This study examined automated multi-class semantic segmentation of Pap smear images used for cervical cancer detection. The effectiveness of existing deep learning methods is often limited due to a lack of labeled data, high morphological variability of cervical cells, overlapping structures, noise, low contrast, and imaging artifacts characteristic of cytology specimens.

In this study, the authors propose a cross-domain transfer learning approach that adapts pre-trained deep neural networks to the task of multi-class Pap smear segmentation. All networks were pre-trained on large-scale natural image datasets. In the experiments, both convolutional neural networks and Transformer-based models, including hybrid configurations, were refined and systematically compared. Network performance was assessed using quantitative metrics (Dice score, IoU, HD95), as well as qualitative visual assessment of segmentation edges and boundaries.

The results obtained from the experiments showed that Transformer-based architectures, in particular SegFormer, significantly outperform convolutional models when processing noisy and heterogeneous cytological data. Using specialized data augmentation strategies developed specifically for medical imaging, SegFormer increased Dice scores to 0.95 across all classes (healthy, unhealthy, rubbish, both cells), as well as improved edge accuracy and robustness to artifacts and cell aliasing.

Multi-scale feature extraction and global context modeling proved essential for accurately identifying cellular structures in data-constrained settings. The results obtained in the study can help in the development of reliable automated diagnostic tools to assist cytopathologists, as well as to improve the overall accuracy and efficiency of cervical cancer screening programs

Keywords: transfer learning, Pap smear, cervical cancer, segmentation, deep learning

UDC 004.8
DOI: 10.15587/1729-4061.2026.352892

IMPROVEMENT OF THE METHOD OF THE MULTICLASS PAP SMEAR IMAGE SEGMENTATION BASED ON CROSS-DOMAIN TRANSFER LEARNING WITH LIMITED DATA

Margulan Nurtay
Master of Technical Sciences*
ORCID: https://orcid.org/0000-0002-0786-6195
Gaukhar Alina
Corresponding author
PhD Student, Master*
E-mail: alinagaukhar@gmail.com
ORCID: https://orcid.org/0000-0002-7697-4667
Ardak Tau
Master of Technical Sciences
ORCID: https://orcid.org/0000-0003-4883-6328
*Department of Information and Computing Systems
Abylkas Saginov Karaganda Technical University
N. Nazarbayev ave., 56, Karagandy,
Republic of Kazakhstan, 100027

Received 09.12.2025
Received in revised form 02.02.2026
Accepted 16.02.2026
Published 27.02.2026

How to Cite: *Nurtay, M., Alina, G., Tau, A. (2026). Improvement of the method of the multiclass Pap smear image segmentation based on cross-domain transfer learning with limited data. Eastern-European Journal of Enterprise Technologies, 1 (9 (139)), 47–55. https://doi.org/10.15587/1729-4061.2026.352892*

1. Introduction

Cervical cancer detection remains a large-scale task, where automation makes practical sense. Pap smear generates large arrays of imaging data that require automated analysis. Its use is associated with several limitations, including subjective interpretation and the risk of false negatives, highlighting the need to improve analysis methods and implement automated technologies [1]. Despite the advances in HPV vaccination and the transition to HPV-based primary screening in a number of developed countries, traditional Pap smear testing remains the primary screening tool in many regions of the world, particularly in developing countries. However, manual Pap smear testing is associated with several significant limitations: subjective interpretation, high reliance on cytologist skills, a significant risk of false-negative results, and low reproducibility. Furthermore, many countries experience a severe shortage of qualified cytologists and pathologists, leading to long waiting lists, diagnostic delays, and reduced screening coverage. These factors contribute to a significant proportion of potentially preventable cervical cancer cases being detected at late stages.

However, automation of Pap smear analysis faces significant challenges, such as high variability in cell morphology, noise, and artifacts in smears. Furthermore, the limited and poorly labeled nature of available medical data complicates segmentation and classification tasks.

Transfer learning is widely used in cytological image analysis as a way to reduce labeled data requirements and stabilize model training. Convolutional networks pretrained on large natural image datasets (e.g., ImageNet) generate generalized low-level representations related to edge, local structure, and texture detection, which can potentially be adapted for medical classification and segmentation tasks [2, 3]. However, transferring such representations is not a universal solution, as significant differences between natural scenes and microscopic cell images lead to domain shifts and possible inappropriate feature transfer. Although several studies in histopathology and cytology demonstrate empirical improvements using transfer learning in data-limited settings [4], transfer performance varies significantly depending on the chosen architecture, fine-tuning strategy, and specific problem setting, suggesting the need for a more systematic analysis of the applicability limits of this approach.

2. Literature review and problem statement

The papers [4, 5] present the research results based on classical methods and approaches. It has been shown that CNN-based approaches suffer from localization limitations, transformers require adaptation to small sets of cytological data with domain shift, and standard ImageNet transfer training shows contradictory results due to domain differences.

However, there are still unresolved issues related to the problems of effectively bridging the gap between preprocessing of natural images and cytological segmentation using cross-domain transfer learning. The reason for this may be insufficient study of CNN hybrid and pure transformer architectures based on limited Pap smear data for multiclass segmentation, through systematic comparison and optimization. Some approaches to segmentation of Pap smears are based on traditional methods such as threshold value, watershed clustering, and k-mean. These methods are less computationally complex and do not even require labeling the data. However, they have one significant drawback: they are ineffective if the data contains noise, low contrast, anti-aliasing, and other defects related to image artifacts. Therefore, they cannot capture the semantic context or any complex relationships that could otherwise be extracted through image analysis.

