MFEM-CIN: A Lightweight Architecture Combining CNN and Transformer for the Classification of Pre-Cancerous Lesions of the Cervix

Peng Chen, *Member, IEEE*, Fobao Liu, Jun Zhang, and Bing Wang, *Senior Member, IEEE*

Abstract—Goal: **Cervical cancer is one of the most common cancers in women worldwide, ranking among the top four. Unfortunately, it is also the fourth leading cause of cancer-related deaths among women, particularly in developing countries where incidence and mortality rates are higher compared to developed nations. Colposcopy can aid in the early detection of cervical lesions, but its effectiveness is limited in areas with limited medical resources and a lack of specialized physicians. Consequently, many cases are diagnosed at later stages, putting patients at significant risk.** ***Methods:*** **This paper proposes an automated colposcopic image analysis framework to address these challenges. The framework aims to reduce the labor costs associated with cervical precancer screening in undeserved regions and assist doctors in diagnosing patients. The core of the framework is the MFEM-CIN hybrid model, which combines Convolutional Neural Networks (CNN) and Transformer to aggregate the correlation between local and global features. This combined analysis of local and global information is scientifically useful in clinical diagnosis. In the model, MSFE and MSFF are utilized to extract and fuse multi-scale semantics. This preserves important shallow feature information and allows it to interact with the deep feature, enriching the semantics to some extent.** ***Conclusions:*** **The experimental results demonstrate an accuracy rate of 89.2% in identifying cervical intraepithelial neoplasia while maintaining a lightweight model. This performance exceeds the average accuracy achieved by professional physicians, indicating promising potential for practical application. Utilizing automated colposcopic image analysis and the MFEM-CIN model, this research offers a practical solution to reduce the burden on healthcare providers and improve the efficiency and accuracy of cervical cancer diagnosis in resource-constrained areas.**

Index Terms**—Cervical cancer, cervical intraepithelial neoplasia,deep learning, CNN, transformer.**

Impact Statement**—The proposed lightweight model achieved an impressive accuracy of 89.2% in identifying cervical intraepithelial neoplasia, which is better than the average professional physician, indicating the promising potential for practical application.**

I. Introduction

Cervical cancer is the fourth most common malignancy in women worldwide and is the only malignancy with a known cause in human tumors. However, there has been no significant decline in cervical cancer incidence and mortality. In 2018, there were approximately 570000 new cases of cervical cancer worldwide, accounting for 3.15% of all malignancy incidences, and approximately 310000 deaths, accounting for 3.26% of all malignancy deaths. Cervical cancer poses a significant global burden [1], with a large proportion of incidence and deaths occurring in developing countries. Surveys indicate that approximately 85% of women diagnosed and 87% of women who die from cervical cancer reside in low- and middle-income countries (LMIC) [2]. In these regions, there is a lack of sufficient knowledge about cervical cancer and inadequate medical resources. Many women in less economically developed areas are not regularly screened for precancerous cervical lesions, putting them at high risk. Improved medical equipment and resources can play a crucial role in reducing the incidence and mortality rates of cervical cancer in these areas.

Screening tests for precancerous lesions include HPV DNA testing, cytology (Pap test), and visual cervical screening. Visual

Manuscript received 21 September 2023; revised 3 December 2023 and 6 January 2024; accepted 5 February 2024. Date of publication 20 February 2024; date of current version 27 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072002, Grant 62172004, and Grant 62273001, in part by Anhui Province Collaborative Innovation Project under Grant GXXT-2022-050 and Grant GXXT-2022-053, in part by the Outstanding Research and Innovation Team Project of Anhui Province under Grant 2022AH010005, and in part by the Special Fund for Anhui Agriculture Research System under Grant 2021-2025. The review of this article was arranged by Editor Amir Amini. *(Corresponding authors: Jun Zhang; Bing Wang.)*

Peng Chen is with the National Engineering Research Center for Agro-Ecological Big Data Analysis and Application, Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology and School of Internet, Anhui University, Hefei 230601, China, and also with the Fin China-Anhui University Joint Laboratory for Financial Big Data Research, Hefei Financial China Information and Technology Company, Ltd., Hefei 230022, China (e-mail: pengchen@ustc.edu).

