Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Two-Phase Transfer Learning for Accurate Multi-Class Cervical Cell Classification from Optical Microscopy Images

Md. Romzan Alom
Jagannath University
Muhammad Aminur Rahaman
aminur@bubt.edu.bd

Bangladesh University of Business and Technology
Md Abu Layek
Jagannath University

Research Article

Keywords: Cervical Cancer, Deep Learning, Two-Phase Training, MobileNet, Explainable AI, SipakMed, Optical Microscopy

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9256261/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable Two-Phase Transfer Learning for Accurate Multi-Class Cervical Cell Classification from Optical Microscopy Images

Md. Romzan Alom[1,2], Muhammad Aminur Rahaman[2,*], and Md Abu Layek[1,*]

[1]Department of Computer Science and Engineering, Jagannath University (JnU), Dhaka, Bangladesh
[2]Department of CSE, Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh
[*]Corresponding authors: layek@cse.jnu.ac.bd, aminur@bubt.edu.bd

ABSTRACT

Cervical cancer remains a significant global health concern, emphasizing the need for accurate and reliable automated cytology screening systems. Although deep learning has demonstrated strong potential for cervical cell analysis, many existing approaches suffer from limited interpretability and inconsistent generalization. This study proposes an explainable two-phase transfer learning framework named Deep Neural Cervical Cancer Detection (DNCCD) for accurate multi-class cervical cell classification from optical microscopy images. The framework employs pretrained convolutional neural networks within a two-phase transfer learning strategy, where initial feature extraction is followed by fine-tuning to improve classification performance and generalization. Comprehensive preprocessing techniques, including image resizing and intensity normalization, are applied to improve data consistency and robustness. Experiments are conducted on the publicly available SIPaKMeD cervical cytology dataset, which is systematically divided into training, validation, and testing subsets to ensure unbiased evaluation. As the core classification engine of DNCCD, the proposed Two-Phase MobileNet (TP-MobileNet) model achieves strong classification performance, achieving 92.03% accuracy, 92.24% F1-score and an AUC of 99%, demonstrating improved performance compared to baseline models. To enhance model transparency, explainable artificial intelligence techniques including Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME), are employed to provide clinically meaningful visual interpretation of prediction decisions. Furthermore, statistical analysis based on the paired t-test validates the robustness and reliability of the proposed framework. In addition, a user-friendly web application is developed to support the practical deployment of the proposed system for the screening of cervical cancer in the real-world.

Keywords: Cervical Cancer, Deep Learning, Two-Phase Training, MobileNet, Explainable AI, SipakMed, Optical Microscopy.

1 Introduction

Cervical cancer is a malignant neoplasm that develops in the cervix, the lower portion of the uterus that connects to the vagina[1,2]. It is predominantly caused by persistent infection with high-risk types of human papillomavirus (HPV), which is responsible for nearly all diagnosed cases[3,4]. Globally, cervical cancer ranks as the fourth most common cancer among women, posing a significant public health burden[5]. According to recent estimates, approximately 660,000 new cases and 350,000 deaths were reported worldwide in 2022[6,7]. The burden of the disease is disproportionately higher in low- and middle-income countries, primarily due to limited access to HPV vaccination, routine screening programs, and effective treatment facilities[8]. These disparities highlight the urgent need for affordable, accurate, and scalable diagnostic solutions. Cervical cancer develops through a series of cellular abnormalities that can be identified by distinct morphological patterns in cervical cytology[9]. Based on cytopathological characteristics, cervical cell images are commonly categorized into five major types: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial cells[10]. Dyskeratotic cells are associated with abnormal keratinization and are often indicative of high-grade squamous intraepithelial lesions or invasive carcinoma, representing a high risk of malignant progression. Koilocytotic cells exhibit perinuclear clearing and nuclear enlargement, which are hallmark features of HPV infection and are considered precancerous, carrying a moderate risk of progression to cervical cancer if persistent infection occurs[11]. Metaplastic cells arise during the transformation of glandular epithelium into squamous epithelium and are generally benign; however, they are clinically significant because the transformation zone is the primary site for cervical cancer development. Parabasal cells are immature epithelial cells typically observed in atrophic conditions or inflammatory states and may indicate abnormal epithelial regeneration, occasionally complicating cancer diagnosis. In contrast, Superficial cells represent mature, healthy squamous epithelial cells and are usually associated with normal cervical cytology and low cancer risk[12]. Accurate differentiation among these cell types is crucial for early detection, risk assessment, and appropriate clinical

management of cervical cancer[13].

Conventional cervical cancer screening methods, such as the Papanicolaou (Pap) smear test, have played a crucial role in reducing cervical cancer mortality through early detection[14,15]. Despite their effectiveness, these techniques suffer from several limitations. The screening process is often time-consuming, labor-intensive, and highly dependent on manual examination by skilled cytopathologists[16]. Moreover, diagnostic accuracy can vary significantly due to subjective interpretation and differences in expertise among clinicians, leading to potential false-positive or false-negative results[17]. These challenges motivate the exploration of automated and objective diagnostic approaches. In recent years, deep learning–based automated systems have emerged as promising alternatives for cervical cancer detection[18,19]. By leveraging convolutional neural networks (CNNs), these systems can analyze optical microscopy images of cervical cells and identify malignant patterns with high accuracy and efficiency[20]. However, despite their impressive performance, most deep learning models operate as "black boxes," offering limited transparency into how predictions are made[21,22]. This lack of interpretability poses a major barrier to clinical adoption, as healthcare professionals require explainable and trustworthy decision-support tools to confidently integrate artificial intelligence into medical practice[23]. To address these challenges, this study proposes a comprehensive decision-support framework named Deep Neural Cervical Cancer Detection (DNCCD) system for multi-class cervical cell classification. The DNCCD framework integrates an explainable two-phase transfer learning strategy to enhance feature learning and improve model robustness. As the core classification engine, a Two-Phase MobileNet (TP-MobileNet) model is developed, employing a two-stage training process consisting of feature extraction followed by fine-tuning. The system is evaluated using the publicly available SIPaKMeD cervical cytology dataset[24] containing optical microscopy images across five diagnostically relevant cell categories. To improve transparency and clinical trust, the DNCCD system incorporates explainable artificial intelligence techniques, including Gradient-weighted Class Activation Mapping (Grad-CAM)[25] and Local Interpretable Model-agnostic Explanations (LIME)[26], enabling visual interpretation of prediction outcomes. In addition, statistical significance analysis is conducted to validate the robustness and reliability of the proposed framework. Furthermore, a user-friendly web application is developed to support practical deployment of the DNCCD system for real-world cervical cancer screening assistance. The main contributions of this study are summarized as follows:

- We propose the DNCCD framework integrating a Two-Phase MobileNet (TP-MobileNet) model for automated multi-class cervical cell classification from optical microscopy images, using a two-stage transfer learning strategy to enhance performance and generalization.
- We incorporate explainable artificial intelligence techniques, including Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME), to provide clinically interpretable visual explanations of model predictions.
- We perform statistical significance analysis using paired t-tests to validate the robustness and reliability of the proposed framework.
- We evaluate the DNCCD system on the publicly available SIPaKMeD dataset and demonstrate improved classification performance compared with baseline models, achieving 92.03% accuracy with strong discriminative capability.
- We develop a user-friendly web-based application to support practical deployment of the DNCCD system for real-world cervical cytology screening.

This study aligns with the United Nations Sustainable Development Goals, particularly SDG 3: Good Health and Well-Being and SDG 9: Industry, Innovation, and Infrastructure. The proposed TP-MobileNet framework supports SDG 3 by enabling early and accurate cervical cancer detection, which can help reduce mortality through timely intervention. Additionally, by integrating transfer learning and explainable AI techniques, this work contributes to SDG 9 by promoting innovation and the development of reliable, scalable, and interpretable intelligent healthcare systems.

The structure of this paper is as follows: Section 2 discusses existing approaches to cervical cancer detection using optical microscopy images. Section 3 outlines the proposed DNCCD framework, including dataset characteristics, preprocessing techniques, and the TP-MobileNet classification model. Section 4 presents the experimental results and evaluates the model's performance. Finally, Section 5 explores the implications of our findings for clinical practice, summarizes the study, and outlines its limitations and directions for future research.

2 Literature Review

In recent years, machine learning (ML) and deep learning (DL) techniques have been extensively explored to enhance diagnostic accuracy and efficiency in cervical cancer screening[27,28]. The growing global burden of cervical cancer has accelerated research toward automated, reliable, and scalable detection systems[29]. In particular, image-based analysis of Pap smear datasets has enabled objective evaluation of cervical cell morphology, significantly improving early diagnosis and reducing dependence on manual cytological examination[30,31].