In the paper [5], the authors used the Otsu threshold value and morphological operations for cell segmentation, but encountered problems related to artifacts and overlays. Marker-controlled watershed in combination with ensemble models demonstrated 98.27% accuracy for binary classification and 94.09% for multiclass classification [6], however, the authors again noted losses due to strong coincidences and artifacts. The reason for this may be that such models do not have the ability to generalize and therefore cannot always be applied in practice.

All this suggests that it is advisable to conduct a study based on the encoder-decoder Architecture with bandwidth connections, which is implemented in the U-Net model and copes well enough with the task of spatial information recovery. This model makes it possible to effectively isolate the cell nucleus, despite the existing noise [7]. The Attention U-Net model [8] allows to focus on more informative areas of the image. The residual U-network simplifies deep learning, and its multiscale variants can take into account different cell sizes.

Transformer models have shifted their focus to self-regulation mechanisms to identify long-term addictions, for example, when it is necessary to simultaneously take into account the local characteristics of cells and the context of tissues. Modifications of Vision Transformers (ViTs), such as UNETR [9], embed ViT encoders into U-Net-like structures for global object perception. Hybrid models such as TransUNet [10] combine the extraction of local features using convolution with global transformer modeling. Swin-UNet and similar models implementing the architecture of the simplest converters are superior to convolutional neural networks and their hybrid counterparts in solving medical problems [11].

Transfer learning solves the problem of a limited amount of labeled medical data, since pre-trained models (VGG, ResNet, and others) can often learn low-level functions in large databases such as ImageNet. For example, ResNet-50 shows 51.9% of pre-trained users versus 8.2% from scratch and demonstrates significant improvements [12]. However, the gap in application between natural and cytological images often reduces the benefits, with some studies showing comparable results from scratch and on pre-prepared models in medical tasks [13, 14].

The analysis of the related works reveals that, despite advancements in specific aspects of pathological image segmentation, the comprehensive problem of multi-class segmentation of Pap smears under conditions of domain shift, noise, artifacts, cell overlaps, and the extremely limited volume of annotated data remains unresolved. Existing approaches [5, 6, 8] either suffer from insufficient generalization and sensitivity to real data defects, fail to undergo systematic comparison and optimization specifically for this task, or produce contradictory results due to domain shift.

Therefore, it can be concluded that it is advisable to conduct a study focused on the development and comparative evaluation of hybrid CNN-transformer segmentation models with optimized cross-domain transfer learning strategies.

3. The aim and objectives of the study

The aim of the study is to improve the method of the multiclass Pap smear image segmentation with using cross-domain transfer learning with limited data. This will allow improving cervical cancer screening by increasing segmentation accuracy on noisy and limited datasets, while reducing reliance on large volumes of labeled medical imaging data and improving detection of abnormal cells.

To achieve this aim, the following objectives were accomplished:

– to evaluate the applicability of models pre-trained on the ImageNet database for multiclass segmentation of Pap smear images and establish baseline performance without data augmentation;

– to investigate the impact of data augmentation techniques on model performance, particularly in overcoming noise, artifacts, cell overlaps, and other common obstacles in cytological images;

– to perform qualitative and quantitative evaluation of the segmentation results, including visual assessment of contours, boundaries, and overall segmentation quality, as well as statistical validation of the proposed method's reliability.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of the study is multiclass pixel-level segmentation of cervical cells in conventional Pap smear cytological images, with a primary focus on addressing the challenge of limited annotated data through cross-domain transfer learning techniques.

The main hypothesis is that the application of cross-domain transfer learning enables achieving significantly higher segmentation accuracy compared to training segmentation models from scratch or using only small in-domain data, while maintaining robustness to visual variability in cell morphology, texture, and background artifacts.

Assumptions made in the study are:

– only conventional Pap smear images containing cervical cells are considered;

– the images primarily feature isolated or mildly overlapping cells suitable for multiclass semantic segmentation into four classes: Healthy, Unhealthy, Rubbish, and BothCells;

– limited annotated data in the target domain is available, while abundant pre-trained weights from large external domains can be effectively adapted;

– expert-validated annotations and standard quantitative metrics such as Dice, IoU, Precision, Recall, HD95 adequately reflect segmentation quality for potential clinical relevance;

– no patient-specific clinical metadata or additional contextual information is incorporated into the segmentation models.

Simplifications adopted in the study

– cell overlapping is handled implicitly by the segmentation network without dedicated post-processing or instance-level separation algorithms;

– variations in staining intensity, illumination, microscope artifacts, and scanner differences are addressed solely through data augmentation and transfer learning, without specialized color normalization or domain adaptation modules beyond fine-tuning;

– input images are uniformly resized to 256 × 256 pixels, ignoring potential benefits of higher or multi-scale resolutions.

4. 2. Dataset description

The dataset used to solve the segmentation problem in this study was taken from the Annotated Pap cell images and smear slices for Cell Classification (APACC) dataset. While the dataset was originally intended for cancer cell classification and detection, its suitability for the present study was confirmed based on its image quality in collaboration with cytomorphology experts.

