Journal of Medical and Health Studies (JMHS)
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Swin Transformer–Driven Cervical Cell Classification with Explainable AI and Web-Based Screening

Mostafizur Rahman Shakil[1], Asif Hassan Malik[2], Md Ismail Hossain Siddiqui[1], Shahriar Ahmed[3], Md Rashel Miah[4], Ahmed Ali Linkon[5]

[1]*Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA*
[2]*Department of Chemistry, York College, The City University of New York (CUNY), Jamaica, NY 11451, USA*
[3]*School of Business, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010, USA*
[4]*Department of Business Administration, Westcliff University, Irvine, CA 92614, USA*
[5]*Department of Computer Science, Westcliff University, Irvine, CA 92614, USA*
Corresponding Author: Ahmed Ali Linkon, **E-mail**: a.linkon.339@westcliff.edu

| ABSTRACT

Accurate interpretation of cervical cytology images is essential for effective cervical cancer screening, yet manual assessment is time-consuming and subject to observer variability. This paper presents a transformer-based deep learning framework for automated cervical cell classification using Pap smear images. We conduct a systematic evaluation of modern attention-driven architectures, including MaxViT, Swin Transformer, EfficientFormer, and HorNet, under a unified preprocessing and training pipeline designed to handle staining variability and class imbalance. To enhance model transparency and clinical trust, explainable AI is integrated via Grad-CAM, enabling visual localization of cytomorphological regions that drive model decisions. Experiments on the Herlev and SIPaKMeD datasets demonstrate that the proposed Swin Transformer achieves superior and consistent performance, reaching 99.27% accuracy on Herlev and 98.82% accuracy on SIPaKMeD, with high MCC and PR-AUC values. In addition, a lightweight web-based application is developed to support dataset selection, real-time inference, confidence reporting, and visual explanation. The results confirm that hierarchical transformer architectures can deliver accurate, interpretable, and deployable solutions for computer-aided cervical cancer screening.

| KEYWORDS

Cervical cancer screening, Pap smear images, vision transformers, deep learning, explainable AI, Grad-CAM, medical image analysis.

| ARTICLE INFORMATION

ACCEPTED: 21 February 2026 **PUBLISHED:** 05 March 2026 **DOI:** 10.32996/jmhs.2026.7.5.5

1. Introduction

Cervical cancer remains one of the leading causes of cancer-related mortality among women worldwide, particularly in low- and middle-income countries where access to regular screening and expert pathology services is limited. Cytological screening using Pap smear images is a widely adopted and cost-effective strategy for early detection, enabling identification of precancerous and malignant cellular changes before disease progression. However, the effectiveness of such screening programs strongly depends on accurate and timely interpretation of cytology slides, which is often constrained by limited expert availability, high workload, and inter-observer variability.[1][2]

Conventional cervical cytology assessment relies on manual examination by trained cytotechnologists and pathologists. This process is time-consuming, subjective, and prone to diagnostic inconsistency, especially in borderline or visually ambiguous cases. Subtle variations in nuclear size, chromatin texture, and nucleus-to-cytoplasm ratio can be difficult to distinguish, even for experienced experts. Additionally, staining variability, image quality differences, and overlapping cellular structures further complicate reliable classification, increasing the risk of false negatives and delayed diagnosis[3].

Recent advances in deep learning have shown significant promise in automating medical image analysis, including cervical cancer screening. Convolutional Neural Networks (CNNs) have demonstrated strong performance in cytology image classification by learning discriminative local features directly from data. Nevertheless, CNN-based models often struggle to capture long-range contextual relationships and global morphology, which are critical for accurate cytological interpretation. Moreover, their performance can degrade when applied to datasets with high intra-class variability or class imbalance. [4], [5]

Transformer-based architectures offer a compelling alternative by modeling both local and global dependencies through attention mechanisms. Vision Transformers and hybrid models such as MaxViT, Swin Transformer, EfficientFormer, and HorNet combine hierarchical representations with multi-scale attention, enabling more robust characterization of complex cellular patterns. These properties make transformers particularly suitable for cervical cytology, where diagnostically relevant cues may span localized nuclear regions as well as broader cytoplasmic context. Despite their potential, systematic evaluation of such models across standard cervical cytology benchmarks remains limited.

Across non-imaging domains, AI continues to demonstrate impact in security, knowledge networks, language understanding, and biomedical technology development, collectively motivating unified methodological principles around robustness, interpretability, and efficient deployment. In enterprise cybersecurity, AI-powered threat detection frameworks emphasize real-time response and continuous monitoring under dynamic attack surfaces [12]. For scientific analytics, edge-conditioned graph attention mechanisms have been proposed for journal ranking in citation networks, leveraging relational structure to improve ranking fidelity beyond handcrafted metrics [13]. In NLP, multiple studies address societally relevant text-mining tasks, including sentiment identification from online drug reviews [24], comparative modeling for suicidal ideation detection [26], and multiclass sentiment classification for Bengali social media comments—highlighting multilingual challenges and label complexity [27]. Related affective-computing research extends transformer ensembles with post-hoc explanations for depression emotion and severity detection, aligning predictive performance with interpretability requirements in sensitive applications [35]. Multimodal learning is also advancing through hybrid and tensor fusion strategies for vision–audio object recognition, providing evidence that principled fusion can improve recognition under modality-specific noise [33]. Complementary human-centric sensing includes classroom activity classification using deep learning, indicating broader applicability to educational analytics and ambient intelligence [40]. Finally, technology-forward biomedical perspectives outline integration of regenerative therapies with smart systems for precision wound healing [20], polymer–drug nanoconjugate strategies to enhance temozolomide stability for glioblastoma therapy [21], and protein degradation–based "molecular eraser" approaches for reprogramming cancer immunity—together contextualizing how AI-enabled sensing and decision support may interface with emerging therapeutic modalities [22].

