TOWARDS IMPROVED CERVICAL CANCER SCREENING: VISION TRANSFORMER-BASED CLASSIFICATION AND INTERPRETABILITY

Khoa Tuan Nguyen[1,2], *Ho-min Park*[1,2], *Gaeun Oh*[1], *Joris Vankerschaver*[1,3], *Wesley De Neve*[1,2]

[1] Center for Biosystems and Biotech Data Science, Department of Environmental Technology, Food Technology, and Molecular Biotechnology, Ghent University Global Campus, Incheon, Korea
[2] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics,Ghent University, Ghent, Belgium

ABSTRACT

We propose a novel approach to cervical cell image classification for cervical cancer screening using the EVA-02 transformer model. We developed a four-step pipeline: fine-tuning EVA-02, feature extraction, selecting important features through multiple machine learning models, and training a new artificial neural network with optional loss weighting for improved generalization. With this design, our best model achieved an F1-score of 0.85227, outperforming the baseline EVA-02 model (0.84878). We also utilized Kernel SHAP analysis and identified key features correlating with cell morphology and staining characteristics, providing interpretable insights into the decision-making process of the fine-tuned model. Our code is available at `https://github.com/Khoa-NT/isbi2025_ps3c`.

***Index Terms*—** Cell Classification, Cervical Cancer, Explainable AI, Vision Transformers

1. INTRODUCTION

Cervical cancer remains a significant global health challenge, ranking as the fourth most common cancer among women with over 600,000 new cases and 300,000 deaths annually [1]. Early detection through Pap smear screening has proven crucial in reducing mortality rates by identifying precancerous lesions. However, traditional analysis methods face major challenges: they are resource-intensive, time-consuming, and heavily dependent on cytologist expertise. To address these challenges, we participated in the Pap Smear Cell Classification Challenge (PS3C) [2, 3], organized as part of the IEEE International Symposium on Biomedical Imaging (ISBI) 2025 Challenge Program to foster the development of automated classification systems for cervical cell images.

In this challenge, we were tasked with developing deep learning models to classify Pap smear cell images into three categories: healthy cells without abnormalities, unhealthy cells indicating potential pathological changes, and unsuitable images due to artifacts or poor quality. The challenge provided a comprehensive training dataset containing four classes (including an additional 'Both cells' category), while the test dataset focused on the three primary categories [2, 3]. Effectiveness was evaluated using the F1-score, calculated for each class and averaged across all classes to account for data imbalance.

2. METHODS

We used a four-step method to achieve high F1-scores. First, we fine-tuned the pre-trained transformer-based image classification model EVA-02 [4]. Then, as shown in Fig. 2, we (A) extracted features from the model, (B) selected important features, and (C) trained a new Artificial Neural Network (ANN) model with the selected features with/without loss weighting for imbalanced labels. Finally, (D) we employed Kernel Shapley Additive Explanations (SHAP) to analyze the decision-making process of the fine-tuned model and provide interpretable insights into how the model makes its classifications [5].

2.1. Fine-tuning EVA-02

We use EVA-02 [4], a transformer-based architecture, as our baseline classification model. The selection of EVA-02 was based on its superior performance among transformer-based image classification models available in PyTorch Image Models[1] [6]. Instead of training from scratch, we fine-tune the pre-trained ImageNet model[2] on the entire training dataset to classify Healthy, Unhealthy, and Rubbish output classes, as shown in Fig. 1. In addition, encouraged by the challenge organizers [3], we found that training with three classes (merging the 'Both cells' category into the 'Unhealthy' category) produced slightly better results than training with four classes. Therefore, we adopt this merging as the default setting. The same configuration as the pre-trained ImageNet model is used for image preprocessing and augmentation. We employ the

[1] According to the benchmark results: `https://github.com/huggingface/pytorch-image-models/blob/main/results/results-imagenet.csv`

[2] Model: 'timm/eva02_base_patch14_448.mim_in22k_ft_in22k_in1k'