Kalbhor *et al.*[32] developed a cervical cancer detection framework utilizing pretrained CNN models for feature extraction combined with machine learning classifiers and fine-tuned transfer learning strategies. Although the approach achieved promising performance on the Herlev dataset, it was limited to binary classification tasks and lacked explainability analysis and large-scale clinical validation. Pacal *et al.*[18] conducted a comprehensive evaluation of numerous CNN and vision transformer models for cervical cell classification using the SIPaKMeD dataset. While ensemble learning improved diagnostic performance, the study primarily emphasized performance comparison and did not address model interpretability or deployment considerations. Lightweight and transfer learning–based architectures have also been investigated to improve computational efficiency. Mehedi *et al.*[33] proposed CCanNet, a compact deep learning model integrating squeeze blocks and residual connections. Although Grad-CAM was employed for interpretability, the framework relied on cropped single-cell images from a single dataset and lacked validation across diverse clinical scenarios. Similarly, Mathivanan *et al.*[34] introduced a hybrid framework combining deep transfer learning feature extractors with traditional machine learning classifiers. Despite achieving high accuracy, the approach focused on binary classification conversion and did not incorporate explainable analysis or multi-center dataset evaluation. Comparative studies of deep transfer learning models have also been reported. Kaur *et al.*[35] analyzed sixteen pretrained architectures and demonstrated strong classification performance. However, the study relied on relatively small datasets and did not propose a novel architecture or deployment-oriented framework.

Beyond classification-only approaches, researchers have explored data augmentation and segmentation-assisted learning strategies. Wubineh *et al.*[36] proposed a RES_DCGAN-based augmentation framework combined with self-attention pretrained models to enhance classification performance. Although improved accuracy was reported, the results were dependent on synthetic image quality and lacked extensive validation on real clinical datasets. In another study, Wubineh *et al.*[37] developed a dual deep learning framework integrating SPP-SegNet for segmentation and SE-DenseNet201 for classification. While strong performance was achieved, the framework was limited to single-cell images and lacked real-world clinical deployment evaluation. Subsequently, Wubineh *et al.*[38] proposed an integrated segmentation–classification architecture combining improved U-Net and ensemble classifiers. Despite high segmentation and classification accuracy, reliance on manually generated masks without expert validation may affect generalizability. Furthermore, Wubineh *et al.*[39] introduced an SE-DeepLabV3+ segmentation model with dynamic atrous convolution and an ensemble classification strategy. Although competitive results were reported, the use of pseudo-labeled masks and limited external testing restrict practical applicability. Deployment-oriented research has also emerged. Zammataro *et al.*[40] proposed CINNAMON-GUI, an open-source digital pathology platform for cervical cancer classification with interactive visualization features. However, performance was constrained by limited dataset diversity and potential overfitting risks. Austin *et al.*[41] developed a CNN-based framework capable of classifying whole-slide cervical cytology images without prior segmentation. Although promising performance was observed, accuracy declined in complex multi-class scenarios, highlighting the need for improved feature learning strategies and larger annotated datasets.

Despite these advancements, several research challenges remain. Many existing studies prioritize classification performance while overlooking interpretability, statistical reliability validation, and real-world deployment considerations. Moreover, conventional single-stage transfer learning strategies may lead to suboptimal generalization when distinguishing visually similar cervical cell types. In addition, limited research has focused on developing integrated screening systems that combine robust deep learning architectures, explainable decision-support mechanisms, and accessible deployment platforms for practical clinical use.

3 Proposed Methodology

The overall workflow of the proposed Deep Neural Cervical Cancer Detection (DNCCD) framework is illustrated in Fig. 1. The framework consists of several sequential stages including dataset acquisition, image preprocessing, multi-model training using a two-phase transfer learning strategy, performance evaluation, and explainable prediction analysis. This structured pipeline enables accurate and interpretable multi-class cervical cell classification from optical microscopy images.

3.1 Dataset Description

This study utilizes the publicly available SIPaKMeD cervical cytology dataset developed by Plissiti *et al.*[24] at the University of Ioannina, Greece. The dataset was collected in a clinical laboratory environment in Greece during routine Papanicolaou (Pap) smear screening procedures, ensuring strong clinical relevance for automated cervical cancer diagnosis research. Initially, 966 high-resolution Pap smear images were captured using an Olympus BX53F optical microscope equipped with a CCD digital camera (Infinity 1 Lumenera), enabling detailed visualization of cellular morphology. Since each smear image may contain multiple overlapping or adjacent cells, expert-guided manual cropping was performed to isolate individual cervical cell samples. This preprocessing step resulted in a total of 4,049 labeled single-cell images suitable for machine learning and deep learning–based classification studies. The dataset consists of five cytomorphological classes: superficial–intermediate, parabasal, metaplastic, koilocytotic, and dyskeratotic cells. These classes represent clinically meaningful diagnostic categories corresponding to normal, benign, and abnormal cellular conditions associated with cervical cancer progression. Such hierarchical labeling enhances the dataset's suitability for multi-class classification and computer-aided diagnostic system development. For experimental evaluation, the dataset is divided into training (70%), validation (15%), and testing (15%)

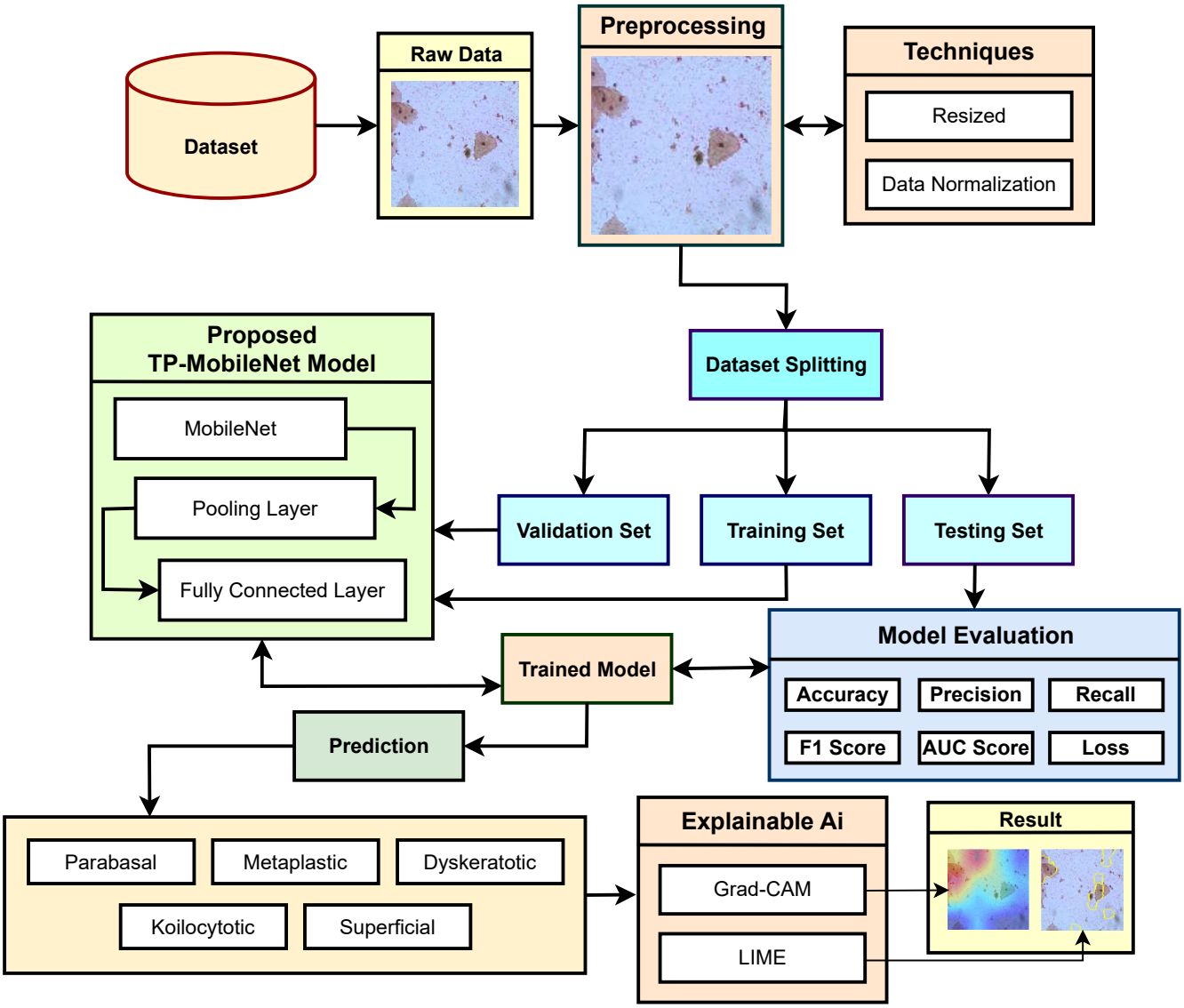
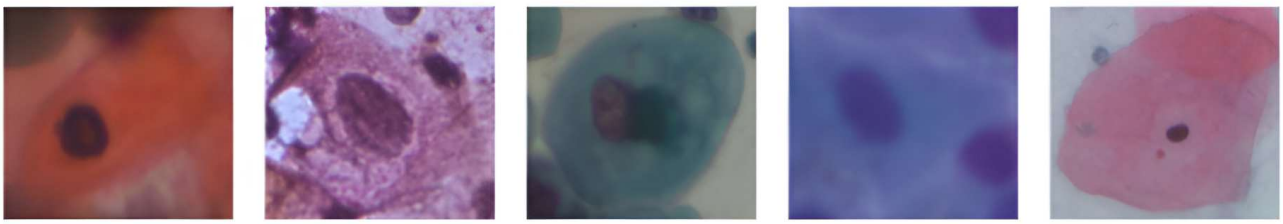


Figure 1. Overall workflow of the proposed DNCCD framework for cervical cytology image classification.

subsets using random sampling to ensure unbiased performance assessment. Furthermore, the SIPaKMeD dataset exhibits significant intra-class and inter-class variability in cellular structure, staining intensity, texture patterns, and imaging conditions, closely reflecting real-world cytological screening scenarios. Representative examples of different cervical cell categories are illustrated in Fig. 2.



