficient in its task, with minimal difference between the two model variants. This suggests that either variant could be used effectively, but mit\_b2 might have a slight edge in accuracy when it comes to object detection and segmentation.

# CONCLUSIONS

Based on the iou testing conducted on the mit\_b2, mit\_b3, and mit\_b4 model variants, it was found that increasing the number of parameters in more complex models, such as mit\_b3 and mit\_b4, provides potential improvements in performance for segmentation tasks. However, the final results still depend on how well the models can learn key features from the dataset used. Overall, variants with more parameters tend to produce higher iou values, indicating that more complex models are better at capturing detailed information, thus improving segmentation accuracy.

# Acknowledgments

The authors would like to acknowledge and also grateful to dr. M. Khalif Anfasa, Sp.OG (K)Onk for their insightful comments and for providing access to the necessary images. finally, we thank the anonymous reviewers for their constructive feedback.

# References

1. J. Ferlay *et al.*, “Cancer statistics for the year 2020: An overview,” *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, 2021.
2. A. Das and A. Choudhury, “A novel humanitarian technology for early detection of cervical neoplasia: ROI extraction and SR detection,” *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 457–460, 2018.
3. X. Zhang, Q. Zeng, W. Cai, and W. Ruan, “Trends of cervical cancer at global, regional, and national level: data from the Global Burden of Disease study 2019,” *BMC Public Health*, vol. 21, no. 1, pp. 1–10, 2021.
4. M. U. Barut *et al.*, “Analysis of sensitivity, specificity, and positive and negative predictive values of smear and colposcopy in diagnosis of premalignant and malignant cervical lesions,” *Med. Sci. Monit.*, vol. 21, pp. 3860–3867, 2015.
5. X. Li, X. Yang, Z. Ma, and J. H. Xue, “Deep metric learning for few-shot image classification: A Review of recent developments,” *Pattern Recognit.*, vol. 138, p. 109381, 2023.
6. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer : Simple and Efficient Design for Semantic Segmentation with Transformers,” pp. 1–18.
7. D. A. Minciună *et al.*, “AnnoCerv: A new dataset for feature-driven and image-based automated colposcopy analysis,” *Acta Univ. Sapientiae, Inform.*, vol. 15, no. 2, pp. 306–329, 2023.
8. J. He, W. Li, and J. Qu, “Shifted Window SegFormer For Remote Sensing Image Segmentation,” *ICIIBMS 2023 - 8th Int. Conf. Intell. Informatics Biomed. Sci.*, vol. 8, pp. 117–121, 2023.
9. A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
10. A. Dosovitskiy *et al.*, “an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale,” *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.
11. A. Mishra, “Contrast Limited Adaptive Histogram Equalization (CLAHE) Approach for Enhancement of the Microstructures of Friction Stir Welded Joints,” pp. 1–28, 2021.