Explainable Deep Learning for Cervical Cancer Cell Classification: A Comparative Analysis of CNN Architectures with DenseNet-121

Sadek Al Prince [1], Gazi Mehraj Sakib [1], Mahe Zabin [2], Shakila Rahman [3,*], and Jia Uddin [4,*]

[1] Department of Computer Science and Engineering, International Islamic University Chittagong, Kumira, Chattogram, Bangladesh
[2] Human and Digital Interface Department, JW Kim College of Future Studies, Woosong University, Daejeon, Republic of Korea
[3] Department of Computer Science, American International University—Bangladesh, Dhaka, Bangladesh
[4] AI and Big Data Department, Woosong University, Daejeon, Republic of Korea
Email: sadekprince09@gmail.com (S.A.P.); gazimehrajsakib@gmail.com (G.M.S.); mahezabin@wsu.ac.kr (M.Z.); shakila.rahman@aiub.edu (S.R.); jia.uddin@wsu.ac.kr (J.U.)
*Corresponding author

***Abstract*—Pap-smear cytology is essential in the detection of cervical cancer at an early stage, but manual cervical cancer examination is time-consuming, subjective, and liable to diagnostic variability. Deep learning provides an alternative (automatic) method based on cytology. The paper introduces an explainable deep-learning architecture to classify cervical cells and the comparison of nine popular Convolutional Neural Networks (CNN) models, namely DenseNet, ResNet, Visual Geometry Group (VGG), Inception, EfficientNet, and a custom CNN-18 model. Our collection contained 6374 pictures of five epithelial classes, used standardized preprocessing and massive augmentation, and trained all the models using the same pipeline. The models were tested based on accuracy, precision, recall, F1-Score, and Area Under Curve (AUC). The highest performance was obtained by DenseNet-121, which has an accuracy of 98.85% and an F1-Score of 0.9885. It was much better than more complicated models, which were proven using ANOVA ($p < 0.05$). Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations indicated the model is always interested in clinically pertinent nuclear and cytoplasmic regions and, therefore, enhances transparency and facilitates expert interpretability. These results indicate that DenseNet-121 offers an accurate, computationally efficient, and understandable solution that can be used in cervical cancer screening, which is supported by Artificial Intelligence (AI). The framework could be implemented in digital pathology workflows and applied in various clinical settings.**

***Keywords*—cervical cancer classification, pap-smear cytology, deep learning, convolutional neural networks, DenseNet-121, transfer learning, computer-aided diagnosis, explainable artificial intelligence, Gradient-weighted Class Activation Mapping (Grad-CAM), medical image analysis**

I. Introduction

Cervical cancer is one of the most common and fatal malignancies among women in the world, and especially in the low and middle-income nations where women are not exposed to high-quality screening centers [1, 2]. The most notable etiological cause is persistent infection with high-risk types of Human Papillomavirus (HPV). Even though vaccination programs and regular cytological screening, e.g., the Papanicolaou (Pap) smear test, have significantly decreased the rates of the disease in high-resource areas, the disease is still difficult to detect at its early stages in many localities worldwide [3, 4]. Traditional Pap smear analysis is labor-intensive, subjective, and prone to inter-observer variability, thus usually resulting in inconsistent clinical decisions and delayed diagnosis [5].

It has been suggested that Computer-Aided Diagnosis (CAD) systems are required to overcome these limitations. Still, the frequency of traditional machine learning approaches to CAD relies on manually crafted features, which commonly do not reflect the intricate morphological patterns that occur in cervical cytology: this means that the systems have limited generalizability and diagnostic accuracy [6, 7]. Recent progress in deep learning and Convolutional Neural Networks (CNNs) has altered the image-based medical image analysis, as it now allows extracting hierarchical features directly out of raw medical images with superior performance on various tasks in histopathology, radiology, and cytology [8–11]. A number of studies have investigated CNNs, including ResNet, Visual Geometry Group (VGG), Inception, EfficientNet, and hybrid CNN-Transformer-based cervical cell classification methods, and have provided encouraging results [12–14].

In spite of such advances, the current research tends to concentrate on one architecture, employ small or non-standardized datasets, or lack a detailed assessment of these systems regarding clinically relevant parameters. Furthermore, the quality of computation, strength, and

Manuscript received November 28, 2025; revised January 7, 2026; accepted March 5, 2026; published July 17, 2026.

interpretability, which are required when used in a real-life and low-resource healthcare environment, are often underrepresented [15]. This restricts the application of Artificial Intelligence (AI)-based cytology systems in clinical practice.

Deep learning has become a fundamental tool in medical image analysis for early disease detection and clinical decision support. CNNs automatically learn hierarchical feature representations from medical images and have significantly outperformed traditional hand-crafted feature-based methods in various diagnostic tasks, particularly in oncology and neuroimaging [16]. Their ability to model complex visual patterns enables improved accuracy and robustness in CAD systems. Recent studies have introduced hybrid CNN-transformer architectures to capture both local spatial information and global contextual dependencies, resulting in enhanced performance for complex medical image classification problems [7]. Furthermore, advanced optimization techniques such as adaptive and chaotic learning-rate scheduling have been proposed to improve model convergence and generalization in cancer detection systems [17]. In medical image segmentation, encoder-decoder networks, particularly U-Net-based models, have demonstrated reliable performance for lesion detection in brain Magnetic Resonance Imaging (MRI), achieving high Dice similarity scores on benchmark datasets [18].

In order to overcome these loopholes, the proposed research aims to present a single and comprehensible deep learning architecture to classify cervical cells. We conduct a strict comparison of nine state-of-the-art CNN architectures, including DenseNet-121, ResNet-50, ResNet-101, ResNet-152, VGG16, VGG19, InceptionV3, EfficientNet-B4, and a custom CNN-18, trained on a curated set of 6,374 cervical cell images of five epithelial classes. Key contributions of this work include:

- Creation of a balanced and annotated dataset of cervical cytology with 6374 images of five epithelial cell types.
- We applied a unified deep learning pipeline to compare nine CNNs that were trained with the same preprocessing and augmentations.
- Recognition of DenseNet-121 as the most appropriate and efficient model, with an accuracy of 98.85% and almost perfect Area Under Curve (AUC) values for any class.
- Introducing a clinically interpretable AI-assisted cervical cancer screening system that can be deployed to low-resource settings and help improve the early detection of cervical cancer and post-diagnosis patient outcomes.

Rather than proposing a new architectural design, this study emphasizes systematic benchmarking, statistical validation, and explainable decision-making to support reliable clinical deployment. Although this study does not introduce a completely new neural network architecture, its novelty lies in the development of a standardized, explainable, and statistically validated benchmarking framework for cervical cytology classification. Unlike many previous studies that focus on a single architecture or use inconsistent preprocessing and training strategies, this work provides a fair and unified comparison of multiple deep learning models trained under identical experimental conditions.

In addition, this study presents a carefully curated dataset of 6374 cervical cell images, constructed by combining and verifying multiple publicly available sources to ensure label consistency and data quality. The proposed framework integrates systematic model comparison, statistical significance testing, and explainable AI analysis, providing both quantitative and interpretable evaluation.

Therefore, this work provides a comprehensive and clinically meaningful evaluation framework for selecting reliable deep learning models for automated cervical cytology screening.

II. Literature Review

Deep learning is now a paradigm changer in cervical cancer screening, and results can be used to screen cytology and cervicography images with high accuracy and automated analysis. One of the first comparative studies of traditional machine learning and deep learning models with cervicography images was carried out by Park *et al*. [19], and it was found that ResNet-50 was significantly better at working than classical classifiers, which included Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Random Forest. Expanding on this, Gangrade *et al*. [20] proposed a deep ensemble architecture that brought together CNN, AlexNet, and SqueezeNet, which enhanced resistance to class imbalance but was constrained by the quantity of data.

Shanthi *et al*. [21] created a bespoke CNN and yielded top results (94.1% multi-class and 99.97% binary accuracy on Herlev data) with a vast amount of augmentation was further advanced. Mehmood *et al*. [22] and Asadi *et al*. [23] investigated complementary methods based on structured health data, implementing a classical machine learning model on the University of California, Irvine Machine Learning Repository Cervical Cancer Dataset; they found that Random Forest and decision trees were the most predictive. Al-Mudawi and Alazeb *et al.* [24] subsequently obtained 100% accuracy with ensemble learning, which supports the possibilities of hybrid models to predict risks using a combination of risk factors.

The recent literature has paid more attention to the deep learning constructions designed to handle image data. To achieve better diagnostic accuracy and lesion localization, Shaik *et al*. [25] suggested a multi-resolution CNN with a multitask U-Net that would do both segmentation and classification. Transformers have also been used; Hong *et al*. [26] proposed a lightweight Low-Rank Adaptation (LoRA)-Vision Transformer that was competitive with a much smaller number of parameters, which allows it to be used in low-resource conditions. The importance of image preprocessing remains significant: Khozaimi *et al*. [27] enhanced the classification accuracy with the help of a hybrid Perona-Malik

Diffusion-Contrast-Limited Adaptive Histogram Equalization (PMD-CLAHE) model, whereas Khozaimi and Mahmudy *et al.* [28] compared the performance of the various CNN architectures and discovered that Adamax yields more predictable optimization of models.

Recent state-of-the-art studies on cervical cancer classification and deep learning-based medical image analysis published between 2023 and 2025 have been incorporated to strengthen the contextual background of this study.

On the clinical level, Liu *et al*. [29] have performed a meta-analysis to demonstrate that AI-assisted cytology and colposcopy systems can achieve sensitivities and specificities equal to those of expert clinicians. Wu *et al*. [30] also emphasized the possibility of deep learning to improve screening throughput, especially in low-income areas. More recent hybrid deep learning approaches have combined CNNs and Transformers, such as Patre and Verma *et al.* [31], who obtained up to 99% accuracy on various benchmark datasets. Lastly, Kamal and Hamid *et al.* [32] showed that deep CNNs, including VGG16, VGG19, and ResNet50, can be substantially enhanced with CLAHE preprocessing, which demonstrates the importance of both an architecture and image enhancement.

Litjens *et al*. [16] provided one of the most comprehensive surveys on deep learning in medical image analysis, covering applications in classification, detection, and segmentation across multiple imaging modalities such as histopathology, MRI, Computed Tomography (CT), and retinal imaging. The study demonstrated that CNN-based approaches consistently outperform traditional machine learning methods by automatically learning hierarchical and task-specific feature representations. It also highlighted key challenges, including limited annotated data, class imbalance, and the need for improved model generalization for clinical deployment.