The dataset initially contained 103,675 annotated images extracted from 107 complete Pap smears. For segmentation, 4,000 images were used and then divided into subsets. The data was divided into training (70%), validation (15%), and test (15%) subsets in a stratified manner, maintaining class balance.

Using QuPath 0.6.0, a medical image annotation and labeling software, the source data was labeled for the cell segmentation task. Annotations were created manually, then automatically exported as multi-class masks (one channel per class: Healthy, Unhealthy, Rubbish, BothCells, with an implicit background) using a Groovy script. The resulting masks and the images themselves were stored in .png format.

All images and corresponding masks were normalized to a fixed spatial resolution of 256 × 256. To improve the generalization capability of the models and alleviate overfitting, data augmentation strategies were applied. The set of applied augmentation operations is detailed in Table 1.

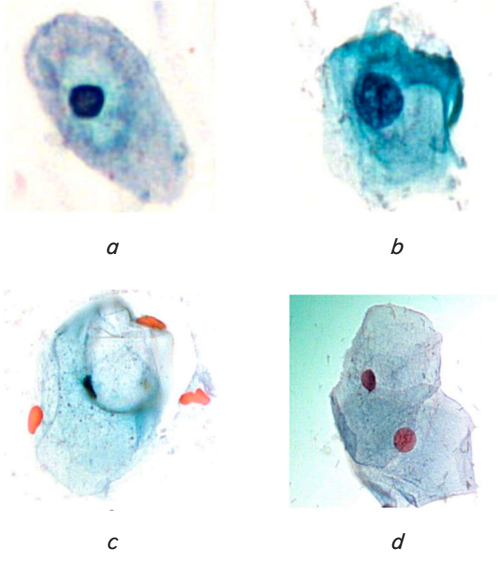
Table 1

Applied data augmentation techniques

Augmentation type	Parameters	Applied to
Random rotation	±15°	Image & Mask
Horizontal flipping	$p = 0.5$	Image & Mask
Vertical flipping	$p = 0.5$	Image & Mask
Random scaling	[0.9, 1.1]	Image & Mask
Translation (shift)	up to 10%	Image & Mask
Brightness adjustment	±10%	Image only
Contrast adjustment	±10%	Image only

Examples of images for each of the four classes are shown in Fig. 1.

As shown in Fig. 1, in terms of cell morphology, the four different classes in the dataset exhibit significant visual variability. Furthermore, these differences extend to cell texture and background artifacts. Healthy and unhealthy cells differ in nuclear size, shape irregularities, and chromatin distribution, while the Rubbish class contains heterogeneous non-cellular structures that may visually overlap with true cell regions. The BothCells class presents additional complexity due to the spatial overlap of multiple cells within a single region.



a *b* *c* *d*

Fig. 1. Representative examples of the four classes from the used dataset: *a* — healthy; *b* — unhealthy; *c* — rubbish; *d* — BothCells

4. 3. Evaluation metrics

The segmentation performance of the proposed models was evaluated using pixel-wise quantitative metrics widely adopted in semantic segmentation studies. In particular, Dice coefficient, Intersection over Union (IoU), Precision, and Recall were used to assess the agreement between the predicted segmentation masks and the ground truth annotations. The evaluation of models was performed on a per-class basis, and the final indicators were obtained by averaging the results over the entire test set.

Precision and Recall quantify the accuracy and completeness of the predicted foreground regions. and are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The Dice coefficient, also known as the Sørensen-Dice index, evaluates the overlap between the predicted and ground truth masks and is calculated as

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Intersection over Union (IoU), also referred to as the Jaccard index, measures the ratio between the intersection and the union of the predicted and ground truth regions and can be mathematically described as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}.$$

In the above equations, TP denotes the number of correctly classified foreground pixels, FP represents the number of background pixels incorrectly classified as foreground, and FN denotes the number of foreground pixels incorrectly classified as background.

In addition to region-based metrics, the Hausdorff Distance (HD) was used to evaluate the boundary accuracy of the segmentation results.

Formally, the Hausdorff Distance between two point sets P and G, representing the boundaries of the predicted and ground truth masks, is defined as

$$HD(P,G)=\max\left(\sup_{p\in P}\inf_{g\in G}p-g,\ \sup_{g\in G}\inf_{p\in P}g-p\right).$$

Since the standard Hausdorff Distance is highly sensitive to outliers, the 95th percentile of the Hausdorff Distance (HD95) was used in this study. HD95 measures the maximum distance between the predicted and ground truth boundaries while excluding extreme deviations and therefore provides a more robust estimation of contour alignment.

5. Results of the enhanced cross-domain transfer learning approach

5. 1. Description of the implemented segmentation models with transfer learning

Existing deep learning segmentation architectures are enhanced by applying cross-domain transfer learning from ImageNet-pretrained encoders to limited annotated Pap smear cytological images.

To assess the effectiveness of Transfer Learning-based segmentation in Pap smear cytological images under limited annotated data conditions, three different segmentation architectures were selected and implemented.