(a) Dyskeratotic (b) Koilocytotic (c) Metaplastic (d) Parabasal (e) Superficial

Figure 2. Representative cervical cell images from the SIPaKMeD dataset across five diagnostic categories.

3.2 Image Preprocessing

To ensure uniform input representation, all images are resized to a spatial resolution of 128×128 pixels. Let an original image be denoted as $I(x,y)$; the resized image $I_r(x,y)$ is obtained through interpolation as defined in Equation 1.

$$I_r = \text{Resize}(I, 128, 128) \tag{1}$$

Pixel intensity normalization is subsequently applied to improve numerical stability during training. Given original intensity values in the range $[0,255]$, normalized images I_n are computed using Equation 2.

$$I_n = \frac{I_r}{255} \tag{2}$$

These preprocessing operations reduce illumination variability and facilitate faster convergence of deep learning models.

3.3 Model Training Strategy

To evaluate the effectiveness of transfer learning for cervical cytology image classification, several state-of-the-art convolutional neural network architectures including MobileNet, VGG16, ResNet18, DenseNet121, EfficientNetB0, and InceptionV3 were trained and analyzed within the proposed DNCCD framework. In addition to their conventional configurations, enhanced variants of these architectures were developed using a two-phase transfer learning strategy. Among the investigated models, the Two-Phase MobileNet (TP-MobileNet) demonstrated superior classification performance and was therefore selected as the core classification module of the DNCCD system.

The architectural workflow of the TP-MobileNet model is illustrated in Fig. 3. As shown in the figure, resized cervical cell images are first provided as input to a pretrained MobileNet backbone network, which functions as a deep feature extractor. During the first phase of training, the convolutional layers of the backbone network are frozen to preserve generic visual representations learned from large-scale image datasets. The extracted feature maps are then passed through a Global Average Pooling (GAP) layer to reduce spatial dimensionality and minimize overfitting. A custom classification head consisting of fully connected layers is subsequently added to learn task-specific discriminative features. The predicted probability distribution over the cervical cell classes is computed using the softmax function as expressed in Equation 3.

$$\hat{y} = \text{Softmax}(W f_\theta(x) + b) \tag{3}$$

The optimization objective during this feature extraction phase is defined using categorical cross-entropy loss, as formulated in Equation 4.

$$\mathscr{L}_1 = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4}$$

In the second phase, all network layers are unfrozen and the entire model is fine-tuned using a smaller learning rate to refine feature representations and improve generalization capability. The fine-tuning optimization objective incorporates an ℓ_2 regularization term to enhance model stability and reduce overfitting, as expressed in Equation 5.

$$\mathscr{L}_2 = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) + \lambda \|\theta\|_2^2 \tag{5}$$

This two-stage learning strategy enables progressive adaptation of pretrained features to cervical cell morphology. As depicted in Fig. 3, the final softmax output layer produces multi-class prediction probabilities corresponding to the five diagnostic cervical cell categories. The trained TP-MobileNet model is subsequently integrated into the DNCCD framework for performance evaluation, explainability analysis, and deployment in a clinical decision-support environment.

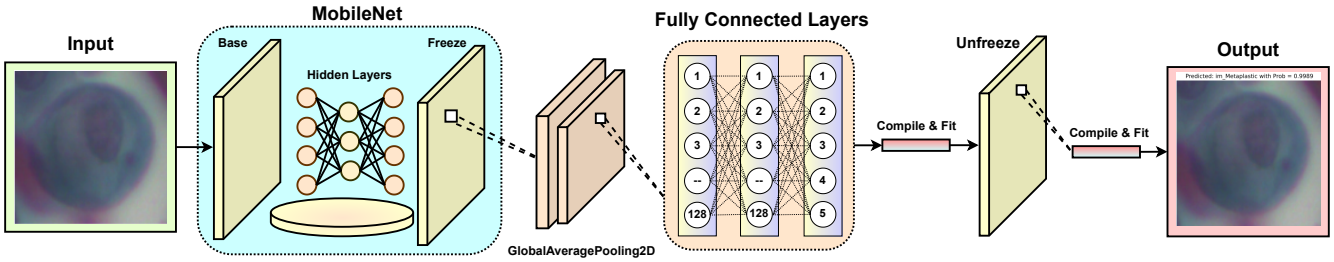


Figure 3. Architecture of the proposed Two-Phase MobileNet (TP-MobileNet) model showing feature extraction using a pretrained MobileNet backbone, Global Average Pooling, fully connected classification layers, fine-tuning stage, and final multi-class prediction output.

3.4 Model Testing and Evaluation

After completing the training phase, the developed models are evaluated using the independent test set to assess their generalization capability on previously unseen cervical cytology images. The testing process is conducted within the DNCCD framework to determine the effectiveness of each model in multi-class cervical cell classification. Model performance is quantitatively measured using several standard evaluation metrics, including classification accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). These metrics collectively provide a comprehensive understanding of the model's predictive performance, class discrimination ability, and robustness across different cervical cell categories. Following performance evaluation, statistical significance analysis using paired t-tests is performed to compare the proposed TP-MobileNet model with baseline deep learning architectures. Based on the overall evaluation results, the model demonstrating the best classification performance and statistical reliability is selected as the final classification module of the DNCCD system for subsequent explainability analysis and real-world deployment.

3.5 Model Explainability Analysis

After completing the classification stage, the next step of the DNCCD framework focuses on improving interpretability to support reliable clinical decision-making. To achieve this, explainable artificial intelligence (XAI) techniques are applied to analyze and visualize the prediction behavior of the selected classification model. Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to generate visual heatmaps that highlight the most influential regions of cervical cell images contributing to classification outcomes. These localization maps enable clinicians to understand where the model focuses while identifying different cytological patterns. In addition, Local Interpretable Model-agnostic Explanations (LIME) is utilized to provide instance-level interpretability by identifying important image features responsible for specific predictions. This approach offers complementary insights into model decision logic and enhances transparency. By integrating Grad-CAM and LIME within the DNCCD system, the proposed framework ensures that classification results are not only accurate but also interpretable. This interpretability-driven analysis supports trustworthy deployment of automated cervical cytology screening systems in real-world clinical environments.

4 Result Analysis

4.1 Experimental Setup

All experiments were conducted on the Kaggle cloud platform, which provides a stable computational environment for deep learning applications. The hardware configuration includes NVIDIA Tesla T4 GPUs, an Intel Xeon CPU, and approximately 13 GB RAM. The DNCCD framework was implemented using Python with TensorFlow and Keras libraries. Model training was performed using a two-phase transfer learning strategy. In the first phase, pretrained backbone networks were used as frozen feature extractors while the newly added classification layers were trained using a higher learning rate. In the second phase, the entire network was unfrozen and fine-tuned using a lower learning rate to improve feature adaptation and generalization performance. To ensure reproducibility, a fixed random seed value was used during data splitting and model training. The key training hyperparameters used for the proposed TP-MobileNet model are summarized in Table 1.

Table 1. Training hyperparameter configuration for the proposed TP-MobileNet model.