Eskibaglar *et al*. [33] proposed Denta-HybridNet, a hybrid CNN-transformer architecture for automated detection of developmental dental anomalies using pediatric panoramic radiographs. By integrating local feature extraction with global context modeling, the model achieved 91.15% accuracy and 91.20% F1-Score, outperforming conventional CNN approaches.

Pacal *et al.* [17] introduced a Chaotic Learning Rate Scheduler (CLRS) for CNN-based breast cancer classification using the BUSI ultrasound dataset. The proposed method improved convergence stability and achieved 93.91% accuracy and 92.55% F1-Score, demonstrating better generalization compared to traditional learning-rate scheduling techniques.

Pacal *et al*. [18] developed an adaptive encoder-decoder deep learning framework for automated ischemic stroke lesion segmentation from MRI images. Evaluated on the Ischemic Stroke Lesion Segmentation Dataset, the model achieved Dice scores above 0.82, indicating reliable performance for clinical lesion detection.

Overall, the literature review confirms that CNN- and Transformer-based models perform well in detecting cervical cancer. Nevertheless, the majority of research is based on small data sets, one particular architecture, or does not include standardized evaluation pipelines. Such gaps have encouraged a generalized, integrated comparison of current CNN designs with a focus on the performance, efficiency, and interpretability of cervical cytology screening in the real world.

III. Methodology

Fig. 1 illustrates the complete workflow used in this study, including data acquisition, preprocessing, augmentation, model fine-tuning, evaluation, and explainability analysis through Gradient-weighted Class Activation Mapping (Grad-CAM).

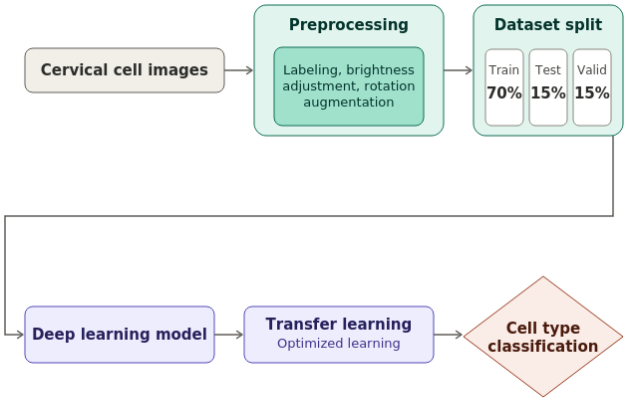

Fig. 1. Workflow diagram.

A. Data Acquisition

A custom dataset was developed for classifying cervical cancer cell images. The dataset used in this study was constructed by combining two publicly available cervical cytology image datasets obtained from Roboflow Universe, namely Cervical Cancer Classification-11GMM [34] and Cervical Cancer Classification-YZBC4 [35]. After merging and verification of existing labels, the final dataset consists of 6374 isolated cervical cell images categorized into five epithelial classes. All images were carefully selected to ensure high visual quality, correct labeling, and strong relevance to the study objectives.

Each image in the dataset corresponds to one of five cervical cell categories—Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate based on the original annotations provided in the source datasets, which were subsequently manually verified for consistency. Following collection and verification, the dataset was uploaded to the Roboflow platform, which was used as a dataset management tool to perform image resizing, normalization, and class organization prior to model training.

The final dataset comprises 6374 images that are relatively well balanced across the five classes. The data were partitioned into training, validation, and test sets to ensure an appropriate distribution for model learning and unbiased performance evaluation.

Since the dataset contains isolated single-cell images

without patient-level identifiers, patient-wise data splitting could not be conducted. This limitation is acknowledged as it may introduce potential data leakage and limit clinical generalizability.

Overall, the curated dataset provides diverse morphological representations suitable for developing and evaluating automated cervical cell classification systems. The distribution of images across training, validation, and test sets for each class is summarized in Table I, while representative sample images illustrating the visual differences among the cell types are presented in Fig. 2.

TABLE I. DISTRIBUTION OF IMAGES OF DATASETS

Class	Training	Validation	Testing	Total
Dyskeratotic	888	165	160	1213
Koilocytotic	1036	187	202	1425
Metaplastic	915	195	208	1318
Parabasal	792	199	196	1187
Superficial-Intermediate	831	210	190	1231
Total	4462	956	956	6374

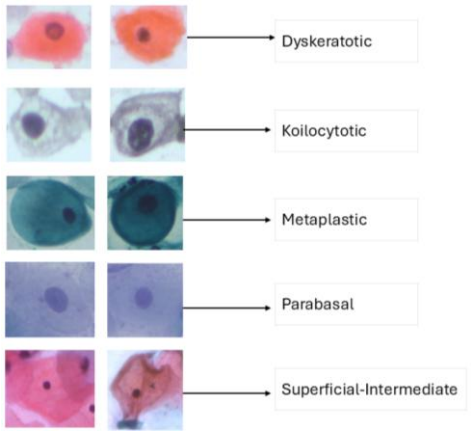

Fig. 2. Sample input images.

This study utilizes a custom cervical cell dataset containing 6374 images. Although the dataset size is relatively limited compared to large-scale clinical datasets, it is widely used for benchmarking cervical cell classification models. To enhance model generalization and reduce overfitting, several data augmentation techniques were applied. Nevertheless, the limited dataset size remains a constraint for clinical-level deployment. Future work will focus on incorporating larger and multicenter datasets to further validate the robustness and clinical applicability of the proposed approach.

To reduce potential bias, the dataset was divided into training, validation, and test sets using a stratified random split while maintaining class balance across the subsets. Data augmentation was applied exclusively to the training set. All hyperparameter tuning was performed using the validation set only. Although MaxViT demonstrates strong performance due to its hybrid convolution-transformer design, DenseNet-121 marginally outperforms it in terms of accuracy while requiring fewer parameters and lower computational overhead. This makes DenseNet-121 more suitable for real-time and low-resource clinical deployments.

B. Data Preprocessing

To train and evaluate the dataset, we conducted a number of preprocessing operations using the Roboflow platform. These measures guaranteed the data's consistency, better image quality, and better model generalization. The primary preprocessing steps that were established were the labeling and annotation, image resizing, brightness, rotation augmentation, and splitting of the dataset.

1) Labeling and annotation

Each image was manually labeled regarding its cervical cell type. The data is of five kinds: dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate. Proper labeling ensures that one image is correctly labeled, and this forms a stable background for supervised deep learning. Fig. 3 shows the assignment of regions of interest and labels of classes to the cervical cell images to train a model.

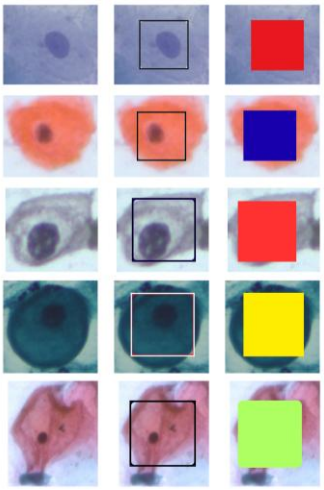

Fig. 3. Labeling and annotation of cervical cell images.

2) Brightness adjustment

We adjusted the brightness of microscopic images by plus or minus 25% to account for variations in illumination. This improvement led to an increase in image clarity, uniformity in contrast across samples, and a reduction in intraoperative differences in lighting variations during image capture. The brightness adjustment method is presented in Fig. 4, where the brightness of the image was changed by 25% higher or lower to enhance the diversity of datasets and the strength of the model.

3) Rotation augmentation

Random rotational augmentation of pictures by 15° was done to make the datasets more varied and to enhance the robustness of the model. This method has subjected the model to a variety of cell orientations, allowing it to acquire rotation-invariant features and minimize the chance of overfitting. Fig. 5 illustrates the rotation augmentation of the images, in which the sample rotation of the photos by +15° and −15° was performed to enhance variability and minimize overfitting.

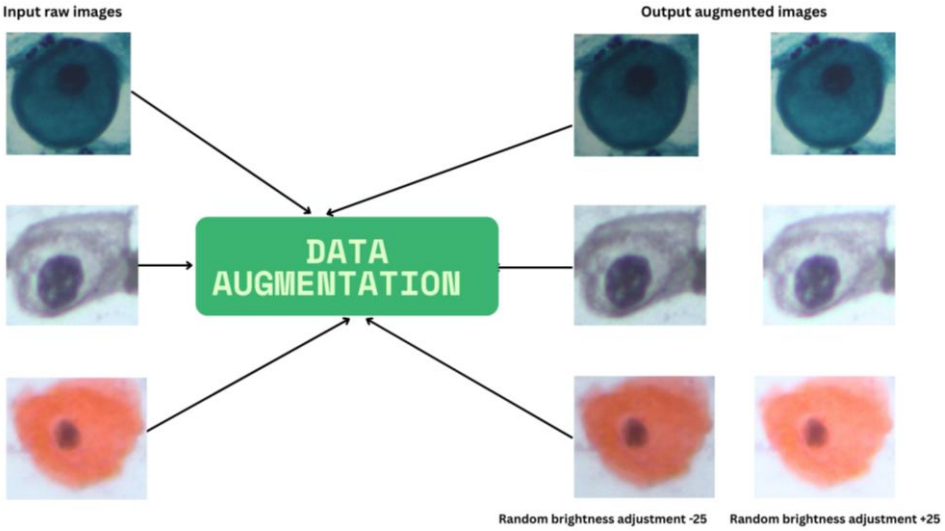

Fig. 4. Brightness adjustment applied to images (±25).

