Embedding Analysis for Assisted Segmentation of Cervical Cytology

Mario Alejandro García[1] and Martín Nicolás Gramática[1]

Universidad Tecnológica Nacional, Facultado Regional Córdoba, GIA
mgarcia@frc.utn.edu.ar
https://www.investigacion.frc.utn.edu.ar/gia/

Abstract. Pap test image segmentation is an important step in the development of computer-aided systems for the early diagnosis of cervical cancer. In this work, we study the quality of embeddings from different depth levels in three neural network architectures by applying them to nucleus and cytoplasm segmentation. We conclude that higher-resolution embeddings (width and height) and those relying on local rather than global information tend to achieve better performance. The analyzed models are ResNet-18, BiomedCLIP, and Swin Transformer (SwinT). SwinT achieved the best performance, whereas BiomedCLIP, despite being trained on medical data, performed the worst.

Keywords: cervical cytology , digital pathology , segmentation , deep learning

Análisis de embeddings para segmentación asistida de citología cervical

Resumen La segmentación de imágenes de Pap test es un paso importante para el desarrollo de un sistema de asistencia al diagnóstico temprano de cáncer de cuello uterino. En este trabajo se estudia la calidad de los embeddings de diferentes niveles de profundidad en tres arquitecturas de redes neuronales mediante su aplicación a la tarea de segmentación de núcleos y citoplasmas. Se concluye que los embeddings con mayor definición (ancho y alto) y que utilizan información local en lugar de global parecen lograr un mayor rendimiento. Los modelos analizados son ResNet-18, BiomedCLIP y Swin Transformer (SwinT). SwinT fue el de mejor rendimiento y BiomedCLIP, a pesar de ser entrenado con datos médicos, el peor

Keywords: citología cervical, patología difital clave, segmentación , aprendizaje profundo

1 Introduction

Cervical cancer is the fourth most frequent cancer among women worldwide. Low- and middle-income countries show the highest incidence and mortality rates due to limited access to effective screening methods (Bray et al., 2024).

The Papanicolaou test (Pap test) is the most widely used method for cervical cancer screening. It consists of collecting, staining, and microscopically analyzing cells from the cervix. Its main purpose is to detect cells with certain types of abnormalities, which enables the early detection of Human Papillomavirus (HPV) and the implementation of preventive treatment (Sankaranarayanan et al., 2001). Visual abnormalities in cells mainly involve changes in the shapes and relative sizes of nuclei and cytoplasms.

An automatic or semi-automatic Pap test image analysis tool could improve the quality and efficiency of the pathologist's work. The development of such a tool faces both general artificial intelligence (AI) challenges in medical practice and task-specific difficulties. The common challenges are mainly the need for explainability and for multicenter datasets annotated by experts (Haug & Drazen, 2023; McGenity et al., 2024; Topol, 2019; Van der Laak et al., 2021).

The use of AI for Pap test analysis is an active research topic (Abrar et al., 2025). The most frequent approaches are whole-image classification, one-stage cell detection and classification, two-stage cell detection followed by classification, and instance segmentation for subsequent classification. Although the general trend is the use of deep learning, which is often advantageous in end-to-end models, the two-stage separation improves explainability and does not necessarily reduce accuracy (Gramática et al., 2026). In that line of work, a group of studies focuses on segmenting cells, or cytoplasms and nuclei, in order to classify them using simple rules based on morphological features, such as the relationship between nucleus and cytoplasm size. Representative examples are discussed below.

Harangi *et al.* present the APACS23 dataset with 37,000 manually segmented cells, on which they perform automatic segmentation using dense (*fully connected*) neural networks (Harangi et al., 2024).

Wubineh and Rusiecki propose a two-stage approach to detect abnormal cells. Segmentation is performed with SPP-SegNet and classification with SE-DenseNet201, both models proposed by the authors (Wubineh et al., 2025).

In the two previous cases, complete cells are segmented without distinguishing cytoplasm from nucleus, which is insufficient for subsequent classification. By contrast, Zhang *et al.* segment only nuclei in cervical cytology using a *Global Context* U-Net based on a DenseNet *backbone* (B. Zhang et al., 2024). According to the systematic review by Rasheed *et al.*, there are at least 78 papers since 2010 aimed at nucleus segmentation in Pap test images (Rasheed et al., 2025).

