ViT-SE_Res: A Hybrid Vision Transformer and ResNet50V2 with Squeeze-and-Excitation Block for Cervical Cell Classification

Betelhem Zewdu Wubineh [1][0000-0002-4790-7449] , Andrzej Rusiecki [1][0000-0003-2239-1076] and Krzysztof Halawa [1][0000-0003-2239-1076]

[1] Wroclaw University of Science and Technology, Faculty of Information and Communication Technology, Wroclaw, Poland

betelhem.wubineh@pwr.edu.pl, andrzej.rusiecki@pwr.edu.pl, krzysztof.halawa@pwr.edu.pl

Abstract. Cervical cancer is the leading cause of mortality among women, and early detection through effective screening is crucial for a better prognosis and treatment. Traditional manual methods, such as the Papanicolaou (Pap) test, are often time-consuming and ineffective. To overcome these limitations, this study proposes a novel hybrid model, Vision Transformer with Squeeze-and-Excitation blocks incorporated into the ResNet50V2 (ViT-SE_Res), for cervical cell classification. The model integrates the local feature extraction capabilities of ResNet50V2 with the global context modeling strength of ViT, effectively capturing both local and global features for improved accuracy. The model was evaluated on two datasets, Pomeranian and SIPaKMeD, achieving an accuracy of 98.80% and 98.51%, respectively, with SIPaKMeD being a binary classification task. The proposed ViT-SE_Res model offers a robust and efficient tool for cervical cancer screening, providing reliable detection of abnormalities to support early intervention.

Keywords: Cervical Cancer, Classification, ResNet50V2, Vision Transformer, Squeeze-and-Excitation

1 Introduction

Cervical cancer is one of the most widespread diseases among women, and its control is, therefore, a major concern worldwide [1][2][3]. The prognosis and treatment of cervical cancer have improved significantly with early detection and diagnosis [4]. Papanicolaou (Pap) test is a widely used screening method to detect cervical cancer and precancerous lesions [5][6][7][8]. However, manual screening of Pap smears is a tedious, inefficient, and expensive process [9][10]. Furthermore, there is a subjective judgment of the decision made by different experts, which may be prone to errors [11][12]. Lastly,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

traditional techniques are very labor-intensive and have a very long lead time, which makes them undesirable. These limitations led to the intent to investigate automatic screening methods.

Deep learning (DL) has emerged as a promising approach for automated cervical cell classification, offering the potential to learn complex patterns in cervical cytopathology images without manual feature engineering [13][14]. Deep learning methods have performed very well in the field of medical imaging [15] [16]. Convolutional neural networks (CNNs) have shown considerable success in the classification of cervical cytopathology images [15]. Nevertheless, classical CNNs may lack the ability to model global aspects of cervical cell extraction [1], while Vision Transformers (ViTs), which have shown great potential in computer vision, may fall short in adequately extracting cervical cell morphology information due to the compact size and unique structure of cervical cytopathology images [1]. As such, it is necessary to investigate techniques that could effectively integrate the advantages of CNN and ViT to improve the performance of cervical cell classification [3].

Several studies have explored deep learning-based methods for cervical cell classification. For example, Hemalatha et al. [13] introduced CervixFuzzyFusion for cervical cancer cell image classification. This study uses a feature fusion method that combines features from DenseNet201 and ViT using shifted patch tokenization and locality self-attention models. Yang et al. [1] proposed a pyramid convolutional mixer (PCMixer) to classify cervical cells. The mixer integrates a pyramidal morphology module and a nuclear spatial mixing block to extract information on cervical cytopathology effectively. Furthermore, Anand and Bachhal [2] used the VGG16 architecture, referred to as Cervical-Net, for the classification of cervical cells. Maurya et al. [14] presented an ensemble approach of combining the Vision Transformer network (ViT) and CNN for cervical cell classification. All these studies utilized the publicly available SIPaKMeD datasets.