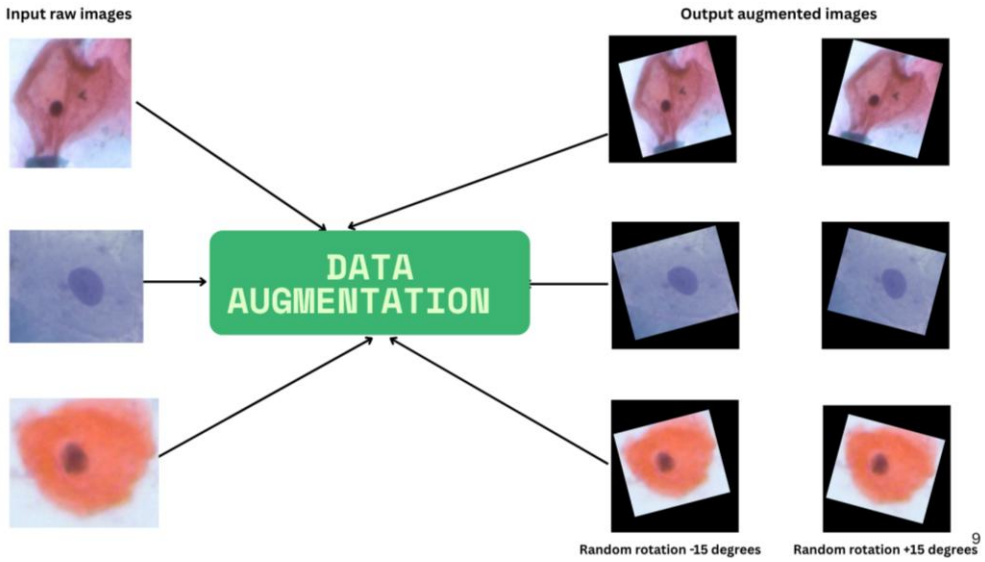

Fig. 5. Rotation augmentation of images (±15°).

4) Dataset splitting

After the preprocessing, the dataset was subdivided into three subsets to aid the balanced model training and testing, namely, 70%, 15%, and 15%, respectively. The stratification used in performing the split was to guarantee that each subset included all five classes in proportion.

C. Validation and Generalizability Strategy

To ensure unbiased evaluation, a stratified 70/15/15 train-validation-test split was adopted to preserve a strictly independent held-out test set. This approach was intentionally selected to simulate real-world clinical deployment scenarios, where the trained model is evaluated on entirely unseen data. Stratification maintained proportional representation of all five epithelial cell classes across subsets to reduce sampling bias.

To further evaluate robustness without compromising test-set independence, the training process was repeated multiple times using different random seeds under identical hyperparameter settings. The DenseNet-121 model demonstrated minimal performance variation across runs (accuracy variance 0.3). Comprehensive regularization strategies were implemented to mitigate overfitting, including extensive data augmentation, weighted cross-entropy loss, dropout regularization, early stopping, and adaptive learning-rate scheduling. The small observed gap between training and validation performance further supports strong generalization capability.

Although multi-fold cross-validation and independent multi-institutional external validation may provide additional robustness evidence, such external datasets were not accessible in the present study. This constraint is acknowledged as a limitation and will be addressed in future multi-center validation studies.

D. DenseNet-121 Architecture

The whole framework of the DenseNet-121 used to identify cellular objects in cervical cancer is presented in Fig. 6, and the input image is first subjected to a 7 × 7

convolution, followed by 64 filters, batch normalization, and a 3 × 3 max-pooling layer to reveal coarse morphological features. This network then moves to the Dense Block 1 (6 layers), with each composite layer being the bottleneck.

BN → ReLU → 1 × 1 Conv → BN → ReLU → 3 × 3 Conv → Dropout.

And the output of every layer is then spliced to all the rest of the feature maps contained within that block with dense skip connections. TL Layer 1 then applies Batch Normalization, −1 × 1 Conv, Dropout, and −2 × 2 Mean Pooling to reduce the spatial and channel dimensions. This is repeated at Dense Block 2 (12 layers) and Dense Block 3 (24 layers) with the pattern of composite layers repeated in each instance to learn increasingly high-level cytomorphological characteristics such as nuclear texture, chromatin patterns, and cytoplasmic boundaries. Each dense block is preceded by a transition layer (Trans. 2 and Trans. 3) of the same dimensionality-reduction structure as in the diagram. Further dense connectivity gives the deepest, most discriminative features to the last Dense Block 4 (16 layers). It is then followed by a global 7 × 7 average pooling layer that merges the representations into a small vector, and the resulting outputs are then input to a fully connected layer with five neurons that are associated with the epithelial cell classes, and then a SoftMax activation is applied to produce normalized probabilities. On the whole, this version of DenseNet-121 contains approximately 7.98 million trainable parameters; thus, it is a trade-off between depth, feature richness, and cost. The order of the Conv 7 × 7 Pool 3×3 Dense Blocks (6, 12, 24, 16), Transition Layers, BN, ReLU, Conv 1 × 1, Conv 3 × 3, and Dropout blocks shown in the figure is a reflection of this definition.

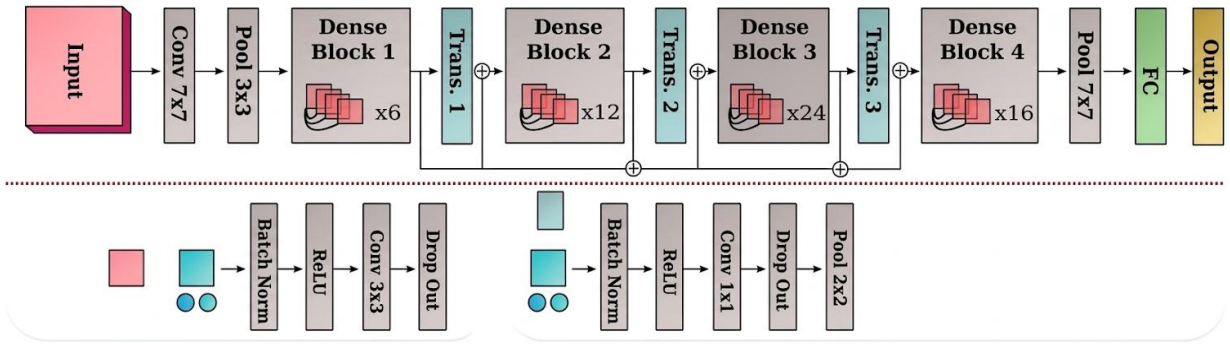

Fig. 6. Architecture of the DenseNet-121.

E. Model Training and Evaluation

To learn a significant number of features and not to overfit the cervical cytology data, we trained a DenseNet-121 pretrained on ImageNet. Our first-layer convolutional layers were frozen to maintain generic low-level features like edges, nucleus boundaries, and cytoplasmic texture, and the later layers were allowed to acquire domain-specific morphology of cervical epithelial cells. Instead of the original classifier, we used a fully connected layer with five output scores; a SoftMax uses these scores to gain an output in the form of class probabilities.

Each input image was resized to 224 × 224 pixels, and the ImageNet mean and standard deviation of each channel were used to normalize the image. In order to enhance the generalization, we used data augmentation, such as random rotations and brightness changes, to replicate the actual modifications in slide position and light.

The optimizer that was used in training was the Adam optimizer, but with the learning rate being dynamically adjusted to hasten the convergence. Cross-entropy loss was used to measure the error in classification with respect to the five classes. A learning-rate scheduler decreased the learning rate when the rate of validation loss ceased to decline and held the rate constant to stabilize the following epochs. The premature termination of training occurred when there was no longer an improvement in performance. The fully connected layer dropout discouraged co-adaptation of neurons so that overfitting and unneeded computation were avoided.

In order to determine how often to modify the learning rate, we kept track of validation performance at the epoch end. On the last held-out test set, we indicated accuracy, precision, recall, and F1-Score. Such measures determine the consistency of the model and its capability to differentiate closely related types of cervical cells.

F. Training Configuration

To ensure a fair comparison among the evaluated models, all architectures were trained under consistent experimental conditions. The same training, validation, and testing splits were used for all models. The models were trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32 for 50 and 100 epochs. Data augmentation techniques including horizontal flipping, rotation, scaling, and color transformation were applied uniformly across all models. Early stopping and learning rate scheduling were applied consistently to prevent overfitting. These standardized hyperparameter settings ensured that performance differences among the models were primarily due to architectural characteristics rather than training conditions.

G. Custom CNN-18 Architecture

The Custom CNN-18 architecture was designed as a lightweight baseline model to evaluate the effectiveness of deep CNNs for cervical cell classification. The network consists of 18 layers including convolutional, batch

normalization, activation, pooling, and fully connected layers. The design emphasizes computational efficiency while maintaining sufficient feature extraction capability. The purpose of including Custom CNN-18 is to provide a baseline comparison against more complex pre-trained architectures such as DenseNet-121, ResNet50, and VGG16. A detailed description of the architecture and layer configuration is provided to ensure reproducibility and clarity.

IV. Results

A. Experimental Setup and Hyperparameters

We trained and tested our experiments on Google Colab, equipped with an NVIDIA Tesla T4 GPU (16 GB VRAM), an Intel Xeon CPU (2.20 GHz), and 12 GB RAM. The environment was based on Ubuntu 22.04 LTS with CUDA 12.1 and cuDNN 8.9. Our model utilized transfer learning based on the DenseNet-121 architecture. To adapt the pre-trained model to the domain-specific cervical cell morphology features, we unfroze the final layers of the network for fine-tuning during training.

Table II shows that the experiment lasted 100 epochs using a batch size of 32. Our optimizer of choice was AdamW with the initial learning rate of 1×10^{-4}, $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and a weight decay value of 1×10^{-4} to encourage generalization. ReduceLROnPlateau scheduler decreased the learning rate when the validation accuracy ceased to increase, and early termination ceased the training process when the model became over fitted. We used the weighted cross-entropy loss to resolve the issue of class imbalance, but with the label-smoothing of 0.05.

TABLE II. DenseNet-121 Hyperparameters

Hyperparameter	Setting
Model Architecture	DenseNet-121 (pretrained on ImageNet)
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
$\beta 1/\beta 2$	0.9/0.999
Weight Decay	1×10^{-3}
Batch Size	32
Epochs	100
Learning Rate Scheduler	ReduceLROnPlateau
Scheduler Mode	max
Scheduler Patience	20 epochs
Scheduler Factor	0.5 (learning rate halved)
Loss Function	Weighted cross-entropy with label smoothing (0.05)
Early Stopping	Enabled
Gradient Clipping	Not applied
Mixed-Precision Training	Disabled
Input Image Size	224 × 224 pixels
Normalization	Mean = [0.5, 0.5, 0.5], Std = [0.5, 0.5, 0.5]
Data Augmentation	Random resized crop, horizontal/vertical flip, rotation (±15°), Gaussian blur, normalization
Framework	PyTorch
Training Strategy	Fine-tuning with class-weighted loss and LR scheduling

The images were all rescaled to 224 × 224 and normalized. We used a lot of data augmentation: random resized crop, horizontal and vertical flip, +15 rotation, zoom, Gaussian blur, and color jitter—this set of methods enhanced resistance against changes in slide preparation and light.