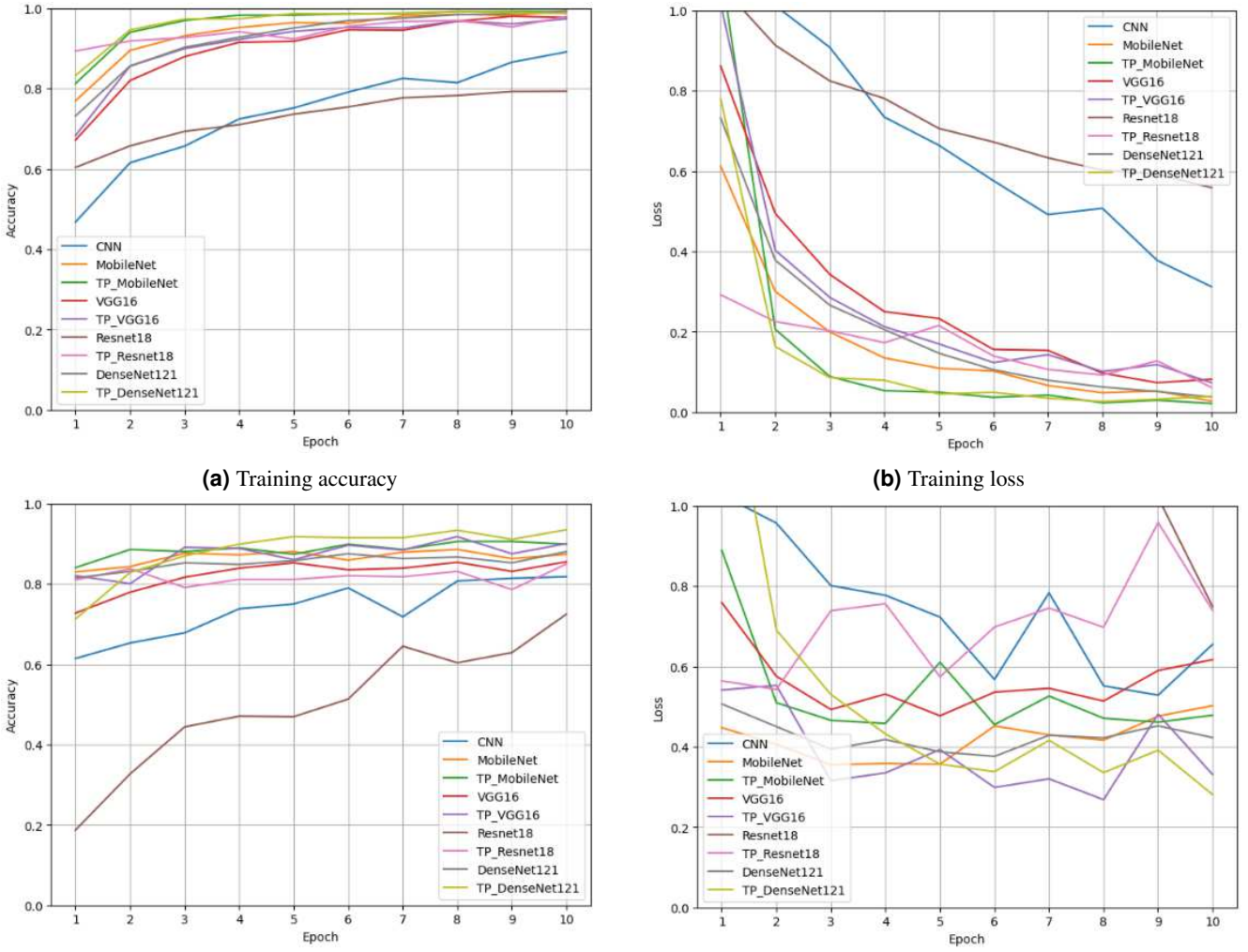
S.No.	Parameter	Value
1	Input Image Size	$128 \times 128 \times 3$
2	Optimizer	Adam
3	Learning Rate (Phase-1)	1×10^{-3}
4	Learning Rate (Phase-2)	1×10^{-4}
5	Batch Size	32
6	Number of Epochs	10
7	Kernel Size	3×3
8	Activation Function	ReLU
9	Output Activation	Softmax
10	Loss Function	Categorical Cross-Entropy
11	Random Seed Value	42

4.2 Performance Measurement Parameters

The performance of the proposed DNCCD framework is evaluated using several standard classification metrics to provide a comprehensive assessment of its effectiveness in multi-class cervical cell classification. A confusion matrix is first employed to summarize the classification outcomes by identifying correctly and incorrectly predicted samples across different cell categories. Based on this analysis, multiple evaluation metrics including accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (AUC) are computed. Accuracy measures the overall correctness of the model's predictions, while precision reflects the reliability of positive class predictions. Recall (sensitivity) evaluates the model's ability to correctly identify actual instances of each cervical cell type. The F1-score provides a balanced measure by combining precision and recall, making it particularly useful for handling class-wise performance variations. In addition, ROC curve analysis is conducted to examine the discriminative capability of the model across different decision thresholds, and the corresponding AUC value is used as an aggregate performance indicator. Together, these performance measurement parameters offer a robust and comprehensive evaluation of the classification capability, reliability, and generalization performance of the proposed DNCCD system.

4.3 Result Discussion

In this study, Fig. 4 presents the training and validation performance curves of all trained models across epochs, including CNN, MobileNet, VGG16, ResNet18, DenseNet121, and their corresponding TP-based variants. The loss curves illustrate the optimization behavior during training and validation, highlighting differences in convergence speed and stability among the models. Similarly, the accuracy curves demonstrate how each architecture learns and generalizes over time. Overall, the figure provides an initial comparative overview of model learning dynamics, revealing the impact of architectural choices and TP integration on convergence behavior, generalization capability, and performance consistency. These observations form the basis for the detailed performance analysis discussed in the following sections. Fig. 5 presents the confusion matrices of different trained models, illustrating the relationship between the actual class labels and the predicted class labels. These visual representations provide an intuitive understanding of the classification behavior of each architecture by highlighting correct predictions along the main diagonal and misclassifications in the off-diagonal regions. Models with stronger performance exhibit a higher concentration of correctly classified samples on the diagonal, whereas weaker models show increased confusion among classes.



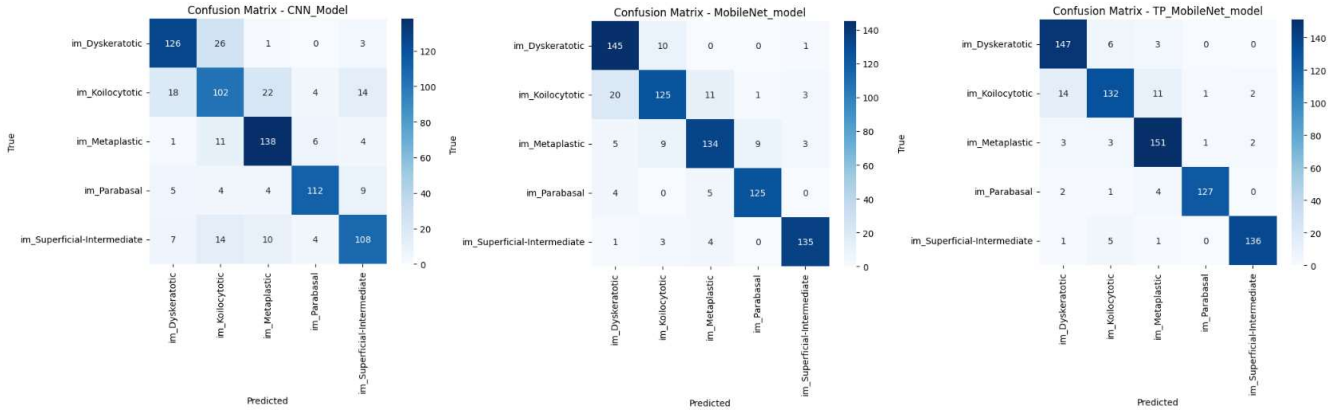
(a) Training accuracy

(b) Training loss

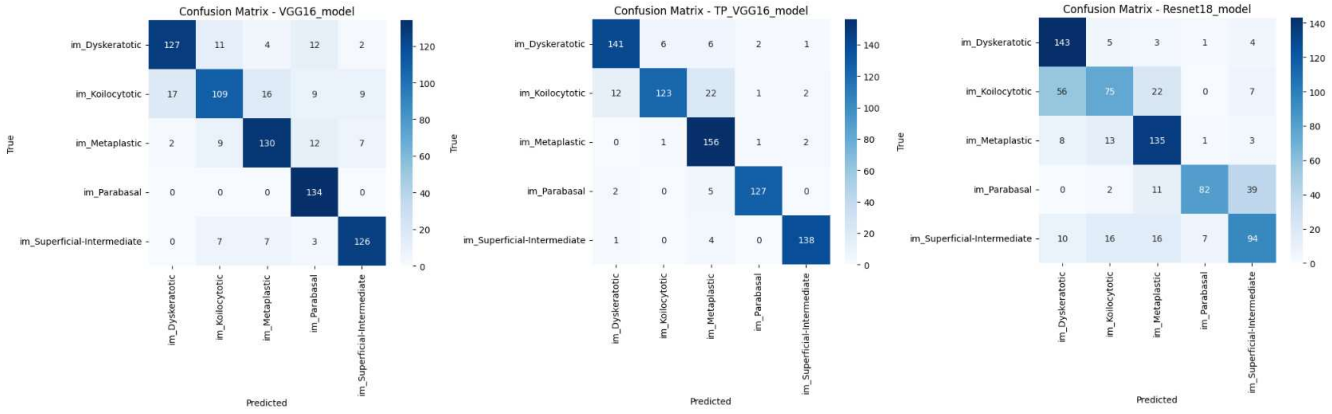
(c) Validation accuracy

(d) Validation loss

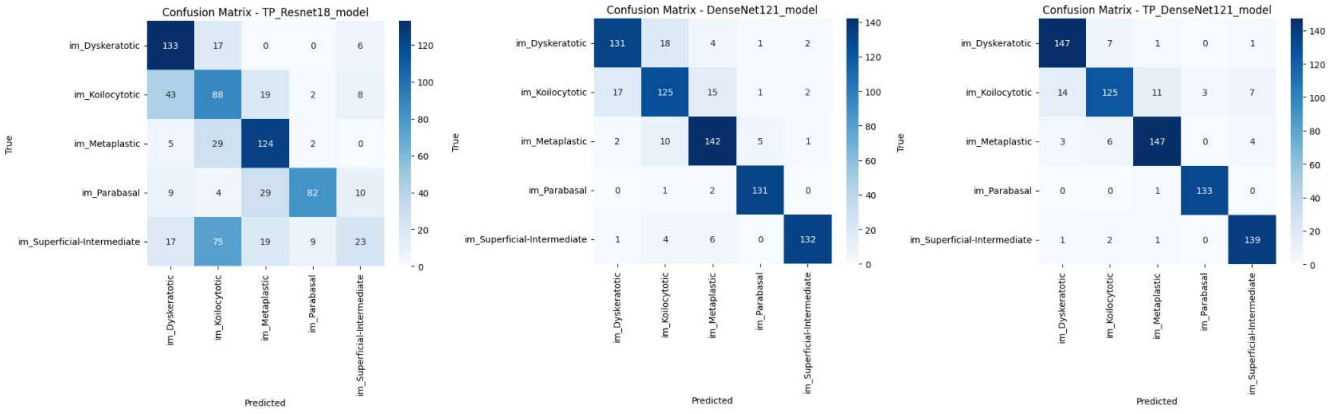
Figure 4. Training and validation performance curves for all trained models. The training accuracy and loss plots demonstrate the learning behavior and convergence trends during optimization, while the validation accuracy and loss curves evaluate the generalization performance on unseen data. These curves collectively provide insight into training stability, convergence characteristics, and the presence of overfitting or underfitting across the models.