Fobao Liu and Jun Zhang are with the National Engineering Research Center for Agro-Ecological Big Data Analysis and Application, Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology and School of Internet, Anhui University, Hefei 230601, China (e-mail: junzhang@ahu.edu.cn).

Bing Wang is with the School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China (e-mail: wangbing@ustc.edu).

Digital Object Identifier 10.1109/OJEMB.2024.3367243

cervical screening is a cost-effective alternative to cytology, which requires complex training [3]. However, HPV testing is more commonly used in clinical practice for primary cervical cancer screening [4]. Therefore, visual cervical screening is indicated in less economically developed areas (LMIC). Visual cervical screening is a method of examining the cervix using acetic acid (VIA) and Lugol's iodine (VILI) to identify precancerous lesions and cervical intraepithelial neoplasia (CIN). During the examination, the colposcopist applies 3–5% VIA, which turns precancerous lesions white (acetowhite) and cervical intraepithelial nodules visible. Furthermore, CIN is a crucial screening method for detecting abnormal cell growth on the surface of the cervix. According to the World Health Organization, CIN is classified into three categories: CIN1, CIN2, and CIN3. Clinically, CIN is divided into two main categories: Low-Grade Squamous Intraepithelial Lesion (LSIL) which corresponds to CIN1, and High-Grade Squamous Intraepithelial Lesion (HSIL) which corresponds to CIN2 and CIN3. CIN1 usually resolves spontaneously with a 60% chance and can be treated conservatively by observation and regular diagnosis. However, HSIL (CIN2 and CIN3) can directly develop into invasive cancer, requiring immediate surgical treatment [5]. There is a shortage of experienced colposcopists in economically disadvantaged areas.

In recent years, deep learning techniques have been applied in the field of complementary medicine, including cervical cancer detection. However, many proposed CNN-based algorithms have obvious drawbacks. This is because the physiological image obtained from colposcopic images cannot easily distinguish between cervical epithelial neoplasia and other types of abnormal growths, such as celiac disease, due to the variable physiological condition of the cervix. The convolution operation in CNN captures only local information, whereas the Vision Transformer (ViT) extracts global information of the image through the multi-headed self-attentive mechanism (MHSA) and has a global perception field. This property allows the model to focus more on the features of cervical epithelial lesions, compensating for the missing property of CNNs. ViT has performed well in various vision tasks, including image classification [5], semantic segmentation [6], and video understanding [7], [8]. Some studies have suggested that ViT's prediction error is closer to human prediction error than that of CNN [9]. These features have made ViT popular in medical imaging and physiological image analysis [10].

However, using the Transformer model on the ground remains challenging due to two main reasons. Firstly, the number of parameters is significantly larger compared to CNN [11]. Secondly, the Transformer lacks spatial induction bias, and spatial information is crucial in image data. VIT researchers introduced absolute position bias to the model, while Swin Transformer researchers added relative position bias. These measures can partially solve the problem of spatial information loss, but they require adjustment when applied to other tasks. Specifically, the relative position encoding may be adjusted as a trainable parameter during training to adapt to the position offset requirements of the new task.

To solve the practical problem, we propose a lightweight structure called MFEM-CIN, which stands for multiscale feature extraction module for CIN. It is based on a mixture of CNN and Transformer. CNN reduces the number of parameters for the model and eliminates position bias. Additionally, the inclusion of CNN can expedite network convergence, resulting in a more stable training process. The Transformer architecture replaces traditional convolution with a global sensing field, enhancing the attention given to key features and improving the model's classification performance. MEFM-CIN is designed to consider both local and global features in practical diagnosis. This includes vascularity, texture, boundary definition, position, and relative size. The aim is to prevent the loss of low-level semantics by paying attention to different scales of features. The model architecture employs the MSFE (multiscale feature extraction module) to extract features at different scales, and the MSFF to fuse the semantics at different scales. This paper is the first to apply Transformer to cervical precancer detection.

The following section outlines previous research on the application of machine learning and deep learning to precancerous lesion detection. The third section introduces the dataset and data preprocessing methods. The fourth section presents the deep learning tools used. The fifth section analyses the experimental results. Finally, the paper concludes by discussing the limitations of the current study and future developments.

II. Related Work

The field of medical diagnosis has seen considerable results and greater impact due to the rapid development of artificial intelligence technology in recent years. Techniques based on deep learning and machine learning have been increasingly applied. Scholars have also made some explorations in the field of automated cervical cancer precancer detection.