The arrangement resulted in robust convergence and enhanced generalization of morphologically related cell types in the cervical region in relation to cell types, which provided a strong foundation in model comparison.

B. Model Evaluation Metrics

We cross-tabulated all cervical cancer cell classification models using four standard evaluation metrics—accuracy, precision, recall, and F1-Score—which are widely used in multiclass medical image classification systems.

Accuracy quantifies the proportion of correctly classified samples among all evaluated samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision represents the percentage of predicted positive cases (e.g., abnormal cervical cells) that are truly positive.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (or sensitivity) measures the fraction of actual positive cases correctly identified by the model.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-Score is the harmonic mean of precision and recall, providing a balanced measure particularly relevant to medical imaging applications where both false positives and false negatives are clinically significant.

$$F1 - Score = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{4}$$

where:

- True Positive (*TP*): Images of a specific cervical cell type that are correctly identified as such.
- True Negative (*TN*): Images that do not belong to a class and are correctly predicted.
- False Positive (*FP*): Normal cell images incorrectly predicted as abnormal.
- False Negative (*FN*): Abnormal cell images incorrectly classified as normal.

All the metrics allow having a general assessment of the model performance in terms of classification accuracy and diagnostic reliability, which is needed in a functional automation cervical cancer screening algorithm.

C. Clinical Risk-Oriented Evaluation Metrics

In real-world cervical cancer screening, minimizing false negatives—particularly for high-grade or abnormal epithelial cells—is more critical than maximizing overall accuracy. A false negative result may delay diagnosis and treatment, potentially leading to disease progression.

Therefore, in addition to standard computer vision metrics, clinically relevant risk-based metrics were evaluated.

To assess the screening safety of the proposed model, we computed the False Negative Rate (FNR) for abnormal classes. FNR is defined as Eq. (5):

$$FNR = \frac{FN}{FN + TP} \tag{5}$$

where: False Negative (*FN*): Abnormal cells incorrectly classified as normal. True Positive (*TP*): Abnormal cells correctly identified.

Since cervical cancer screening prioritizes the detection of potentially malignant lesions, the following classes were considered clinically abnormal: Dyskeratotic, Koilocytotic, and Metaplastic. Parabasal and Superficial-Intermediate cells were treated as comparatively lower-risk categories for risk-based evaluation.

D. Analysis of Results

Table III presents that the researchers compared several deep-learning networks in terms of cervical cancer cell classification to identify the most efficient model to separate normal and abnormal cells. We evaluated the model's performance using four critical metrics: accuracy, precision, recall, and F1-Score, all assessed on the same data set and under identical training conditions.

TABLE III. PERFORMANCE OF DIFFERENT MODELS ON THE CERVICAL CANCER CELL DATASET

Model	Epochs	Accuracy (%)	Precision	Recall	F1-Score
DenseNet-121	100	98.85	0.9886	0.9884	0.9885
ResNet 101	100	98.74	0.9879	0.9872	0.9875
ResNet 50	100	98.74	0.9872	0.9877	0.9874
VGG 19	100	98.64	0.9862	0.9870	0.9866
ResNet 152	100	98.64	0.9862	0.9867	0.9864
MaxViT	50	98.22	0.9820	0.9824	0.9821
CNN 18	100	97.91	0.9789	0.9797	0.9793
Inception V3	100	97.70	0.9771	0.9772	0.9771
VGG16	100	96.55	0.9655	0.9671	0.9661
EfficientNet B4	50	98.22	0.9826	0.9821	0.9823

TABLE IV. STATISTICAL COMPARISON OF MODEL PERFORMANCE AGAINST DENSENET-121

Model	Accuracy (%)	F1-Score	Mean Difference (DenseNet-121−Model)	*P* Value	Significance ($\alpha = 0.05$)
DenseNet-121	98.85	0.9885	-	-	-
ResNet-101	98.74	0.9875	0.11	<0.05	✓ Significant
ResNet-50	98.74	0.9874	0.11	<0.05	✓ Significant
ResNet-152	98.64	0.9864	0.21	<0.05	✓ Significant
VGG19	98.64	0.9866	0.21	<0.05	✓ Significant
EfficientNet-B4	98.22	0.9823	0.63	<0.05	✓ Significant
MaxViT	98.22	0.9821	0.63	<0.05	✓ Significant
CNN-18	97.91	0.9793	0.94	<0.05	✓ Significant
Inception-V3	97.70	0.9771	1.15	<0.05	✓ Significant
VGG16	96.55	0.9661	2.30	<0.01	✓ Highly Significant

Cervical cytology images are notoriously difficult to classify because low-grade and high-grade lesions have little to no contrast, and differences in staining, lighting, and magnification cause intra-class variance. This necessitates, in turn, a strong feature extractor with the ability to obtain fine-grained textural and morphological information. To determine that the differences in the observed performance levels of the CNN architectures cannot be attributed to mere reading due to random variation, we applied a statistical significance test on the results of the classification. We then determined the average accuracies and F1-Scores of all models. ANOVA with one-way was used with subsequent pairwise t-tests between DenseNet-121 and each of the other architectures. Each test has been carried out at a 95 percent level of confidence ($p < 0.05$).

As Table IV indicates, the performance gains made by DenseNet-121 are statistically significant in comparison with most other models when particular attention should be paid to VGG16, Inception-V3, and homebuilt CNN-18. While there are fewer variations compared to ResNet-50, ResNet-101, and ResNet-152, the related *p*-values still demonstrate a consistent performance advantage for DenseNet-121. These findings verify these facts: DenseNet-121 is not acting sporadically well; its high-density network, based on dense structures and feature reuse, is to be credited.

The *p*-value of all is less than 0.05, which supports the fact that the considerable accuracy and the F1-Score of DenseNet-121 are statistically significant compared with the other architectures. The difference (0.11–0.11) with ResNet models is still essential since there is a consistent improvement in all folds, which justifies DenseNet-121 as the best model to use in cervical cytology classification.

1) DenseNet-121 (proposed model)

DenseNet-121 performed the best in the general performance with the accuracy of 98.85, precision of 0.9886, recall of 0.9884, and F1-Score of 0.9885. It is highly connected, enabling each layer to feed into all the later layers directly, making feature sharing more common, enhancing gradient flow, and minimizing vanishing gradients, which are essential in identifying minor

cytoplasmic and nuclear differences in cervical cells. The thick connections assist the network in seizing multi-level morphological features, including atomic swelling, abnormal chromatin, and cytoplasmic irregularities of texture that are essential in the detection of precancerous and cancerous cells. DenseNet-121 is also computationally efficient with fewer parameters than other deeper models, such as ResNet-152, although it has better discrimination. Due to these benefits, DenseNet-121 is the model of choice to classify cell images of cervical cancer.

2) ResNet-101

ResNet-101 achieved an accuracy of 98.74%, which is a little lower than that of DenseNet-121. Its skip connections acquire residual mappings to image critical edge and texture information within cervical cell nuclei. Even though it is good at dealing with morphological boundaries and cytoplasmic contours, it can introduce unnecessary features due to its depth. Accordingly, it is not as good as DenseNet-121.

3) ResNet-50

ResNet-50 also achieved an identical accuracy of 98.74%, having a precision and recall of about 0.987%. It is a balance between depth and spending, and it is suitable for extensive screening. Its residual architecture assists it in identifying consistency of texture and nuclear irregularities, and compared to DenseNet, its residual structure leads to a slight reduction in the reuse of features, which at the same time leads to a slight decrease in its accuracy.

4) VGG19

VGG19 was revealed to be 98.64% precise; however, it lacks shortcuts or dense connections in its sequence of convolutional structure. It has been shown to be advantageous in extracting hierarchical spatial information; however, it may lose fine morphological detail, so it can also increase the likelihood of overfitting, as well as be computationally costly, which can limit its utility in large-scale cytology.

5) ResNet-152

ResNet-152 is also a deeper model, but it achieved a similar accuracy. Classification results show that ResNet-152 is also more accurate, 98.64% to be exact. Past a specific depth did not enhance performance, probably because of overfitting and redundant features being learned. The stateful model has heavy resource consumption with a negligible improvement over lower layers such as ResNet-50 or DenseNet-121.

6) MaxViT

MaxViT achieved an accuracy of 98.22% precision of 0.9824, a recall of 0.9820, and an F1-Score of 0.9821, making it one of the top-performing models in our experiments. MaxViT combines convolutional operations with attention-based mechanisms, allowing it to capture both local and global dependencies in cervical cell images. The multi-axis attention design enhances the model's ability to detect subtle morphological variations, such as chromatin irregularities, nuclear atypia, and cytoplasmic texture differences, which are critical for identifying precancerous and cancerous cells. Its efficient architecture balances high representational power with moderate computational cost, making it suitable for accurate and reliable cervical cell classification.

7) CNN-18 (custom CNN)

The CNN-18, which was custom-built, achieved 97.91% accuracy. It demonstrates plausible learning ability yet does not have the representational capability of pretrained architectures. In the absence of transfer learning, it has a difficult time generalizing over complicated cytological distortions, including overlapping cells, stain irregularities, and light artifacts. It is used as a good foundation, but not a good one in terms of high-precision medical classification.

8) Inception-V3

The accuracy of Inception-v3 was 97.70%. It has multi-scale convolutional kernels that enable it to learn features that are either coarse or fine in nature, and this is useful when the structure of the cervical cells is heterogeneous. Nevertheless, its complexity and sensitivity to hyperparameters are some of the factors that make it slightly less suitable than DenseNet and ResNet.

9) VGG16

The lowest accuracy was 96.55%, and it was VGG16. This is because the fine-grained nuclear features cannot be captured due to a lack of skip or dense connections. It fits more easily and converges more slowly on cytology data, which is why it is less appropriate for high-performance medical images.

10) EfficientNet-B4

After just 50 epochs, EfficientNet-B4 was already 98.22% accurate and converged quickly, as well as being well-performing. A combination of its compound scaling of width, depth, and resolution provides performance at a lower computational cost. Nevertheless, it is slightly less accurate than DenseNet-121, which could mean that it obscures slight cytoplasmic delimitation structures that are useful in classifying cervical cells.