The segmentation model follows the Encoder-Decoder architecture. The encoder is based on the ResNet-50 convolutional neural network [15] (Fig. 2), which consists of four residual stages producing feature maps at spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input image size, with corresponding channel dimensions of 256, 512, 1024, and 2048. The encoder is initialized with ImageNet-pretrained weights [16], followed by fine-tuning on a dataset of 4,000 Pap smear images to adapt the model to the cytological segmentation task.

The decoder adopts a symmetric U-shaped structure composed of four upsampling blocks, as originally proposed in U-Net [7]. Each block performs bilinear upsampling followed by a 3×3 convolution and concatenation with the corresponding encoder feature map via skip connections. This design allows the model to preserve fine-grained spatial information essential for accurate cell boundary delineation in low-contrast cytological images.

The final segmentation layer consists of a 1×1 convolution followed by a softmax activation function, producing a multi-class segmentation mask with four foreground classe.

The segmentation model consists of a hybrid architecture combining a Vision Transformer (ViT) encoder with a DeepLabV3+ segmentation head (Fig. 3).

The encoder is based on the ViT-B/16 architecture [17], where the input image is partitioned into non-overlapping patches of size 16×16. Each patch is linearly embedded into a 768-dimensional feature vector. The transformer encoder comprises 12 layers with multi-head self-attention using 12 attention heads. The Vision Transformer encoder is initialized with ImageNet-pretrained weights, while the DeepLabV3+ segmentation head is trained from scratch on the Pap smear image dataset.

The segmentation head follows the DeepLabV3+ design [18] and incorporates an Atrous Spatial Pyramid Pooling (ASPP) module with dilation rates of 6, 12, and 18. The ASPP output is upsampled to the original image resolution to produce a dense prediction.

The final output layer applies a 1×1 convolution followed by a softmax activation function, generating a multi-class segmentation mask with four classes.

The hierarchical transformer encoder and the MLP-based decoder of the SegFormer model are illustrated in Fig. 4 [19].

SegFormer is a fully transformer-based semantic segmentation model composed of a hierarchical transformer encoder (Mix Transformer backbone) and a lightweight multilayer perceptron (MLP) decoder. In this study, the SegFormer-B2 variant is used.

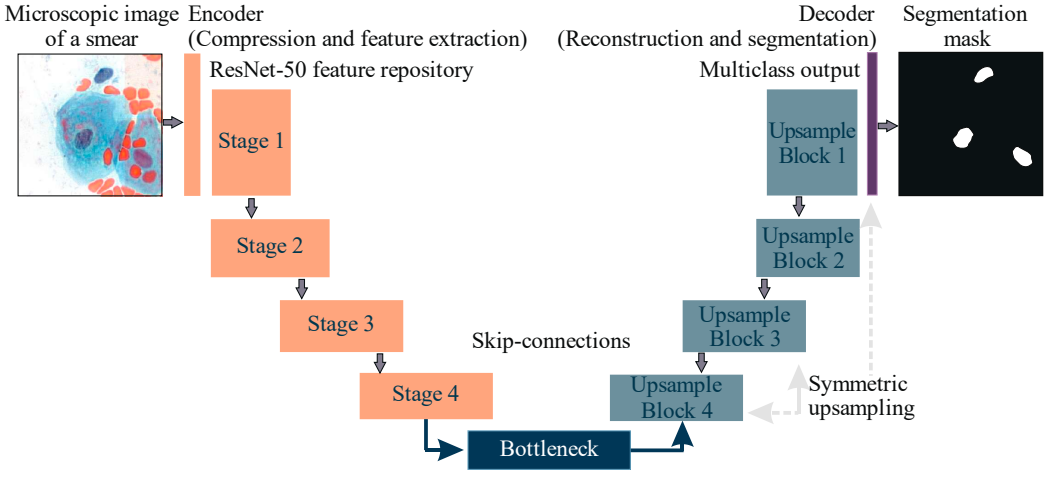


Fig. 2. U-Net architecture with a ResNet-50 encoder pretrained on ImageNet for cytological image segmentation

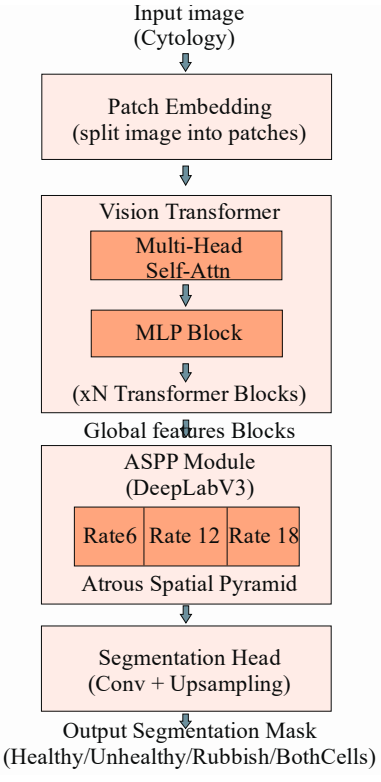


Fig. 3. Hybrid DeepLabV3 model with a Vision Transformer encoder pretrained on ImageNet