Although previous studies have explored architectures such as DenseNet201, VGG16, and ensemble methods involving CNNs and Vision Transformers, the potential of combining the ResNet family with ViT for cervical cell classification remains unexplored. This study addresses this gap by leveraging the residual learning of ResNet50V2 with Squeeze-and-Excitation (SE) block and the ViT attention mechanisms to classify cervical cells. ResNet50V2 has shown promising performance in cervical cell classification [17]. The SE block enables the network to focus more on the important features, thereby improving the performance [18]. Additionally, Vision Transformer has demonstrated potential in image classification tasks [5]. By combining the local feature extraction capabilities of ResNet50V2 and the global feature extraction capabilities of ViT [19], this study aims to achieve more accurate and robust classification results. The key contribution of this study is as follows:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

- Incorporating SE blocks into the residual block of ResNet50V2 to focus on the important feature and improve performance.
- Proposing a novel hybrid model of SE-ResNet50V2 and Vision Transformer (ViT-SE_Res) for cervical cell classification.

This study aims to develop an effective hybrid deep-learning model for classifying cervical cells to help diagnose cervical cancer. The rest of the paper is organized as follows. Section 2 describes the materials and methods, Section 3 discusses the results and findings, and Section 4 presents the concluding remarks, followed by the references.

2 Materials and Methods

In this section, we discuss the dataset and its preprocessing technique, the proposed method, and the evaluation metrics.

2.1 Dataset and Preprocessing Technique

The datasets used in this study are from the Pomeranian Medical University in Szczecin, Poland. This dataset contains a total of 419 cervical images, divided into 268 training images, 67 validation images, and 84 testing images. The dataset includes three categories: high squamous intra-epithelial lesion (HSIL), low squamous intra-epithelial lesion (LSIL), and normal squamous intra-epithelial lesion (NSIL) [20]. Most of the images have a resolution of 1130 x 1130 pixels. The second dataset is SIPaKMeD [21], which contains a total of 4049 single-cell images. These are divided into 2,591 training images, 648 validation images, and 810 testing images. This dataset comprises five categories: dyskeratotic, koilocytotic, metaplastic, parabasal, and supra-intermediate. Table 1 provides the details of the datasets.

Table 1 Details of the dataset used for the study

Dataset	Category	No. of image	Category	Training	Validation	Testing
Pomeranian	HSIL	124				
	LSIL	61		268	67	84
	NSIL	234				
SIPaKMeD	Dyskeratotic	713	Abnormal			
	Koilocytotic	725				
	Metaplastic	693		2591	648	810
	Parabasal	687	Normal			
	Superficial-Intermediate	731				

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

During data preprocessing, we resized the images to a uniform size of 224 x 224 pixels. Then, we normalized the images by scaling the pixel values to the range of [0, 1]. To further optimize input for the Vision Transformer (ViT-B16), the resized images were divided into non-overlapping patches of 16x16 pixels, ensuring compatibility with the model architecture. Sample images from both the Pomeranian and the SIPaKMeD dataset are shown in Fig. 1.