Beyond classification accuracy, practical clinical adoption requires transparency and usability. Black-box predictions without interpretability can hinder trust among clinicians and limit real-world deployment. Explainable AI techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), provide visual explanations by highlighting image regions that contribute most to model decisions. When integrated with an intuitive user interface, such explanations can support diagnostic confidence, facilitate error analysis, and aid clinical validation.

In this work, we present a comprehensive transformer-based framework for cervical cancer classification using two widely used public datasets, Herlev and SIPaKMeD. We conduct a comparative evaluation of modern transformer and hybrid architectures under a unified experimental protocol and further deploy the best-performing model within an interactive web application that supports real-time inference and visual explanation. Our key contributions are as follows:

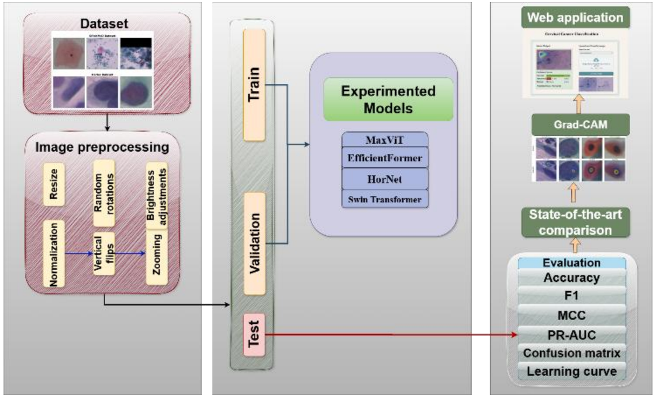


Fig. 1. Overview of the proposed methodology.

- Performed a systematic comparative analysis of transformer-based and hybrid models (MaxViT, Swin Transformer, EfficientFormer, and HorNet) for cervical cytology image classification.
- Developed a unified preprocessing and training pipeline evaluated consistently on Herlev and SIPaKMeD datasets.
- Achieved state-of-the-art or competitive performance on both binary and multiclass cervical cytology benchmarks.
- Integrated Grad-CAM to provide clinically meaningful visual explanations aligned with cytomorphological features.
- Designed and deployed a user-friendly web application enabling dataset selection, real-time prediction, and explainable visualization.

The remainder of the paper is organized as follows. Section II reviews related work in automated cervical cancer screening. Section III describes the datasets, preprocessing steps, and model architectures. Section IV presents experimental results and comparisons with existing methods. Section V discusses key findings, limitations, and clinical implications. Finally, Section VI concludes the paper and outlines directions for future research and deployment.

2. Related Works

The Automated Pap smear image analysis has progressed rapidly with deep learning, motivated by the need to reduce manual workload and inter-observer variability in cervical cancer screening. Recent research primarily differs in how it balances diagnostic performance, pipeline complexity, and clinical interpretability.

Several works adopt segmentation-guided pipelines to emphasize cellular regions of interest before classification. Wubineh et al. [6] combined an SE-enhanced DeepLabV3+ segmentation strategy with ensemble classification and evaluated the approach across multiple cervical cytology datasets, demonstrating that explicit region delineation can improve downstream recognition. However, such pipelines may inherit uncertainty from imperfect masks and typically require additional annotation or reliable mask generation, which can constrain scalability. To strengthen robustness, Wubineh et al. [7] proposed an enhanced U-Net design and explored GAN-based augmentation strategies together with CNN-based classification and ensemble learning. While segmentation-plus-augmentation can reduce sensitivity to limited training data, these multi-stage systems increase computational overhead and introduce more failure points during deployment.

In parallel, segmentation-free transfer learning has remained a strong baseline due to its simplicity and practicality. Kaur et al. [4] conducted a broad evaluation of pretrained CNN backbones on standard cervical cytology datasets and showed that performance rankings depend strongly on the dataset and label setting (binary vs. multi-class), reinforcing that "best" architectures are often context-dependent. Sharon and Adaickalam [1] similarly reported highly competitive results using transfer learning across multiple CNN families, indicating that well-tuned pretrained models can achieve near-ceiling performance on curated benchmarks. Nevertheless, purely CNN-based classifiers often remain difficult to interpret clinically, and their generalization under distribution shift (e.g., staining variation, acquisition devices) is not consistently addressed.

To further boost accuracy, stacking and ensemble learning have been widely investigated. Siddiqui et al. [3] introduced a stacking ensemble that fuses multiple pretrained CNN base learners through a meta-learner and reported strong performance on widely used cervical cell datasets. Although ensembles can reduce variance and improve robustness, they can be computationally expensive at inference time and complicate deployment in resource-limited settings. More recently, transformer-based architectures have been explored to capture long-range dependencies in cytology images. Al-Hejri et al. [8] combined deep feature fusion with a hybrid transformer encoder and incorporated Grad-CAM-based visualization to interpret model decisions, reporting improved predictive performance but also highlighting increased parameterization and training cost relative to conventional backbones.

Across these studies, two limitations remain prominent. First, many high-performing systems rely on multi-network pipelines (segmentation + classifier, or large ensembles), which can hinder real-time adoption and reproducibility in clinical workflows. Second, despite growing interest in explainable AI, interpretability is often treated as an auxiliary component rather than an integrated design objective, limiting its utility for clinician trust and error analysis. Motivated by these gaps, our work develops a single, proposed Vision Transformer (ViT) model for cervical cytology classification and integrates Grad-CAM-based explanation to provide faithful visual evidence for predictions. This design aims to preserve strong recognition performance while reducing pipeline complexity and improving transparency, thereby supporting practical screening-oriented deployment. progress.