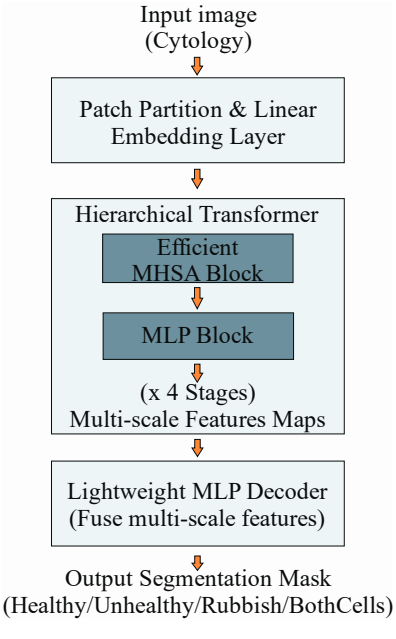


Fig. 4. SegFormer: fully transformer-based architecture for semantic segmentation of cytological images

The encoder consists of four stages that progressively reduce the spatial resolution while increasing the feature dimensionality, with embedding dimensions of 64, 128, 320, and 512, respectively. Each stage employs efficient self-attention mechanisms (without explicit positional encodings) to extract multi-scale feature representations.

The decoder is implemented as a lightweight multilayer perceptron that aggregates multi-scale features from all encoder stages without the use of explicit upsampling or convolutional operations.

The final segmentation layer applies a 1×1 convolution followed by a softmax activation function, producing a multi-class segmentation mask with four classes.

The SegFormer encoder is initialized with ImageNet-pretrained weights [19], followed by fine-tuning on the Pap smear image dataset.

5. 2. The impact of data augmentation techniques on model performance

In this section, let's evaluate the segmentation performance of the models without data augmentation. Quantitative results without augmentation are presented in Table 2.

Table 2

Quantitative segmentation results without augmentation

Model	Class	Dice	IoU	Precision	Recall	HD95
U-Net	Healthy	0.78	0.68	0.76	0.80	14.2
U-Net	Unhealthy	0.74	0.64	0.73	0.76	18.7
U-Net	Rubbish	0.71	0.61	0.70	0.72	20.1
U-Net	BothCells	0.73	0.63	0.71	0.75	17.9
DeepLabV3 + ViT	Healthy	0.88	0.80	0.87	0.90	10.6
DeepLabV3 + ViT	Unhealthy	0.85	0.77	0.84	0.87	13.2
DeepLabV3 + ViT	Rubbish	0.81	0.73	0.80	0.83	14.8
DeepLabV3 + ViT	BothCells	0.83	0.75	0.82	0.84	12.9
SegFormer	Healthy	0.92	0.86	0.91	0.93	8.4
SegFormer	Unhealthy	0.89	0.82	0.88	0.90	10.1
SegFormer	Rubbish	0.85	0.77	0.84	0.86	11.6
SegFormer	BothCells	0.87	0.79	0.86	0.88	9.8

The SegFormer model demonstrates the best baseline performance across all metrics and classes, with Dice coefficients exceeding 0.85 and low HD95 values. The DeepLabV3 + ViT and U-Net models perform slightly worse, especially when identifying the Rubbish and BothCells classes.

To assess, how each architecture responds to increased data variability, all three models were retrained using a data augmentation pipeline. The resulting quantitative metrics are reported in Table 3.

Table 3

Quantitative segmentation results with data augmentation

Model	Class	Dice	IoU	Precision	Recall	HD95
U-Net	Healthy	0.79	0.69	0.75	0.83	15.1
U-Net	Unhealthy	0.69	0.59	0.66	0.74	22.4
U-Net	Rubbish	0.66	0.56	0.64	0.70	24.0
U-Net	BothCells	0.68	0.58	0.65	0.72	21.2
DeepLabV3 + ViT	Healthy	0.92	0.85	0.91	0.94	7.9
DeepLabV3 + ViT	Unhealthy	0.89	0.82	0.88	0.91	9.6
DeepLabV3 + ViT	Rubbish	0.85	0.78	0.84	0.87	11.2
DeepLabV3 + ViT	BothCells	0.88	0.81	0.87	0.89	9.1
SegFormer	Healthy	0.94	0.88	0.93	0.95	7.6
SegFormer	Unhealthy	0.90	0.84	0.89	0.92	9.4
SegFormer	Rubbish	0.86	0.79	0.85	0.87	11.0
SegFormer	BothCells	0.88	0.81	0.87	0.89	9.2

The quantitative results in Tables 2, 3 illustrate that data augmentation affects segmentation models in different ways. The classical U-Net demonstrates limited robustness to augmented data. While a slight improvement is observed for the Healthy class, the performance for Unhealthy, Rubbish, and BothCells declines. This trend is reflected not only in Dice and IoU scores but also in the Hausdorff Distance. The increase in HD95 values indicates less accurate object boundaries and a higher sensitivity to image distortions.

DeepLabV3 + ViT shows a clear and consistent improvement after data augmentation. Dice and IoU scores increase across all classes. At the same time, HD95 values decrease substantially, which indicates more precise boundary localization. The strongest improvements are observed for the Unhealthy and BothCells classes. This behavior suggests that the combination of multi-scale atrous convolutions and global context modelling enables better adaptation to increased data variability.

SegFormer achieves the best overall performance in both experimental settings. Without augmentation, it already produces low Hausdorff Distance values, which indicates accurate contour prediction. After augmentation, only minor improvements are observed in Dice, IoU, and HD95. This suggests that SegFormer has strong inherent generalization capability, and data augmentation mainly improves robustness rather than absolute accuracy.