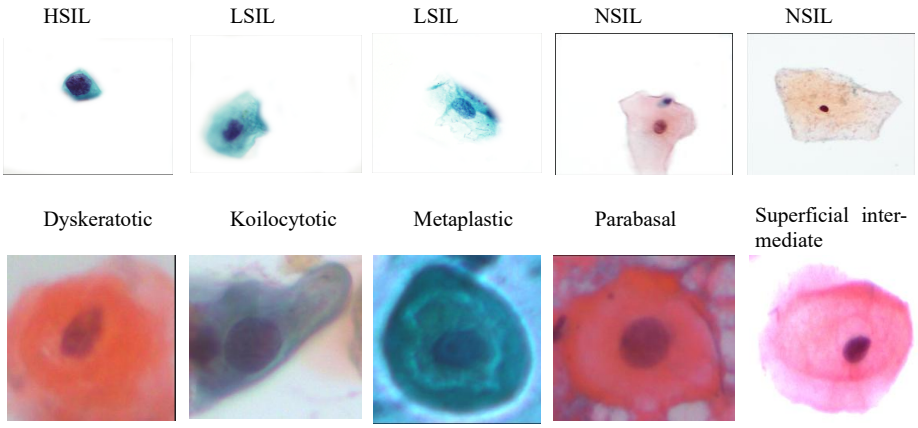


Fig. 1. Sample images from the Pomeranian and SIPaKMeD dataset

As shown in Fig. 1, the first row represents samples from the Pomeranian, while the second row corresponds to the SIPaKMeD dataset. In the Pomeranian dataset, HSIL refers to more serious abnormalities characterized by moderate to severe dysplasia, which carry a higher risk of progressing to cervical cancer if left untreated. LSIL denotes non-cancerous cervical abnormalities where cells exhibit mild dysplasia or slight irregularities. NSIL represents normal cervical cells without precancerous changes. In the SIPaKMeD dataset, dyskeratotic cells are defined by abnormal keratinization and typically appear in dense clusters with orangeophilic cytoplasm and vesicular nuclei, often indicating HPV infection. Koilocytotic cells are characterized by large perinuclear halos, hyperchromatic and irregular nuclei and are pathognomonic for HPV infection, frequently appearing as binucleated or multinucleated cells. Metaplastic cells resemble parabasal cells but exhibit more uniform size and shape, well-defined round cytoplasm, and darker staining; they are often associated with high-grade lesions (HSIL). Parabasal cells are small, immature squamous cells with cyanophilic cytoplasm and large vesicular nuclei, and they can be difficult to distinguish from metaplastic cells. Finally, superficial-intermediate cells are the most abundant type, with large polygonal cytoplasm and small pycnotic or vesicular nuclei, which may display morphological changes in the presence of severe lesions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

2.2 Proposed Method

In this paper, we propose a hybrid cervical cell classification model using ResNet50V2 with SE block and Vision transformer (ViT) architectures. ResNet50V2 is a pre-trained CNN on ImageNet [22], used to extract rich hierarchical spatial features, leveraging its deeper residual structure [15] to improve performance. In our approach, Squeeze-and-Excitation blocks are added following each ResNet block to improve the model's capability to adjust channel-wise feature responses. This allows the network to dynamically focus on more relevant features [23]. It applies global average pooling to squeeze spatial dimensions, followed by two fully connected layers to excite and weight each channel, enhancing the network's focus on important features. On the other hand, ViT, a transformer-based model, captures global contextual relationships and long-range dependencies [19].

Both models are adapted by excluding their classification heads and freezing their pre-trained weights to retain their feature extraction capabilities and prevent overfitting. The input images are processed through these two branches independently, with features of ResNet50V2 aggregated using global average pooling and SE-enhanced features, and features of ViT flattened into a one-dimensional representation. The outputs from both branches are fused via feature concatenation, creating a unified feature vector. This fused representation is passed through a fully connected classification head, which includes dense layers with ReLU activations, dropout for regularization, and a final softmax layer for class predictions.

By leveraging the complementary strengths of convolutional networks with SE-enhanced residual learning and transformers, this hybrid model can effectively capture both local and global features, providing a robust and efficient framework for cervical cell classification. The overall framework of the proposed model is illustrated in Fig. 2. In our case, we used ViT-B16, which divides the input image of size 224x224 into patches of 16x16 pixels. This patch size allows the model to capture local information to process the ViT model effectively.

The ViT-B16 architecture used in our model consists of 12 transformer encoder blocks, each comprising a multi-head self-attention mechanism with 12 attention heads and a feed-forward network (FFN) with a hidden size of 3072 [24]. The embedding dimension for each input patch is 768. Each block includes Layer Normalization and residual connections, followed by a GELU activation in the MLP layers. Position embeddings are added to the input patch embeddings to retain spatial information. A dropout rate of 0.1 is applied throughout the transformer blocks to mitigate overfitting.

2.3 Evaluation Metrics

To qualitatively evaluate the proposed model's performance, we used the accuracy, precision, recall, and F1 scores. These metrics provide a comprehensive assessment, with

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

accuracy reflecting overall performance and precision, recall, and F1 score providing information on the model's effectiveness in an imbalanced class [19].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (1)$$