Aaseegha and Venkataramana propose a hybrid model combining SeNet, ResNet-50, and DeiT (*data-efficient image transformer*) to segment nuclei and cytoplasms (Aaseegha & Venkataramana, 2025).

The models proposed in the papers mentioned so far must be trained, which implies the need for large, annotated, multicenter datasets. Other approaches

use "black-box" models, such as that of Almohimeed *et al.*, who propose ViT-PSO-SVM, integrating ViT (the *vision transformer* introduced by Google (Dosovitskiy et al., 2020)) to create features, *particle swarm optimization* (PSO) to reduce dimensionality, and SVM to classify cells. Although it is not an end-to-end model, it also lacks segmentation and none of its stages provides explainability (AlMohimeed et al., 2024).

Despite some promising metrics, we are not aware of satisfactory solutions either for assisted diagnosis or for Pap test segmentation. We have also found little justification in the reviewed literature for the choice of the neural architectures that were used.

1.1 Objective

This work is part of an ongoing research process that seeks to develop techniques to support Pap test analysis.

Segmenting nuclei and cytoplasms is a key step within the defined process. To achieve this, we propose starting from feature embeddings generated by neural networks.

The objective of this work is to improve our understanding of the information provided by different neural architectures, at different depths, when they are used as encoders for Pap test images.

Neural networks used in computer vision recognize simple patterns in early layers and complex patterns, formed as combinations of simpler ones, in higher layers. It is not clear at which level or levels of complexity the useful patterns for recognizing nuclei and cytoplasms in cytology are found. Useful patterns might include textures, edges, or complex shapes. Therefore, not only the choice of model is important, but also the level from which the embedding is taken.

2 Materials and Methods

During the experiments, a set of embeddings is defined, an assisted segmentation process is carried out based on each embedding in the set, and the results are compared.

2.1 Data

To our knowledge, no public cervical cytology dataset simultaneously (1) contain separate nucleus and cytoplasm annotations and (2) provide annotations for all cells appearing in the image.

In addition, Pap test images very frequently contain elements that are visually similar to epithelial cell nuclei, but smaller and with very little cytoplasm, which gives them the appearance of "isolated nuclei." These elements are usually leukocytes and are not annotated in the available datasets. Although the objective of this work is not to recognize "isolated nuclei," it is necessary to account for their existence when evaluating the results. At this point in the pipeline of

the segmentation process, it is expected that these elements will be considered nuclei rather than cytoplasm or background. They can later be distinguished from true nuclei through additional processing.

The data used consist of ten images taken from the CCEDD dataset (J. Liu et al., 2022). CCEDD was annotated by experts and separates nuclei from cytoplasms, but not all epithelial cells are annotated and the "isolated nuclei" are not annotated either. To address these limitations, the project members manually annotated the missing elements as "unconfirmed cytoplasms," "unconfirmed nuclei," and "isolated nuclei." Finally, "unconfirmed nuclei" were treated as nuclei and "unconfirmed cytoplasms" as cytoplasms, while "isolated nuclei" remained in their own category until the evaluation stage. Figure 1 shows one image with all annotations.

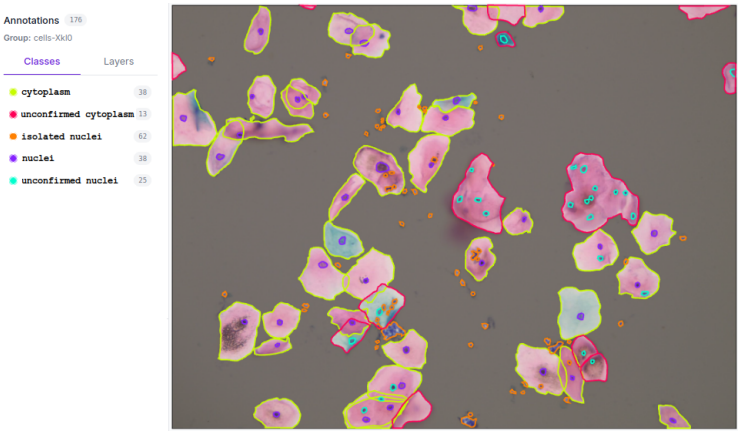

Fig. 1. Annotated image from the CCEDD dataset after manual completion of missing labels. Tool: Roboflow (https://universe.roboflow.com/)

.