Xu et al. [12] trained CNNs on almost 1000 patient images, each with a self-reported cervical cytology and HPV test result. The final model achieved 88.91% accuracy in identifying LSIL+. Ma et al. [13] designed a network architecture based on a feature pyramid network (FPN) with a lightweight booster (CCDB), which was trained on 4107 cervical smears and finally achieved a specificity of 95.14% for detecting abnormal squamous cells. Zhang et al. [14] employed convolutional neural networks (ConvNets) to distinguish abnormally growing cervical cells. The model was first pre-trained on a natural image dataset and then subsequently fine-tuned on a cervical cell dataset. Finally, the trained model was evaluated on a Pap smear and LBC dataset, achieving a classification accuracy of 98.3%. Li et al. [15] proposed a deep learning model based on RestNet-101 with E-GCN, trained on a dataset of 7668 delayed colposcopies, and finally this model achieved an accuracy of 78.33% for the detection of LSIL+. The model achieved an accuracy of 78.33% for the detection of LSIL+, surpassing that of a group of specialists who conducted a competition to identify LSIL+ under the same conditions.

Buiu et al. [16] proposed MobileNetV2 using multiple colposcopic images of the same patient as input. These images consist of five consecutive images of acetic acid action, one image through a green lens, and one image after the action of iodine solution. The final model achieved a recognition accuracy of 83.33% for multivariate classification on a dataset of over 3000 images. Although the model approach demonstrated good performance, there are concerns regarding its clinical application due to the need to input too many images. In poorer areas, it may be unlikely to obtain colposcopic images in three different states required for the model. Additionally, the diagnostic process may become too burdensome, which should be avoided. Saini et al. [17] developed the ColpoNet network with a DenseNet backbone and tested it on 800 images, achieving an accuracy of 81.35% for the binary classification task. The dataset used in their experiments was limited in its robustness and did not include special cases that may be encountered in a clinical setting.

Xue et al. [18] aimed to remove specular reflections, which often appear on the surface of vaginal mirrors, and identify acetyl white in the ROI region. The team used two main steps to achieve this. The first step involved identifying specular reflective regions, which were determined by the presence of high bright spots in pixels with high luminance (I) and low colour saturation (S) values near high gradients. The second step is to fill these regions. The paper describes two methods of filling: average color fill and weighted color fill. Similar to Xue et al. [18], Yue's team [19] also made significant contributions to the study of specular reflections in vaginoscopic images. The cervigram was converted to HIS colour space and thresholds were set on the S (saturation) and I (intensity) channels to obtain ROIs. These were then filled by searching for patches with the greatest similarity to the pixels around the ROIs.

In [20], the authors trained a convolutional neural network model on a colposcopic dataset of 330 patients. The final model identified HISL+ with an accuracy of 94.1%, compared to 84.3% for competing gynaecologic oncologists in the same condition. In this experiment, the automated screening method outperformed the manual method. There are few publicly available datasets of reliable quality. Li et al. [21] collected colposcopic images from 8604 patients to construct a dataset of colposcopic images of cervical intraepithelial malignancies with fine-grained lesions. The dataset was labelled based on the anterior texture of the supraacetabular white skin and the appearance of blood vessels. It is anticipated that this dataset will be accessible to additional researchers in the future, thereby contributing to more automated diagnostic studies.

Models that combine CNNs and Transformers are becoming increasingly popular due to their ability to capture both local and global information. This combination enhances the model's ability to capture contextual information in the image, ultimately leading to improved performance. The choice of which model to use depends on the specific needs and task at hand. Chen et al. [22] introduced SleepZzNet, a model that combines CNN and Transformer architectures to classify EEG sleep stages. Similarly, Hong et al. [23] used a transformer-CNN network with a dual coding system to segment multi-organ CT maps among different organs and achieved superior results.