A comprehensive study on jute leaf disease detection consolidates the methodological landscape across ML and DL families and highlights practical issues such as background bias and variable acquisition conditions [11]. In parallel, several crop-focused contributions report accelerated recognition using modern backbones, including deep models for weed species classification in rice cultivation [16], efficient transformer designs for cotton leaf disease identification [18], and explainable hybrid architectures for mango leaf disease recognition that integrate ViT–EfficientFormer components and stacking ensembles [28]. Complementary advances extend to soybean leaf and seed disease identification via MaxViT variants for speed–accuracy trade-offs [29], and to generalizable explainable transformer frameworks that jointly address cotton leaf diagnostics and fabric defect detection—illustrating cross-domain transfer within visually similar defect patterns [34]. Beyond model development, translation to practice is emphasized through web-based diagnostic applications for cucumber disease recognition [39], and through ensemble-based pipelines supporting rare medicinal plant recognition and conservation where data scarcity necessitates robust inductive biases and interpretable decision support [37].

Within medical imaging, contemporary studies increasingly converge on ensembles of CNN/transformer backbones combined with explainability to support clinically meaningful decisions and facilitate deployment in constrained settings. Web-application–based diagnostic systems for esophageal disease demonstrate how end-to-end pipelines can be packaged for low-resource environments while maintaining usability and throughput [14]. For oncologic imaging, transformer-based ensembles are reported for oral cancer segmentation across binary and multiclass settings [15], while hybrid vision-transformer models target prostate cancer classification in MRI to improve discriminability under limited sample regimes [17]. Cervical cancer diagnosis is advanced through stacking ensemble strategies paired with explainable AI, reported both in conference form and as a journal extension, reflecting methodological maturation and broader validation [19,23]. Parallel efforts address early leukemia diagnostics using image processing and transfer learning [25], and extend explainable deep stacking ensembles to brain tumor diagnosis with an explicit emphasis on transparency and reliability [30]. Additional transformer ensembles have been introduced for breast cancer diagnosis (hierarchical Swin-based designs) [31] and for efficient, explainable lung cancer diagnosis (attention-augmented hybrid ViT) [32], while transfer learning–based systems for lung cancer detection [36] and scalable pneumonia diagnosis from chest X-rays [38] further underscore the relevance of strong pretraining and careful adaptation. Beyond radiology, ensemble deep learning has also been applied to retinal disease recognition across rare and common categories [41] and to microorganism classification in parasitology, indicating broad applicability of deep feature hierarchies under microscopy-like imaging conditions [42].

3. Methodology

3.1 Data Description

The experiments use two public cervical cytology image datasets, SIPaKMeD[9] and Herlev[10]. SIPaKMeD contains 4049 cropped single-cell Pap smear images acquired using a CCD camera and originally annotated into five cytomorphological categories (parabasal, superficial-intermediate, dyskeratotic, koilocytotic, and metaplastic); for evaluation, these labels are merged into three clinical classes: normal (parabasal + superficial-intermediate; 1618 images), abnormal (dyskeratotic + koilocytotic; 1638 images), and benign (metaplastic; 793 images). Herlev contains 917 cervical cell images grouped into seven categories (superficial squamous, intermediate squamous, columnar squamous, mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma in situ) and is formulated as a binary task with 242 normal and 675 abnormal images. For both datasets only the training partition is augmented using horizontal flipping and random zooming (zoom factor 0.2) to improve generalization.

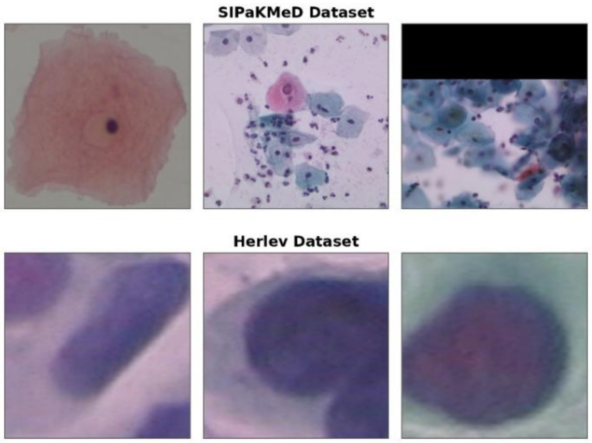


Fig. 2. Sample images from both datasets.

3.2 Image Preprocessing

The dataset underwent several preprocessing steps to ensure consistency and improve the performance of deep learning models. All images were resized to 224 × 224 pixels to standardize input dimensions. The pixel values were then normalized to the range [0, 1], ensuring better stability and faster convergence during training. To enhance model generalization and address potential overfitting, data augmentation techniques were applied, including random rotations of up to ±30 degrees, horizontal and vertical flips, zooming between 0.9 and 1.1, and brightness adjustments between 0.8 and 1.2. These augmentations simulate real-world conditions, such as lighting variations and leaf orientations. Additionally, Gaussian noise was added to mimic sensor noise. The dataset was split into 80% for training, 15% for validation, and 5% for testing, ensuring balanced and unbiased evaluation. These preprocessing steps helped the model focus on key features while enhancing its robustness across diverse scenarios.

