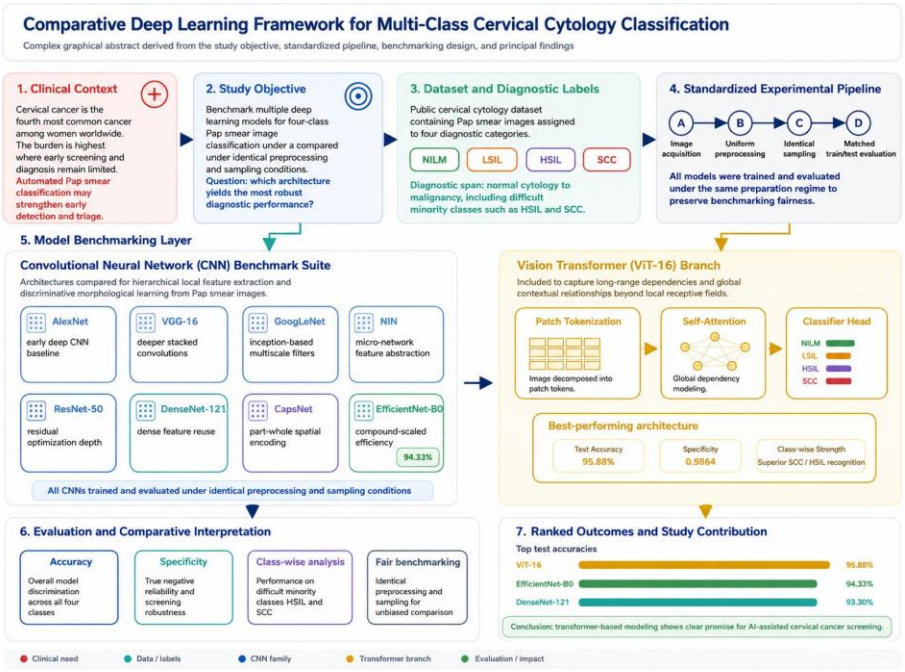
Cervical Cytology Classification Using Multiple CNN Architectures with Transformer-Based Feature Enhancement

Mehreen Sirshar[1*], Omama Shakeel[2], Nayab Asim[3], Fakeeha Jafari[4],
Hani Almoamari[5], Adnan Nadeem[6], Mohammad Zubair Khan[7*], Ibrahim Aljubayri[8]
Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan [1, 2, 3]
Department of Software Engineering, National University of Modern Languages, Pakistan [4]
Faculty of Computer and Information System, Islamic University of Madinah, Madinah 42351, Saudi Arabia [5, 6, 7]
Department of Computer Science and Information,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia [8]

***Abstract*—Cervical cancer is the fourth most common cancer among women worldwide and remains a major public health challenge, particularly in regions with limited access to early screening and diagnosis. Accurate classification of cervical cytology images is critical for early detection of cervical cancer, which remains a major health burden in low- and middle-income countries. This study presents a comprehensive evaluation of multiple deep learning architectures for the automated classification of Pap smear images into four diagnostic categories: Negative for Intraepithelial Lesion or Malignancy (NILM), Low-Grade Squamous Intraepithelial Lesion (LSIL), High-Grade Squamous Intraepithelial Lesion (HSIL), and Squamous Cell Carcinoma (SCC). We systematically compare eight Convolutional Neural Network (CNN) architectures: AlexNet, VGG-16, GoogLe-Net, Network-in-Network (NIN), ResNet-50, DenseNet-121, Capsule Networks, and EfficientNet-B0 on a publicly available cervical cytology dataset. To enhance feature representation and capture long-range dependencies, we additionally incorporate a Vision Transformer (ViT-16) model. All models are trained and evaluated under identical preprocessing and sampling conditions to ensure fair benchmarking. Experimental results demonstrate that ViT-16 achieves the highest test accuracy of 95.88% and an overall specificity of 0.9864, outperforming all CNN counterparts. EfficientNet-B0 and DenseNet-121 also showed strong performance, achieving 94.33% and 93.30% accuracy, respectively. Notably, ViT-16 provided superior classification outcomes for challenging minority classes such as SCC and HSIL. The findings highlight the growing potential of transformer-based models in cytopathology and underscore the importance of architectural design in developing robust diagnostic tools. This work contributes a comparative foundation for future research in AI-assisted cervical cancer screening systems.**

***Keywords*—Cervical cancer; CNN architectures; vision transformer**



Graphical abstract.

*Corresponding author.

I. Introduction

The cervix is the lower, narrow portion of the uterus that connects the uterine cavity to the vaginal canal. Functionally, it plays a vital role in female reproductive health by facilitating the passage of sperm, acting as a barrier to infections, and serving as the birth canal during labor. Despite its biological importance, the cervix is susceptible to malignant transformations that lead to cervical cancer.

Cervical cancer arises primarily due to persistent infection with high-risk human papillomavirus (HPV) types. According to Delgado 2022 [1], cervical cancer accounted for approximately 604,000 new cases and 342,000 deaths worldwide, making it the fourth most common cancer among women globally. The disease burden is disproportionately higher in low- and middle-income countries, where access to early screening and preventive healthcare remains limited. These statistics highlight the urgent need for reliable and scalable diagnostic systems to support early detection and intervention. The disease progresses through precancerous stages, including low-grade squamous intraepithelial lesions (LSIL) and high-grade squamous intraepithelial lesions (HSIL), before developing into invasive squamous cell carcinoma (SCC). Early detection and classification of these lesions are crucial for improving survival rates and reducing treatment complexity. Fig. 1 shows an overview of cervical cytology classes: (A) HSIL, (B) LSIL, (C) NILM, and (D) SCC Pap smear images representing different diagnostic categories.

Traditionally, cervical cytology diagnosis is performed using the Papanicolaou (Pap) smear test, which relies on expert microscopic examination of cervical epithelial cells. Although effective, manual screening is labor-intensive, subjective, and prone to inter-observer variability. In response to these challenges, artificial intelligence (AI) and deep learning approaches have gained traction for automated image-based diagnostics. Convolutional Neural Networks (CNNs), in particular, have demonstrated superior performance in feature extraction and classification tasks across medical imaging domains [2, 3].