2.2 Models

Three neural network architectures representing different design paradigms were used as feature extractors: (1) ResNet-18 (a relatively small CNN), (2) the ViT that serves as the encoder in BiomedCLIP, and (3) SwinT, a transformer with a hierarchical structure that, like CNNs, resembles the hierarchical processing of the visual cortex.

The ResNet-18 and SwinT models used were trained on ImageNet, whereas BiomedCLIP was trained on medical data, such as microscope images and radiographs.

A (He et al., 2016; Z. Liu et al., 2021; S. Zhang et al., 2023)

2.3 Embeddings

For each model, several intermediate layers were selected, whose activations are the embeddings used in the experiment. The architectures of the three models have layers grouped into blocks. Initially, for each network, the behavior of the early layers, the outputs of the blocks, and some internal block layers was analyzed. For simplicity, the internal block layers are not shown in the results, since in no case do they outperform the final output of the block. Table 1 contains the selected embeddings and their sizes for an input image of 2016×1344 pixels.

Table 1. Embeddings. Source layer or block and embedding size (height $\times$ width $\times$ channels) for an image of 2016×1344 pixels.

ResNet-18		**BiomedCLIP**		**SwinT**	
Layer	Size	Layer	Size	Layer	Size
1st Conv.	$672 \times 1008 \times 64$	Conv.	$84 \times 126 \times 768$	Conv.	$336 \times 504 \times 96$
1st ReLU	$672 \times 1008 \times 64$	Bl. 0	$84 \times 126 \times 768$	Bl. 3-4	$336 \times 504 \times 96$
Block 1	$336 \times 504 \times 64$	Bl. 1	$84 \times 126 \times 768$	Bl. 3-5	$336 \times 504 \times 96$
Block 2	$168 \times 252 \times 128$	Bl. 2	$84 \times 126 \times 768$	Bl. 3-8	$168 \times 252 \times 192$
Block 3	$84 \times 126 \times 256$	Bl. 3	$84 \times 126 \times 768$	Bl. 3-9	$168 \times 252 \times 192$
Block 4	$42 \times 63 \times 512$	Bl. 4	$84 \times 126 \times 768$	Bl. 3-12	$84 \times 126 \times 384$
		Bl. 5	$84 \times 126 \times 768$	Bl. 3-13	$84 \times 126 \times 384$
		Bl. 6	$84 \times 126 \times 768$	Bl. 3-14	$84 \times 126 \times 384$
		Bl. 7	$84 \times 126 \times 768$	Bl. 3-15	$84 \times 126 \times 384$
		Bl. 8	$84 \times 126 \times 768$	Bl. 3-16	$84 \times 126 \times 384$
		Bl. 9	$84 \times 126 \times 768$	Bl. 3-17	$84 \times 126 \times 384$
		Bl. 10	$84 \times 126 \times 768$	Bl. 3-20	$42 \times 63 \times 768$
		Bl. 11	$84 \times 126 \times 768$	Bl. 3-21	$42 \times 63 \times 768$

The input size of the three models is $224 \times 224 \times 3$. To process larger images without resizing them, the image is divided into 224×224 pixel patches, embeddings are computed independently for each patch, and they are then combined while preserving the original location of each patch.

Patch-based processing, at least when performed without overlap (that is, with $stride = 224 \times 224$), favors the appearance of artifacts. These artifacts, generally grid-shaped, originate from (1) padding in CNNs and (2) the non-uniform distribution of attention weights (edges vs. center) in transformers. According to our observations, the effect of artifacts is stronger in transformers. Figure 2 shows an example for BiomedCLIP.

The artifacts in Fig. 2 are reduced by decreasing the *stride* and taking only the central region of the embedding for each patch position. As a compromise between artifact magnitude and efficiency, this work uses a 50% overlap, that is, $stride = 122 \times 122$. Figure 3 graphically shows the process for an image of 448×448 pixels and an activation of size $8 \times 8 \times ch$ for each patch. In this example, the resulting embedding size would be $16 \times 16 \times ch$.