TABLE I
DISTRIBUTION OF DATA IN THE TRAINING AND VALIDATION SETS

	LSIL	HSIL	Total
Training	2017	1248	3265
Validation	504	312	816

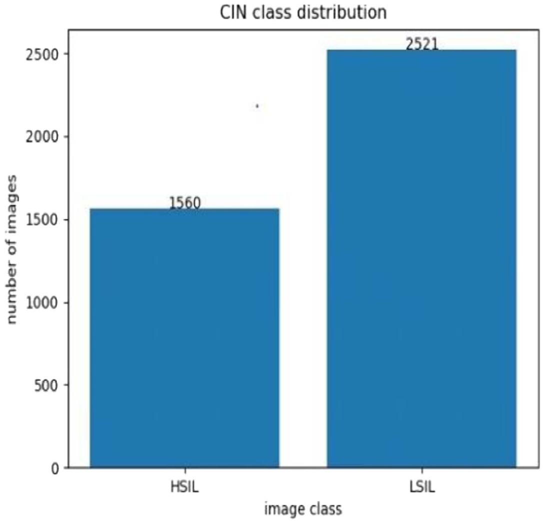

Fig. 1. Classification distribution of the CIN data set.

III. DATASET AND PREPROCESSING

A. Experimental Dataset

The dataset used in our experiment comprises colposcopic images collected from Wanan Medical College. The dataset includes images of over 4000 patients, each taken after a VIA procedure performed by a specialized colposcopist. Clinically relevant conditions such as bleeding, polyps, erosions, displaced cervix, and foreign body obscuring the lens are included. The dataset for CIN grade was graded by several professional and experienced gynecologists. The images are all uniformly sized at 960 x 640. Normal specimens are described as CIN0.

In our experiments, we categorized CIN0 and CIN1 as LSIL and CIN2 and CIN3 as HSIL. The training and validation sets were randomly divided in an 8:2 ratio, as shown in Table I. Fig. 1 displays the corresponding histogram distribution.

The CIN dataset includes a small proportion of LSIL and HSIL class samples, as determined by the following formula: where num_{max} is the number of samples from the largest class, and num_{min} is the number of samples from the smallest class. Therefore, the effect of sample imbalance does not need to be considered at this time.

$$\frac{num_{max}}{num_{min}} \leq 10 \tag{1}$$

B. Experimental Environment

Our experimental environment consists of an Intel Gold 512 CPU, two Nvidia Geforce GTX 1080Ti GPUs (8GB each), and

TABLE II
DATA PRE-PROCESSING

Preprocessing
Image enhancement
Specular reflection and inpaint
Resize：224×224pixels
Horizontal flip
ToTensor
Normalization：mean = [0.485, 0.456, 0.406] std = [0.229, 0.224, 0.225]

we use the open-source deep learning framework Pytorch on Ubuntu 16.04 LTS.

C. Preprocessing

Prior to the official experiment, we dedicated significant effort to preprocessing the images. The methods employed are listed in Table II, and are explained in detail below:

1) Histogram Equalization: The complexity of the content in the colposcopic images was due to the physiological state of the patient. Additionally, some samples did not reflect distinctive lesion features after the action of acetyl. Therefore, we performed image enhancement to improve the details for all samples. This improved the classification accuracy of the model and sped up the convergence rate. Contrast enhancement is used to achieve detail enhancement. This is because an image with too much concentration of grayscale in the grayscale histogram will look blurred and have too little detail. Histogram equalization can make the distribution of grayscale more even [24]. However, using the normal histogram equalization method to enhance contrast can cause problems such as areas becoming brighter or darker and losing detail, which should be avoided [25]. The contrast enhancement was limited to histogram equalization. The process involved mapping the image into LAB space and running the CLAHE algorithm on each of the three channels L, A, and B. The resulting three channels were merged and then mapped back into RGB space.

$$image_{RGB} \xrightarrow{trannsfer} image_{LAB} \tag{2}$$

$$C_{L,}C_{A},C_{B} = F_{spilt}\left(image_{LAB}\right) \tag{3}$$

$$C'_{L,}C'_{A,}C'_{B} = F_{CLAHE}\left(C_{L,}C_{A},C_{B}\right) \tag{4}$$

$$image = F_{merge}\left(C'_{L,}C'_{A,}C'_{B}\right) \tag{5}$$

$$image_{LAB} \xrightarrow{trannsfer} image_{RGB} \tag{6}$$

where $image_{RGB}$ denotes the colposcopy image in RGB representation, $image_{LAB}$ denotes the colposcopy image under the LAB presentation, $C_{L,}$ C_{A}, C_{B} denote L, A, and B channels, respectively, $C'_{L,}$ $C'_{A,}$ C'_{B} denote L, A and B channels after Clahe enhancement. Moreover, F_{spilt} splits an image represented by multi-channel into multiple single channels, F_{CLAHE} denotes enhanced images by Clahe algorithm, and F_{merge} merges multi-channels.