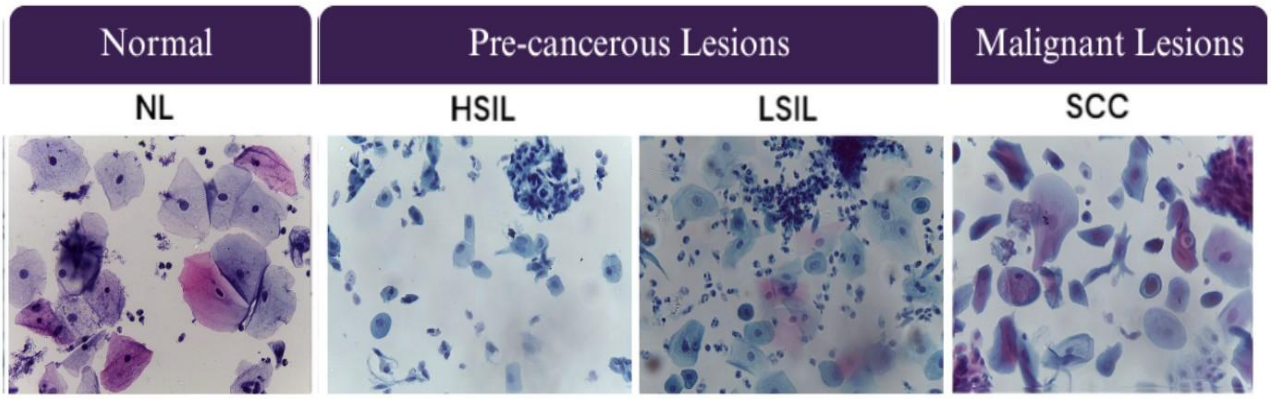


Fig. 1. Overview of cervical cytology classes: (A) HSIL, (B) LSIL, (C) NILM, and (D) SCC pap smear images representing different diagnostic categories.

Recent studies have introduced various CNN-based architectures for cervical cancer detection. For instance, Liu et al. [4] proposed a hybrid CNN and Transformer model achieving high accuracy in classifying Pap smear images. Similarly, Gangrade et al. [5] and Huang et al. [3] utilized ensemble and fine-tuned deep learning techniques to improve classification outcomes. These models outperform conventional machine learning algorithms by reducing manual feature engineering and enhancing generalizability [6].

Nevertheless, issues such as class imbalance, image variability, and staining artifacts continue to challenge model robustness and reliability [8, 9]. To address these, researchers have incorporated data augmentation [6] and attention-based models [10] to enhance feature learning. Despite such advancements, the evaluation of different CNN architectures on a balanced and diverse dataset remains a key research gap.

Cervical cytology classification via deep options has been an active area of research. Many studies rely primarily upon transferring knowledge from other domains (transfer learning) and focus on establishing high levels of performance within ideal conditions. Very few published studies systematically examine multiple convolutional neural network (CNN) architectures and/or transformers under standardized training conditions and/or real, imbalanced, cervical cytology datasets. Therefore, some controlled comparative experiments will be performed to determine the impacts of architectural designs on robustness of the classification under the constraints of realistic clinical data.

The remainder of this study is organized as follows: Section II presents the related work with a focus on recent advances in cervical cancer classification using deep learning. Section III details the Materials and Methods, model architectures, and Dataset. Section IV discusses the experimental setup, Methodology, Training, and Implementation Details, and Section V provides performance comparisons among the evaluated models. Finally, Sections VI and VII present the Discussion and Conclusion that outline potential directions for future research.

II. Related Work

Cervical cancer remains a leading cause of mortality among women worldwide, emphasizing the urgent need for early and accurate diagnostic tools. Deep learning techniques, particularly convolutional neural networks (CNNs) and transformer-based models, have demonstrated strong potential in automating

cervical cytology image classification. This section critically reviews recent studies, analyzing their methodological innovations, performance outcomes, and relevance to real-world medical applications.

Nirmala et al. [2] introduced an adaptive system combining CNNs with adaptive vision transformers, achieving 97.43% accuracy through refined segmentation and classification strategies. Similarly, Huang et al. [3] employed AF-SENet with fine-tuned CNNs to enhance the detection of LSIL, pushing the accuracy to 95.33%. Liu et al. [4] proposed the CVM-Cervix framework, which synergizes CNNs, visual transformers, and MLPs, attaining 91.72% accuracy and exemplifying the advantage of hybrid architectures.

Several researchers explored ensemble and multi-model strategies. Gangrade et al. [5] combined CNN, AlexNet, and SqueezeNet, achieving 94% accuracy in five-class classification, while Tripathi et al. [6] applied majority voting across CNN models to improve classification reliability. Benyes et al. [6] used AE-CNN on SurePath and ThinPrep datasets, reporting 96.54% accuracy and demonstrating effective domain adaptation.

Signal processing integration was explored by Palanisamy et al. [7], who combined dual-tree complex wavelet transform (DTCWT) with CNNs, achieving 99% accuracy across four Pap smear categories. Complementarily, Sellamuthu et al. [7] developed a CNN pipeline with augmentation and pooling, achieving a Pap smear detection index of about 99%.

Transfer learning techniques also showed promise. Tan et al. [8] tested 13 pretrained CNNs and identified DenseNet-201 as the most effective on the Herlev dataset. Novitasari et al. [9] applied ResNet-50 to classify colposcopy images into five stages with 100% accuracy, while Wulandari et al. [9] highlighted the stability of ResNets in mitigating vanishing gradient problems.

Whole-slide image (WSI) analysis was addressed by Cheng et al. [10], who proposed a progressive lesion recognition system using dual-resolution WSIs, achieving 95.1% sensitivity and 93.5% specificity. Kaur et al. [11] compared 16 CNN models and reported 95% accuracy with ResNet-50 and 99.95% with VGG16 for binary classification, emphasizing the strength of pre-trained models.

Alsubai et al. [12] integrated PCA and GWO for feature selection in a CNN-based classifier on SIPaKMeD data, achieving 91.13% accuracy and strong precision. De La Cruz Paucar et al. [13] improved generalization by employing dropout in a two-stage CNN, where ResNet-50 outperformed competing models. Finally, Zhao et al. [14] tackled data imbalance by generating cervical images using taming transformers, thereby enhancing classification accuracy and dataset diversity. Recent advances in medical image classification and deep learning architectures, including transformer-based models and conventional CNN frameworks, have demonstrated strong potential for improving automated diagnostic systems in medical imaging applications [15–22]. Recent studies have also explored optimization-driven and hybrid deep learning approaches for medical image analysis and disease classification tasks [31–33].

These studies collectively underline the transformative impact of deep learning on cervical cancer screening, while also revealing ongoing challenges related to class imbalance, model generalizability, and clinical-scale deployment. Table I shows a comparative summary of related studies on cervical image analysis.

III. Materials and Methods

This section outlines the dataset, preprocessing techniques, model architectures, and training strategies used in this study. The proposed methodology is designed to classify cervical cytology images into four diagnostic categories using multiple CNN architectures and a transformer-based feature enhancement stage. The process includes image preprocessing, CNN-based feature extraction, integration of a transformer model for contextual learning, and final classification.

A. Dataset

This study employs the publicly available Mendeley cervical cancer image dataset [23] (DOI:10.17632/zddtpgzv63.4), which includes high-resolution cytology slide images stained using the Papanicolaou (Pap) method. The dataset is composed of 963 images representing four clinically annotated classes: Negative for Intraepithelial Lesion or Malignancy (NILM), Low-grade Squamous Intraepithelial Lesion (LSIL), High-grade Squamous Intraepithelial Lesion (HSIL), and Squamous Cell Carcinoma (SCC). Images were captured at 400× magnification (using a Leica ICC50 HD attached to a microscope) and sourced from 460 patients in real-world clinical settings. Each image in the dataset has been labeled by expert cytologists. The dataset presents real-world variability in terms of staining artifacts, cell overlap, and illumination differences, making it a suitable benchmark for evaluating the robustness of deep learning models. However, it suffers from class imbalance, with the NILM class being overrepresented and SCC and LSIL being underrepresented. This imbalance poses a challenge for classifiers and is addressed through class-weighting and sampling strategies in training.