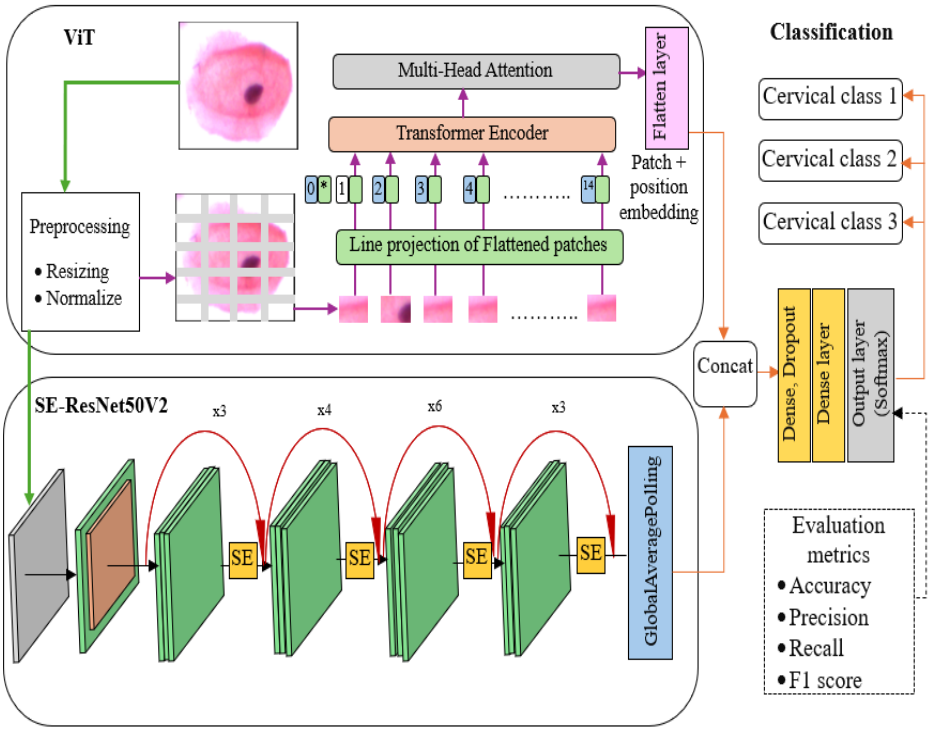


Fig. 2. The ViT-SE_Res proposed method for cervical cell classification

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

3 Results and Discussion

In this study, to evaluate the proposed model, we used the Pomeranian and SIPaKMeD datasets. The dataset comprising 419 and 4049 images, respectively, were divided into training, validation, and testing sets, as detailed in Table 1. The experiment was conducted using Python programming with the TensorFlow and Keras frameworks. The tuning procedure for the study is described as follows: we used the Adam optimizer with the ReLU activation for the hidden layers and Softmax for the final classification layer, with a learning rate of 0.002. A batch size of 18 was used to process the data before updating the model's weights. The model was trained for 100 epochs, allowing it to pass through the entire training dataset, with early stopping to prevent overfitting. The results of the study using the Pomeranian and the SIPaKMeD dataset are illustrated in Table 2.

Table 2 Results of the study using the Pomeranian and SIPaKMeD dataset

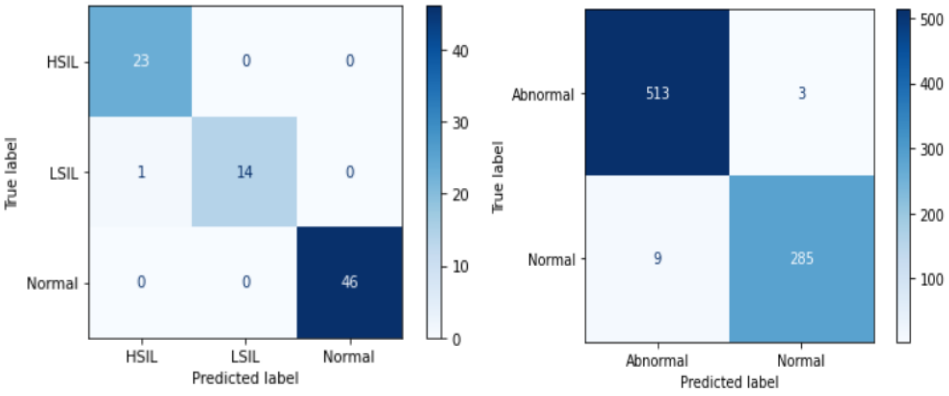
Dataset	Method	Classification type	Accuracy	Precision	Recall	F1 score
Pomeranian	ViT: SE-ResNet50V2	Multiclass	98.80%	98.88%	98.80%	98.81%
SIPaKMeD		Multiclass	94.69%	94.72%	94.69%	94.66%
		Binary	98.51%	98.53%	98.51%	98.52%

As shown in Table 2, the proposed method achieves 98.80% accuracy on the Pomeranian dataset, indicating that the combination of ResNet50V2 with SE and ViT is highly effective in classifying cervical cells. The high precision, recall, and F1 score further demonstrate the robustness of the model for cervical cancer diagnosis. On the SIPaKMeD dataset, the proposed method also performs well, achieving 94.69% and 98.51% accuracy for multiclass and binary classification, respectively. These results highlight the versatility and strong performance of the model across both types of classification tasks. In particular, the model is effective at identifying different types of cervical cells, showing its potential for precise abnormality detection. The high scores in binary classification also underscore its ability to distinguish between normal and abnormal cells, which is crucial in clinical settings for accurate diagnosis.