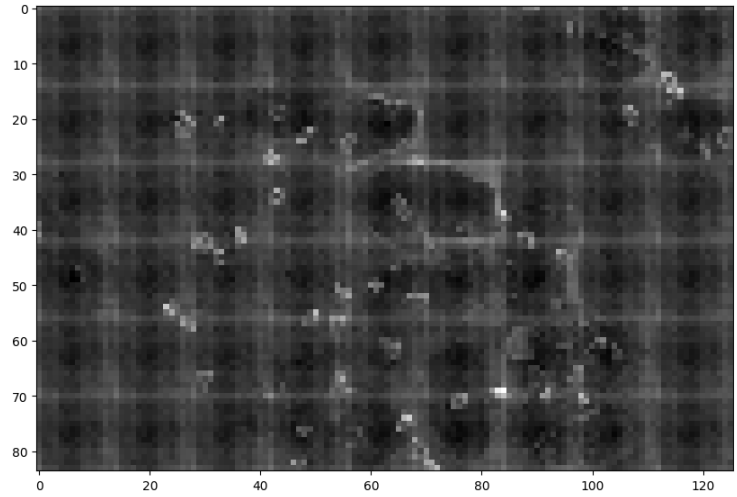

Fig. 2. Norm of the output of BiomedCLIP block 1 with *stride* 224×224 for patch extraction.

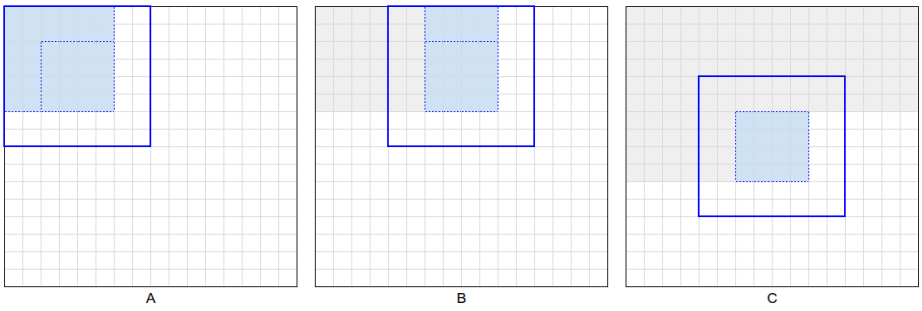

Fig. 3. Computation of an embedding for the full image. Grids A, B, and C represent the embedding at different times. In A, the solid blue square shows the position of the activation and the light blue filled region indicates the part of the activation that is taken to form the embedding. In B, the activation moves four positions (50%) to the right and completes the light blue region. The gray area shows the region loaded in previous steps. In C, the 5th step is shown. Note that in A and B, besides the central square, the borders are also taken.

2.4 Segmentation

The $1 \times 1 \times ch$ elements of the embedding (each small cell in the grids of Fig. 3) are denoted by e_{ij}. Each e_{ij} represents a set of pixels in the input image. For example, if the image size is reduced by half in width and height in the embedding, each e_{ij} will represent a set of 4 pixels.

In this experiment, assisted segmentation is performed, where some examples of each class must be known so that the method can classify the rest. Specifically, the coordinates of 50 pixels from each class (cytoplasm, nucleus, and background) are used.

The procedure is as follows. First, one cytoplasm and one nucleus are randomly selected (belonging or not to the same cell) from the original dataset annotations. From the polygons of the selected annotations, masks M_c and M_n are created for cytoplasm and nucleus, respectively. Subsequently, the regions of M_c that overlap with nucleus or isolated-nucleus annotations are removed. A mask M_f is also created for the background as the negation of the union of all annotations. Once the three masks are created, 50 pixels are randomly selected from each one, defining three sets of coordinates, one for each class. These sets are then projected onto the embedding through a scale transformation, and each element is associated with the corresponding e_{ij}, obtaining a set of reference vectors for each class. In the next step, the class mean of the e_{ij} vectors is computed: $\overline{e_c}$, $\overline{e_n}$, and $\overline{e_f}$. Finally, the cosine distance between each e_{ij} in the embedding and the means representing the classes is calculated. For each e_{ij}, the smallest of the three distances is selected, the corresponding class is assigned, and the coordinates are projected back to the original image, labeling the pixels covered by the projection with the assigned class. Figure 4 shows one segmentation example before projection to the image size.