B. Computing Infrastructure

The Experiments are conducted in Google Colab using a standard Linux-based runtime environment. The operating system is a 64-bit Linux distribution provided by the Colab virtual machine. Model development and training were implemented in Python using the PyTorch deep learning framework with supporting scientific libraries (NumPy, Matplotlib, and scikit-learn). Experiments are executed under Colab's managed hardware configuration, and where GPU acceleration was available, training leveraged an NVIDIA CUDA-enabled GPU provided by the platform. In cases where GPU is not enabled, training is performed on the Colab CPU backend. Data are accessed from cloud storage (Google Drive) and loaded using a standard folder-based dataset structure compatible with ImageFolder. All pre-processing (resizing, normalization, and tensor conversion) and evaluation (confusion matrix, ROC/AUC, and classification report) are performed within the same environment to ensure consistency. The same runtime setup is used across all model variants to maintain comparable experimental conditions.

TABLE I. COMPARATIVE SUMMARY OF RELATED STUDIES ON CERVICAL IMAGE ANALYSIS

Study	Year	Techniques	Application	Strengths	Weaknesses
Nirmala et al. [2]	2025	CNN + Adaptive ViT	Pap smear classification	High accuracy (97.43%); robust segmentation and classification	Sensitive to data imbalance and segmentation quality
Huang et al. [3]	2020	AF-SENet + fine-tuned CNN	LSIL recogni-tion	Improved LSIL detection; effective deep feature fusion	Requires careful tuning; limited generalizability
Liu et al. [4]	2022	CNN + ViT + MLP (CVM-Cervix)	Pap smear classification	Hybrid synergy; solid accuracy (91.72%)	Complex model; dataset-dependent benefits
Gangrade et al. [5]	2025	Ensemble (CNN, AlexNet, SqueezeNet)	Five-class cervical classification	Ensemble outperforms single models (94%)	Higher inference cost; requires model aggregation
Tripathi et al. [6]	2022	Multiple CNNs + majority voting	Multiclass Pap smear	Voting improved reliability	Still sensitive to class imbalance
Benyes et al. [6]	2022	AE-CNN + domain adaptation	Pap smear (SurePath/Thin-Prep)	Strong accuracy (96.54%); cross-prep robustness	Dataset-specific adaptation; limited minority class detail
Palanisamy et al. [7]	2022	DTCWT + CNN	Pap smear (4 categories)	Very high accuracy (99%); DL + signal synergy	Overfitting risk; preprocessing complexity
Sellamuthu et al. [7]	2022	CNN + augmentation + pooling	Pap smear detection index	Robust with augmentation; simple pipeline	Limited innovation; weak for rare classes
Tan et al. [8]	2024	TL with 13 pretrained CNNs (DenseNet-201 best)	Pap smear (Her-lev)	Effective under data scarcity	Dependent on pre-train domain
Novitasari et al. [9]	2022	ResNet-50	Colposcopy, 5 stages	Reported 100% accuracy; strong residual learning	Possible overfitting; domain-specific
Wulandari et al. [9]	2022	Residual networks (multi-stage)	Colposcopy	Mitigates vanishing gradients	Requires high compute; dataset dependency
Cheng et al. [10]	2021	Progressive lesion recognition (WSIs)	Whole-slide imaging	Clinically relevant; high sensitivi-ty/specificity	Compute-heavy; pipeline complexity
Kaur et al. [11]	2025	16 pretrained CNNs (ResNet-50, VGG16)	Pap smear 7	Strong baselines (VGG16: 99.95% binary)	Binary results may not generalize to multiclass
Alsubai et al. [12]	2023	PCA + GWO + CNN	SIPaKMeD dataset	Good accuracy (91.13%); feature reduction helps	Risk of losing rare-class fea-tures
De La Cruz Paucar et al. [13]	2024	Two-stage CNN + dropout (ResNet-50 best)	Pap smear classification	Dropout improved generalization	Limited external validation
Zhao et al. [14]	2022	Taming transformers (data generation)	Cervical image augmentation	Tackles imbalance; boosts accuracy/diversity	Synthetic–real domain gap; validation needed

IV. EXPERIMENTAL SETUP

This section outlines the practical implementation of the proposed methodology, including model training protocols, evaluation metrics, and analysis tools. Each component of the setup, from data preprocessing to final evaluation, is described below.

A. Training and Implementation Details

All of the models have been created with the same dataset structure and preprocessing pipeline. The data preprocessing for images involved resizing all of the images to one common size matching each of the architectures, converting all images to tensors, and normalizing all images with the mean and standard deviation of ImageNet. To keep the experiments reproducible, we pre-split our dataset into three parts (70% training, 15% validation, and 15% testing) using a stratified split. The stratified split preserved the class representation in the training, validation, and testing datasets for all four classes of diagnosis (NILM, LSIL, HSIL, SCC). Random number generator seed 42 was used to split the dataset and create and initialize the models to ensure consistent conditions in these experiments. To handle the great majority of the imbalanced dataset, class weightings were determined based on the number of occurrences of each class in the training set, and we utilized a WeightedRandomSampler during training to create balanced mini-batches of samples. Balanced mini-batches allowed the SCC class to receive better representation in terms of class size and ultimately provided better representation of the SCC class' actual proportion of the study sample.

The Adam optimizer is used for each of the models at a learning rate of 0.0001 and a batch size of 32. The weighted cross-entropy loss function is utilized to incorporate class weights into the optimization process. Each model is trained for 10 epochs, with training and validation loss and accuracy being used to monitor model performance.

The preliminary experiments showed that model validation performance plateaus around 8–10 epochs, indicating early convergence given the moderate size of the dataset. Training models beyond these epochs did not show large increases in model performance and created a higher risk of overfitting; therefore, all models were trained for 10 epochs to maintain consistency and to not over-optimize the models. After training, the best-performing model checkpoint (based on validation accuracy) is evaluated on the independent test set. All experiments are conducted under the same computational environment to ensure comparability across architectures.

B. Evaluation Metrics

The following equations define the metrics used to evaluate model performance:

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

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

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

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

Area Under the ROC Curve (AUC): Measures the model's ability to distinguish between classes across all thresholds.

Specificity Analysis: Class-wise specificity is calculated to assess each model's ability to correctly identify negative samples. Overall specificity is derived by averaging across classes.

In addition to class-wise metrics, the final test accuracy and average specificity are computed to summarize overall model effectiveness. These metrics provide both class-wise and overall assessments of the model's classification ability, particularly highlighting performance on underrepresented classes such as *SCC*.