The observed difference in performance between multiclass and binary classification on the SIPaKMeD dataset can be attributed to the inherent complexity of the multiclass task. In multiclass classification, the model must distinguish between several visually similar cervical cell types, which introduces a higher level of difficulty and potential for misclassification, especially between classes with subtle morphological differences. In contrast, binary classification simplifies the task by reducing it to identifying whether

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

a sample is normal or abnormal, making it less prone to inter-class confusion. Additionally, class imbalance and overlapping features among certain cell types in the multiclass scenario may further contribute to the lower performance. The confusion matrix of the proposed model is shown in Fig. 3.



a b

Fig. 3. Confusion matrix: a) Pomeranian and b) SIPaKMeD binary

In the confusion matrix, only one image was incorrectly classified in the Pomeranian dataset, and 12 images were in the binary classification of the SIPaKMeD dataset. This shows how the proposed method effectively classified cervical cells in different datasets. The comparison of our study with the previous work is shown in Table 3.

Table 3 Comparison of our study with previous work

Ref	Method	Dataset	Accuracy	Precision	Recall	F1 score
[1]	PCMixer	SIPaKMeD	96.21	95.70	95.60	95.30
[2]	VGG16	SIPaKMeD	84.33%	-	-	-
[13]	CervixFuzzyFusion	SIPaKMeD	93.36%	-	-	-
[14]	ViT-CNN Max vote	SIPaKMeD	97.6%	99.54	97.65	98.58
[14]	ViT-CNN Stacking	SIPaKMeD	94.11%	-	-	-
[25]	SA-ResNet50V2	SIPaKMeD	92%	92%	92%	92%
Ours	ViT-Res	SIPaKMeD multi	94.69%	94.72%	94.69%	94.66%
Ours	ViT-Res	SIPaKMeD binary	98.51%	98.53%	98.51%	98.52%
Ours	Vit-Res	Pomeranian	**98.80%**	98.88%	98.80%	98.81%

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

The results in Table 3 show a comparison of the previous works using the SIPaKMeD multiclass dataset. Among these, ViT-CNN demonstrates the highest accuracy, which is 97.6%, outperforming other models such as PCMixer at 96.21%, CervixFuzzyFusion at 93.36%, and ResNet50V2 at 92%. In contrast, traditional deep learning models like VGG16 show significantly lower accuracy at 84.33%, indicating limited performance in complex datasets like SIPaKMeD.

Our proposed ViT-SE_Res hybrid model, however, is distinct from [14], as it integrates the feature extraction capabilities of ResNet50V2 with the attention mechanisms of ViT, providing both local and global feature extraction. On the other hand, [14] ViT-CNN ensembles the predictions of CNN and MobileNet via Max voting, achieving an accuracy of 97.63%, while using the stacking ensemble method results in an accuracy of 94.11%. By concatenating features from ResNet50V2 with SE and ViT, our study performs better than the stacking method but slightly lower than the Max voting ensemble.

This combination improves accuracy, achieving 98.80% on the Pomeranian and 98.51% on the SIPaKMeD binary datasets, and 94.69% accuracy on the SIPaKMeD multiclass dataset, surpassing the model in [13], which achieved an accuracy of 93.36%. Additionally, [25] uses a residual deep convolutional generative adversarial network (Res_DCGAN) for augmentation and adds a self-attention layer at the top of ResNet50V2, achieving an accuracy of 92% which is lower than ours. As a result, our study suggests that the integration of ResNet and ViTs offers significant advantages over standalone approaches, delivering high accuracy and balanced performance metrics across various datasets and configurations.