2.5 Evaluation and Metrics

The obtained segmentation is compared with the expected output for the corresponding image. Evaluation is performed on a pixel-wise basis. The number of compared pixels is the number of pixels in the image minus the number chosen as reference, $2016 \times 1344 - 150 = 2709504$.

Table 2 shows an example of the confusion matrix. Predictions of the *nucleus* class for *isolated nuclei* pixels are considered correct.

Since the metrics originally used were defined for classification or binary segmentation, and the confusion matrix is not square due to the merging of the *nucleus* and *isolated nucleus* classes, the metrics must be redefined. Table 3 shows the names used for each position of the confusion matrix.

The first metric is *accuracy*, defined as

$$Accuracy = \frac{c\_c + n\_n + ns\_n + f\_f}{C + N + NS + F}.$$

The *cytoplasm*, (*nucleus* + *isolated nucleus*), and *background* classes are not balanced. On the one hand, background usually covers more area than cytoplasm, although this may change in images with many cells. On the other hand,

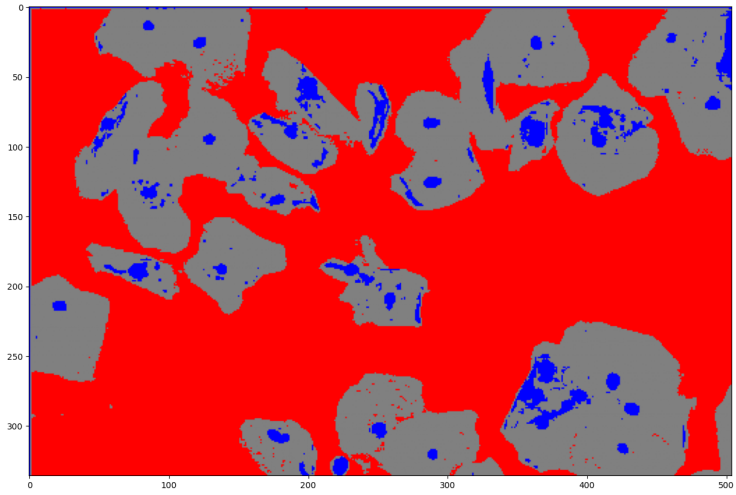

Fig. 4. Segmentation of the *Block 3-4* embedding of the SwinT model for image 0. Gray, blue, and red correspond to the *cytoplasm*, *nucleus*, and *background* classes, respectively.

Table 2. Confusion matrix for image 0, BiomedCLIP *Conv.* embedding.

		Predicted			
		Cytoplasm	**Nucleus**	**Background**	**Total**
True	**Cytoplasm**	906991	157958	48725	1113674
	Nucleus	3708	28897	0	32605
	Isol. N.	557	1665	147	2369
	Background	75056	1388	1484262	1560706

Table 3. Confusion matrix notation for metric definitions.

		Predicted			
		Cytoplasm	**Nucleus**	**Background**	**Total**
True	**Cytoplasm**	c_c	c_n	c_f	C
	Nucleus	n_c	n_n	n_f	N
	Isol. N.	ns_c	ns_n	ns_f	NS
	Background	f_c	f_n	f_f	F
	Total	$\_C$	$\_N$	$\_F$	

the area of nuclei is always smaller than that of cytoplasms and background. To avoid favoring models biased toward majority classes, the *balanced accuracy* metric is also used, defined as

$$Balanced\,Accuracy = \frac{1}{3}\left(\frac{c\_c}{C} + \frac{n\_n + ns\_n}{N + NS} + \frac{f\_f}{F}\right).$$

Finally, *Intersection-over-Union* (IoU), also known as the *Jaccard index*, is computed. In this metric, each class contributes the ratio between the number of correctly classified pixels and the number of pixels that truly belong to the class or that have been classified as if they belonged to it; therefore, false negatives and false positives for the class are included in the denominator. Adapted to Table 3, it is computed as

$$IoU = \frac{1}{3}\left(\frac{c\_c}{U_C} + \frac{n\_n + ns\_n}{U_N} + \frac{f\_f}{U_F}\right)$$

where

$$\begin{aligned} U_C &= C + \_C - c\_c \\ U_N &= N + NS + \_N - n\_n - ns\_n \\ U_F &= F + \_F - f\_f \end{aligned}$$

3 Results

The results of executing the process described above on the 10 images in the dataset are presented below. The tables and plots in this section show the mean performance for each model/embedding as well as the individual metrics so that variability can be appreciated.