C. Methodology

The proposed methodology is designed to classify cervical cytology images into four diagnostic categories using multiple CNN architectures and a transformer-based feature enhancement stage. The process includes image preprocessing, CNN-based feature extraction, integration of a transformer model for contextual learning, and final classification.

1) Deep learning models employed: In this study, we systematically evaluate eight distinct CNN architectures—AlexNet, VGGNet, GoogLeNet, Network-in-Network (NIN), ResNet, DenseNet, Capsule Networks, and EfficientNet—for cervical cytology classification. The objective is to compare their classification performance and assess their capability in addressing limitations such as computational inefficiency, poor spatial representation, and vanishing gradients. By using a publicly available cytology image dataset, we aim to provide a comprehensive benchmark and insights into architecture selection for real-world AI-assisted cervical cancer diagnostics.

AlexNet (2012) serves as a foundational CNN model, comprising five convolutional layers and three fully connected layers, totaling approximately 61 million parameters. Though relatively shallow, it provides a strong baseline for evaluating improvements in newer models [24]. Advantages: Simple and efficient for moderate datasets. Limitations: Limited depth reduces capability for complex feature learning.

VGGNet (2014) introduces a deeper network with about 138 million parameters using uniform 3×3 convolution filters. This architecture improves hierarchical feature learning but is computationally expensive [30]. Advantages: Effective in learning fine-grained features. Limitations: High training time and memory demand. Motivation: To analyze the benefit of increased depth over AlexNet's simpler design.

GoogLeNet (Inception) (2014) features inception modules for multi-scale feature extraction and contains roughly 6.8 million parameters. Its parallel convolution paths make it both deep and computationally efficient [26]. Advantages: Learns hierarchical and multi-scale features. Limitations: Complex to tune and modify. Motivation: Introduced to evaluate multi-path processing and spatial efficiency.

Network in Network (NIN) (2014) replaces dense layers with micro neural networks, namely MLPs, for increased abstraction [29]. It utilizes about 29 million parameters. Advantages: Improved local abstraction through MLPs. Limitations: Lacks the expressiveness of deeper residual connections. Motivation: Explores local abstraction techniques in contrast to global dense features.

ResNet (2015) addresses the vanishing gradient problem through skip connections. ResNet-50, with 25.6 million parameters, enables training of very deep networks [25]. Advantages: Residual learning enables depth without degradation. Limitations: Computationally demanding despite a moderate parameter count. Motivation: Evaluates the effectiveness of residual connections in classification performance.

DenseNet (2017) connects each layer to all subsequent ones to encourage feature reuse and gradient flow [27]. It contains around 8 million parameters. Advantages: Feature reuse, reduced parameters, and strong gradient propagation. Limitations: Memory usage can increase due to feature map concatenation. Motivation: Introduced to compare dense connectivity with residual learning.

EfficientNet (2019) balances depth, width, and resolution using compound scaling. With only 5.3 million parameters, EfficientNet-B0 offers high accuracy at low computational cost [28].

Advantages: Lightweight and efficient architecture. Limitations: May underperform on extremely complex or high-resolution inputs. Motivation: Explores whether efficient models can achieve performance comparable to deeper networks.

Vision Transformer (ViT) (2021) enhances feature representation by applying transformer architecture to image patches [16]. Each image is split into fixed-size patches, linearly embedded, and processed by a stack of transformer encoders using multi-head self-attention. We fine-tuned a pretrained ViT-B/16 model on our dataset. Advantages: Captures long-range dependencies and global context. Limitations: Requires large datasets or pretraining for optimal performance. Motivation: Introduced as a transformer-based alternative to CNNs for global feature learning in cytology images. Fig. 2 shows an overview of the CNN architectures used in this study. Fig. 3 shows the architecture of the Vision Transformer (ViT). The image is divided into fixed-size patches, which are embedded and processed through multiple transformer encoder blocks with multi-head self-attention before final classification.