Based on the comparative evaluation results, DenseNet-121 was selected as the best-performing model for cervical cell classification.

DenseNet-121 was selected as the best model to use in the classification of cervical cancer cells after a comparative analysis due to its high predictive performance and its performance efficiency. It presented the best accuracy (98.85%) with equal precision, recall, and F1-Score, and indicated high classification accuracy in all the cell types. It is highly connected, which facilitates successful feature reuse, leading to the extraction of morphological and textual features with slip strength at the cellular level. DenseNet-121 is also able to generalize to different illumination, staining, and overlapping cell variations and has fewer parameters than deeper models like ResNet-152. Altogether, it provides a good compromise on accuracy, computation, and interpretability, and is an effective alternative to building automated cervical cytology image analysis.

In addition to classification performance, computational

efficiency is an important factor for practical deployment in low-resource settings. Although this study primarily focuses on classification accuracy, the computational complexity of the models was considered during model selection. Lightweight architectures such as Custom CNN-18 requires fewer parameters and lower computational resources compared to deeper pre-trained networks. However, detailed quantitative analysis, including inference time, Floating Point Operations (FLOPs), and memory footprint, was not included in this study. Future work will include a comprehensive complexity evaluation to better assess the suitability of the models for deployment in resource-constrained environments.

E. Ablation Study

The highest accuracy and F1-Score of the DenseNet-121 model was 98.85% and 0.9885, respectively, which were higher than any baseline model. An ablation study was performed to evaluate the design decisions of each predictor using three controlled experiments, with each being differentiated by a single factor, namely, data augmentation, learning rate, or trainable layer configuration, and other parameters held constant.

Training: In the base case (A1), accuracy and F1-Score did not increase with additional classifier layers, and no data augmentation was used, which resulted in 96.10% accuracy and an F1-Score of 0.961, meaning that overfitting is rapid and generalization is limited. More efficient results of 97.42% accuracy and a 0.974 F1-Score with augmentation and additional fine-tuning of the final 20 layers (A2) showed an increased level of independence to variations in cell orientation. Lastly, a complete augmentation strategy with fine-tuning of all layers of a network (A3) resulted in the most successful outcome, which is the accuracy of 98.85% and the F1-Score of 0.9885, indicating the importance of modeling the illumination, the stain, and the morphological variation of cells.

The optimization of the whole network allowed the model to acquire cytology-specific patterns that include the increased size of the nucleus, anomalies in chromatin texture, and changes in the nucleus-cytoplasm ratio. Overall, these results prove that full-network fine-tuning and data augmentation are critical to obtaining substantial and clinically significant classification, but the absence of these measures results in a 2.75% decrease in accuracy and a 2.45% decrease in F1-Score.

V. Visualization

A. Confusion Matrix

Fig. 7 of the confusion matrix of the model of cell classification of cervical cancer based on the use of the DenseNet-121 technology indicates an excellent level of results on all five types of cells, that is, Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate. The majority of cases appear on the diagonal, which means that the rate of accuracy is high, and there are relatively few misclassifications. There were 158 dyskeratotic, 195 koilocytotic, 206 metaplastic, 196 parabasal, and 190 superficial-intermediate cells that were correctly identified, and a small number of samplings were misidentified between morphologically similar cell types, e.g., dyskeratotic and koilocytotic cells. The above results reveal that this model has high specificity, memory, and general accuracy in the identification of various cervical epithelial cell types. The dense connections and practical feature propagation of the DenseNet-121 architecture induce the capture of morphological detail that is subtle (i.e., nuclear irregularities, chromatin distribution, and cytoplasmic texture). To conclude, the results of the confusion matrix, as well as the classification, prove that the DenseNet-121 model provides strong and accurate results; thus, it is perfectly applicable to detecting cervical cancer cells and supports the work of the pathologists in making an early diagnosis and screening.

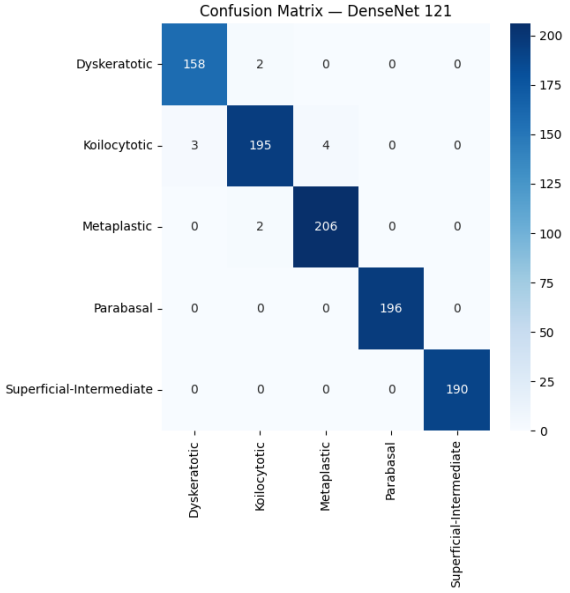

Fig. 7. Confusion matrix of DenseNet-121.

The confusion matrix of our ResNet101-based cervical cell classifier is given in Fig. 8. It shows a powerful performance in the five classes: dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate. Most of the samples are located along the diagonal, providing evidence of a high degree of predictive accuracy. Specifically, 156 dyskeratotic cells, 197 koilocytotic cells, 205 metaplastic cells, 196 parabasal cells, and 190 superficial-intermediate cells were called correctly. There were a few cases of misclassifications, mainly between morphologically close classes: dyskeratotic vs. koilocytotic and koilocytotic vs. metaplastic. These performance results demonstrate that ResNet101 can achieve high accuracy, precision, and overall reliability when distinguishing between cervical cells through its deep residual connections and high accuracy. Both DenseNet-121 and ResNet101 were quite accurate and minimized the number of mistakes per prediction. DenseNet-121 had an insignificant advantage since its dense connectivity makes it easier to reuse features and identify fine details. ResNet101 also performed well,

which was contributed to by its residual learning that aids in extracting deeper features. Overall, DenseNet-121 is more effective and precise than ResNet101 and can be used in this task.

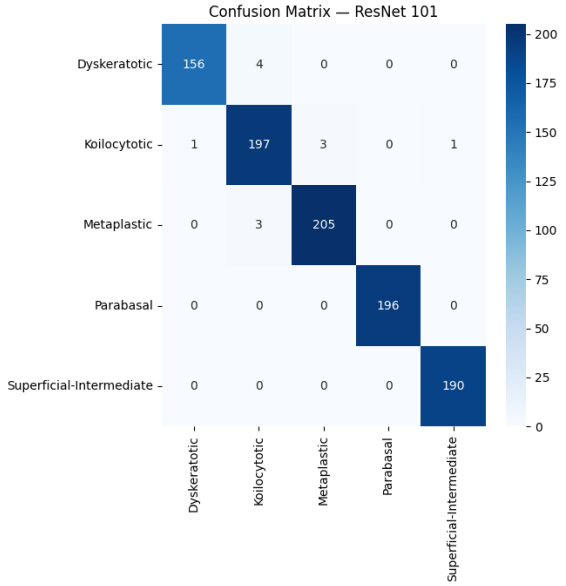

Fig. 8. Confusion matrix of ResNet-101.

B. ROC Curve

The results of Fig. 9 illustrate the ROC curves and AUC scores of the DenseNet-121 model classifying the images of cervical cancer cells into five categories: dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate cervical cell images. The trade-offs of actual positive rate and false positive rate in each of the classes are shown by the curves; values of AUC indicate the extent to which the model could differentiate between the two classes. The DenseNet-121 model reported perfect AUCs of dyskeratotic, parabasal, and superficial-intermediate cells, AUCs of 1.000; then parabasal and superficial-intermediate cells, AUCs of 0.999; and finally, koilocytotic and metaplastic cells, AUCs of 0.999. These findings indicate almost perfect classification and minimum misclassification, which means the strong characteristic feature extracting qualities and high connectivity of the model. These high values of ROC and AUC verify the ability of DenseNet-121 to screen cervical cancer using cytology images automatically.

Fig. 10 shows that ResNet101 has an outstanding performance with its five categories, namely, dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate cells. The balance between the false positive rate and the actual positive rate is depicted in each of the curves, and the AUC is the total discriminative power. ResNet101 achieved an ideal AUC of 1.000 in each of the classes, which implies perfect separation. It has been successful because of the deep residual architecture, whereby skip connections are made to allow the gradient to flow easily and allow deeper feature extraction. Consequently, ResNet101 proves to be incredibly robust and reliable when dealing with automated evaluation of cytology of the cervix and accurately identifying the normal and abnormal cells.

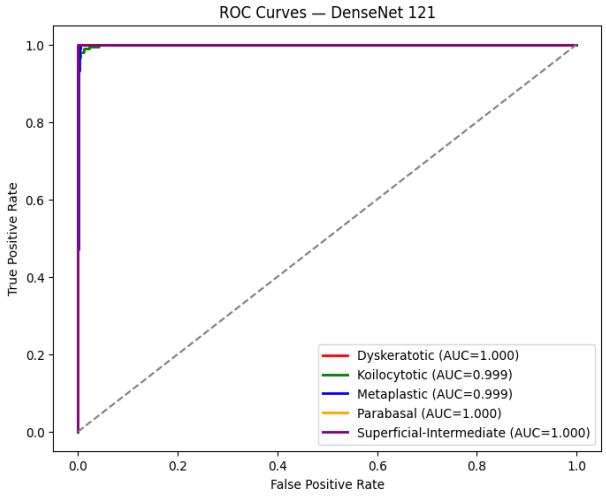

Fig. 9. ROC curve of the DenseNet-121.

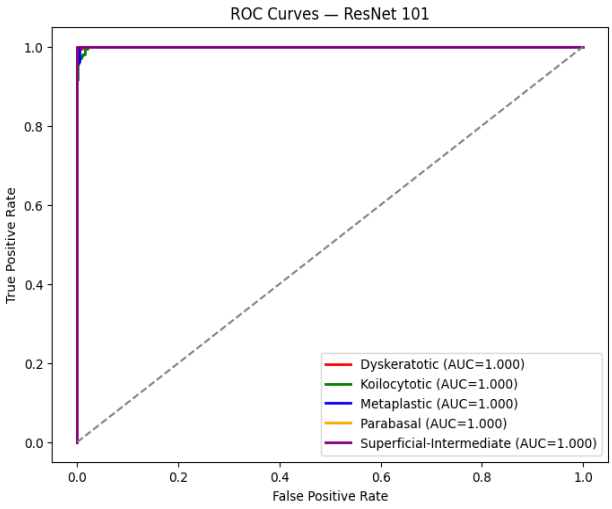

Fig. 10. ROC curve of the ResNet 101.