Figures 5, 6, and 7 graphically show the metrics obtained for ResNet-18, BiomedCLIP, and SwinT, respectively. Figure 8 shows, for the same metrics, a comparison of the results from the best layer/block of each model.

4 Discussion

The results show that *accuracy* > *balanced accuracy* > *IoU* consistently for all embeddings. For the first two metrics, this behavior was expected. In the case of *IoU*, a deeper analysis indicates that the low score is mainly due to cytoplasm pixels that were classified as nucleus. Since the true number of nucleus pixels is low, false positives have a strong influence.

An unexpected behavior, for which we do not have an explanation, is the poor performance of the embedding from the convolutional layer of ResNet. Strikingly, it improves after the ReLU, and in the other models the convolutional layer is one of the best points from which to take information.

Considering the best embedding of each model, SwinT was the best model, followed by ResNet-18 and BiomedCLIP. See Fig. 8.

For all models and metrics, the last blocks are a poor choice. Performance worsens toward the output. Two factors could explain this situation: (1) the loss of detail (embedding size) in the last layers and (2) the fact that the information needed for cell segmentation, with several cells per tile, is local.

The two factors in the previous paragraph could explain the poor performance of BiomedCLIP, which from the beginning (except for the convolutional

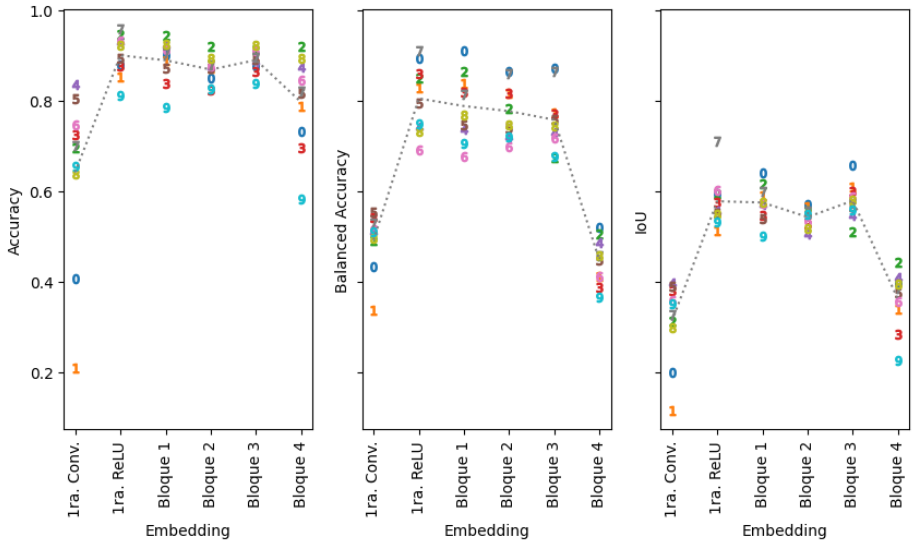

Fig. 5. Metrics by image and embedding for the ResNet-18 model. Each marker is labeled with the corresponding image ID. The dotted lines show the mean for each embedding.