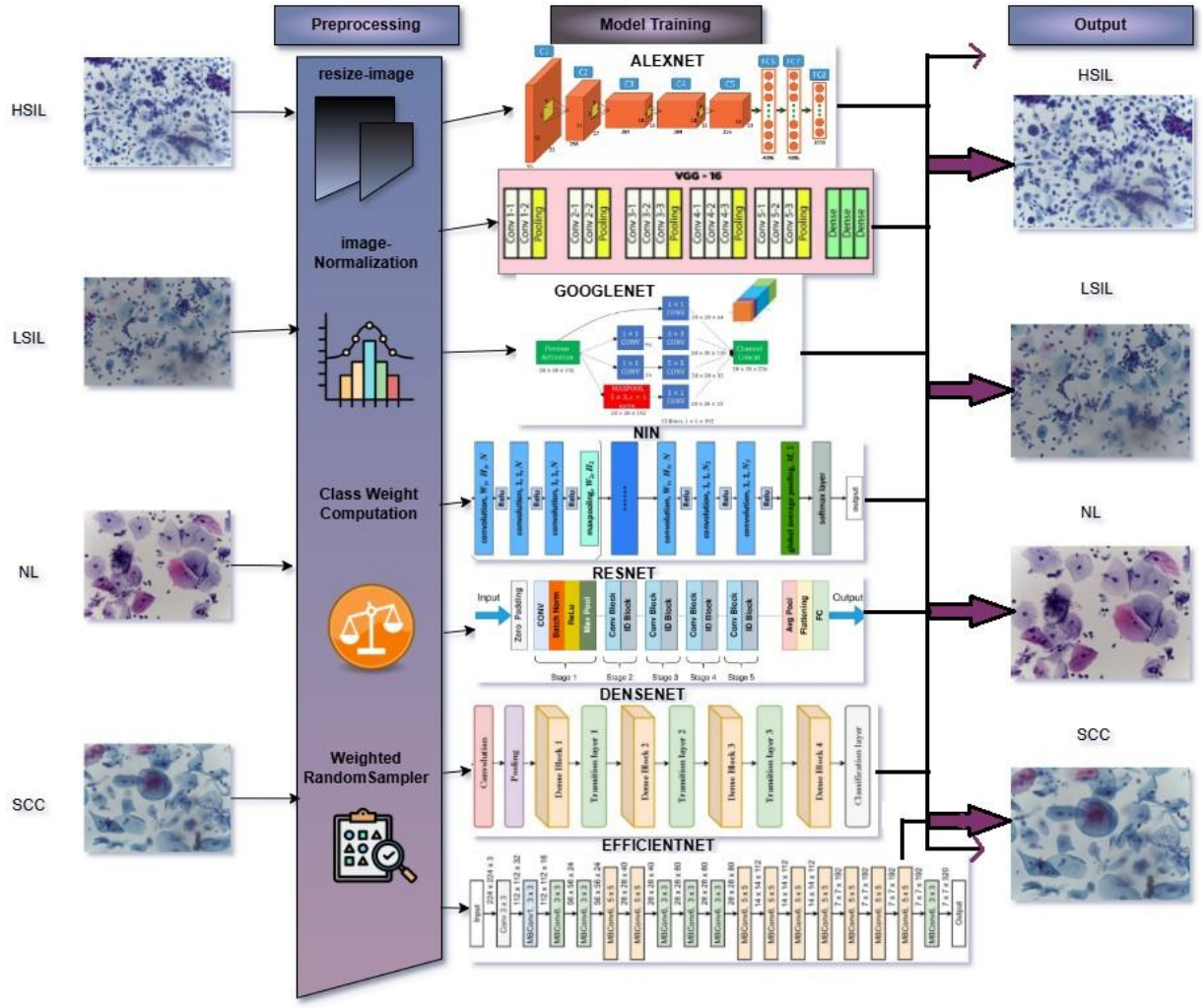


Fig. 2. Overview of CNN architectures used in this study: (A) AlexNet [24], (B) VGGNet [30], (C) GoogLeNet [26], (D) Network-in-Network (NIN) [29], (E) ResNet [25], and (F) DenseNet [27]. These architectures were evaluated to classify cervical cytology images into four diagnostic categories.

V. Evaluation Method

In this section, we present and analyze the classification performance of eight CNN architectures trained without fine-tuning over 10 epochs using a batch size of 32. Each model is evaluated on the same cervical cytology dataset using multiple performance metrics, including accuracy, precision, recall, F1-score, and specificity across four diagnostic categories: HSIL, LSIL, NMIL, and SCC. The dataset is split into training, validation, and test sets, with test results forming the basis of this evaluation to ensure consistent comparison across models.

A. Model-Wise Performance Analysis

The comparative performance analysis of the models shows varying levels of effectiveness. AlexNet achieved a final test accuracy of 66.49% with a specificity of 0.9018, performing relatively well for HSIL recall (0.94) but showing low effectiveness for SCC (recall: 0.07). ResNet-50 improved the results with an accuracy of 81.96% and specificity of 0.9499, offering strong recall for SCC (0.93) along with balanced performance across the other classes. GoogLeNet reported 81.44% accuracy, providing high precision and recall for LSIL and NMIL but moderate performance for SCC. DenseNet-121 reached 93.30% accuracy and 0.9748 specificity, performing consistently across all categories, including the minority class SCC. EfficientNet-B0 achieved 94.33% accuracy with a specificity of 0.9832, maintaining stable precision and recall across diagnostic groups. Network-in-Network (NIN) showed relatively lower performance, with 69.59% accuracy and 0.8237 specificity, and had difficulties with LSIL and SCC. VGG-16 performed with 90.72% accuracy and 0.9725 specificity, showing reliable results for HSIL, LSIL, and NMIL but lower recall for SCC. Finally, Vision Transformer (ViT16) obtained the highest accuracy of 95.88% and specificity of 0.9864, with improved recall for SCC and consistent overall classification. Fig. 4 and 5 show confusion matrices and ROC curves for CNN models. Table II present Comparative performance of pretrained deep learning models on the cervical cytology dataset. All models were evaluated under identical training and preprocessing conditions. Table III presents per-class Precision (P), Recall (R), and F1-score for all evaluated pretrained models on the cervical cytology dataset. Fig. 6 shows ROC curves of the ViT-16 model for four-class cervical cytology.

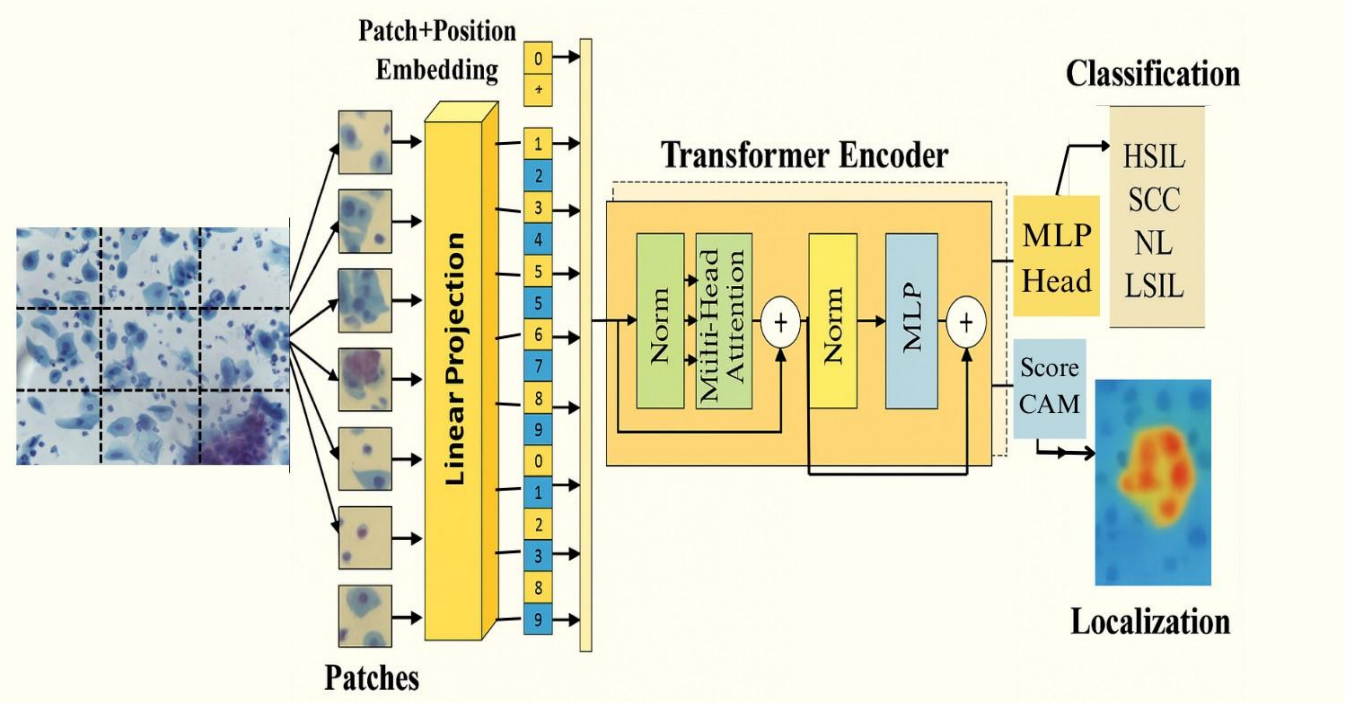


Fig. 3. Architecture of the Vision Transformer (ViT). The image is divided into fixed-size patches, which are embedded and processed through multiple transformer encoder blocks with multi-head self-attention before final classification.