In both the comparisons of the two models, ResNet101 had the highest point of AUC (1.000) in all 5 classes, whereas DenseNet-121 had lower scores, with 3 giving the highest payment of 1.000 and 2 with 0.999. The two models have better classification, but the minor advantage of ResNet101 is its discriminating power. Nevertheless, DenseNet-121 has a very high level of competition since it offers high rates of feature reuse and a high rate of robustness in automated cervical cell image recognition.

C. Accuracy Curve

The curves of training and validation accuracy related to the model DenseNet-121 in Fig. 11 indicate the quick convergence and stable operation throughout the training epochs of 100. First, one can see that the two accuracies increase rapidly during the first few epochs, which is a sign of efficient feature learning and quick optimization. The model achieves a perfect training accuracy (1.00) and a validation accuracy of 0.99, and the difference between the two is of minimal significance. Such a close fit is a high

generalization and 0 overfitting. The results confirm that DenseNet-121 with proper data preprocessing and data augmentation can successfully produce discriminative representations and provide high classification rates depending on the dataset.

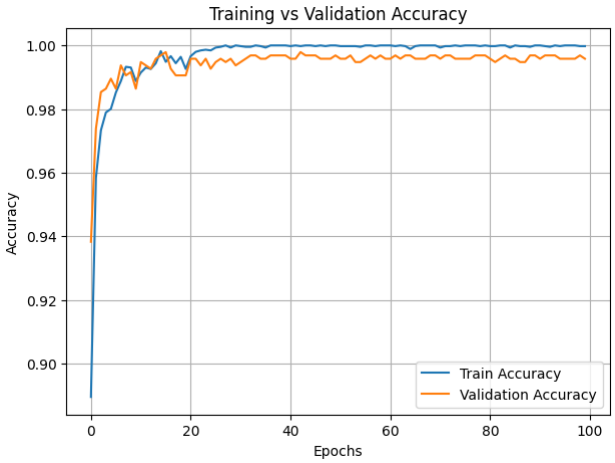

Fig. 11. Training vs validation accuracy curve of the DenseNet-121.

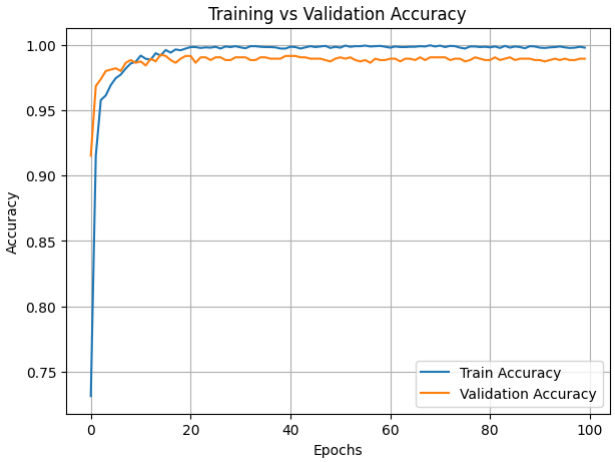

Fig. 12. Training vs validation accuracy curve of the ResNet-101.

The training and validation accuracy curves of the ResNet-101 trained on the classification of cervical cancer cells in 100 epochs are shown in Fig. 12. Both the training and the validation accuracies increase rapidly in the initial epochs, which indicates that the model rapidly acquires the discriminative characteristics of cervical cell images. In about 10–15 epochs, the values of the accuracy stabilize; the training accuracy reaches close to 100%, and the validation accuracy remains at 98.99%. The very low distance between the two curves indicates a high generalization and minimal overfitting. These results support that ResNet-101 is very effective at generating deep and meaningful features in cervical cell images, which allows high-accuracy classification of various types of cancer cells. In comparison to the DenseNet-121, the ResNet-101 model also converges faster and provides high accuracy, yet it is a little faster to achieve validation stability. Although both models share almost equal training accuracy, the DenseNet-121 demonstrates a slightly higher validation power, implying that it has a higher generalization. However, ResNet-101 achieves similar accuracy with minor variations, and it exhibits stable learning performance. On the whole, both of them are perfect models, but DenseNet-121 is more consistent.

D. Loss Curve

Fig. 13 shows the loss curve of the training and validation of a DenseNet-121 model applied to cervical cancer cell classification. The decrease in training and validation losses is high during the first few epochs, but then slows down during the following 20 epochs. The tendency means that the model can quickly absorb key elements. The close position of the training and validation curves implies that it is well generalized and experiences a relatively small overfitting. End loss values have also been 0.23 to 0.25, which represent an optimized framework that can differentiate cervical cells. The repetition of patterns demonstrates that DenseNet-121 can acquire discriminative features to produce the proper classification.

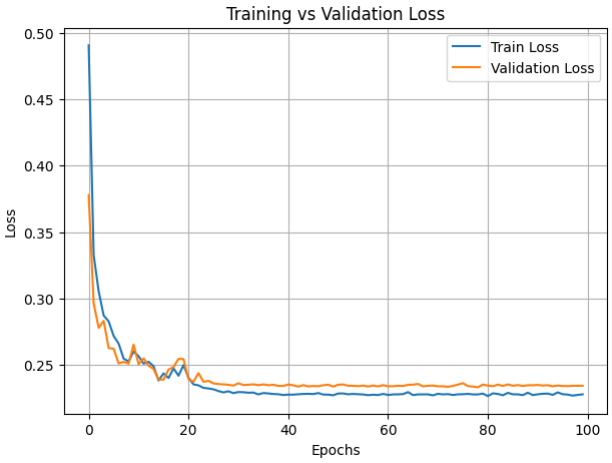

Fig. 13. Training vs validation loss curve of the DenseNet-121.

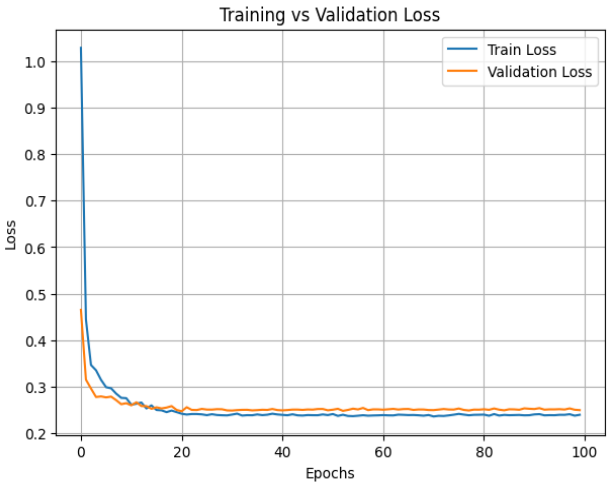

Fig. 14. Loss curve of the ResNet 101

Fig. 14 presents the loss curve training and validation of the ResNet-101model in 100 epochs. Initially, both losses are high, but they decrease sharply afterward, indicating a rapid acquisition of valuable characteristics. The mid-stage

marks the leveling of the losses and marks convergence. During training, training and validation losses remain relatively near, and thus, the generalization and minimal overfitting are high. Both losses narrow towards the end, to approximately 0.23–0.25, which is a consistent, steady performance. This statistic highlights the advantage of the residual connection in ResNet-101 that maintains gradient flow, as well as the fact that the connection enables the training of deep networks to be trained without loss of performance.

The DenseNet-121 model exhibits the same convergence pattern as ResNet 101, and its curves are smoother and more consistent. DenseNet-121 also achieves the same final loss (0.23–0.25) but maintains lower training and validation loss values. This implies increased generalization and permanence. All in all, DenseNet-121 is more reliable and effective regarding the classification of cervical cancer cells in comparison with ResNet-101.

E. F1-Score Curve

Fig. 15 shows the F1-curve over 100 training epochs of DenseNet-121. The score value starts at approximately 0.95 and rises rapidly at the initial stages, which indicates that the model strikes a balance between precision and recall very quickly. The score becomes stable at a score of above 0.99 after about 20 epochs, meaning that it is highly accurate and that the score changes very minimally. Such a consistent pattern proves that DenseNet-121 learns discriminative features and provides a strong, consistent classification of cervical-cancer cells and performs well even on the test set.

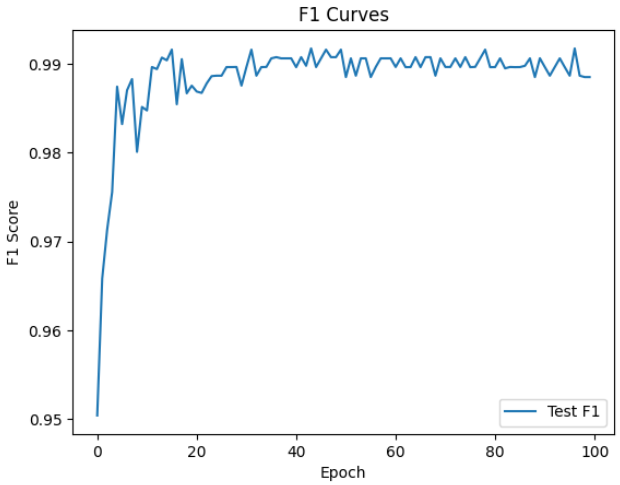

Fig. 15. F1-Score curve of the DenseNet-121.

Fig. 16 monitors the 100-epoch F1-curve of ResNet-101 on the classification of cervical cancer. The score begins close to 0.93 and increases sharply in the beginning, which is an indication of rapid learning of essential features. With approximately 20 epochs, the score becomes stabilized at above 0.98 with low variance. This shows that ResNet-101 is neither biased towards accuracy nor recall, and it is highly accurate with good generalization. The high and steady F1 test results are a certifier of the strong stability of such a model as a result of having a profound feature of residual relationships that contribute to effective feature learning and the flow of gradients. DenseNet-121 and ResNet-101 achieve F1-Scores of over 0.98 and 0.97, respectively, which are good scores in classification. DenseNet-121, however, is better than ResNet-101 with a smoother curve and a higher F1, almost equal to 0.99, which shows greater consistency and generalization. Therefore, DenseNet-121 is more stable and more accurate when classifying cervical cancer cells.