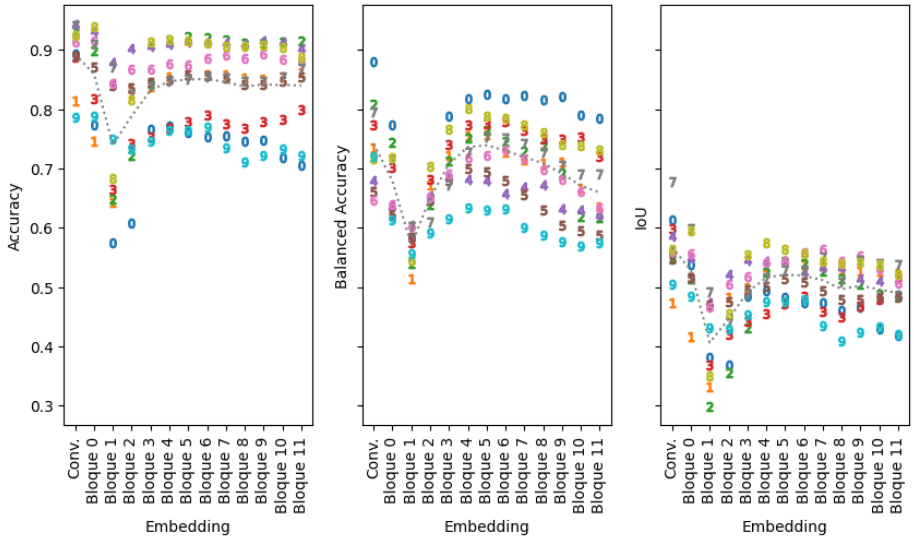

Fig. 6. Metrics by image and embedding for the BiomedCLIP model. The markers are the digit identifiers of the images. The dotted lines show the mean for each embedding.

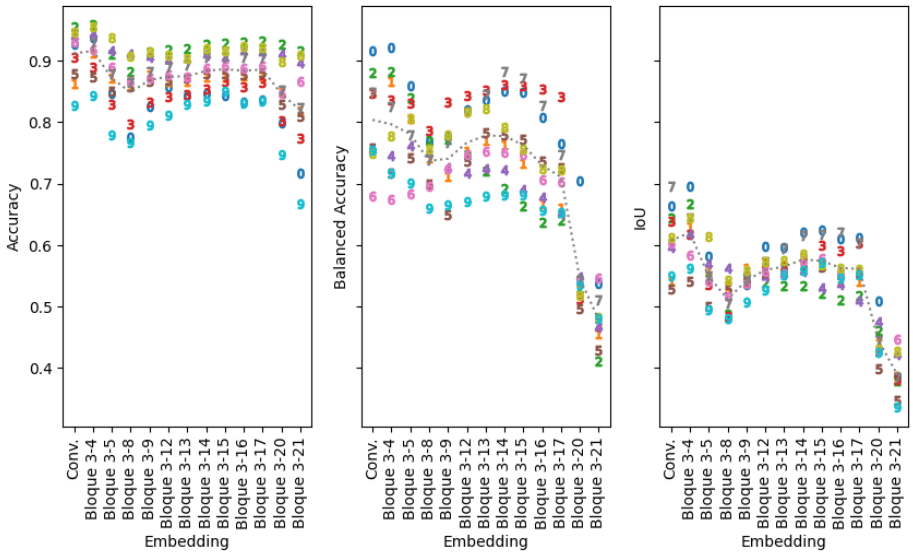

Fig. 7. Metrics by image and embedding for the SwinT model. The markers are the digit identifiers of the images. The dotted lines show the mean for each embedding.

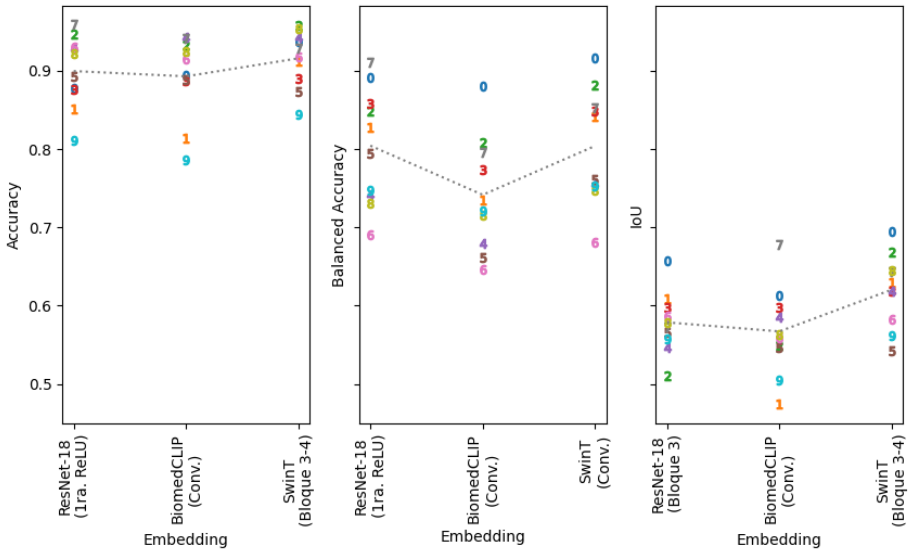

Fig. 8. Metrics by image for the best embedding of each model. The markers are the digit identifiers of the images. The dotted lines show the mean for each embedding.

layer) uses all the information in the tile to compute each embedding position and, moreover, has low resolution in all its layers.