(a) CNN **(b)** MobileNet **(c)** TP-MobileNet



(d) VGG **(e)** TP-VGG **(f)** ResNet



(g) TP-ResNet **(h)** DenseNet **(i)** TP-DenseNet

Figure 5. Confusion matrices of different trained models showing the actual labels versus the predicted labels.

In this research, Table 2 provides a detailed performance comparison between conventional deep learning models and their TP-enhanced variants, evaluated on the test dataset. Among the baseline architectures, MobileNet and DenseNet121 exhibit relatively strong performance, achieving accuracies of 0.8818 and 0.8778, respectively. In contrast, CNN and ResNet18 show lower accuracy values paired with higher loss, indicating their limited generalization capabilities. These trends are further illustrated in the confusion matrices presented in Fig.5, where weaker models demonstrate increased off-diagonal misclassifications. The introduction of the proposed TP enhancement leads to significant performance improvements for most architectures. Notably, TP-MobileNet shows a considerable and consistent enhancement over standard MobileNet across all evaluation metrics. Specifically, its accuracy rises from 0.8818 to 0.9203, precision from 0.8851 to 0.9245, recall from 0.8850 to 0.9220, and F1-score from 0.8841 to 0.9224, accompanied by a reduction in loss. These improvements are visually supported by the clear diagonal dominance and decreased class confusion observed in the TP-MobileNet confusion matrix depicted in

Fig.5. Similar performance enhancements are noted for TP-VGG16 and TP-DenseNet121, both of which surpass their baseline counterparts by achieving higher accuracy, F1-score, and AUC values while significantly lowering loss. However, the TP enhancement does not uniformly benefit all architectures. For instance, TP-ResNet18 demonstrates degraded performance compared to the original ResNet18, as evidenced by lower accuracy, reduced AUC, increased loss in Table 2, and greater class confusion in Fig. 5. This outcome suggests that the effectiveness of the proposed TP enhancement may vary depending on the underlying network architecture and its ability to learn features effectively.

Table 2. Performance comparison of different trained models on the test dataset

Model	Accuracy	Precision	Recall	F1-score	AUC	Loss
CNN[18]	0.7782	0.7825	0.7798	0.7805	0.9518	0.7248
MobileNet[33]	0.8818	0.8851	0.8850	0.8841	0.9832	0.4868
TP-MobileNet	0.9203	0.9245	0.9220	0.9224	0.9900	0.4180
VGG16[42]	0.8313	0.8325	0.8378	0.8315	0.9711	0.6530
TP-VGG16	0.9097	0.9186	0.9121	0.9119	0.9890	0.3612
ResNet18[34]	0.7025	0.7194	0.6997	0.6951	0.9311	0.8078
TP-ResNet18	0.5976	0.6115	0.5901	0.5739	0.8668	1.0457
DenseNet121[39]	0.8778	0.8823	0.8818	0.8818	0.9823	0.4376
TP-DenseNet121	0.9177	0.9191	0.9214	0.9192	0.9926	0.3110

Additionally, a statistical significance analysis was performed to assess the comparative performance of the proposed TP-MobileNet model against other deep learning architectures. A paired t-test was conducted using five classification performance metrics, namely accuracy, precision, recall, F1-score, and AUC, as paired observations. The statistical comparison results presented in Table 3 demonstrate that the proposed TP-MobileNet model achieved statistically significant performance improvements over most baseline and transfer learning models, including CNN, MobileNet, VGG16, TP-VGG16, ResNet18, TP-ResNet18, and DenseNet121, as all corresponding p-values were below the significance level of 0.05. In contrast, the comparison between TP-MobileNet and TP-DenseNet121 did not reveal a statistically significant difference ($p = 0.2440$), indicating that both models provide comparable classification performance on the test set. Overall, the statistical findings in Table 3 further validate the effectiveness and robustness of the proposed TP-MobileNet framework for multi-class cervical cell classification.

Table 3. Statistical significance analysis of the proposed TP-MobileNet model compared with competing models using paired t-test based on classification performance metrics.

Comparison Model	**t-value**	**p-value**	**Significance**
TP-MobileNet vs CNN	5.8392	0.0043	Significant
TP-MobileNet vs MobileNet	5.0700	0.0071	Significant
TP-MobileNet vs VGG16	5.3235	0.0060	Significant
TP-MobileNet vs TP-VGG16	4.0766	0.0151	Significant
TP-MobileNet vs ResNet18	5.8108	0.0044	Significant
TP-MobileNet vs TP-ResNet18	6.9233	0.0023	Significant
TP-MobileNet vs DenseNet121	5.1322	0.0068	Significant
TP-MobileNet vs TP-DenseNet121	1.3648	0.2440	Not Significant

Further insight into the discriminative capabilities of the models is illustrated by the ROC curves presented in Fig. 6. These curves allow for a comparative analysis of the models' performance across various classification thresholds. The ROC curve visualizes the trade-off between the true positive rate and the false positive rate; a curve that approaches the top-left corner indicates improved classification performance. As seen in Fig. 6, the TP-MobileNet model consistently outperforms the other architectures, achieving a higher true positive rate at lower false positive rates. This superior performance is further validated by its high Area Under the Curve (AUC) value, which suggests strong class separability and effective decision-making. In contrast, baseline models such as CNN and ResNet18 display relatively weaker ROC characteristics, with curves that lie closer to the random-guess diagonal. The enhanced ROC performance of TP-MobileNet corresponds with the quantitative results shown in Table 2 and the confusion matrix analysis in Fig. 5. This reinforces TP-MobileNet's effectiveness, establishing it as the most reliable and accurate model proposed in this study.

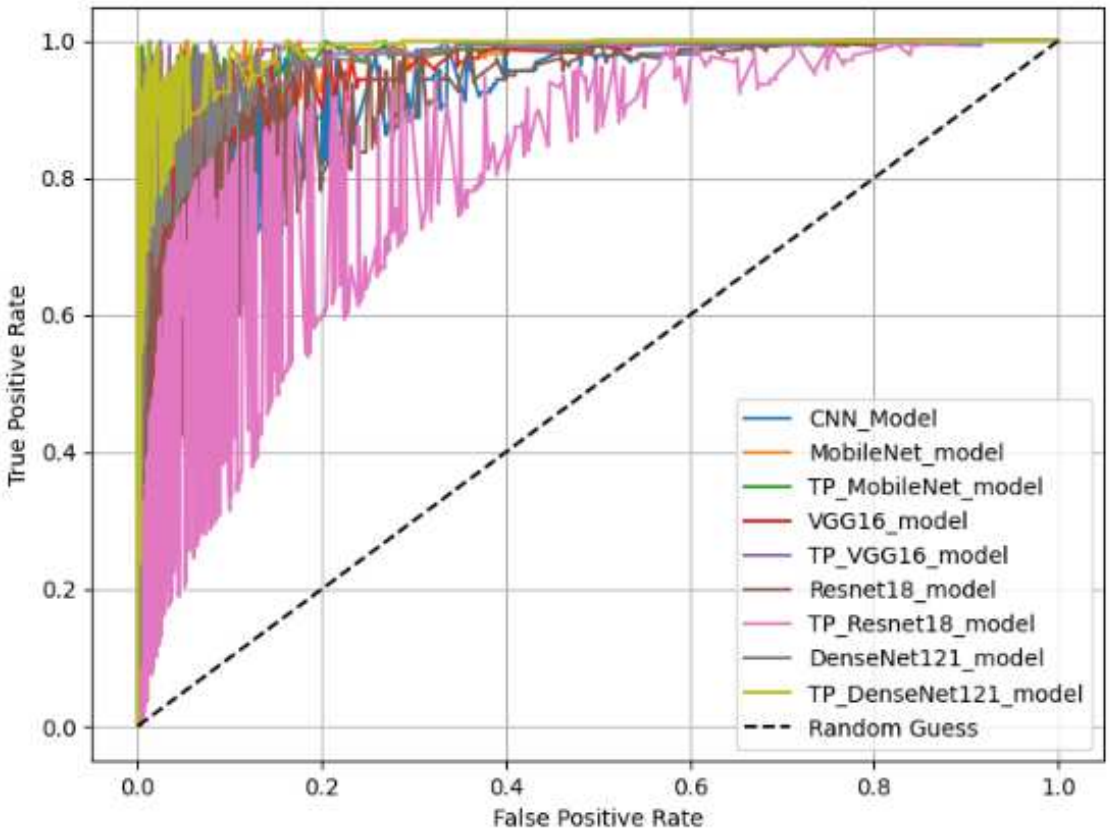


Figure 6. ROC curves of different trained models on the test dataset.