5. 3. Qualitative evaluation of segmentation results

To complement the quantitative analysis, a visual inspection of model predictions was performed on representative images from each class. The qualitative comparison is presented in Fig. 5. In each subfigure, images are arranged from left to right as follows: original image, ground truth, U-Net (ResNet50), DeepLabV3 (ViT), SegFormer.

The visual examples displayed in Fig. 5 are in line with the quantitative results. In our experiments, U-Net tends to produce blurred object boundaries and a noticeable number of false positives, particularly in noisy regions assigned to the Rubbish class. DeepLabV3 + ViT captures the overall cellular structure more reliably, and it is observed fewer spurious detections compared to U-Net. However, in cases where cells overlap, the model occasionally fails to resolve fine structural details, which affects the BothCells class.

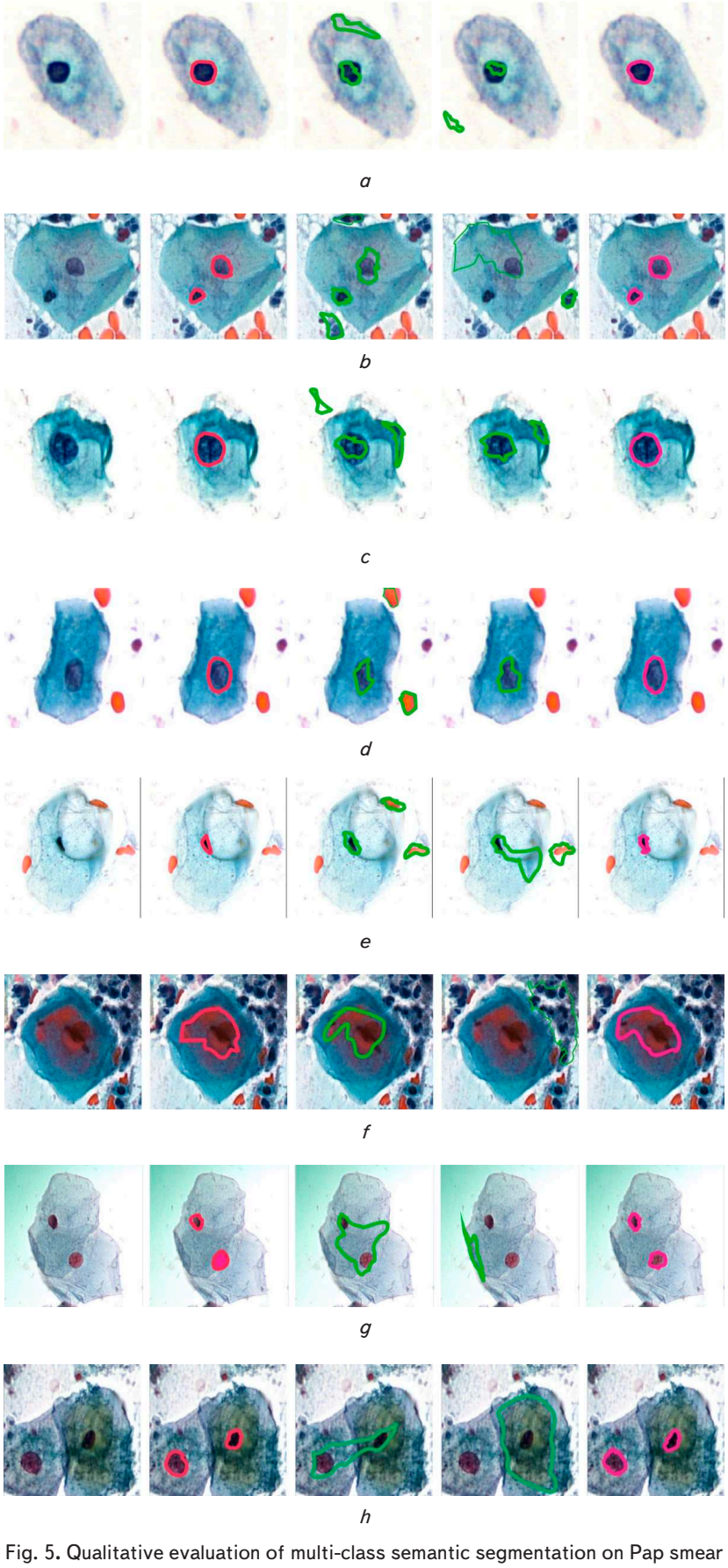


Fig. 5. Qualitative evaluation of multi-class semantic segmentation on Pap smear images: *a, b* — healthy cells; *c, d* — unhealthy cells; *e, f* — debris (rubbish); *g, h* — bothcells

6. Analysis of model performance and architectural factors affecting segmentation accuracy

The results of experiments confirm the effectiveness of the enhanced cross-domain transfer learning approach for multiclass segmentation of Pap smear cytological images under limited annotated data conditions. Without data augmentation (Table 2), SegFormer already demonstrates the highest Dice coefficients (0.92–0.85 across classes) and the lowest HD95 values. After augmentation (Table 3), SegFormer further improves to Dice values of 0.94 (Healthy), 0.90 (Unhealthy), 0.86 (Rubbish), and 0.88 (BothCells), with HD95 reduced to 7.6–11.0 pixels. DeepLabV3 + ViT also shows consistent gains, particularly on Unhealthy and BothCells classes (Dice increase from 0.85 to 0.89 and from 0.83 to 0.88, respectively; HD95 decrease from 13.2 to 9.6 and from 12.9 to 9.1). In contrast, U-Net with ResNet-50 encoder exhibits performance degradation on complex classes after augmentation.