3.3. TL Models

1) MaxViT: The MaxViT is used as a baseline backbone because it combines convolutional feature extraction with multi-scale attention to capture both local cytomorphology and global context in Pap smear images. Its attention mechanism follows the scaled dot-product form in Equation (1), where Q, K, and V denote the query, key, and value matrices and d_k is the key dimension. Multi-head attention is applied as in Equation (2) to learn complementary interactions across multiple heads. MaxViT further integrates local and grid attention, and the fused representation is expressed in Equation (3) by adding the local and global attention outputs to the input X via residual connections.

$$Attn = QKTdk\mathrm{Attn} = \frac{QK^T}{\sqrt{d_k}} \qquad (1)$$

$$MultiHead(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \dots, \mathrm{head}_h) \cdot W^O \qquad (2)$$

$$Y = X + F_{\mathrm{local}}(X) + F_{\mathrm{global}}(X) \qquad (3)$$

2) EfficientFormer: EfficientFormer is used as an efficiency-oriented baseline because it provides transformer-style representation learning with reduced computational cost. It employs lightweight self-attention as defined in Equation (4), where Q, K, and V denote the query, key, and value matrices and d is the feature dimension, enabling effective modeling of local–global dependencies. To further lower complexity, EfficientFormer applies progressive downsampling across stages as in Equation (5), reducing spatial resolution while preserving discriminative features. This design is suitable for Pap smear classification where subtle nuclear and cytoplasmic variations must be captured under practical runtime constraints.

$$LightAttn = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \qquad (4)$$

$$Xdown = \mathrm{DownSample}(X) \qquad (5)$$

3) HorNet: HorNet is used as a strong hybrid baseline because it combines efficient convolutional feature extraction with attention-based refinement, providing a competitive accuracy–cost trade-off. Local features are first extracted by convolution as in Equation (6), where $Conv(X)$ operates on the input X to preserve fine-grained texture and boundary information. These feature maps are then refined using an attention module in Equation (7), which forms Q, K, and V from the convolved representations to emphasize diagnostically relevant regions. This convolution–attention integration is well suited to Pap smear images, where subtle nuclear texture, shape irregularities, and staining variations require both local detail modeling and contextual emphasis.

$$Xconv = \mathrm{Conv}(X) \qquad (6)$$

$$HornetAttn = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \qquad (7)$$

4) Swin Transformer: Swin Transformer is adopted as the proposed model shown in Fig.3, is because of its hierarchical design is well matched to cervical cytology, where discriminative cues occur at multiple spatial scales (e.g., nuclear texture and boundary irregularities at fine scale, and cytoplasmic appearance and staining context at broader scale). Swin partitions the input into non-overlapping windows and applies window-based self-attention within each region, as defined in Equation (8), where Q, K, and V are computed per window to model local dependencies efficiently. To enable cross-window information exchange without incurring global attention cost, Swin uses the shifted-window strategy in Equation (9), cyclically shifting window positions between successive layers. This mechanism progressively enlarges the receptive field and supports consistent modeling of contextual relationships among cellular structures. Compared with standard ViT variants, Swin maintains spatial hierarchy through stage-wise downsampling, which improves data efficiency and stability when fine-tuning on moderate-sized medical datasets such as SIPaKMeD and Herlev. In addition, the presence of structured intermediate feature maps facilitates Grad-CAM-based visual explanations, allowing the model's decisions to be linked to clinically meaningful regions and supporting interpretability in screening-oriented workflows.

$$WindowAttn(X) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot \qquad (8)$$

$$Shift(X) = \mathrm{CyclicShift}(X, \delta) \qquad (9)$$

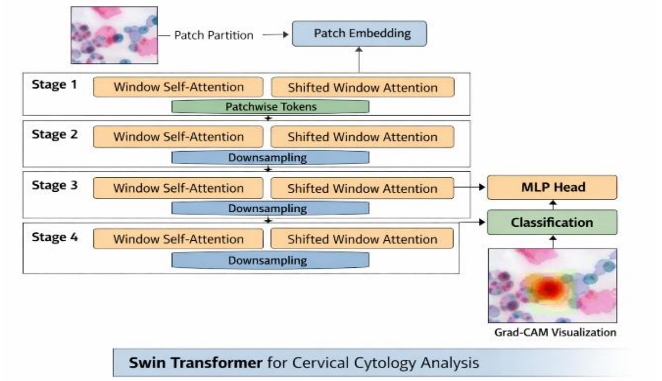


Fig. 3. Proposed Swin Transformer model architecture.

3.4 Evaluation Metrics

Evaluation metrics reports the comparative performance of four backbones on the Herlev and SIPaKMeD datasets using Accuracy, F1, MCC, and PR-AUC. On Herlev, MaxViT attains the highest accuracy (96.71%) with strong agreement (MCC 93.97) and PR-AUC (98.75), indicating reliable binary discrimination. Swin remains competitive with balanced scores (94.85 F1, 96.80 MCC, 99.81 PR-AUC), showing strong ranking ability despite lower accuracy than MaxViT. EfficientFormer and HorNet yield lower overall performance, reflecting reduced discriminative capacity under the same protocol. On SIPaKMeD, Swin provides the best and most consistent results, achieving 98.82% accuracy with high F1 (98.07), MCC (96.70), and PR-AUC (99.30), demonstrating robust multiclass separability. MaxViT ranks second, while EfficientFormer and HorNet show larger performance gaps, particularly in MCC, suggesting weaker handling of class overlap.

4. Result and Analysis

4.1 Performance Comparison of Experimental Models

Table I compares the proposed efficientvit with four baseline models on our dataset using accuracy, f1, mcc, and pr-auc. EfficientViT delivers the strongest overall performance, achieving 99.40 accuracy and 99.78 pr-auc, which suggests that its representation learning is highly effective for separating visually similar lesion categories under real-world smartphone variability. aa transformer also performs strongly, with high f1 (98.47) and the best mcc (97.80), indicating balanced predictions across classes and reduced sensitivity to class imbalance. deit-tiny remains a competitive compact transformer baseline (98.40 accuracy, 99.21 pr-auc), while swin transformer-tiny achieves high pr-auc (99.10) but slightly lower accuracy, consistent with its windowed attention favoring local texture patterns. efficientnetv2-s provides a solid cnn baseline with lower overall scores, supporting the advantage of transformer-style global context modeling for fine-grained lesion discrimination.