Moreover, it is important to visualize which parts of the images (features) are crucial for the classification of cervical cancer. This can be achieved through explainable AI (XAI) techniques, which highlight the specific regions of the image that contribute to the model's classification decision. Gradient-weighted class activation mapping (Grad-CAM) is a model-specific XAI method designed for use with DL models that incorporate CNN, where the spatial relationships in the data remain intact following its passes through the convolutional layers [26]. Fig. 4 shows the original images alongside their Grad-CAM visualizations, highlighting the area that contributes the most to the model's classification decision.

In Fig. 4, the first and the second column shows images from the Pomeranian dataset and corresponding Grad-CAM, while the third and fourth column displays images from the SIPaKMeD dataset with corresponding Grad-CAM. This visualization helps identify the area of the images that is important for the classification model. For the Pomeranian dataset, the center of the cell background near the cell is significant, whereas in the SIPaKMeD dataset, the center of the cell plays a key role.

In this study, Grad-CAM was employed to visualize and interpret the regions of cervical cell images that contributed most to the model's predictions. Since Grad-CAM is designed specifically for convolutional neural networks, the visualizations were

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

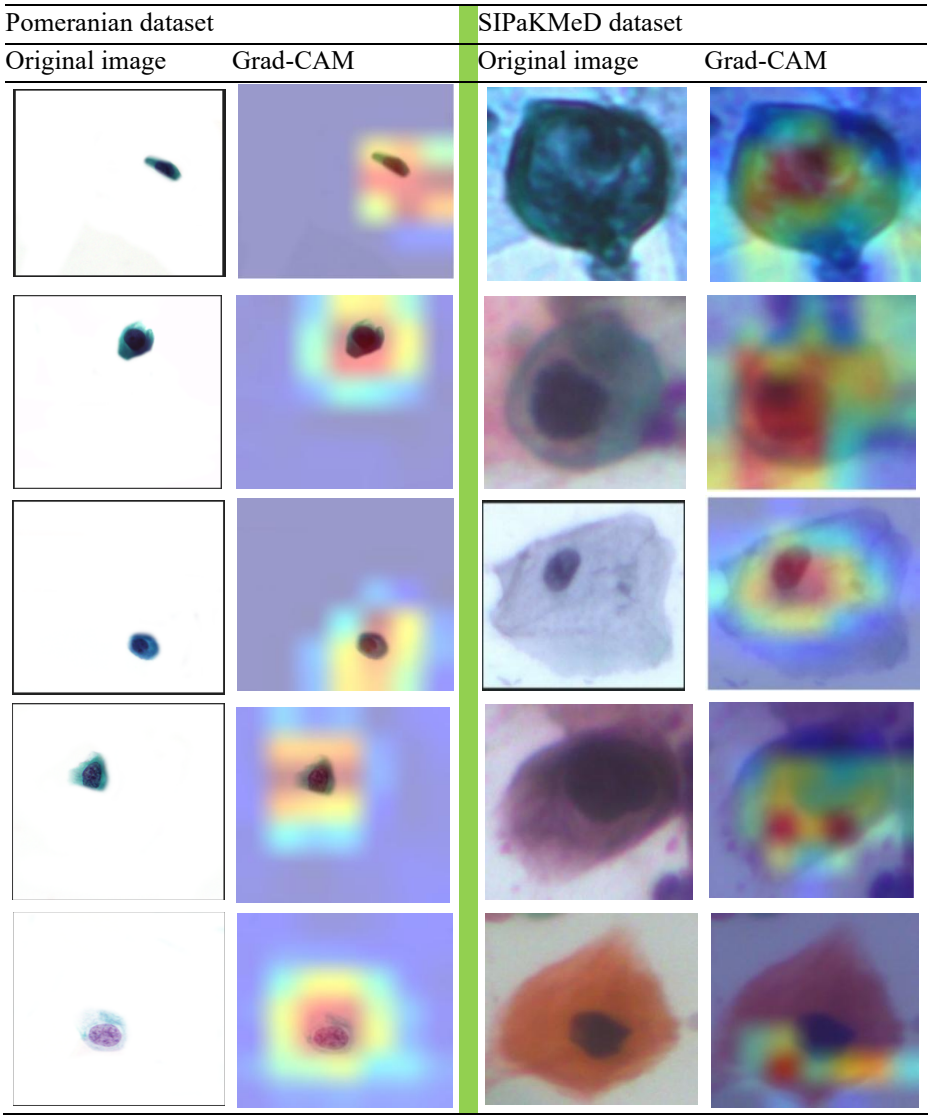


Fig. 4. Visualization of images using Grad-CAM

generated using the feature maps from the ResNet50V2 branch of our hybrid model. The Vision Transformer (ViT) branch does not retain the spatial hierarchies required for traditional Grad-CAM, as it operates on flattened patch embeddings and lacks convolutional layers. Therefore, Grad-CAM was not applied to the ViT component. While the transformer branch plays a critical role in the overall classification performance through feature fusion, its interpretability requires alternative methods, such as