In addition, Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11 illustrate the explainable classification outcomes of the proposed deep learning model for Parabasal, Metaplastic, Dyskeratotic, Koilocytotic, and Superficial–Intermediate cervical cell types, respectively. For each figure, the first image presents the original cervical cell image used as input to the model, while the second image shows the corresponding predicted class, confirming the model's classification capability. The third image provides the Grad-CAM visualization, which highlights the most discriminative and spatially relevant regions within the cell image that significantly influenced the model's decision-making process. Furthermore, the fourth image displays the LIME-based explanation, offering a localized, feature-level interpretation by identifying the image regions that contributed positively or negatively to the final prediction. Collectively, Fig. 7–Fig. 11 demonstrate not only the classification performance of the proposed approach but also its transparency and reliability by visually validating that the model focuses on clinically meaningful and diagnostically relevant regions across different cervical cell categories.

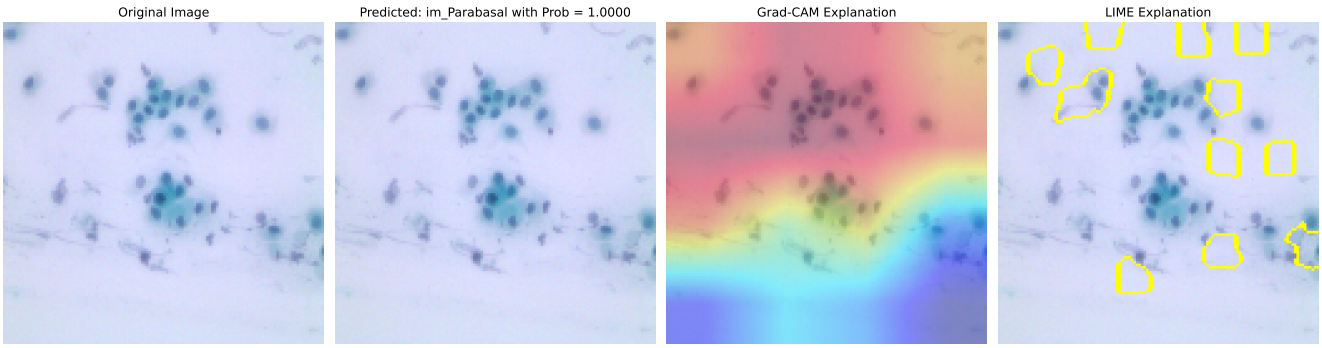


Figure 7. Parabasal cell classification results: the first image shows the original input image, the second image presents the predicted output identifying the Parabasal cell type, the third image illustrates the Grad-CAM visualization highlighting the regions contributing to the model's decision, and the fourth image displays the LIME explanation indicating feature importance.

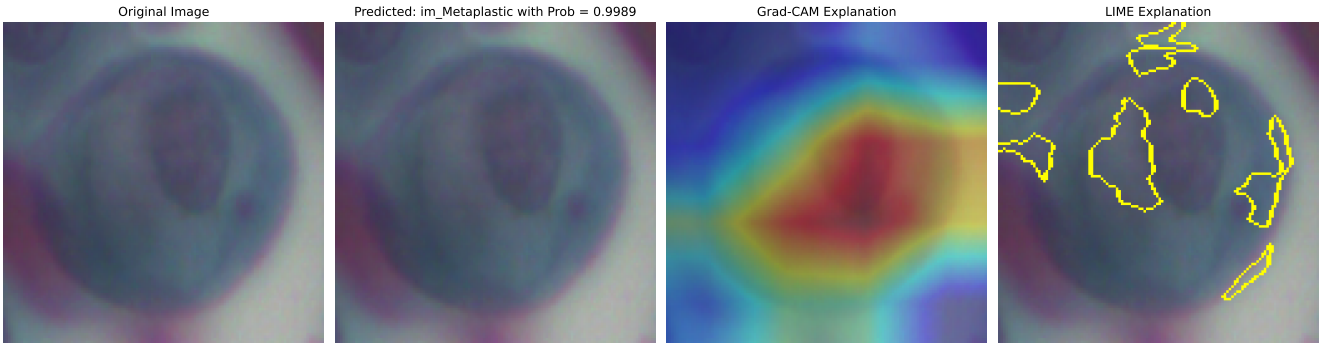


Figure 8. Metaplastic cell classification results: the first image represents the original input image, the second image shows the predicted output classified as Metaplastic, the third image provides the Grad-CAM explanation emphasizing discriminative regions, and the fourth image presents the LIME-based interpretability visualization.

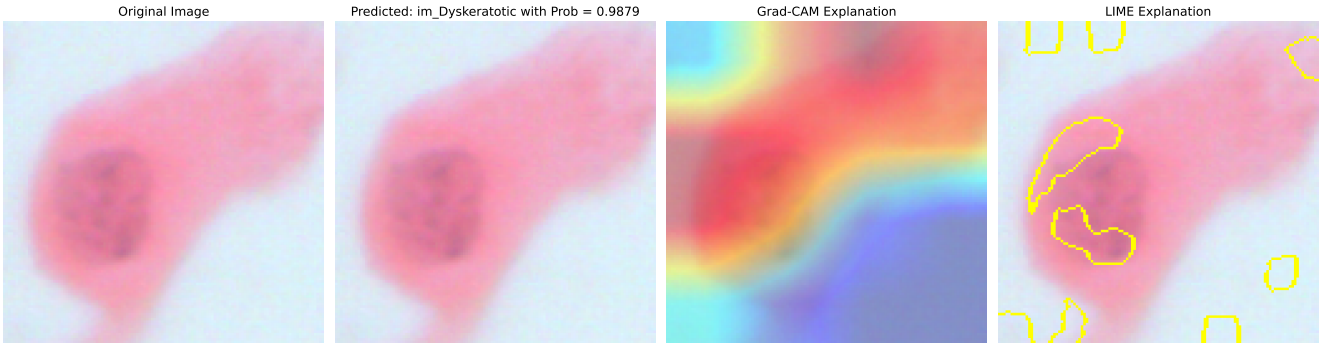


Figure 9. Dyskeratotic cell classification results: the first image is the original input image, the second image displays the predicted Dyskeratotic cell classification, the third image demonstrates the Grad-CAM heatmap explaining the model's focus areas, and the fourth image shows the LIME explanation highlighting influential features.

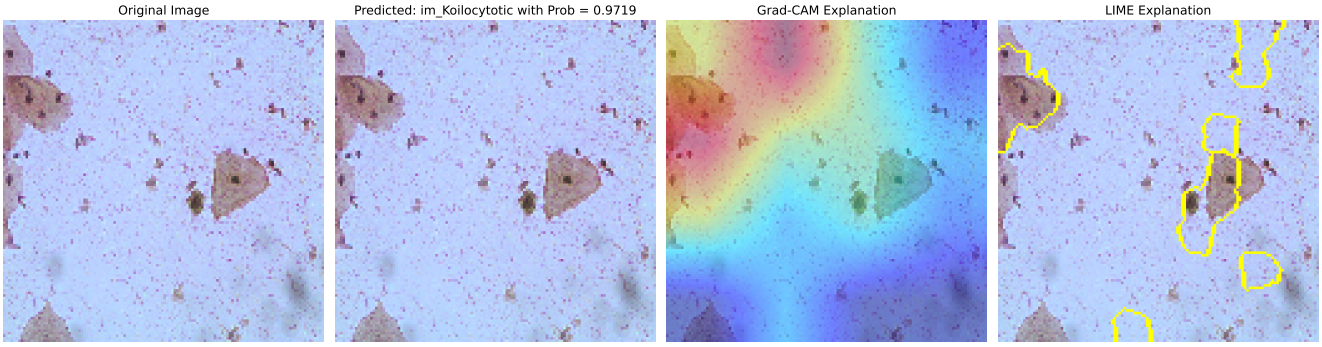


Figure 10. Koilocytotic cell classification results: the first image illustrates the original input image, the second image shows the predicted Koilocytotic cell type, the third image presents the Grad-CAM visualization for model interpretability, and the fourth image provides the LIME explanation of feature contributions.