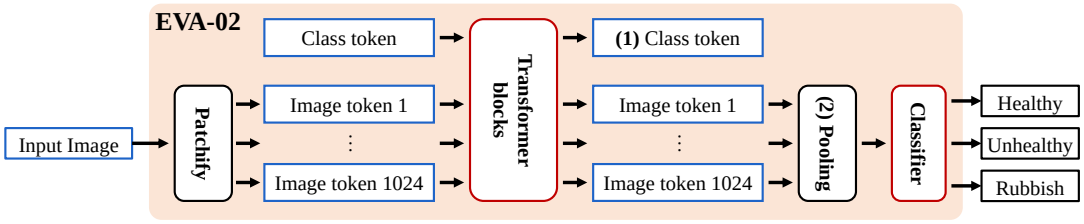


Fig. 1: A simple illustrative workflow of the EVA-02 model. The input cell image is divided into patches, creating Image tokens. A sequence of transformer-based blocks processes a trainable Class token along with these Image tokens. The output Image tokens are then average pooled before being fed into the classifier to predict Healthy, Unhealthy, and Rubbish classes. The Class token is extracted at (1), while the Image tokens are extracted at (2).

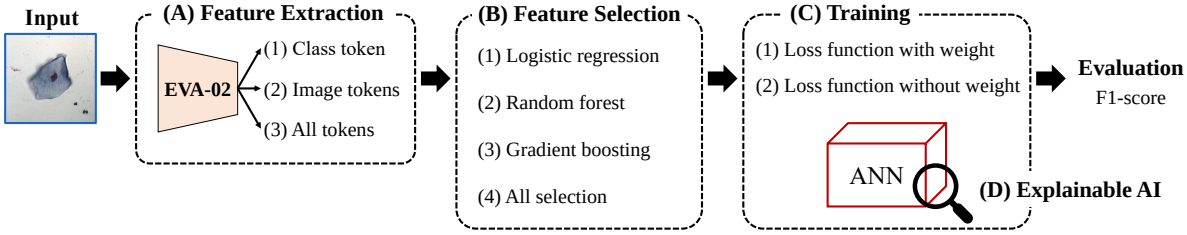


Fig. 2: Overview of the experiment.

AdamW optimizer (weight decay $= 0.05$) with a learning rate of $1e{-}5$, a ReduceLROnPlateau scheduler (patience $= 2$, factor $= 0.5$), and CrossEntropyLoss from PyTorch [7]. We fine-tune the model for 20 epochs and select the checkpoint with the lowest loss value.

2.2. (A) Feature extraction

After completing the training, we extracted image features using three options: Class token, Image tokens, and All tokens. The Class token, shown as (1) in Fig. 1, is an additional token designed to capture comprehensive information from the images in the dataset [8]. The Image tokens, which can be seen as (2) in Fig. 1, are generated by dividing the input image into patches, with each token representing local features containing both positional and visual information from its corresponding patch. The 'All tokens' option includes both the Class token and Image tokens, providing the most comprehensive feature representation by combining both global and local information.

2.3. (B) Feature selection

The features from EVA-02, namely Class token, Image tokens, and All tokens, share the same format shape (N, L, E), where N is the number of images in the training dataset, L represents the token length, and E denotes the embedding size. By averaging along the token length axis (with $L_{\text{Class token}} = 1$, $L_{\text{Image tokens}} = 1024$, and $L_{\text{All tokens}} = 1025$), each feature extracted from an image obtains a universal dimensionality of $E = 768$. Given the limited amount of training data available relative to the image size, we trained these features with their corresponding labels (Rubbish, Healthy, Unhealthy) using three machine learning models (Logistic regression, Random forest, Gradient boosting) to reduce the risk of overfitting and eliminate less substantial features that could potentially introduce noise [9].

We extracted the importance of each feature from each trained model. For instance, in Logistic regression, the absolute value of coefficients served as the importance measure for the corresponding features. The importance values were averaged across the three classes to generate rankings, and feature selection was performed by applying different thresholds for each model:

- Logistic regression: Determined through manual inspection, applied a manual threshold of $1e{-}16$ (31.12% filtered)
- Random forest: Applied a threshold of $3e{-}6$, determined by manually identifying the point where the number of data points and their values decreased sharply in the distribution (1.95% filtered)
- Gradient boosting: Excluded all data points with an importance value of 0 (40.10% filtered)

This feature selection process enabled us to maintain model effectiveness while reducing the risk of overfitting and facilitating more efficient training. For comparison purposes, we also included an 'All selection' option where all feature dimensions were used without any selection process.

2.4. (C) Training a new ANN model

To integrate the extracted features into a classification model, we constructed an ANN with three hidden layers (1024, 512,

Table 1: F1-scores ↑ (higher is better) for different combinations of feature extraction methods and machine learning models under weighted and unweighted loss conditions. The notable scores are shown in **bold**.