TABLE II. COMPARATIVE PERFORMANCE OF PRETRAINED DEEP LEARNING MODELS ON THE CERVICAL CYTOLOGY DATASET. ALL MODELS WERE EVALUATED UNDER IDENTICAL TRAINING AND PREPROCESSING CONDITIONS

Model	Accuracy (%)	Macro F1	SCC Recall	Overall Specificity	Best For
AlexNet	66.49	0.57	0.07	0.9018	HSIL Recall
ResNet-50	81.96	0.70	0.93	0.9499	SCC Detection
GoogLeNet	81.44	0.76	0.53	0.9463	LSIL Classification
DenseNet-121	93.30	0.86	0.73	0.9748	Balanced Performance
EfficientNet-B0	**94.33**	0.84	0.40	0.9832	High Overall Accuracy
Network-in-Network (NiN)	69.59	0.35	0.00	0.8237	HSIL/NILM Detection
VGG-16	90.72	0.75	0.13	0.9725	LSIL/NILM Classification
ViT-B/16	**95.88**	**0.89**	**0.87**	**0.9864**	Overall Best Performance

TABLE III. PER-CLASS PRECISION (P), RECALL (R), AND F1-SCORE FOR ALL EVALUATED PRETRAINED MODELS ON THE CERVICAL CYTOLOGY DATASET

Model		HSIL			LSIL			NILM			SCC	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AlexNet	0.52	0.94	0.67	0.68	1.00	0.81	0.99	0.60	0.75	0.04	0.07	0.05
ResNet-50	0.73	0.33	0.46	0.72	1.00	0.84	1.00	0.90	0.95	0.39	0.93	0.55
GoogLeNet	0.58	0.91	0.71	1.00	1.00	1.00	1.00	0.79	0.88	0.36	0.53	0.43
DenseNet121	0.87	0.82	0.84	1.00	0.87	0.93	0.98	1.00	0.99	0.65	0.73	0.69
EfficientNetB0	- 0.78	0.94	0.85	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.40	0.52
Network-in-Network (NIN)	0.54	0.58	0.56	0.00	0.00	0.00	0.73	0.94	0.82	0.00	0.00	0.00
VGG-16	0.69	1.00	0.81	0.96	1.00	0.98	1.00	0.96	0.98	0.50	0.13	0.21
ViT-B/16	**0.93**	**0.85**	**0.89**	**1.00**	0.96	**0.98**	0.99	**1.00**	**1.00**	**0.72**	**0.87**	**0.79**

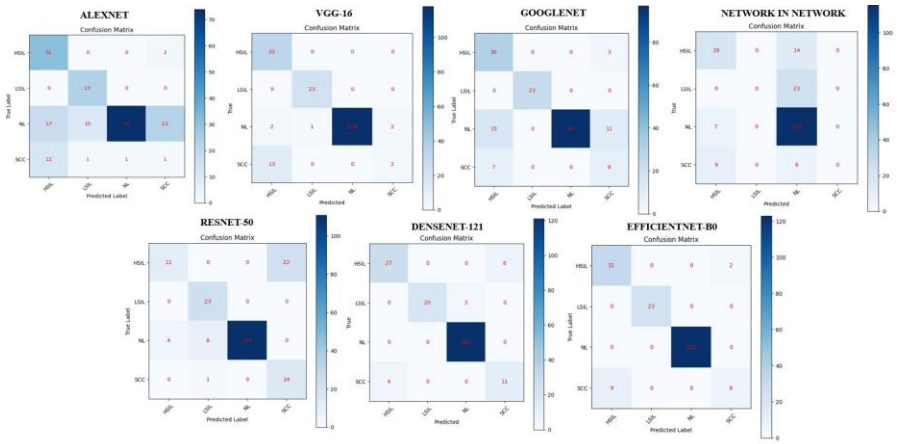


Fig. 4. Confusion matrices for CNN models including DenseNet-121, NIN, AlexNet, GoogLeNet, ResNet-50, VGG-16, and EfficientNet-B0 on the cervical cytology dataset.

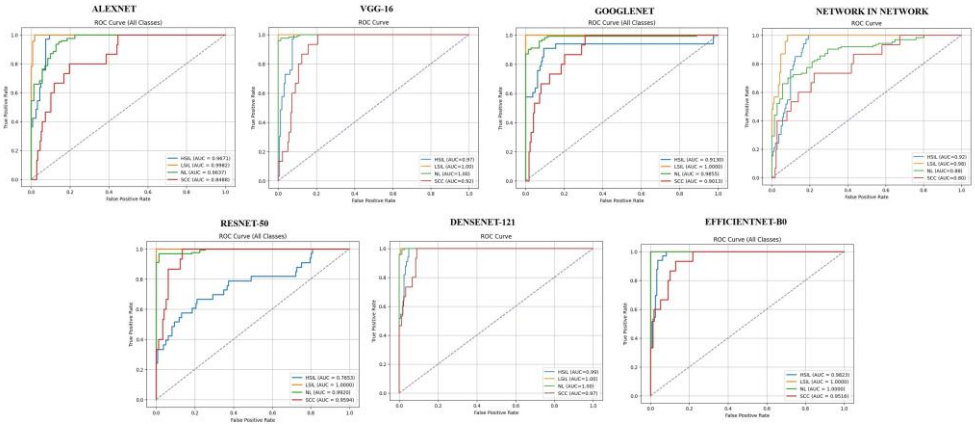


Fig. 5. ROC curves for CNN models including AlexNet, VGG-16, GoogLeNet, NIN, ResNet-50, DenseNet-121, and EfficientNet-B0 on the cervical cancer dataset.

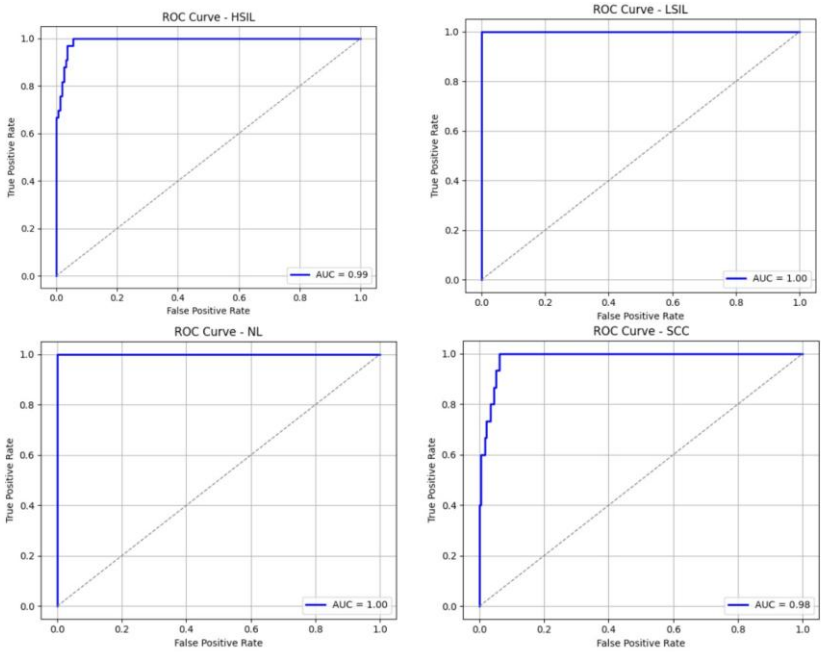


Fig. 6. ROC curves of the ViT-16 model for four-class cervical cytology classification: HSIL, LSIL, NILM, and SCC. The model achieved high AUC values of 0.99, 1.00, 1.00, and 0.98, respectively.