The qualitative results presented in Fig. 5 confirm and visually illustrate these quantitative findings. The figure displays representative examples for each class, with columns showing the original image, ground truth mask, U-Net prediction, DeepLabV3 + ViT prediction, and SegFormer prediction. SegFormer consistently provides the most accurate and stable cell boundaries, effectively resolving overlapping structures in the BothCells class and minimizing fragmentation and false positives in low-contrast Unhealthy regions and noisy Rubbish areas. In contrast, U-Net frequently produces blurred contours, over-segmentation in debris regions, and higher error rates in overlapping cells. DeepLabV3 + ViT captures overall cellular structure more reliably than U-Net with fewer spurious detections, but occasionally fails to resolve fine details in overlapping regions. These visual differences highlight the advantages of hierarchical multi-scale feature extraction and lightweight MLP decoder in SegFormer, as well as the global context modeling provided by Transformer encoders.

Primarily, the obtained results are explained by the specific architectural features of the models and their ability to capture context at different scales. As shown in Tables 2, 3, the U-Net model with the ResNet-50 encoder is less accurate in segmenting Unhealthy, Rubbish, and BothCells images due to the lack of a built-in mechanism for global context and long-term dependencies – a critical limitation when dealing with cytological images characterized by high noise, low contrast, cell overlap, and significant shape variability.

Unlike the classic U-Net, DeepLabV3 + ViT combines a transformer encoder with ASPP, resulting in a significant leap in segmentation performance. Global features enable better understanding of the entire image, leading to increased Dice and IoU scores across all classes (Table 2), with particularly noticeable improvements for Healthy and Unhealthy cells.

SegFormer takes this approach further by employing a hierarchical transformer encoder and multi-scale feature aggregation combined with a lightweight MLP decoder instead of a traditional convolutional one. This design allows for more accurate cell boundary detection and reduced errors in areas of ambiguity or overlap, explaining why SegFormer's Dice metrics are significantly higher than those of the other models (Tables 2, 3).

The results obtained in this study are consistent with and extend those presented in related works on medical image segmentation. Classical U-Net-based approaches [7] have demonstrated high baseline performance on biomedical segmentation tasks; however, they are often limited in capturing long-term dependencies. Our results confirm that Transformer architectures can address this issue: the DeepLabV3 + ViT configuration achieves a Dice score of 0.90 for the Unhealthy class, while SegFormer, originally proposed in [19] for natural image segmentation, extends its state-of-the-art performance to cytological cell segmentation with an average Dice score of 0.89–0.90 (Table 3).

Compared with existing methods for cell segmentation in Pap smear images, classical U-Net-based models [7] and their modifications, including Attention U-Net [8], UNETR [9], TransUNet [10], and architectures similar to Swin-UNet [11], typically achieve Dice scores ranging from 0.90 to 0.96 for nuclei or cytoplasm segmentation. Such performance is more common in two-class tasks or when using additional techniques such as hybrid blocks or pretraining on large datasets. In this study, comparable accuracy (average Dice 0.89 across all classes and 0.90 for the Unhealthy class using SegFormer) was obtained in the context of full multiclass segmentation without ensembles, specialized augmentations, or multi-stage hybrid constructions.

Unlike fully convolutional architectures [7], which exhibit noticeable metric reductions when working with complex objects (abnormal cells, occlusions, artifacts, low-contrast areas), transformer-based solutions – in particular ViT in DeepLabV3+ and the hierarchical transformer in SegFormer [19] – significantly improve the capture of global context and long-term spatial dependencies. As a result, substantial increases in Dice and IoU are observed compared to the baseline U-Net model with ResNet-50 encoder, particularly for classes with high shape and boundary variability (Tables 2, 3).

The main limitations of the study include the use of a single dataset, the absence of cross-validation, a fixed augmentation strategy, and evaluation under a single set of imaging conditions and staining protocols. Additional methodological constraints involve a fixed training configuration, lack of systematic hyperparameter optimization, and absence of detailed ablation studies to isolate individual architectural contributions. The evaluation was conducted under a single experimental protocol without cross-dataset validation, which limits the robustness analysis of the proposed approach. These constraints may be addressed in future research through structured ablation studies, automated hyperparameter search, and multi-dataset benchmarking.

Future developments include integration of multi-task learning (segmentation + classification), weakly and semi-supervised approaches to reduce labeling dependency, testing on multiple external datasets (Herlev, SIPaKMeD, etc.), and clinical validation. Potential challenges may include overfitting of transformers on small datasets and significant domain shift between different staining protocols and scanners.