(A) Extraction method	(B) Selection	(C-1) Weighted loss	(C-2) Unweighted loss
Class token	Gradient boosting	0.85007	0.85033
	Random forest	0.85020	0.85112
	Logistic regression	0.85031	0.84921
	All selection	0.85039	0.85015
Image tokens	Gradient boosting	0.85069	0.85037
	Random forest	0.85070	0.85166
	Logistic regression	0.85078	0.85141
	All selection	0.85154	0.85108
All tokens	Gradient boosting	0.85056	0.85044
	Random forest	**0.85225**	0.85008
	Logistic regression	0.85004	0.85072
	All selection	0.85169	**0.85227**
EVA-02 baseline		0.84170	**0.84878**

and 256 neurons, respectively), choosing these values empirically based on their practical performance. Compared to the EVA-02 baseline classifier, which consists of a single linear layer (input to output: $768 \rightarrow 3$), this deeper architecture allows for more expressive feature learning.

The ANN was trained for 100 epochs using cross-entropy loss, and we selected the checkpoint with the lowest validation loss for evaluation.

Loss Weighting. Instead of treating all categories equally in the cross-entropy loss, we propose adding weights to the loss function for each category based on the number of data points, as shown in Table 2. We also fine-tune the baseline model with loss weighting for comparison.

The weighted cross-entropy loss function is defined as follows:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(p_c),$$

where:

- C is the number of classes.
- y_c is the binary indicator (0 or 1) for whether class label c is the correct classification for the current instance.
- p_c is the predicted probability of class c.
- w_c is the weight for class c, calculated as:

$$w_c = \frac{\max(\text{num\_data\_points})}{\text{num\_data\_points}_c}$$

where:

- $\max(\text{num\_data\_points})$ is the maximum number of data points across all classes.
- num_data_points_c is the number of data points in class c.

2.5. (D) SHAP-based feature analysis

While achieving high effectiveness is important, understanding model decisions is equally crucial, particularly in medical applications where explainability directly relates to patient trust and rights [10]. We employed Kernel SHAP [5] to analyze the decision-making process of our best model. This analysis identifies the most influential feature among the input features and examines how variations in the values of this feature correlate with the actual image characteristics. We compared images where this important feature showed extremely low versus high values, providing qualitative insights into what visual patterns or characteristics this feature represents.

Class	# of datapoints	Weights
Rubbish	50,371	1.000
Healthy	28,895	1.743
Unhealthy + Both cells	5,814	8.664

Table 2: Weight of each class.

3. RESULTS AND DISCUSSION

3.1. (A) Impact of feature extraction

Our analysis indicates that the choice of feature extraction method had a notable impact on the effectiveness of the model. As shown in Table 1, among the three extraction approaches (Class token, Image tokens, and All tokens), the All tokens method generally achieved the highest F1-scores, with a maximum effectiveness of 0.85227 obtained using 'All selection' under unweighted loss conditions. This suggests that the combination of both global (Class token) and local (Image tokens) information provides the most comprehensive representation for classification.

3.2. (B) Feature selection effectiveness

Our feature selection approach, as reflected in the results of Table 1, which filtered out less significant features based

on model-specific thresholds (40.10% for Gradient boosting, 1.95% for Random forest, and 31.12% for Logistic regression), proved effective in maintaining model effectiveness while reducing computational complexity. The competitive effectiveness of models with filtered features suggests that our selection criteria successfully identified the most relevant features for classification.

3.3. (C) Effect of loss weighting

The impact of loss weighting on model effectiveness showed notable patterns in Table 1. While the differences between weighted and unweighted loss scenarios were relatively small (differences typically less than 0.002), there were consistent patterns across different combinations of the extraction (A) and selection (B) methods. For instance, with Class token extraction, weighted loss generally led to more stable F1-scores across different machine learning models (range: 0.85007-0.85039) compared to unweighted loss (range: 0.84921-0.85112).