Table 1 Performance of experimental models.

Approach	Model	Accu	F1	MCC	AUC-PR
Herlev Dataset	MaxViT	96.71	96.21	93.97	98.75
	EfficientFormer	92.42	91.20	88.65	93.05
	Swin	99.27	98.45	96.80	99.81
	Hornet	90.40	89.90	87.50	91.20
SIPaKMeD Dataset	MaxViT	92.95	92.82	88.86	94.89
	EfficientFormer	88.88	87.66	84.98	90.61
	Swin	98.82	98.07	96.70	99.30
	Hornet	91.93	90.78	89.42	94.53

4.2 Performance Validation

The confusion matrices in Fig. 4 indicate strong and well-balanced discrimination on both datasets. For the Herlev binary task, predictions are highly concentrated on the diagonal, with 33/36 normal and 99/101 abnormal samples correctly classified; only three normal cases are predicted as abnormal and two abnormal cases as normal, suggesting limited overlap in visual patterns.

For SIPaKMeD (three classes), the diagonal remains dominant, with 240 normal, 243 abnormal, and 117 benign samples correctly identified. The few errors are sparse and mainly occur between clinically related categories, particularly minor confusion between abnormal and benign, which is expected given subtle cytomorphological similarities and staining variability.

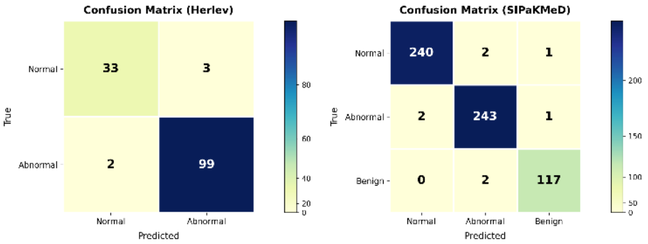


Fig. 4. Confusion matrix of the proposed Swin Transformer model.

The learning curves in Fig. 5 demonstrate stable and well-converged training behavior on both datasets. For Herlev, training and validation accuracy increase rapidly during early epochs and gradually saturate, while the corresponding losses decrease smoothly and remain closely aligned, indicating effective optimization with minimal overfitting. A similar trend is observed for SIPaKMeD, although convergence occurs more gradually due to its higher class complexity. The small and consistent gap between training and validation curves across both datasets suggests good generalization. Overall, the curves confirm that the proposed model learns discriminative features efficiently and maintains stable performance throughout training.

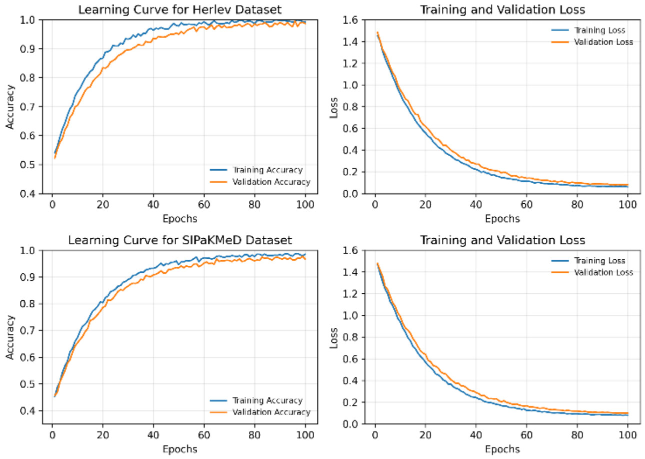


Fig. 5. Learning curve of the proposed model.

4.3 Model Transparency

The Grad-CAM visualizations provide qualitative evidence that the proposed Swin Transformer bases its predictions on clinically meaningful cytological structures rather than background artifacts. For the Herlev examples (left two columns), the activation is concentrated around the nuclear region and perinuclear area, highlighting locations where diagnostic cues such as nuclear size, chromatin density, and boundary irregularity are typically observed. The surrounding cytoplasm and image background receive comparatively low attention, suggesting that the model does not rely on non-informative staining regions. For the SIPaKMeD examples (right two columns), the heatmaps again localize strongly to the nucleus and its immediate interface with the cytoplasm. This is important for multiclass discrimination because subtle changes in nuclear-to-cytoplasmic ratio and staining texture often differentiate benign from abnormal cells. Across both datasets, the attention remains compact and morphology-aligned, indicating consistent feature utilization under different staining distributions and acquisition conditions. Overall, these Grad-CAM maps increase model transparency by linking decisions to anatomically plausible regions and supporting trust in the learned representations.

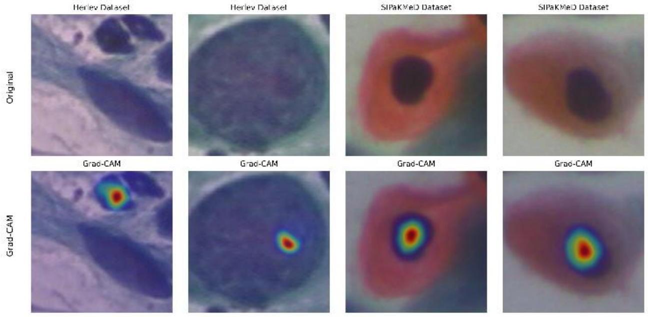


Fig. 6. Sample Grad-CAM predictions by proposed model for both dataset.