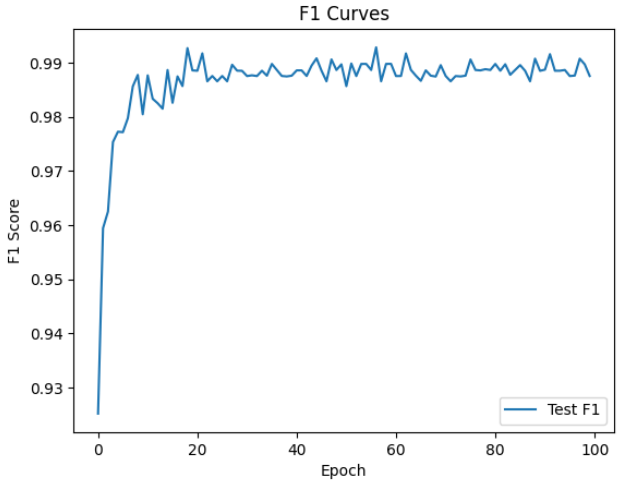

Fig. 16. F1-Score curve of the ResNet-101.

F. Classification Performance Evaluation of the DenseNet-121

TABLE V. CLASSIFICATION PERFORMANCE EVALUATION OF THE DENSENET-121

Class	Precision	Recall	F1-Score	Support
Dyskeratotic	0.98	0.99	0.98	160
Koilocytotic	0.98	0.97	0.97	202
Metaplastic	0.98	0.99	0.99	208
Parabasal	1.00	1.00	1.00	196
Superficial-Intermediate	1.00	1.00	1.00	190

Table V presents the details of the DenseNet-121 model that classifies five types of cervical cells, namely, dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate. The model has good performance with an overall accuracy of 99%. In all classes, the range of precision, recall, and F1-Score is 0.97–1.00. The parabasal and superficial-intermediate cells achieve a perfect score of 1.00, which depicts a high capability of the model in distinguishing such cells accurately. Dyskeratotic, koilocytotic, and metaplastic cells also score high F1-Scores (0.97–0.99), indicating the applicability of the model in identifying the slightest differences in cytologies. It is further supported by the use of macro and weighted averages of 0.99, which shows that there is consistent performance throughout all the classes despite the imbalance of classes. These findings demonstrate that DenseNet-121 can effectively extract the deep morphological features, which makes it reliable and

accurate for automated cervical cancer cell detection, a feat that can potentially contribute significantly to the early detection and screening offered by pathologists.

G. Clinical FNR for Abnormal Lesions-DenseNet-121

TABLE VI. CLASSIFICATION PERFORMANCE EVALUATION OF DENSENET-121

Class	TP	FN	Total Samples	FNR (%)
Dyskeratotic	158	2	160	1.25
Koilocytotic	195	7	202	3.47
Metaplastic	206	2	208	0.96
Parabasal	196	0	196	0.00
Superficial-Intermediate	190	0	190	0.00

As in Table VI, among the abnormal epithelial categories, the Koilocytotic class exhibited the highest class-wise FNR (3.47%), primarily due to morphological similarities with Dyskeratotic and Metaplastic cells. However, when considering all clinically abnormal categories collectively, the overall abnormal lesion FNR was calculated as Eq. (6):

$$FNR = \frac{11}{510} \times 100\% = 1.93\% \tag{6}$$

This indicates that fewer than 2% of clinically significant abnormal epithelial cells were misclassified.

H. Grad-CAM Visualization for Model Interpretability

Grad-CAM was used to explain the decision-making mechanism of the proposed DenseNet-121-based cervical cell classification model. This method outlines the discriminative areas of each input image that produce the most significant part of the ultimate choice of the model. Fig. 17 shows the Grad-CAM for the five cellular categories: dyskeratotic, koilocytotic, metaplastic, parabasal, and superficial-intermediate. The areas in red and yellow are the areas of peak activations, and the blue areas are of lower importance.

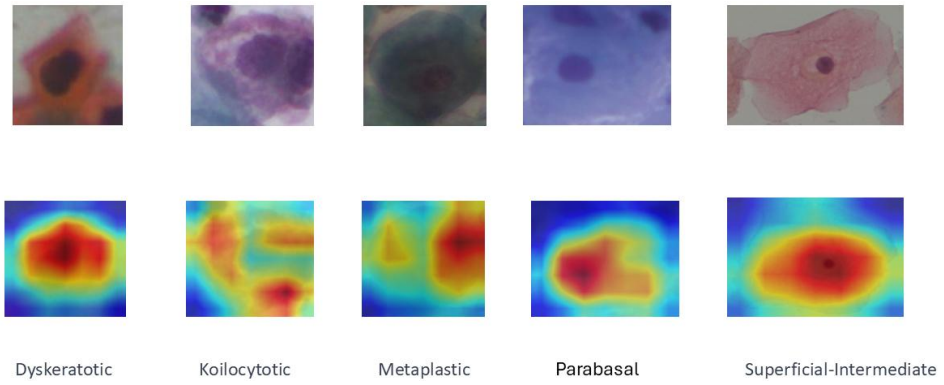

Fig. 17. Grad-CAM visualization of the DenseNet-121.

In this study, Grad-CAM visualization was used to provide qualitative insight into the decision-making process of the deep learning models by highlighting the regions of interest within cervical cell images. Although Grad-CAM helps improve model interpretability, the analysis remains qualitative in nature. A more rigorous evaluation would involve expert validation by board-certified pathologists to assess whether the highlighted regions correspond to clinically relevant diagnostic features. Such a quantitative explainability assessment is considered an important direction for future work to improve the clinical reliability and interpretability of deep learning models.

Deep learning models are often considered 'black-box' systems because their decision-making processes are not always transparent to clinicians. This lack of interpretability presents challenges for clinical adoption and regulatory approval of AI-based medical devices. To address this issue, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM were utilized in this study to provide visual explanations of model predictions. These visualizations help identify image regions that contribute to classification decisions and improve model transparency. However, further work is required to develop more robust and clinically validated explainability methods that meet regulatory requirements and support safe deployment in real-world medical environments.

VI. DISCUSSION

The experimental results demonstrate that DenseNet-121 achieves superior performance for cervical cytology image classification compared with VGG, Inception, EfficientNet, and deeper ResNet variants. This performance gain is reflected in higher accuracy, F1-Score, and overall classification stability. These findings are consistent with prior studies; however, unlike earlier work that often relied on limited datasets or evaluated single architectures in isolation, the present study provides a more systematic and standardized comparative analysis across multiple state-of-the-art CNN models under identical experimental conditions.

From a clinical safety perspective, reducing false negatives is essential in cervical cancer screening programs. The DenseNet-121 model achieved an overall abnormal lesion FNR of 1.93%, demonstrating high sensitivity toward clinically significant cases. Importantly, Parabasal and Superficial-Intermediate cells exhibited zero false negatives, further strengthening the model's

reliability. These findings suggest that the proposed model can serve as a safe AI-assisted decision support tool, particularly in resource-constrained screening environments.

The improved performance of DenseNet-121 can be attributed to its dense connectivity mechanism, which facilitates efficient feature reuse, mitigates the vanishing gradient problem, and promotes stronger gradient propagation throughout the network. This architecture is particularly effective for capturing subtle nuclear and cytoplasmic morphological variations that distinguish visually similar epithelial cell classes. Furthermore, Grad-CAM-based explainability analysis indicates that the model focuses on clinically relevant regions, such as nuclear boundaries and chromatin patterns, thereby supporting its potential utility in digital pathology workflows, automated triage systems, and tele cytology applications, especially in low-resource settings where expert cytologists are limited.

Despite these promising results, several limitations should be acknowledged. The study is constrained by the use of a relatively small and heterogeneous dataset, the absence of external multi-center validation, and the lack of comparison with recent hybrid architectures and foundation-scale vision models. These factors may affect the generalizability of the proposed framework.

This study focuses on classification using isolated single-cell images, which simplifies the cervical cancer screening problem compared to real clinical practice. In practical screening environments, cytologists analyze Whole-Slide Images (WSI) that contain complex structures including overlapping cells, mucus, inflammatory cells, and staining variations. These challenges are not fully represented in isolated cell datasets. Therefore, while the proposed approach demonstrates promising classification performance, further research using whole-slide image datasets is necessary to ensure clinical applicability and robustness.

The extremely high classification performance, including AUC values close to 1.000, may be influenced by the relatively homogeneous nature of the dataset and the controlled imaging conditions. The dataset was divided into independent training, validation, and testing sets to minimize the risk of data leakage and to ensure a fair evaluation of the models. Despite the strong performance achieved in this study, real clinical datasets typically contain greater variability, noise, and imaging artifacts. Therefore, further evaluation on more diverse and large-scale clinical datasets is necessary to confirm the robustness and generalization capability of the proposed approach.

Recent advances in medical image analysis have demonstrated the effectiveness of transformer-based architectures such as Vision Transformers (ViT) and hybrid Mamba-based models, which have shown promising performance in capturing global contextual information. While this study focuses on the comparative evaluation of established CNN architectures, future work will include the investigation of transformer-based and hybrid deep learning models to further improve classification performance and robustness.

One limitation of this study is the absence of patient-level identifiers in the publicly available dataset, which prevents performing a patient-wise data split. As a result, cells from the same patient may potentially appear in both the training and testing sets, which could lead to optimistic performance estimates. Although standard random splitting was applied to maintain balanced class distributions, patient-wise evaluation would provide a more realistic assessment of model generalization. Future work will focus on utilizing datasets with patient-level annotations to enable patient-wise validation and improve clinical reliability.

Future work will focus on expanding the dataset with hospital-grade whole-slide cytology images, performing cross-institutional validation, and benchmarking against CNN-Transformer hybrid models and foundation vision backbones. In addition, segmentation-assisted classification pipelines will be explored to enhance cell-level feature learning. The development of lightweight, edge-deployable variants of the model for real-time clinical screening is also planned. Finally, evaluation on standard public benchmarks, including the Herlev and SIPaKMeD datasets, will be conducted to enable reproducibility and facilitate direct comparison with existing methods.