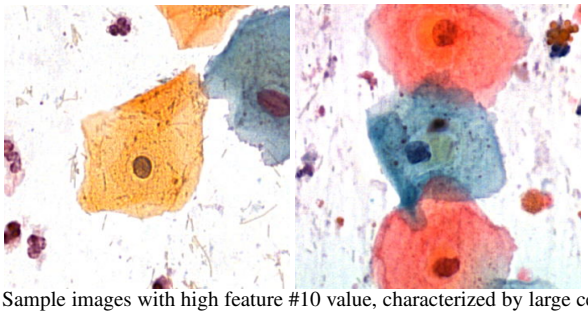
3.4. Comparison with baseline EVA-02

The results in Table 1 demonstrate that our proposed feature extraction and selection pipeline consistently outperformed the baseline EVA-02 model (F1-score: 0.84878). Our best performing configuration, using All tokens, All selection, and unweighted loss, achieved an F1-score of 0.85227, representing a 0.35% improvement over the baseline.

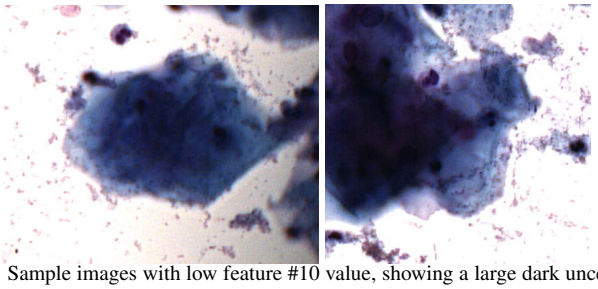
3.5. (D) SHAP-based feature analysis

We analyzed the relative importance of each feature among the 768-dimensional input features using our best model (All tokens + All selection + Unweighted loss). This analysis demonstrated that feature #10 emerged as the most influential feature in the decision-making process of the model.

Fig. 3 provides visual insights into the characteristics of feature #10. Fig. 3(a) displays images with high feature #10 values, while Fig. 3(b) shows images with low values. The visual characteristics exhibit substantial differentiation: images with high feature #10 values are characterized by large cells with rich cytoplasm and intense red/blue staining with clear nuclear boundaries, whereas images with low feature #10 values show dark regions with indistinct cells.



(a) Sample images with high feature #10 value, characterized by large cells with rich cytoplasm and intense red/blue staining with clear nuclear boundaries.



(b) Sample images with low feature #10 value, showing a large dark uncertainty region.

Fig. 3: Comparison of cell images based on feature #10 values, demonstrating distinct morphological and staining characteristics.

4. CONCLUSION

In this study, we proposed an innovative approach for cervical cell image classification to aid in early cervical cancer detection. Our pipeline, which combines feature extraction based on the EVA-02 transformer model, feature selection, and a deeper ANN classifier, achieved higher F1-scores compared to the EVA-02 baseline. A notable finding is that the All tokens approach, which utilizes both global information (Class token) and local information (Image tokens), demonstrated the highest effectiveness. Additionally, we confirmed that through the feature selection process, we could maintain effectiveness while reducing model complexity. Furthermore, through Kernel SHAP, we identified feature #10 as having the most impact on the decision-making of the model. Particularly, through image analysis, we were able to confirm that cell size, cytoplasmic characteristics, and staining intensity demonstrate notable influence in classification.

The main contribution of our research findings lies in presenting a method that satisfies both the high accuracy and interpretability required in medical settings for the automation of cervical cancer screening. However, a limitation of this study is that the difference between the minimum and maximum effectiveness is 1.06% (minimum: 0.8417, maximum: 0.85227), and the final results are not yet confirmed as the evaluation has only been conducted on the public leaderboard.

5. REFERENCES

[1] World Health Organization, "Cervical Cancer: Key Facts," 2024, Accessed: 2025-02-03.

[2] David Kupas, Andras Hajdu, Ilona Kovacs, Zoltan Hargitai, Zita Szombathy, and Balazs Harangi, "Annotated Pap cell images and smear slices for cell classification," *Scientific Data*, vol. 11, no. 1, pp. 743, 2024.

[3] Balázs Harangi, András Hajdu, Dávid Kupás, Ilona Kovács, Péter Kovács, and Nicolai Spicher, "Pap Smear Cell Classification Challenge (PS3C)," https://kaggle.com/competitions/pap-smear-cell-classification-challenge, 2024.

[4] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao, "Eva-02: A visual representation for neon genesis," *Image and Vision Computing*, p. 105171, 2024.

[5] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[6] Ross Wightman, "PyTorch Image Models," https://github.com/huggingface/pytorch-image-models, 2019.

[7] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. Apr. 2024, ACM.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[9] Jundong Li, Ke Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 94:1–94:45, 2017.

[10] Bryce Goodman and Seth Flaxman, "European Union regulations on algorithmic decision-making and a right to explanation," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.