The image is enhanced in detail, as shown in Fig. 2. The white areas are believed to be the effect of acetyl on the lesion. This step's significance will be discussed in detail in the following experimental discussion.

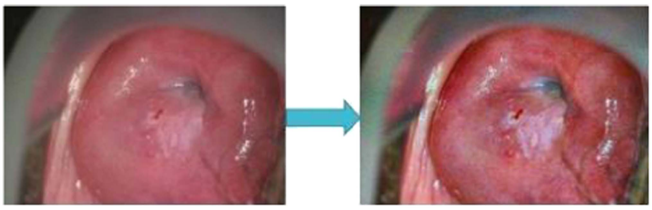

Fig. 2. Colposcopic image enhancement.

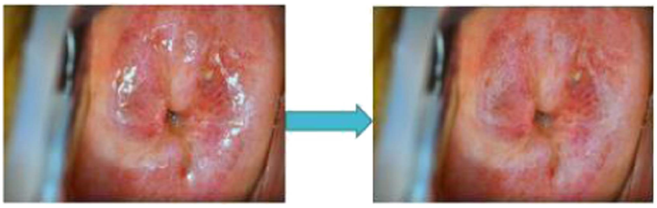

Fig. 3. Mirror reflection removal effect.

2) Mirror Reflection Removal: To improve the accuracy of classification, we removed and filled in the specular reflections from the endoscopic images in the dataset. This was done by setting a suitable threshold to detect the specular reflection areas, based on the fact that high bright spots tend to appear at low saturation and high intensity [26]. Secondly, based on the size of the identified areas, we fill them one by one using an algorithm that searches for the most similar patches within the same image sample. The resulting rendering can be seen in Fig. 3

3) Image Enhancement: Additionally, to avoid overfitting, we expanded the dataset to include random horizontal flipping and cropping of images.

IV. EXPERIMENTAL METHOD

A. Experimental Model

1) General Overview of Model Architecture: The following section describes the architecture of MFEM-CIN.

Initially, the model convolves the image input to extract low-level features and downsamples it to remove redundant data. This module comprises a convolution kernel size of 3, a step size of 2, batch normalization (BN), and a SiLu activation function. The second part of the MobileNetV2 network consists of the inverse residual structure. Within this structure, the data is first down-sampled by a 1x1 convolution kernel, followed by a deep separable convolution with a 3x3 convolution kernel, and finally by a 1x1 convolution up-sampling. The output data is then concatenated with the input of this structure, as shown in [27]. The data is then down-sampled by the MV structure in layer 2 and processed by the multi-scale feature extraction (MSFE) component.

The MSFE component has two parallel branches, each with nodes that process data using MV2 and MFViT blocks with different scale features from each layer. The multi-scale feature fusion (MSFF) module collects and merges features from different scales in the previous module. Finally, the high-level semantics are extracted, dimensionally adjusted, globally pooled, and flattened. They are then fed into the fully connected layer for decision-making to obtain the predicted CIN risk level.