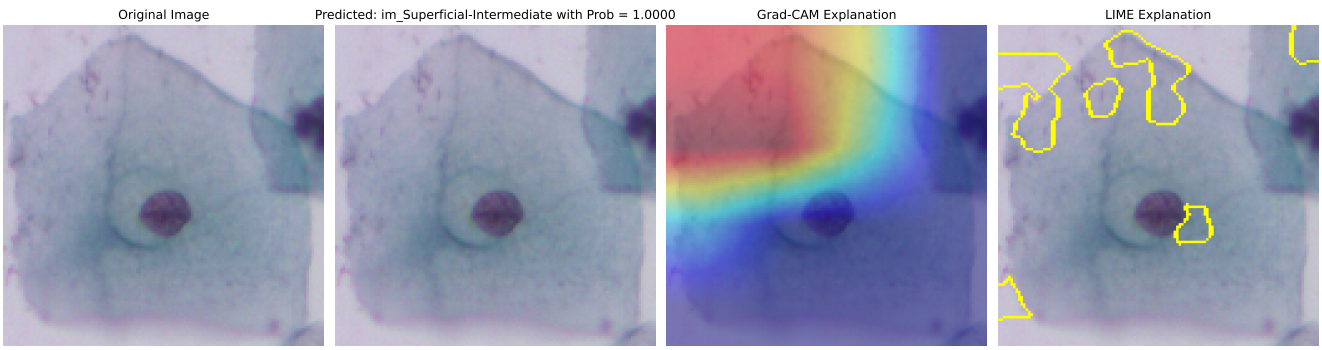


Figure 11. Superficial–Intermediate cell classification results: the first image depicts the original input image, the second image represents the predicted Superficial–Intermediate cell type, the third image displays the Grad-CAM explanation highlighting key regions, and the fourth image shows the LIME-based interpretability analysis.

In addition to classification performance and explainability analysis, the practical applicability of the proposed DNCCD framework is demonstrated through the development of a web-based decision support tool, as illustrated in Fig. 12. The interface enables users to upload cervical cytology images and obtain real-time classification results along with the predicted diagnostic category. This deployment highlights the potential of the proposed system for assisting early cervical cancer screening in clinical and remote healthcare settings. The web application provides an accessible platform that integrates automated prediction and visual interpretability, thereby facilitating user interaction, supporting clinical workflow, and promoting the translational impact of the DNCCD framework in real-world scenarios.

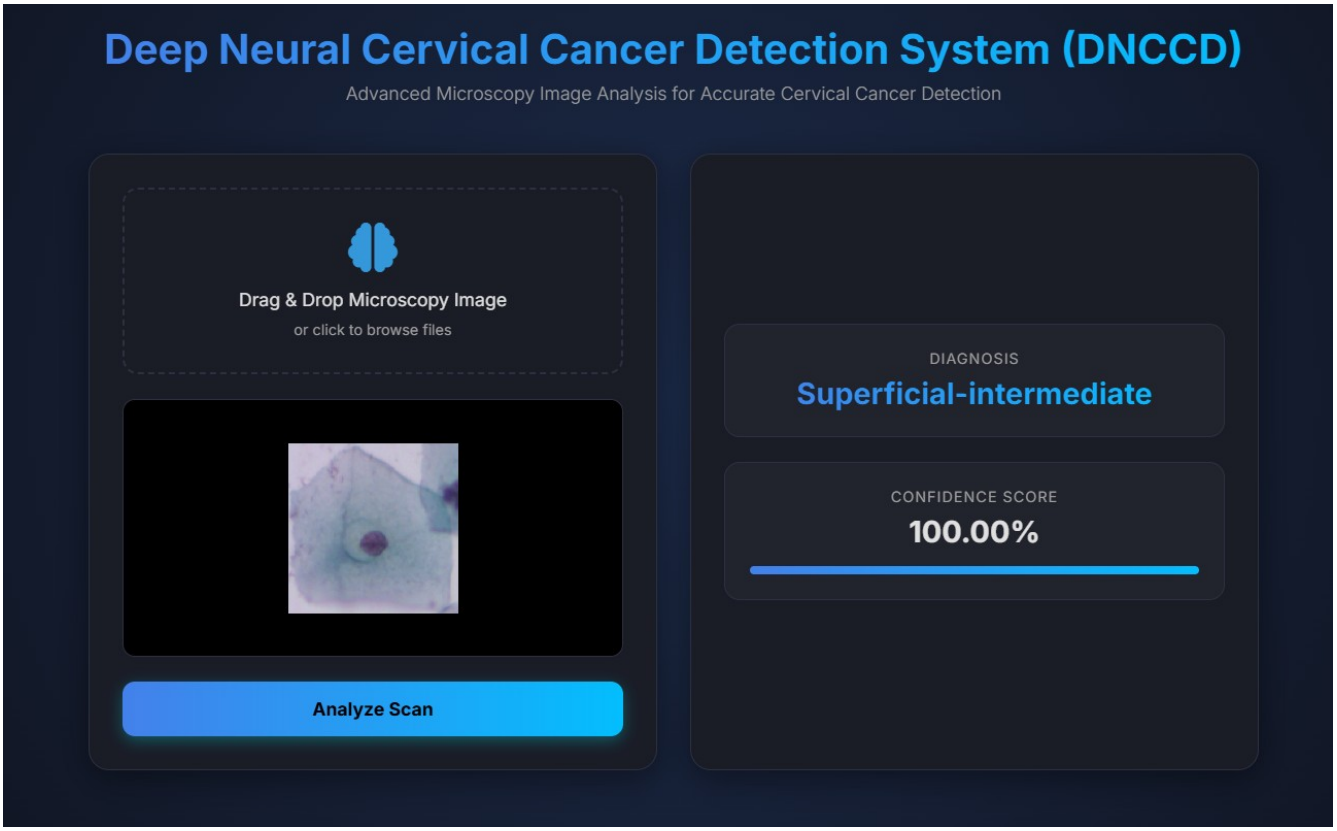


Figure 12. Web application using the DNCCD system where the tool works very well accordingly.

Finally, Table 4 summarizes and critically compares recent deep learning–based cervical cancer detection studies with the proposed DNCCD framework. Most existing methods focus primarily on improving classification accuracy through transfer learning, hybrid deep learning–machine learning models, segmentation-guided pipelines, or data augmentation strategies. However, these studies generally lack rigorous statistical validation, comprehensive explainability mechanisms, and practical

Table 4. Comparison of existing cervical cancer detection studies with the proposed DNCCD framework.

Author	Method	Training Strategy	Statistical Validation	Explainable AI	Deployment Tool
Kalbhor et al.[32]	Pretrained CNN feature extraction + ML classifiers; transfer learning with CNNs	Feature-based learning and fine-tuning	No	No	No
Pacal et al.[18]	Large-scale comparison of CNN and ViT models with ensemble learning	Transfer learning with data augmentation and ensemble training	No	No	No
Mehedi et al.[33]	Lightweight CNN (CCanNet) with squeeze blocks, residual blocks, and skip connections	End-to-end training with augmentation	No	Grad-CAM + Guided Grad-CAM	No
Mathivanan et al.[34]	Hybrid DL feature extraction + ML classifiers using pretrained networks	Transfer learning-based feature extraction	No	No	No
Kaur et al.[35]	Comparative analysis of multiple pre-trained transfer learning models	Fine-tuning of pretrained CNNs	No	No	No
Wubineh et al.[36]	RES_DCGAN augmentation + self-attention integrated pretrained CNNs	Conventional training with synthetic augmentation	No	No	No
Wubineh et al.[37]	SPP-SegNet segmentation + SE-DenseNet201 classification	Multi-stage segmentation and classification training	No	No	No
Wubineh et al.[39]	SE-DeepLabV3+ segmentation + weighted ensemble classification	Multi-stage training with segmentation-guided classification	No	Grad-CAM + Saliency Map	No
Wubineh et al.[38]	Improved U-Net segmentation + ensemble classification with RES_DCGAN	Multi-stage training with augmentation before/after segmentation	No	No	No
Zammataro et al.[40]	CNN-based GUI platform for Pap smear image classification	Conventional CNN training with configurable augmentation	No	Feature map visualization	Yes
Austin et al.[41]	Custom CNN for direct whole-slide/cell image classification	End-to-end training	No	No	No
Proposed DNCCD	TP-MobileNet + explainable diagnostic framework	**Two-Phase Transfer Learning**	**Paired t-test**	**Grad-CAM + LIME**	**Web-based tool**

deployment solutions for real-world clinical use. In contrast, the proposed DNCCD framework integrates a two-phase transfer learning strategy with paired t-test–based statistical analysis, incorporates Grad-CAM and LIME for model interpretability, and provides a web-based diagnostic platform. These combined contributions enhance both the reliability and practical applicability of automated cervical cancer detection systems.

5 Conclusion

In this study, we proposed an explainable deep learning–based cervical cancer screening system named Deep Neural Cervical Cancer Detection (DNCCD) for multi-class classification of cervical cytology images. The system integrates a Two-Phase MobileNet (TP-MobileNet) model as the core classification module, where pretrained features are first utilized for feature extraction and subsequently fine-tuned to improve task-specific representation learning and generalization capability. Standard preprocessing procedures were applied to ensure consistent data quality and stable model training. Experimental findings demonstrate that the proposed TP-MobileNet model consistently surpasses baseline deep learning architectures, with statistical significance analysis confirming the robustness and reliability of the proposed framework. The incorporation of explainable artificial intelligence techniques, including Grad-CAM and LIME, enhances model transparency by providing clinically meaningful visual interpretation of prediction outcomes. To support real-world applicability, a user-friendly web-based application was also developed as part of the DNCCD system, enabling practical deployment for automated cervical cytology screening. Overall, the proposed framework highlights the potential of explainable two-phase transfer learning as an effective and trustworthy decision-support solution for early cervical cancer detection. Despite these promising findings, the study has certain limitations. The SIPaKMeD dataset presents class imbalance across diagnostic categories, which may affect model robustness under diverse clinical conditions. Moreover, the current framework is limited to optical microscopy images and has not yet been validated on multi-institutional or multimodal medical datasets. Future work will focus on expanding the dataset with more diverse samples, investigating advanced imbalance-handling strategies, and extending the DNCCD system to support multimodal learning by incorporating histopathology images and clinical metadata. Additionally, further optimization of the web-based platform for clinical integration and large-scale deployment will be explored.