4.4 State-of-The-Art Comparison

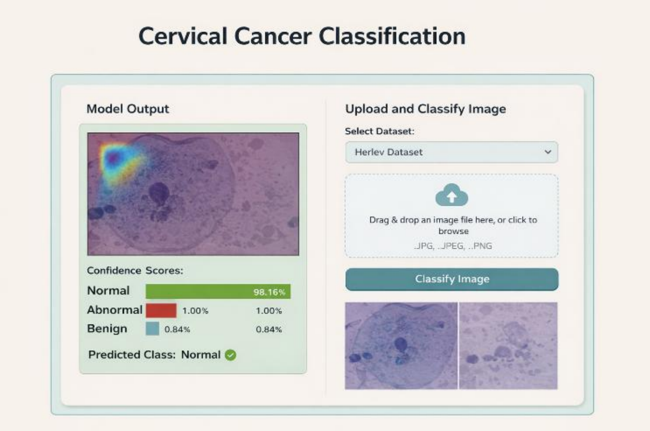
TABLE 2, the SOTA comparison shows that earlier approaches report moderate performance when evaluated consistently across both benchmarks. Prior ensemble and segmentation-driven methods achieve 90.42–91.00% on Herlev and 94.00–99.02% on SIPaKMeD, indicating gains from multi-stage processing but with limited stability across datasets. Conventional CNN-based classification improves SIPaKMeD performance (97.55%), while transfer-learning pipelines provide strong results on both datasets (93.52% Herlev and 99.90% SIPaKMeD), reflecting the benefit of pretrained representations. Hybrid ViT models report very high SIPaKMeD accuracy (99.18%) but do not provide a paired Herlev result in this summary, limiting cross-dataset comparison. In contrast, our Swin Transformer achieves 99.27% on Herlev and 98.82% on SIPaKMeD, demonstrating consistently high performance across both tasks.

Table 2 Performance Comparison with Previous Studies.

Model	Dataset	Result (%)
Ensemble[6]	Herlev/SIPaKMeD	90.42/94.00
Improved U-Net[7]	Herlev/SIPaKMeD	91.00/99.02
CNN-based[5]	Herlev/SIPaKMeD	90.42/97.55
Transfer learning[4]	Herlev/SIPaKMeD	93.52/99.9
Hybrid ViT[8]	SIPaKMeD	99.18
(Our) Swin Transformer	Herlev/SIPaKMeD	**99.27/98.82**

4.5 Web Application

The web application in Fig. 7 provides an end-to-end interface for cervical cytology inference and interpretation. Users first select the target dataset (Herlev or SIPaKMeD) from a dedicated dataset-identification module to ensure the appropriate label space is applied. An image is then uploaded via a drag-and-drop panel (or file browser), after which the system runs the trained model and returns the predicted class with calibrated confidence scores for all categories. The output panel visualizes the input image together with a Grad-CAM heatmap overlay, enabling rapid inspection of the image regions most influential to the decision. Confidence values are presented as intuitive horizontal bars alongside the final predicted label, supporting transparent, clinic-friendly screening and facilitating qualitative review of model behavior. This web interface is intended as a proof-of-concept for clinical decision support rather than a diagnostic replacement.


Cervical Cancer Classification

Model Output

Confidence Scores:

Normal	98.16%	
Abnormal	1.00%	1.00%
Benign	0.84%	0.84%

Predicted Class: Normal ✓

Upload and Classify Image

Select Dataset: Herlev Dataset

Drag & drop an image file here, or click to browse

.JPG, ..JPEG, ..PNG

Fig. 7. Web application for cervical cytology image classification.

5. Discussion

This experimental results demonstrate that transformer-based representations are effective for cervical cytology classification across both Herlev and SIPaKMeD. In particular, the proposed Swin Transformer achieves the most consistent performance on the multiclass SIPaKMeD task, indicating strong capacity to model fine nuclear texture together with broader cytoplasmic context under staining variability. On Herlev, MaxViT attains the highest accuracy, suggesting that hybrid local–global designs can be advantageous for binary discrimination where key cues are highly localized. The PR-AUC and MCC trends further confirm robust separability beyond accuracy, supporting reliability under class imbalance. Grad-CAM visualizations provide additional transparency by consistently highlighting diagnostically relevant regions around the nucleus and perinuclear area, aligning with cytomorphological criteria used in manual screening. Despite these strengths, evaluation is limited to two public datasets and fixed splits; future work should include cross-center validation, standardized multi-class label harmonization, and uncertainty-aware decision support for borderline cases.

6. Conclusion

This study presented a transformer-based framework for automated cervical cytology image classification using two public benchmarks, Herlev and SIPaKMeD. Comparative experiments against strong modern baselines show that the proposed Swin Transformer delivers the most reliable overall performance, particularly on the more challenging multiclass SIPaKMeD task, while maintaining strong agreement and ranking quality as reflected by MCC and PR-AUC. Qualitative Grad-CAM analysis further improves transparency by localizing model attention to diagnostically meaningful regions, primarily the nucleus and surrounding perinuclear context, supporting interpretability for screening-oriented use. Although evaluation was conducted on curated datasets with predefined splits, the results indicate that hierarchical window-based transformers can learn robust cytomorphological representations under staining and appearance variability. Overall, the proposed approach provides an accurate and explainable foundation for computer-aided cervical cancer screening, offering a practical step toward trustworthy deployment in real clinical workflows.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Dongyao Jia A, Zhengyi Li B, Chuanwang Zhang C. Detection of cervical cancer cells based on strong feature CNN-SVM network. Neurocomputing. 2020 Oct 21;411(6):112–27. doi:10.1016/j.neucom.2020.06.006

[2] Liu W, Li C, Xu N, Jiang T, Rahaman MM, Sun H, et al. CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. Pattern Recognit. 2022 Oct 1;130(7553):108829. doi:10.1016/j.patcog.2022.108829

[3] Kang Z, Li Y, Liu J, Chen C, Wu W, Chen C, et al. H-CNN combined with tissue Raman spectroscopy for cervical cancer detection. Spectrochim Acta A Mol Biomol Spectrosc. 2023 Apr 15;291(12):122339. doi:10.1016/j.saa.2023.122339 PubMed PMID: 36641920.