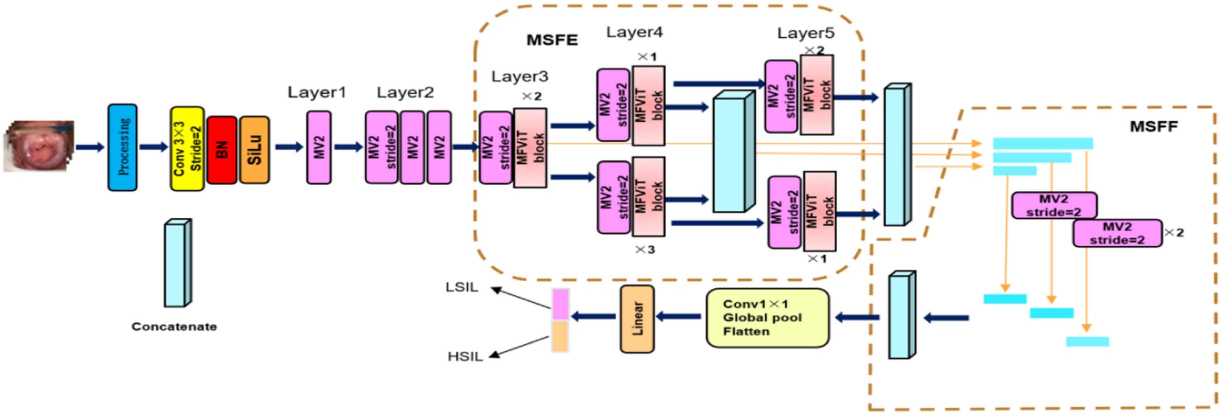

Fig. 4. Flowchart of the model structure.

TABLE III
INFORMATION RELATING TO THE STRUCTURE OF THE MODEL

Flatten	One-dimensionalization of multi-dimensional data, and horizontal stitching in the direction.
Filter size	Module 1 uses convolution kernels of size 3x3, while the remaining modules use an MV2 structure consisting of a 1x1 convolution kernel to downscale the data, a 3x3 kernel to extract features, and a 1x1 kernel to upscale and concatenate with the original input. The convolution structure in the Transformer block is similar.
Downsampling	The aim of this process is to filter sensory features and select representative ones, reducing the number of parameters required for the model.
Batch Normalization	The use of Batch Normalization (BN) has a remarkable effect on the process. Firstly, it stabilizes the input data at each layer, leading to faster convergence. Secondly, it mitigates the vanishing gradient problem, allowing the use of saturated activation functions. Finally, it prevents overfitting by making all the samples in the batch related, ensuring the network does not learn from one sample alone.
Dropout	The model sets the parameter p in the dropout to 0.1. This causes the neural network to stop propagating with a 0.1 probability of neural units during the learning process. This approach has the advantage of reducing the network's dependence on any particular feature, which can prevent overfitting.

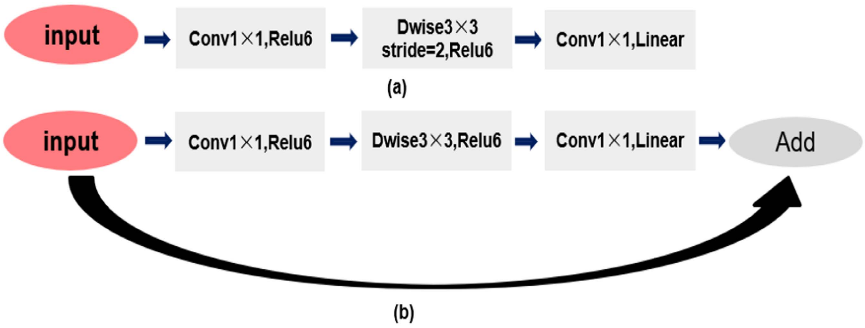

Fig. 5. MV2 structure. (a) The MV2 structure with stride = 2. (b) The MV2 structure with stride = 1.

The model is a hybrid architecture of CNN and Transformer, designed to extract multi-scale high-level semantics, which are then fused and fed into the fully connected layer for classification. Fig. 4 illustrates the model's structure. Table III explains some of the basic operations in the model.

2) Multi-Scale Feature Extraction Module: The paper presents a MSFE module that is divided into two branches. The features extracted from different branches but the same layer are fused and interacted. This fusion is necessary because the feature extraction components on each node have differences, which can provide different information. The feature extraction components used are MV2 and MFViT Block. The MV2 structure comprises of a 1x1 convolution and DW convolution, which reduces computational intensity while retaining high-dimensional information with a low loss rate.

Another MV2 structure can be found in Fig. 5, as presented by Sandle et al. The MFViT module was designed to learn both local and global information with an effective perceptual field of $H \times W$.