The resolution factor can also be analyzed from another point of view. The three models have embeddings of the same size if channels are not taken into account, 84×126. See Table 1. When performance is compared for that subset, BiomedCLIP still has the worst metrics, which strengthens the hypothesis that local information is needed for this task.

Finally, Figs. 5, 6, and 7 show that the best- and worst-segmented images tend to remain similar in all cases. A visual analysis of these images showed that the visually least clear images were those that achieved low performance.

5 Conclusions

Our results indicate that for all models, the quality of the information obtained to segment nuclei and cytoplasms decreases toward the output of the network. This deterioration is not linear, but it occurs in all three cases.

Three factors are proposed to explain this behavior: the size (width and height) of the embeddings, the use of local versus global information, and the training data.

To determine with confidence the degree to which each factor influences the results, new experiments are needed, since architectural details and training data are not independent in the combination used here. Even so, the results seem to indicate that local information and larger embedding size favor good performance, whereas training on medical data did not provide a significant advantage for BiomedCLIP.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Aaseegha, M. D., & Venkataramana, B. (2025). Segresdeit: A hybrid segnet–resnet-50–deit framework for automated cervical cancer segmentation and classification. *BMC Medical Imaging*, *25*(1), 469. https://doi.org/10.1186/s12880-025-02001-8

Abrar, S. S., Isa, S. A. M., Hairon, S. M., Ismail, M. P., & Ab Kadir, M. N. B. N. (2025). Recent advances in applications of machine learning in cervical cancer research: A focus on prediction models. *Obstetrics & gynecology science*, *68*(4), 247–259.

AlMohimeed, A., Shehata, M., El-Rashidy, N., Mostafa, S., Samy Talaat, A., & Saleh, H. (2024). Vit-pso-svm: Cervical cancer predication based on integrating vision transformer with particle swarm optimization and support vector machine. *Bioengineering*, *11*(7), 729.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gramática, M. N., Pianzola, G., Alfici, H. D., & García, M. A. (2026). Relationship between object detection and image classification stages in pap test. In *Advances in image processing, reliability, and artificial intelligence* (pp. 275–291). Elsevier.

Harangi, B., Bogacsovics, G., Toth, J., Kovacs, I., Dani, E., & Hajdu, A. (2024). Pixel-wise segmentation of cells in digitized pap smear images. *Scientific Data*, *11*(1), 733.

Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. *New England Journal of Medicine*, *388*(13), 1201–1208.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Liu, J., Fan, H., Wang, Q., Li, W., Tang, Y., Wang, D., Zhou, M., & Chen, L. (2022). Local label point correction for edge detection of overlapping cervical cells. *Frontiers in neuroinformatics*, *16*, 895290.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

McGenity, C., Clarke, E. L., Jennings, C., Matthews, G., Cartlidge, C., Freduah-Agyemang, H., Stocken, D. D., & Treanor, D. (2024). Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. *NPJ digital medicine*, *7*(1), 114.

Rasheed, A., Shirazi, S. H., Khan, P., Aseere, A. M., & Shahzad, M. (2025). Techniques and challenges for nuclei segmentation in cervical smear images: A review. *Artificial Intelligence Review*, *58*(10), 295.

Sankaranarayanan, R., Budukh, A. M., & Rajkumar, R. S. (2001). Effective screening programmes for cervical cancer in low- and middle-income developing countries. *Bulletin of the World Health Organization*, *79*(10), 954–962.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature medicine*, *25*(1), 44–56.

Van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: The path to the clinic. *Nature medicine*, *27*(5), 775–784.

Wubineh, B. Z., Rusiecki, A., & Halawa, K. (2025). Spp-segnet and se-densenet201: A dual-model approach for cervical cell segmentation and classification. *Cancers*, *17*(13), 2177.

Zhang, B., Jiang, X., & Zhao, W. (2024). An enhanced mask transformer for overlapping cervical cell segmentation based on detr. *IEEE Access*, *12*, 176586–176597. https://doi.org/10.1109/ACCESS.2024.3505616

Zhang, S., et al. (2023). Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.