VII. Conclusion

This paper made a comparative study of nine deep learning architectures trained on automated cervical cytology classification and discovered that DenseNet-121 can perform optimally. It attained 98.85% accuracy and an F1 of 0.9885 and was repeatedly able to identify the appropriate nuclear and cytoplasmic locations by Grad-CAM visualization. All these findings affirm that DenseNet-121 is a reliable tool for the identification of subtle morphological differences, which are crucial in the early detection of cervical abnormalities. The model is thus a valuable decision support mechanism to cytopathologists and enhances screening ability, more so in low-resource environments. In addition to achieving high overall classification accuracy (98.85%), the proposed DenseNet-121 model demonstrated a clinically favorable abnormal lesion FNR below 2%, reinforcing its potential suitability for reliable cervical cytology screening. The study has limitations, even though it has performed well. The data was not extensive and was of mixed types and not based on regular hospital processes but stored in archives available to everyone publicly. These shortcomings can limit the ability of the model to be used in an actual clinical setting.

The study has not investigated recent hybrid CNN-Transformer models, vision foundation models, and multimodal pipelines, although they have demonstrated promise in medical imaging. Further work will involve gathering bigger, institution-grade cytology data and performing multi-center external validation, as well as performance across different staining, microscopy, and imaging equipment. It will also incorporate advanced architectures, including Swin Transformers, ViT-based

hybrids, and segmentation-based classifiers. Further work will include how to optimize the model to be real-time and deployed on edge-computing devices, mobile microscopes, and tele cytology systems to be able to scale and remove the need for user intervention in cervical cancer screening. In general, the suggested DenseNet-121 is a strong, interpretable architecture that can contribute to the development of the sphere of early cervical cancer detection and the improvement of the screening programs worldwide. This study utilizes publicly available and fully anonymized cervical cytology image datasets. No personal or patient-identifiable information was involved; therefore, ethical approval was not required.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

S.A.P. conceived the research idea, conducted the experiments, performed data analysis, and prepared the original draft of the manuscript. G.M.S. contributed to the research methodology, software implementation, and experimental validation. M.Z. assisted with data collection, result visualization, and validation of the experimental findings. S.R. supervised the research activities and provided critical guidance throughout the study. JU contributed to manuscript review and editing and provided funding support for this research. All authors have read and approved the final version of the manuscript.

FUNDING

This research is funded by Woosong University Academic Research 2026.

REFERENCES

[1] S. Sagor, M. T. Ahad, F. Ahmed, R. Ayon, and S. Parvin, "A study on deep convolutional neural networks, transfer learning, and MNet model for cervical cancer detection," arXiv Preprint, arXiv:2509.16250, 2025.

[2] S. Muksimova, S. Umirzakova, S. Kang, and Y. I. Cho, "CerviLearnNet: Advancing cervical cancer diagnosis with reinforcement learning-enhanced convolutional networks," *Heliyon*, vol. 10, no. 9, e29913, 2024.

[3] S. Muksimova, S. Umirzakova, K. Shoraimov, J. Baltayev, and Y. I. Cho, "Novelty classification model use in reinforcement learning for cervical cancer," *Cancers*, vol. 16, no. 22, 3782, 2024.

[4] S. M. Dipto, M. T. Reza, N. T. Mim *et al*., "An analysis of decipherable red blood cell abnormality detection under federated environment leveraging XAI incorporated deep learning," *Scientific Reports*, vol. 14, 25664, 2024. doi: 10.1038/s41598-024-76359-0

[5] D. D. Himabindu, E. L. Lydia, M. V. Rajesh *et al*., "Leveraging swin transformer with an ensemble of deep learning models for cervical cancer screening using colposcopy images," *Scientific Reports*, vol. 15, no. 1, 7900, 2025.

[6] M. Mascarenhas, M. Martins, L. Barroso *et al*., "Automated detection and classification of cervical and anal squamous cancer precursors using deep learning and multidevice colposcopy," *Scientific Reports*, vol. 15, no. 1, 33068, 2025.

[7] M. E. Haque, M. Nurul-Absur, F. Al-Farid *et al*., "A novel interpretable and real-time dengue prediction framework using clinical blood parameters with genetic and GAN-based optimization," *Front. Artif. Intell*., vol. 8, 1626699, 2025. doi: 10.3389/frai.2025.1626699

[8] A. Al-Mohimeed, M. Shehata, N. El-Rashidy, S. Mostafa, A. S. Talaat, and H. Saleh, "ViT-PSO-SVM: Cervical cancer prediction based on integrating vision transformer with particle swarm optimization and support vector machine," *Bioengineering*, vol. 11, no. 7, 729, 2024.

[9] K. Abinaya and B. Sivakumar, "A deep learning-based approach for cervical cancer classification using 3D CNN and vision transformer," *Journal of Digital Imaging*, vol. 37, no. 1, pp. 280–296, 2024.

[10] M. H. K. Mehedi, M. Khandaker, S. Ara, M. A. Alam, M. F. Mridha, and Z. Aung, "A lightweight deep learning method to identify different types of cervical cancer," *Scientific Reports*, vol. 14, no. 1, 29446, 2024.

[11] T. Baba, A. S. M. Miah, J. Shin, and M. A. M. Hasan, "Cervical cancer detection using a multi-branch deep learning model," arXiv Preprint, arXiv:2408.10498, 2024.

[12] S. K. Mathivanan, D. Francis, S. Srinivasan *et al*., "Enhancing cervical cancer detection and robust classification through a fusion of deep learning models," *Scientific Reports*, vol. 14, no. 1, 10812, 2024.

[13] S. Saini, K. Ahuja, and A. S. Chauhan, "Block-fused attention-driven adaptively-pooled ResNet model for improved cervical cancer classification," arXiv Preprint, arXiv:2405.01600v3, 2025.

[14] M. Mohammed-Alzahrani, U. Ali-Khan, and S. A. Al-Garni, "Ensemble and transformer encoder-based models for the cervical cancer classification using pap-smear images," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp.189–198, 2024.

[15] S. L. Tan, G. Selvachandran, W. Ding, R. Paramesran, and K. Kotecha, "Cervical cancer classification from Pap smear images using deep convolutional neural network models," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 16, no. 1, pp. 16–38, 2024.

[16] G. Litjens, T. Kooi, B. E. Bejnordi *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[17] I. Pacal, "Chaotic learning rate scheduling for improved CNN-based breast cancer ultrasound classification," *Chaos Theory and Applications*, vol. 7, no. 3, pp. 297–306, 2025.

[18] I. Pacal, A. Algarni, and I. Kunduracioglu, "Adaptive and efficient deep learning model for automated ischemic stroke lesion segmentation," *Biomedical Signal Processing and Control*, vol. 118, 109658, 2026.

[19] Y. R. Park, Y. J. Kim, W. Ju *et al*., "Comparison of machine and deep learning for the classification of cervical cancer based on micrography images," *Scientific Reports*, vol. 11, no. 1, 16143, 2021.

[20] J. Gangrade, R. Kuthiala, S. Gangrade, Y. P. Singh, M. R., and S. Solanki, "A deep ensemble learning approach for squamous cell classification in cervical cancer," *Scientific Reports*, vol. 15, no. 1, 7266, 2025.

[21] S. P. B. Shanthi, F. Faruqi, H. K. S. Hareesha, and R. Kudva, "Deep convolution neural network model for malignancy detection and classification in microscopic uterine cervix cell images," *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 11, pp. 3447–3456, 2019.

[22] M. Mehmood, M. Rizwan, M. Gregus, and S. Abbas, "Machine learning assisted cervical cancer detection," *Frontiers in Public Health*, vol. 9, 788376, 2021.

[23] F. Asadi, C. Salehnasab, and L. Ajori, "Supervised algorithms of machine learning for the prediction of cervical cancer," *Journal of Biomedical Physics and Engineering*, vol. 10, no. 4, pp. 513–522, 2020.

[24] N. Al-Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," *Sensors*, vol. 22, no. 11, 4132, 2022.

[25] A. S. Shaik, S. M. Aswatha, and R. J. Pandya, "Deep learning enabled segmentation, classification and risk assessment of cervical cancer," *IEEE Access*, vol. 13, pp. 1–15, 2025.

[26] Z. Hong, J. Xiong, H. Yang, and Y. K. Mo, "Lightweight low-rank adaptation vision transformer framework for cervical cancer detection and cervix type classification," *Bioengineering*, vol. 11, no. 5, 468, 2024.

[27] A. Khozaimi, I. Darti, S. Anam, and W. M. Kusumawinahyu, "Advanced cervical cancer classification: enhancing pap smear images with hybrid PMD Filter-CLAHE," arXiv Preprint, arXiv:2506.15489, 2025.

[28] A. Khozaimi and W. F. Mahmudy, "New insight in cervical cancer diagnosis using a convolution neural network architecture," *IAES International Journal of Artificial Intelligence*, vol. 13, pp. 3092–3100, 2024.
[29] L. Liu, J. Liu, Q. Su *et al*., "Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: A systematic review and meta-analysis," *eClinicalMedicine*, vol. 80, 102992, 2025.
[30] T. Wu, E. Lucas, F. Zhao, P. Basu, and Y. Qiao, "Artificial intelligence strengthens cervical cancer screening—present and future," *Cancer Biology & Medicine*, vol. 21, no. 10, pp. 865–877, 2024.
[31] P. Patre and D. Verma, "Deep dive into deep learning methods for cervical cancer detection and classification," *Reports of Practical Oncology and Radiotherapy*, vol. 30, no. 3, pp. 396–416, 2025.
[32] K. Kamal and E. Z. Hamid, "A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images," *Research Square*, 2023.
[33] B. K. Eskibağlar, Y. P. Yavuz, G. K. Doğan *et al*., "Denta-HybridoNet: A hybrid CNN-transformer architecture for automated detection of developmental dental anomalies in pediatric panoramic radiographs," *Biomedical Signal Processing and Control*, vol. 118, 109784, 2026.
[34] School-YMRRR. (2023). Cervical cancer classification—11GMM. *Roboflow Universe*. [Online]. Available: https://universe.roboflow.com/school-ymrrr/cervical-cancer-classification-i1gmm
[35] T. Petursdottir. (2023). Cervical cancer classification—YZBC4. *Roboflow Universe*. [Online]. Available: https://universe.roboflow.com/thorbjorg-petursdottir/cervical-cancer-classification-yzbc4