Fig. 6 displays the fundamental structure of the MFViT module. Initially, the number of channels is established using the standard 3 × 3 and 1 × 1 convolution kernels, and the tensor size is set from $\alpha \in X^{H \times W \times C}$ to $\alpha_l \in X^{H \times W \times d}$. Subsequently, the unfolding module is divided α_l into N non-overlapping

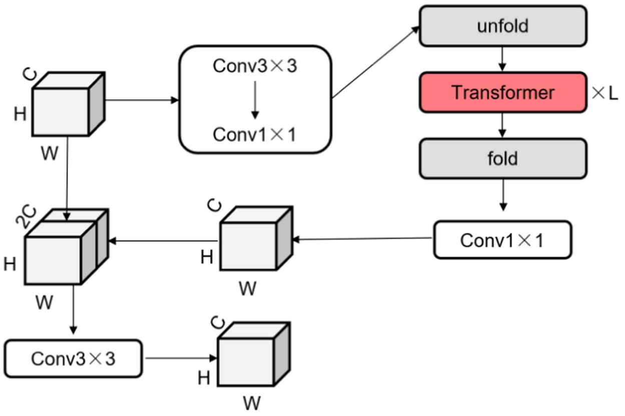

Fig. 6. MFViT structure.

sequences $\alpha_u \in X^{P\times N\times d}$, where $N = HW/wh$,$P = hw$, h and w represent the size of a patch.

Fig. 7 shows that the patch size is 2 × 2 and the input is divided into 9 patches. Each patch in the same color region is stitched together to form a sequence. These sequences are then entered into the Transformer module for global relational modeling. Finally, the different tokens are collapsed into the original feature map, preserving the location information of each region. As demonstrated in (7), each token is self-attended by the token of the same colour, reducing parameter calculation.

$$\alpha_g\left(p\right) = Transformer\left(\alpha_u\left(p\right)\right), 1 \leq p \leq P \tag{7}$$

Following the Transformer computation, $\alpha_g \in X^{P\times N\times d}$ is collapsed into $\alpha_f \in X^{H\times W\times d}$, and the number of channels is adjusted to C by a 1 × 1 convolution kernel. The output $\alpha \in X^{H\times W\times C}$ is obtained by concatenation with the original input $\alpha_f \in X^{H\times W\times C}$.

3) Multi-Scale Feature Fusion Module: The MSFF module is built to collect the multi-scale modules generated by the previous feature extraction module. The MV2 module is used to downsample the features from different scales for better fusion. The logical details within the module are shown in the following equation. The MV2 structure has been introduced in the previous section, and Concat represents the connection of the different features of the input. Here, X_i is the feature of the same scale processed by MV2 structure. X_{sum} is the advanced semantics after fusion of different scales. The expectation is that these high-level semantics will provide richer and more hierarchical information, which is crucial for the model to accurately focus on key features and improve classification accuracy.

$$\begin{aligned} X_i &= MV2\left(x_i\right)\ \ (i \in \{1,2,3\}) \\ X_{sum} &= Concat\left(X_i\right) \end{aligned} \tag{8}$$

B. Model Forward Propagation

This section presents the number of channels and the variation of the data length and width ($C \times H \times W$) of the sample data as they pass through our model are presented to provide a more intuitive view of the forward propagation process.

TABLE IV
FORWARD PROPAGATION PROCESS

Process	Output(C×H×W)
Input	3 x 224 x 224
Conv(3×3)	32 x 112 x 112
Layer1	32 x 112 x 112
Layer2	48 x 56 x 56
Layer3	64 x 28 x 28
Layer4	80 x 14 x 14
Layer5	96 x 7 x 7
Conv(1×1)	384 x 7 x 7
Classifier	2

The data first passes through the 3 × 3 convolution at the beginning of the model, where the stride is set to 2. As a result, the size of the tensor becomes 112 × 112, and the number of channels increases to 32. The data is processed through a 3x3 convolution layer in the MV2 structure of module 1 with padding=(*kernel_size*-1)/2, resulting in a constant H and W. The output after the module is 32 × 112 × 112. The output of this module is 48 × 56 × 56. Module 2 contains three MV2 structures, with one having a stride of 2 while convolving and the other two being the same as the structure in module 1, which does not affect the data size. The three layers in the next branch differ only in the number of transformer encoders used. This module does not affect the size, only the MV2 structure. As a result, the output of the next three modules are 64 × 28 × 28, 80 × 14 × 14 and 96 × 7 × 7. Please refer to Table IV for a summary of this section.