B. Best Model Overview

Among all models evaluated, EfficientNet-B0 and Vision Transformer (ViT-16) demonstrated the highest classification performance. EfficientNet-B0 achieved a final test accuracy of 94.33% with an overall specificity of 0.9832. ViT-16 further surpassed these results with a test accuracy of 95.88% and specificity of 0.9864. Table IV presents ViT-16's class-wise metrics.

TABLE IV. PERFORMANCE METRICS BY CLASS FOR VIT-16

Class	Precision	Recall	F1-score	Specificity
HSIL	0.93	0.85	0.89	0.9876
LSIL	1.00	0.96	0.98	1.0000
NMIL	0.99	1.00	1.00	0.9859
SCC	0.72	0.87	0.79	0.9721

VI. DISCUSSION

The in-depth study here provides information about the experimental results, connections between models, dataset characteristics, and challenges faced in evaluating results. This section goes into deeper detail about the important findings from the comparative study of CNN and transformer-based models on the classification of cervix cytology. The cervix cytology dataset used in this study consists of four cervix cytology classifications: HSIL high-grade squamous intraepithelial lesion, LSIL low-grade squamous intraepithelial lesion NILM (Negative for intraepithelial lesion or malignancy), and SCC (Squamous cell carcinoma). Each classification is different in how many total images are available for training, as well as how complex the images are visually. The cervix cytology dataset also presented several challenges, such as imbalanced class distributions, variability between images, inconsistent staining, and subtle visual differences between the HSIL and LSIL classes. The SCC class had the lowest number of training sample images when compared to the other three cervix cytology classifications, and therefore, many of the models had difficulty generalizing well to the SCC. With only 963 total images in this study, transfer learning with pre-trained models was used to reduce the chance of overfitting and to increase generalization ability for features. Weighted class sampling was also used during the training phase along with applying class-balanced loss to address issues related to imbalanced class distributions. The experiments also show that EfficientNet-B0 and ViT-B16 performed well enough in terms of both recall and specificity on the SCC class while demonstrating a reasonable amount of resilience against the data scarcity issue. The preprocessing steps like cropping, normalizing, and augmenting datasets are critical to ensure that all models are trained on the same, equalized, and enhanced dataset. Additionally, using a WeightedRandomSampler helped balance the number of samples from minority classes by creating balanced mini-batches for training and increasing SCC detection accuracy. These results provide evidence that using class balancing approaches improves learning stability and sensitivity for minority-class samples. The hyperparameter settings used to fine-tune the models are identical to one another and included, using pretrained ImageNet weights, the same preprocessing methods for all models, a batch size of 32, Adam optimizer (learning rate 0.0001), weighted cross-entropy loss, and a maximum number of 10 epochs of training. Using standardized hyperparameter settings provided a consistent basis for an unbiased comparison of all models. Initial experiments showed that validation accuracy converged around epochs 8-10 for the majority of models, and further training yielded little or no improvement in the model's classification accuracy, with an increased probability of overfitting.

All models differ in their respective performance based on the same conditions of training, with substantially different performances. The least productive performers in regard to both recall and specificity (regardless of which dataset) are the shallow networks, including both AlexNet and Network-in-Network (NIN) networks, with especially poor performance seen with respect to the SCC class. One potential reason for their limitations includes fewer feature extractions and less ability to capture a complex spatial representation. Conversely, as the models become deeper and more connected (DenseNet-121 and EfficientNet-B0), both the amount of dense feature reuse and compound scaling contributed to increased performance through 1) dramatically improving feature representation and 2) increasing classification consistency among all diagnostic categories. However, as the models become deeper, they also require greater computational complexity and money to train, which could limit their ability to deploy in real-time clinical environments that are limited in available resources. Among the models analysed, the transformer-based ViT-B16 had the highest overall classification score out of all models on all datasets. The self-attention mechanism of this model provided the capability to effectively model long-range contextual dependencies and global structural information, allowing the model to be more effective at distinguishing visually similar lesion classes. Additionally, the strong performance displayed by ViT-B16 when trained with a very limited amount of training data is likely due to factors such as transfer learning, standardisation of preprocessing, balanced sampling, and the representational power of the attention mechanism of transformers. These findings indicate the emerging promise of transformer-based methods in the medical imaging field. Multiple metrics were assessed, including accuracy, precision, recall, F1 score, specificity, and ROC-AUC analysis, and while overall accuracy provides a high-level representation of how well a model performs, it can sometimes lead to inaccurate conclusions about models trained on imbalanced medical datasets. For this reason, class-level metrics were used to measure the ability of the models to detect the minority class. The importance of specificity was emphasized as it measures the model's ability to predict negative results while attempting to detect positive results in clinical screening situations where false positives present an additional complication to the costly and sometimes dangerous process of diagnosing patients. EfficientNet-B0 and ViT-B16 were able to generate consistently good specificity estimates across all classes, including SCC, indicating excellent diagnostic utility. Both macro and weighted F1 scores supported that the models were generating consistent predictions across majority and minority classes. EfficientNet-B0 and ViT-B16 demonstrated superior performance on all of these metrics, which would indicate they would be able to generalise better than the other models considered and have balanced learning behaviour. The evaluation of ROC-AUC

analysis provided additional support for the robust discriminatory ability of the models across all eleven diagnostic classifications.

Architectural depth, dense connectivity, transfer learning and attention mechanisms are important in enhancing performance for the classification of cervical cytology specimens. Although traditional convolutional neural networks (CNN) architectures are still valid benchmark models, ViT (Vision Transformer) architectures such as ViT-B16 have demonstrated greater accuracy for classifying complex feature relationships as well as identifying minority classes. Specifically, for the squamous cell carcinoma (SCC) class (which was the least represented in the dataset), ViT-B16 achieved a recall of 87% and a precision of 72%. Despite promising results based on the proposed comparative framework, there are still a number of limitations. The comparative analysis was conducted using one publicly available dataset with no external cross-dataset validation, therefore limiting its generalizability to other clinical settings. Furthermore, the impact of staining artefacts (severe artefacts) as well as noise variation and imaging artefacts due to differences in laboratory settings were not systematically assessed. Therefore, there will be future studies that will explore cross-dataset validation, larger multi-centre datasets, analysis of computational efficiency, and hybrid CNN/transformers to further enhance robustness and clinical applicability in real-world scenarios.

VII. Conclusion

This study compared eight convolutional neural networks and a vision transformer model on a cervical cytology image dataset. The Vision Transformer (ViT-16) obtained the highest classification performance, while EfficientNet-B0 and DenseNet-121 also showed relatively strong results, including on minority classes such as SCC. Older models like AlexNet and NIN had more difficulty with class imbalance and deeper feature learning.