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

attention map visualization, Attention Rollout, or Transformer Attribution techniques. Incorporating these transformer-specific explainability approaches will be considered in future work to provide a more comprehensive understanding of the hybrid model's decision-making process.

4 Conclusions

This study successfully demonstrates the efficacy of a new hybrid approach that combines SE with ResNet50V2 and Vision Transformer (ViT) for cervical cell classification. By leveraging both the local feature extraction capabilities of ResNet50V2 and the global attention mechanisms of ViT, the proposed ViT-Res model effectively handles the challenges of cervical cell classification and achieves remarkable accuracy and performance across different datasets. The findings highlight the advantage of integrating ResNet50V2 and ViT architectures, which collectively enable improved feature extraction and classification. Furthermore, Grad-CAM was used to identify the important features of the images for decision-making in the classification model. Further studies could explore fine-tuning the model and expanding its application to larger and more diverse datasets, ultimately contributing to the automation and enhancement of cervical cancer screening processes. Additionally, validating the models with the independent dataset would help ensure the generalizability of the model.

References

[1] T. Yang, H. Hu, X. Li, M. Qing, L. Chen, and Q. Huang, "A pyramid convolutional mixer for cervical pap-smear image classification tasks," *Biomed. Signal Process. Control*, vol. 99, no. February 2024, p. 106789, 2025, doi: 10.1016/j.bspc.2024.106789.

[2] V. Anand and P. Bachhal, "Cervical Net: An Effective Convolution Neural Network for Five-class Classification of Cervical Cells," *Proc. - 2nd IEEE Int. Conf. Device Intell. Comput. Commun. Technol. DICCT 2024*, pp. 51–55, 2024, doi: 10.1109/DICCT61038.2024.10532902.

[3] M. Fang, M. Fu, B. Liao, X. Lei, and F. X. Wu, "Deep integrated fusion of local and global features for cervical cell classification," *Comput. Biol. Med.*, vol. 171, no. February, p. 108153, 2024, doi: 10.1016/j.compbiomed.2024.108153.

[4] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "Cervical Cell Segmentation and Classification Using U-Net and Hybrid VGG19-AlexNet Architecture," *2024 Int. Conf. Inf. Commun. Technol. Dev. Africa, ICT4DA 2024*, no. November, pp. 37–42, 2024, doi: 10.1109/ICT4DA62874.2024.10777181.

[5] C. Zhao, R. Shuai, L. Ma, W. Liu, and M. Wu, *Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT*, vol. 81, no. 17. 2022. doi: 10.1007/s11042-022-12670-0.

[6] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "Segmentation and Classification Techniques for Pap smear Images in Detecting Cervical Cancer: A Systematic Review,"

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

IEEE Access, vol. 12, no. August, pp. 118195–118213, 2024, doi: 10.1109/ACCESS.2024.3447887.

[7] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "Segmentation of Cytology Images to Detect Cervical Cancer Using Deep Learning Techniques," in *International Conference on Computational Science*, Switzerland: Springer Nature, 2024, pp. 270–278.

[8] E. Hussain, L. B. Mahanta, C. R. Das, M. Choudhury, and M. Chowdhury, "A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in Pap smear images," *Artif. Intell. Med.*, vol. 107, no. October 2019, p. 101897, 2020, doi: 10.1016/j.artmed.2020.101897.

[9] Kurnianingsih *et al.*, "Segmentation and classification of cervical cells using deep learning," *IEEE Access*, vol. 7, pp. 116925–116941, 2019, doi: 10.1109/ACCESS.2019.2936017.

[10] A. Desiani, M. Erwin, B. Suprihatin, S. Yahdin, A. I. Putri, and F. R. Husein, "Bi-path Architecture of CNN Segmentation and Classification Method for Cervical Cancer Disorders Based on Pap-smear Images," *IAENG Int. J. Comput. Sci.*, vol. 48, no. 3, pp. 1–9, 2021.