Declarations

Competing interests

The author(s) declare no competing interests.

Funding

No external funding was received for this study.

Ethics declaration
Not applicable.

Data availability statement
The dataset used in this study is the publicly available SIPaKMeD cervical cytology image dataset developed by Plissiti *et al.*[24] at the University of Ioannina, Greece. It can be accessed from the Kaggle repository at: https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed.

Code availability
The code supporting this study is available at https://github.com/romzanalom/Explainable_AI_Cervical_Cancer_Detection_TP-MobileNet and can be provided upon request.

Authors contribution
Md. Romzan Alom contributed to formal analysis, conceptualization, methodology, visualization, validation, implementation, writing the original draft, and reviewing and editing. Md. Abu Layek and Muhammad Aminur Rahaman were involved in formal analysis, conceptualization, investigation, resources, methodology, supervision and reviewing and editing. All authors have read and agreed to the published version of the manuscript.

AI usage declaration
During the preparation of this work, the author(s) used AI tools (e.g., ChatGPT, Grammarly) solely to improve language and readability. No AI tools were used to develop the main contents, including the methodology, results, discussion, figures, or data. All contents that are refined with AI assistance were thoroughly reviewed and edited by the author(s), who take full responsibility for the final version of the work.

References

1. Ray, A. Human papillomavirus and other relevant issues in cervical cancer pathogenesis. *Int. J. Mol. Sci.* **26**, 5549 (2025).
2. Mahajan, P., Kaur, P., Singh, K. & Cheema, A. K. Comparative analysis for classification of cervical cancer cells using binary red deer algorithm and resnet framework. *Biomed. Signal Process. Control.* **113**, 109107 (2026).
3. Nie, F., Si, H., Qin, H., Wang, Y. & Wang, Y. Cervical and vaginal metastases from ascending colon cancer: a case report and literature review. *BMC Women's Heal.* **25**, 1–12 (2025).
4. Zhao, M. *et al.* Characteristic of persistent human papillomavirus infection in women worldwide: a meta–analysis. *PeerJ* **11**, e16247 (2023).
5. Perkins, R. B., Wentzensen, N., Guido, R. S. & Schiffman, M. Cervical cancer screening: a review. *Jama* **330**, 547–558 (2023).
6. World Health Organization (WHO). Cervical cancer. WHO Fact Sheet (2024). Accessed July 2025.
7. World Health Organization. Cervical cancer. WHO Health Topics webpage (2025). Accessed 28 July 2025.
8. Singh, D. *et al.* Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the who global cervical cancer elimination initiative. *The lancet global health* **11**, e197–e206 (2023).
9. Jain, S., Jain, A. & Jangid, M. Mgwo-cnn: hyperparameter optimization of cnn classifier for cervical cancer detection using modified grey wolf optimizer. *Sci. Reports* (2025).
10. Mohanta, A., Roy, S. D., Nath, N., Datta, A. & Bhowmik, M. K. A comprehensive survey on diagnostic microscopic imaging modalities, challenges, taxonomy, and future directions for cervical abnormality detection and grading. *IEEE Transactions on Artif. Intell.* (2025).
11. Dogan, Y. Autoefffusionnet: A new approach for cervical cancer diagnosis using resnet-based autoencoder with attention mechanism and genetic feature selection. *IEEE Access* (2025).
12. Rodriguez, I., Fernandez-Nestosa, M. J., Sanchez, D. & Cubilla, A. L. Pathological features, differential diagnosis, prognosis, and diagnostic challenges in the classification of penile squamous neoplasia. *Histopathology* **88**, 265–300 (2026).
13. Niu, S., Zhang, L., Wang, L., Zhang, X. & Liu, E. Hybrid feature fusion in cervical cancer cytology: a novel dual-module approach framework for lesion detection and classification using radiomics, deep learning, and reproducibility. *Front. Oncol.* **15**, 1595980 (2025).
14. Tewari, K. S. Cervical cancer. *New Engl. J. Medicine* **392**, 56–71 (2025).

15. Bedell, S. L., Goldstein, L. S., Goldstein, A. R. & Goldstein, A. T. Cervical cancer screening: past, present, and future. *Sex. medicine reviews* **8**, 28–37 (2020).

16. Caruso, G. *et al.* Cervical cancer: a new era. *Int. J. Gynecol. Cancer* **34**, 1946–1970 (2024).

17. Rezende, M. T., Bianchi, A. G. & Carneiro, C. M. Cervical cancer: Automation of pap test screening. *Diagn. cytopathology* **49**, 559–574 (2021).

18. Pacal, I. & Kılıcarslan, S. Deep learning-based approaches for robust classification of cervical cancer. *Neural Comput. Appl.* **35**, 18813–18828 (2023).

19. Nirmala, G., Nayudu, P. P., Kumar, A. R. & Sagar, R. Automatic cervical cancer classification using adaptive vision transformer encoder with cnn for medical application. *Pattern Recognit.* **160**, 111201 (2025).

20. Shanthi, P., Hareesha, K. *et al.* Automated detection and classification of cervical cancer using pap smear microscopic images: a comprehensive review and future perspectives. *Eng. Sci.* **19**, 20–41 (2022).

21. Hassija, V. *et al.* Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* **16**, 45–74 (2024).

22. Alom, M. R. *et al.* An explainable ai-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. *Sci. Reports* **15**, 17531 (2025).

23. Chau, M., Rahman, M. & Debnath, T. From black box to clarity: Strategies for effective ai informed consent in healthcare. *Artif. Intell. Medicine* 103169 (2025).

24. Plissiti, M. E. *et al.* Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE international conference on image processing (ICIP)*, 3144–3148 (IEEE, 2018).

25. Ennab, M. & Mcheick, H. Advancing ai interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-cam models. *Mach. Learn. Knowl. Extr.* **7**, 12 (2025).

26. Prusty, S. *et al.* Predicting cervical cancer risk probabilities using advanced h20 automl and local interpretable model-agnostic explanation techniques. *PeerJ Comput. Sci.* **10**, e1916 (2024).

27. Dixit, S., Kumar, A. & Srinivasan, K. A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions. *Diagnostics* **13**, 1353 (2023).

28. de Lima, R. C. & Quaresma, J. A. S. Exploring deep learning model opportunities for cervical cancer screening in vulnerable public health regions. *Computers* **14**, 202 (2025).

29. Hakim, R. U., Amin, T. & Ul Islam, S. B. Advances and challenges in cervical cancer: From molecular mechanisms and global epidemiology to innovative therapies and prevention strategies. *Cancer Control.* **32**, 10732748251336415 (2025).

30. Abdulazeem, Y. *et al.* Challenging the status quo: Why artificial intelligence models must go beyond accuracy in cervical cancer diagnosis. *Biomed. Signal Process. Control.* **105**, 107620 (2025).

31. Muhammad, D., Salman, M., Keles, A. & Bendechache, M. All diagnosis: can efficiency and transparency coexist? an explainble deep learning approach. *Sci. Reports* **15**, 12812 (2025).

32. Kalbhor, M. M. & Shinde, S. V. Cervical cancer diagnosis using convolution neural network: feature learning and transfer learning approaches. *Soft Comput.* 1–11 (2023).

33. Mehedi, M. H. K. *et al.* A lightweight deep learning method to identify different types of cervical cancer. *Sci. Reports* **14**, 29446 (2024).

34. Mathivanan, S. K. *et al.* Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. *Sci. Reports* **14**, 10812 (2024).

35. Kaur, H., Sharma, R. & Kaur, J. Comparison of deep transfer learning models for classification of cervical cancer from pap smear images. *Sci. Reports* **15**, 3945 (2025).

36. Wubineh, B. Z., Rusiecki, A. & Halawa, K. Classification of cervical cells from the pap smear image using the res_dcgan data augmentation and resnet50v2 with self-attention architecture. *Neural Comput. Appl.* **36**, 21801–21815 (2024).

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

38. Wubineh, B. Z., Rusiecki, A. & Halawa, K. Deep learning-based automatic segmentation and classification for cervical cancer detection using an improved u-net and ensemble methods. *Sci. Reports* (2026).

39. Wubineh, B. Z., Rusiecki, A. & Halawa, K. Se-deeplabv3+: Cervical cell segmentation and classification using a novel se-based deeplabv3+ and ensemble method. *IEEE Access* (2025).

40. Zammataro, L. Cinnamon-gui: Revolutionizing pap smear analysis with cnn-based digital pathology image classification. *F1000Research* **13**, 897 (2024).

41. Austin, R. & Parvathi, R. Cnn based method for classifying cervical cancer cells in pap smear images. *Sci. reports* **15**, 23936 (2025).

42. Pramanik, R., Banerjee, B. & Sarkar, R. Msenet: Mean and standard deviation based ensemble network for cervical cancer detection. *Eng. Appl. Artif. Intell.* **123**, 106336 (2023).