Overall, the findings suggest that models incorporating attention mechanisms or compound scaling provide advantages for medical image classification under imbalanced conditions. Future work may involve fine-tuning transformer models, exploring ensemble approaches, or integrating domain knowledge to improve classification reliability. For future directions, cross-dataset validation and the development of larger, multi-centre datasets will be pursued

A. Limitations

This study has several constraints. All models were trained from scratch without pretrained weights, which may affect generalization. The dataset was limited in size and imbalanced, especially for SCC, which may have influenced performance. In addition, results were obtained from a single dataset, and external validation would be useful for assessing robustness. Finally, computational efficiency was not the main focus, which could affect practical deployment in clinical environments.

Authors' Contribution

Dr. Mehreen Sirshar and Dr. Fakeeha Jaferi have supervised this work. Ibrahim Aljubayri and Abdulrahman Alahmadi prepared the figures. Omama Shakeel and Nayab Asim prepared the manuscript. Muhammad Zubair Khan reviewed the manuscript.

Non-financial Competiting interest. No, I declare that the authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this study.

Funding Declaration: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

This research is funded by the Deanship of Scientific Research, Islamic University of Madinah, Madinah, Saudi Arabia.

References

[1] Delgado Soler, D. (2022). Cervical cancer: a systematic review.

[2] G. Nirmala, P. P. Nayudu, A. R. Kumar, R. Sagar, "Automatic cervical cancer classification using adaptive vision transformer encoder with CNN for medical application," Pattern Recognition, vol. 160, p. 111201, 2025.

[3] P. Huang, X. Tan, C. Chen, X. Lv, Y. Li, "AF-SENet: Classification of cancer in cervical tissue pathological images based on fusing deep convolution features," Sensors, vol. 21, no. 1, p. 122, 2020.

[4] W. Liu, C. Li, N. Xu, T. Jiang, M. M. Rahaman, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, et al., "CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, Visual Transformer and Multilayer Perceptron," Pattern Recognition, vol. 130, p. 108829, 2022.

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

[6] Y. Karasu Benyes, E. C. Welch, A. Singhal, J. Ou, A. Tripathi, "A comparative analysis of deep learning models for automated cross-preparation diagnosis of multi-cell liquid Pap smear images," Diagnostics, vol. 12, no. 8, p. 1838, 2022.

[7] V. Sellamuthu Palanisamy, R. K. Athiappan, T. Nagalingam, "Pap smear based cervical cancer detection using residual neural networks deep learning architecture," Concurrency and Computation: Practice and Experience, vol. 34, no. 4, p. e6608, 2022.

[8] S. L. Tan, G. Selvachandran, W. Ding, R. Paramesran, 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.

[9] D. C. R. Novitasari, P. Wulandari, D. Z. Haq, "Cervical cancer diagnosis system using convolutional neural network ResidualNet," International Journal of Computing, vol. 21, no. 1, pp. 61–68, 2022.

[10] S. Cheng, S. Liu, J. Yu, G. Rao, Y. Xiao, W. Han, W. Zhu, X. Lv, N. Li, J. Cai, et al., "Robust whole slide image analysis for cervical cancer screening using deep learning," Nature Communications, vol. 12, no. 1, p. 5639, 2021.

[11] H. Kaur, R. Sharma, J. Kaur, "Comparison of deep transfer learning models for classification of cervical cancer from Pap smear images," Scientific Reports, vol. 15, no. 1, p. 3945, 2025.

[12] S. Alsubai, A. Alqahtani, M. Sha, A. Almadhor, S. Abbas, H. Mughal, M. Gregus, "Privacy preserved cervical cancer detection using convolutional neural networks applied to Pap smear images," Computational and Mathematical Methods in Medicine, vol. 2023, no. 1, p. 9676206, 2023.

[13] F. Paucar, C. Bojorque, I. Reyes-Chac´on, P. Vizcaino-Imacan˜a, M. MorochoCayamcela, "Towards accurate cervical cancer detection: Leveraging two-stage CNNs for Pap smear analysis," Proc. Int. Conf., pp. 219–227, 2024.

[14] C. Zhao, R. Shuai, L. Ma, W. Liu, M. Wu, "Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT," Multimedia Tools and Applications, vol. 81, no. 17, pp. 24265–24300, 2022.

[15] E. Lima, J. Smith, and R. Kumar, "A novel approach to medical image classification," IEEE Transactions on Medical Imaging, vol. 42, no. 3, pp. 456–465, 2024.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021.

[17] AlexNet, https://cdn.techscience.cn/ueditor/files/iasc/TSPIASC-36-1/TSPIASC30125/T-SPIASC30125.pdf, accessed: 2025-04-14.

[18] VGG, https://media.geeksforgeeks.org/wp-content/uploads/20200219152327/convlayers-vgg16-1024x450.jpg, accessed: 2025-04-14.

[19] GoogLeNet, https://miro.medium.com/v2/resize:fit:4800/format:webp/0*W5DczbPesqqmEWO.jpg, accessed: 2025-04-14.

[20] Network in Network https://www.researchgate.net/figure/fig3/AS:915542533750784159529 3758115

[21] /The-structure-of-the-NIN-model-used-in-our-proposed-method-The-definitionsof-all.png, accessed: 2025-04-14.

[22] ResNet, https://towardsdatascience.com/wp-content/uploads/2022/08/0tH9evuOFqk8F41FG2048x660.png, accessed: 2025-04-14.

[23] DenseNet, https://media.springernature.com/full/springer-static/image/art, accessed: 2025-04-14.

[24] E. Hussain, "Liquid based cytology Pap smear images for multi-class diagnosis of cervical cancer," Mendeley Data, 2019.

[25] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, vol. 25, 2012.

[26] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning," Image Recognition, vol. 7, no. 4, pp. 327–336, 2015.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

[28] G. Huang, Z. Liu, K. Q. Weinberger, "Densely connected convolutional networks," CoRR, abs/1608.06993, 2016. [ONMILine]. Available: http://arxiv.org/abs/1608.06993

[29] M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, PMLR, pp. 6105–6114, 2019.

[30] M. Lin, Q. Chen, S. Yan, "Network in network," arXiv preprint arXiv:1409.1556, 2014.

[31] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[32] Kumar, G. V., Bellary, M. I., Reddy, T. B. (2022). Prostate cancer classification with MRI using Taylor-Bird squirrel optimization based deep recurrent neural network. The Imaging Science Journal, 70(4), 214-227.

[33] Kumar, A., Ahmad, F., Alam, B. (2025). Hybrid bio-inspired computing in medical image data analysis: A review. Intelligent Decision Technologies, 19(1), 473-488.

[34] Bohmrah, M. K., Kaur, H. (2025). Advanced hybridization and optimization of DNNs for medical imaging: A survey on disease detection techniques. Artificial Intelligence Review, 58(4), 122.