[11] S. Madathil, M. Dhouib, Q. Lelong, A. Bourassine, and J. Monsonego, "A multimodal deep learning model for cervical pre-cancers and cancers prediction: Development and internal validation study," *Comput. Biol. Med.*, vol. 186, no. January, 2025, doi: 10.1016/j.compbiomed.2025.109710.

[12] B. Zewdu Wubineh, Ł. Jeleń, and A. Rusiecki, "DCGAN-based Cytology Image Augmentation for Cervical Cancer Cell Classification," *IEEE Trans. Med. Imaging*, vol. xx, no. 50, pp. 1003–1011, 2020, doi: 10.1016/j.procs.2025.02.206.

[13] H. K, V. V, M. Dhandapani, and A. G. A., "CervixFuzzyFusion for cervical cancer cell image classification," *Biomed. Signal Process. Control*, vol. 85, no. April, p. 104920, 2023, doi: 10.1016/j.bspc.2023.104920.

[14] R. Maurya, N. Nath Pandey, and M. Kishore Dutta, "VisionCervix: Papanicolaou cervical smears classification using novel CNN-Vision ensemble approach," *Biomed. Signal Process. Control*, vol. 79, no. P2, p. 104156, 2023, doi: 10.1016/j.bspc.2022.104156.

[15] M. A. Talukder, M. A. Layek, M. Kazi, M. A. Uddin, and S. Aryal, "Empowering COVID-19 detection: Optimizing performance through fine-tuned EfficientNet deep learning architecture," *Comput. Biol. Med.*, vol. 168, no. August 2023, 2024, doi: 10.1016/j.compbiomed.2023.107789.

[16] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "Data Augmentation Techniques to Detect Cervical Cancer Using Deep Learning: A Systematic Review," in *International Conference on Dependability of Computer Systems*, Cham: Springer Nature Switzerland, 2024, pp. 325–336.

[17] L. Wong, A. Ccopa, E. Diaz, S. Valcarcel, D. Mauricio, and V. Villoslada, "Deep Learning and Transfer Learning Methods to Effectively Diagnose Cervical Cancer from Liquid-Based Cytology Pap Smear Images," *Int. J. online Biomed. Eng.*, vol. 19, no. 4, pp. 77–93, 2023, doi: 10.3991/ijoe.v19i04.37437.

[18] U. Sani, E. Suryani, W. Widiarto, and U. Salamah, "Residual Network with Squeeze-And-Excitation Block for White Blood Cell Classification in Acute Myeloid Leukemia," *Proc. - Int. Conf. Informatics Comput. Sci.*, pp. 239–244, 2024, doi: 10.1109/ICICoS62600.2024.10636919.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2

[19] X. Jiang, S. Wang, and Y. Zhang, "Vision transformer promotes cancer diagnosis: A comprehensive review," *Expert Syst. Appl.*, vol. 252, no. PA, p. 124113, 2024, doi: 10.1016/j.eswa.2024.124113.

[20] Ł. Jeleń, I. Stankiewicz-Antosz, M. Chosia, and M. Jeleń, "Optimizing Cervical Cancer Diagnosis with Feature Selection and Deep Learning," *Appl. Sci.*, vol. 15, no. 3, pp. 1–21, 2025, doi: 10.3390/app15031458.

[21] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images," *Proc. - Int. Conf. Image Process. ICIP*, no. October, pp. 3144–3148, 2018, doi: 10.1109/ICIP.2018.8451588.

[22] Future Machine Learning, "The Significance of F1 Score in Evaluating Algorithms," Future Machine Learning.

[23] N. Shahadat, A. Nguyen, and R. Lama, "Squeeze and Hypercomplex Networks on Leaf Disease Detection," in *Pattern Recognition, ICPR 2024. Lecture Notes in Computer Science*, 2025.

[24] A. Dosovitskiy *et al.*, "An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[25] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "Classification of cervical cells from the Pap smear image using the RES_DCGAN data augmentation and ResNet50V2 with self-attention architecture," *Neural Comput. Appl.*, vol. 0123456789, 2024, doi: 10.1007/s00521-024-10404-x.

[26] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP," *Appl. Energy*, vol. 353, no. PA, p. 122079, 2024, doi: 10.1016/j.apenergy.2023.122079.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_2