7. Conclusion

1. The applicability of ImageNet pre-trained models for multiclass segmentation of Pap smear images was confirmed. It was demonstrated that transfer learning significantly improves segmentation quality compared to training from scratch, with SegFormer achieving Dice scores of 0.92, 0.89,

0.85, and 0.87 for the Healthy, Unhealthy, Rubbish, and BothCells classes, respectively, while the baseline U-Net with a ResNet-50 encoder achieved 0.78, 0.74, 0.71, and 0.73. Unlike classical CNN-based approaches, transformer-based encoders showed better adaptation to cytological data with high noise, low contrast, and overlapping structures.

2. The influence of domain-specific data augmentation on segmentation performance in cytological images was systematically evaluated. The results showed that augmentation improves the performance of transformer-based models, particularly for complex classes, where Dice for DeepLabV3 + ViT increased from 0.85 to 0.89 for the Unhealthy class and from 0.83 to 0.88 for BothCells, while HD95 decreased from 13.2 to 9.6 and from 12.9 to 9.1, respectively. In contrast, the U-Net model demonstrated performance degradation under augmentation for several classes. This indicates that augmentation effectiveness depends on architectural properties and is most beneficial for models with multi-scale feature aggregation and global contextual modeling.

3. A comparative analysis of fine-tuning strategies and model architectures (U-Net with ResNet-50, DeepLabV3 + ViT, and SegFormer) on small cytology datasets showed that transformer-based and hybrid architectures provide superior segmentation accuracy and boundary localization. After augmentation, SegFormer achieved the best overall performance, with Dice values of 0.94, 0.90, 0.86, and 0.88 and the lowest HD95 values of 7.6, 9.4, 11.0, and 9.2 for Healthy, Unhealthy, Rubbish, and BothCells classes, respectively. In comparison with conventional CNN models reported in related studies, the proposed fine-tuning strategy demonstrates improved robustness on noisy and morphologically heterogeneous cytological images. Such advantage is explained by hierarchical multi-scale feature extraction and efficient global context modeling using a lightweight decoder structure.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The article was prepared within the framework of realization of the grant «Saginov Jas Zertteyshileri» of NPJSC «Karaganda Technical University named after Abylkas Saginov» for young scientists, agreement №6 from 27.06.2025.

Data availability

Data cannot be made available for reasons disclosed in the data availability statement.

Use of artificial intelligence

In this study, the Claude Sonnet 4.5 model (Anthropic) was employed to support grammatical and stylistic corrections in the Introduction and partially in the Literature Review sections. The AI tool was used specifically to suggest improvements in sentence structure, clarity, and language consistency. All suggested edits were carefully reviewed and manually approved by the authors to ensure accuracy and maintain the intended scientific meaning. The AI-assisted corrections served solely as a language support tool and did not influence the research methodology, data analysis, or final conclusions of the study.

Authors' contributions

Margulan Nurtay: Conceptualization; Methodology; Investigation; Software; Data curation; Formal analysis; Writing – original draft; **Gaukhar Alina**: Methodology; Validation; Formal analysis; Visualization; Writing – review & editing; **Ardak Tau**: Supervision; Project administration; Resources; Writing – review & editing.

References

1. Mustafa, W. A., Ismail, S., Mokhtar, F. S., Alquran, H., Al-Issa, Y. (2023). Cervical Cancer Detection Techniques: A Chronological Review. Diagnostics, 13 (10), 1763. https://doi.org/10.3390/diagnostics13101763
2. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M. et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005
3. Fang, M., Liao, B., Lei, X., Wu, F.-X. (2024). A systematic review on deep learning based methods for cervical cell image analysis. Neurocomputing, 610, 128630. https://doi.org/10.1016/j.neucom.2024.128630
4. Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S. et al. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging 2014: Digital Pathology, 9041, 904103. https://doi.org/10.1117/12.2043872
5. Pati, P., Jaume, G., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A. M., Scognamiglio, G. et al. (2022). Hierarchical graph representations in digital pathology. Medical Image Analysis, 75, 102264. https://doi.org/10.1016/j.media.2021.102264
6. Win, K. Y., Choomchuay, S. (2017). Automated segmentation of cell nuclei in cytology pleural fluid images using OTSU thresholding. 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), 14–18. https://doi.org/10.1109/icdamt.2017.7904925
7. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
8. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K. et al. (2018). Attention U-Net: Learning Where to Look for the Pancreas. arXiv. https://doi.org/10.48550/arXiv.1804.03999

9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B. et al. (2022). UNETR: Transformers for 3D Medical Image Segmentation. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1748–1758. https://doi.org/10.1109/wacv51458.2022.00181
10. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y. et al. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv. https://doi.org/10.48550/arXiv.2102.04306
11. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D. et al. (2021). Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 548–558. https://doi.org/10.1109/iccv48922.2021.00061
12. Xu, C., Li, M., Li, G., Zhang, Y., Sun, C., Bai, N. (2022). Cervical Cell/Clumps Detection in Cytology Images Using Transfer Learning. Diagnostics, 12 (10), 2477. https://doi.org/10.3390/diagnostics12102477
13. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. arXiv. https://doi.org/10.48550/arXiv.1902.07208
14. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35 (5), 1299–1312. https://doi.org/10.1109/tmi.2016.2535302
15. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/cvpr.2016.90
16. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115 (3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929
18. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Computer Vision – ECCV 2018, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
19. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv. https://doi.org/10.48550/arXiv.2105.15203