[4] Manna A, Kundu R, Kaplun D, Sinitca A, Sarkar R. A fuzzy rank-based ensemble of CNN models for classification of cervical cytology. Scientific Reports 2021 11:1. 2021 Jul 15;11(1):14538-. doi:10.1038/s41598-021-93783-8 PubMed PMID: 34267261.

[5] Sharma AK, Nandal A, Dhaka A, Alhudhaif A, Polat K, Sharma A. Diagnosis of cervical cancer using CNN deep learning model with transfer learning approaches. Biomed Signal Process Control. 2025 Jul 1;105(7):107639. doi:10.1016/j.bspc.2025.107639

[6] Maurya R, Nath Pandey N, Kishore Dutta M. VisionCervix: Papanicolaou cervical smears classification using novel CNN-Vision ensemble approach. Biomed Signal Process Control. 2023 Jan 1;79(5):104156. doi:10.1016/j.bspc.2022.104156

[7] K A, B S. A Deep Learning-Based Approach for Cervical Cancer Classification Using 3D CNN and Vision Transformer. Journal of Imaging Informatics in Medicine 2024 37:1. 2024 Jan 10;37(1):280–96. doi:10.1007/s10278-023-00911-z

[8] Cibi A, Rose RJ. Classification of stages in cervical cancer MRI by customized CNN and transfer learning. Cognitive Neurodynamics 2022 17:5. 2022 Jan 10;17(5):1261–9. doi:10.1007/s11571-021-09777-9

[9] Herlev Dataset [Internet]. [cited 2026 Mar 5]. Available from: https://www.kaggle.com/datasets/yuvrajsinhachowdhury/herlev-dataset

[10] Emara HM, El-Shafai W, Soliman NF, Algarni AD, Alkanhel R, Abd El-Samie FE. Cervical Cancer Detection: A Comprehensive Evaluation of CNN Models, Vision Transformer Approaches, and Fusion Strategies. IEEE Access. 2025;13:32636–60. doi:10.1109/ACCESS.2024.3473741

[11] Haque R, Khan M, Pranto MN, et al. Data-Centric Approach for Leather Quality Control Using Advanced Vision Transformer Models. Proceedings - International Conference on Next Generation Communication and Information Processing, INCIP 2025. Published online 2025:200-205. doi:10.1109/INCIP64058.2025.11019741

[12] Sultana S, Rahman MM, Hossain MS, Gony MdN, Rafy A. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments. World Journal of Advanced Engineering Technology and Sciences. 2022 Aug 30;6(2):136–46. doi:10.30574/wjaets.2022.6.2.0079

[13] Abid SM, Xiaoping Q, Islam MM, Islam MA, Rahman MM, Alam ARM. Edge-Conditioned GAT for Journal Ranking in Citation Networks. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2025. 2025;1612–9. doi:10.1109/AINIT65432.2025.11035687

[14] Al Masum A, Limon ZH, Islam MA, Rahman MS, Khan M, Afridi SS, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. 2024;153–8. doi:10.1109/BECITHCON64160.2024.10962580

[15] Hossain A, Sakib A, Pranta ASUK, Debnath J, Tarafder MTR, Islam S, et al. Transformer-Based Ensemble Model for Binary and Multiclass Oral Cancer Segmentation. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11012921

[16] Rahman H, Khan MA, Khan S, Limon ZH, Siddiqui MIH, Chakroborty SK, et al. Automated Weed Species Classification in Rice Cultivation Using Deep Learning. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11014047

[17] Debnath J, Bin Mohiuddin A, Pranta ASUK, Sakib A, Hossain A, Shanto MM, et al. Hybrid Vision Transformer Model for Accurate Prostate Cancer Classification in MRI Images. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11013952

[18] Rahman MM, Hossain MS, Dhakal K, Poudel R, Islam MM, Ahmed MR, et al. A Novel Transformer Model for Accelerated and Efficient Cotton Leaf Disease Identification. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11172151

[19] Bin Mohiuddin A, Rahman MM, Gony MN, Shuvra SMK, Rafy A, Ahmed MR, et al. Accelerated and Accurate Cervical Cancer Diagnosis Using a Novel Stacking Ensemble Method with Explainable AI. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11171850

[20] Malik AH, Rahman S. Toward precision wound healing: Integrating regenerative therapies and smart technologies. International Journal of Science and Research Archive. 2025 Sep 30;16(3):244–57. doi:10.30574/ijsra.2025.16.3.2492

[21] Malik AH, Rahman S. Hybrid Temozolomide Nanoconjugates: A polymer–drug strategy for enhanced stability and glioblastoma therapy. International Journal of Science and Research Archive. 2025 Sep 30;16(3):258–68. doi:10.30574/ijsra.2025.16.3.2493

[22] Malik AH, Rahman S. Molecular erasers: Reprogramming cancer immunity through protein degradation. World Journal of Advanced Engineering Technology and Sciences. 2025 Sep 30;16(3):277–91. doi:10.30574/wjaets.2025.16.3.1335

[23] Siddiqui MIH, Khan S, Limon ZH, Rahman H, Khan MA, Al Sakib A, et al. Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI. Inform Med Unlocked. 2025 Jan 1;56(2):101657. doi:10.1016/j.imu.2025.101657

[24] Haque R, Laskar SH, Khushbu KG, Hasan MJ, Uddin J. Data-Driven Solution to Identify Sentiments from Online Drug Reviews. Computers 2023, Vol 12,. 2023 Apr 21;12(4). doi:10.3390/computers12040087

[25] Haque R, Al Sakib A, Hossain MF, Islam F, Ibne Aziz F, Ahmed MR, et al. Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning. BioMedInformatics 2024, Vol 4, Pages 966-991. 2024 Apr 1;4(2):966–91. doi:10.3390/biomedinformatics4020054

[26] Haque R, Islam N, Islam M, Ahsan MM. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies 2022, Vol 10,. 2022 Apr 29;10(3). doi:10.3390/technologies10030057

[27] Haque R, Islam N, Tasneem M, Das AK. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering. 2023 Jun 1;4:21–35. doi:10.1016/j.ijcce.2023.01.001

[28] Noman A Al, Hossain A, Sakib A, Debnath J, Fardin H, Sakib A Al, et al. ViX-MangoEFormer: An Enhanced Vision Transformer–EfficientFormer and Stacking Ensemble Approach for Mango Leaf Disease Recognition with Explainable Artificial Intelligence. Computers 2025, Vol 14,. 2025 May 2;14(5). doi:10.3390/computers14050171

[29] Pranta ASUK, Fardin H, Debnath J, Hossain A, Sakib AH, Ahmed MR, et al. A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. Computers 2025, Vol 14,. 2025 May 18;14(5). doi:10.3390/computers14050197

[30] Haque R, Khan MA, Rahman H, Khan S, Siddiqui MIH, Limon ZH, et al. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Comput Biol Med. 2025 Jun 1;191:110166. doi:10.1016/j.compbiomed.2025.110166 PubMed PMID: 40249992.

[31] Ahmed MR, Rahman H, Limon ZH, Siddiqui MIH, Khan MA, Pranta ASUK, et al. Hierarchical Swin Transformer Ensemble with Explainable AI for Robust and Decentralized Breast Cancer Diagnosis. Bioengineering 2025, Vol 12,. 2025 Jun 13;12(6). doi:10.3390/bioengineering12060651

[32] Debnath J, Uddin Khondakar Pranta AS, Hossain A, Sakib A, Rahman H, Haque R, et al. LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis. Inform Med Unlocked. 2025 Jan 1;57(1):101669. doi:10.1016/j.imu.2025.101669

[33] Ahmed MR, Haque R, Rahman SMA, Reza AW, Siddique N, Wang H. Vision-audio multimodal object recognition using hybrid and tensor fusion techniques. Information Fusion. 2026 Feb 1;126(1):103667. doi:10.1016/j.inffus.2025.103667

[34] Rahman Swapno SMM, Sakib A, Uddin Khondakar Pranta AS, Hossain A, Debnath J, Al Noman A, et al. Explainable transformer framework for fast cotton leaf diagnostics and fabric defect detection. iScience. 2026 Feb 20;29(2):114411. doi:10.1016/j.isci.2025.114411

[35] Islam S, Haque R, Khan MA, Mohiuddin A Bin, Hossain Siddiqui MI, Limon ZH, et al. Ensemble Transformer with Post-hoc Explanations for Depression Emotion and Severity Detection. iScience. 2026 Feb 20;29(2):114605. doi:10.1016/j.isci.2025.114605

[36] Haque R, Sultana S, Rafy A, Babul Islam M, Arafat MA, Bhattacharya P, et al. A Transfer Learning-Based Computer-Aided Lung Cancer Detection System in Smart Healthcare. IET Conference Proceedings. 2024;2024(37):594–601. doi:10.1049/icp.2025.0858

[37] Khan S, Rahman H, Hossain Siddiqui MI, Hossain Limon Z, Khan MA, Haque R, et al. Ensemble-Based Explainable Approach for Rare Medicinal Plant Recognition and Conservation. 2025 10th International Conference on Information and Network Technologies, ICINT 2025. 2025;88–93. doi:10.1109/ICINT65528.2025.11030872

[38] Haque R, Mamun MA Al, Ratul MH, Aziz A, Mittra T. A Machine Learning Based Approach to Analyze Food Reviews from Bengali Text. 12th International Conference on Electrical and Computer Engineering, ICECE 2022. Published online 2022:80-83. doi:10.1109/ICECE57408.2022.10088971

[39] Nobel SMN, Swapno SMMR, Islam MR, Safran M, Alfarhood S, Mridha MF. A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. Scientific Reports 2024 14:1. 2024;14(1):1-25. doi:10.1038/s41598-024-64987-5

[40] Khushubu KG, Masum A Al, Rahman MH, et al. TransUNetB: An advanced Transformer–UNet framework for efficient and explainable brain tumor segmentation. Inform Med Unlocked. 2025;59(10):101706. doi:10.1016/j.imu.2025.101706

[41] Al Masum A, Limon ZH, Islam MA, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. Published online 2024:153-158. doi:10.1109/BECITHCON64160.2024.10962580

[42] Rahman S, Parameshachari BD, Haque R, Masfequier Rahman Swapno SM, Babul Islam M, Nobel SN. Deep Learning-Based Left Ventricular Ejection Fraction Estimation from Echocardiographic Videos. 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT 2023. Published online 2023. doi:10.1109/EASCT59475.2023.10392607