C. Loss Fuction

The paper addresses a binary classification problem and employs the cross-entropy loss function, as shown in (9) below. y_{pre} represents the probability of the network arriving at a positive class (HSIL), andy_{act} representsthe actual label value (0/1).

$$loss = -\left(y_{act}\log\left(y_{pre}\right) + \left(1 - y_{act}\right)\log\left(1 - y_{pre}\right)\right) \tag{9}$$

V. EXPERIMENTAL RESULT AND ANALYSIS

In Section V-A, the metrics used to evaluate the experimental results are presented. Section V-B discusses the differences in CIN classification performance between the underlying networks and our proposed network. The time taken by each model to infer a single sample is also analyzed, and the heat map of our proposed model for key feature attention is presented. In Section V-C, the performance of the CIN classification framework is compared with that of other researchers.

A. Evalution Metrics

In the experimental section, we conducted several sets of experiments. One set summarised the effect of different model parameters on the classification accuracy of the model. Another

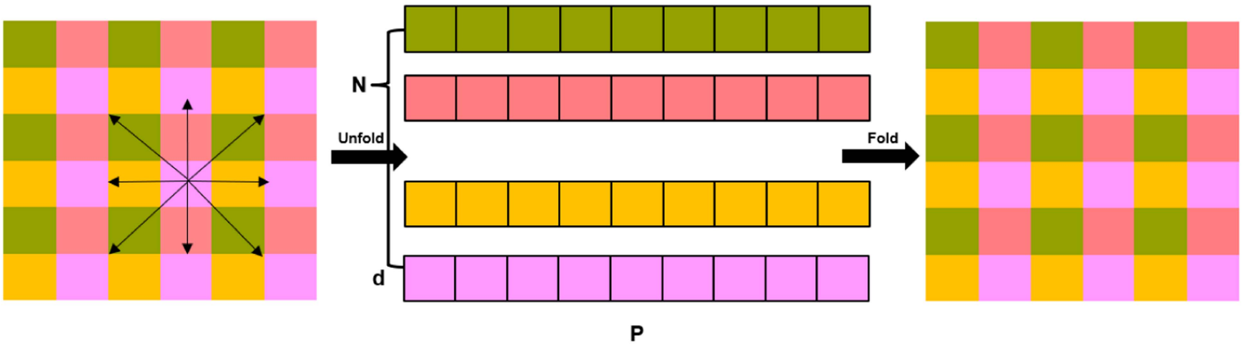

Fig. 7. Flowchart of unfold and fold operations.

set illustrated the improvement of model performance after processing the dataset with our preprocessing method. In the third experiment, we compare the performance of the existing model with our proposed model on the preprocessed dataset to highlight the superiority and necessity of the preprocessing method.

To assess the performance of a classification model, evaluation metrics such as accuracy, precision, sensitivity, and specificity can be used. These metrics are commonly employed to evaluate the performance of classification models. Accuracy is an important metric as it represents the proportion of correctly predicted samples out of all predicted samples. The accuracy metric formula is expressed as follows: TP represents the number of true positive classes classified as positive, TN represents the number of negative classes correctly classified as negative, FP represents the number of positive classes incorrectly classified as negative, and FN represents the number of negative classes incorrectly classified as positive. To simplify matters, the negative class referred to above is actually the LSIL, while the positive class is the HSIL.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{15}$$

B. Performance Comparison With Different Networks

This experiment evaluates the effectiveness of the proposed network compared to alternative networks, including AlexNet, VggNet, MobileNetV2, ResNet50, and ShufflenNet, for CIN classification on the same preprocessed dataset. All networks were pre-trained on ImageNet, which has the advantage of faster convergence during training and the model not having to learn image features from scratch. Table V shows that our CNN and Transformer-based models outperform the previously mentioned deep neural networks on all metrics, demonstrating their superior performance. Our models combine high classification performance with low weight.

To consider practical applications, we conducted an experiment comparing the inference time required to complete a single image for each model.

Colposcopic images from the test set were selected for the experiment, and Table VI shows the average single-sample inference time for each model. The proposed model completes inference quickly and achieves speeds far beyond manual diagnosis in clinical use.

In our experiments, we found that colposcopy images contain noise, which can negatively impact the network's classification accuracy. To evaluate the model's attention to key features and its ability to filter out noise after extensive training, we utilized Grad-Cam [28]. The Grad-Cam is computed by weighting the output of the final model layer with the results of the backpropagation. The first row of Fig. 8 illustrates this. The model successfully avoids interference and focuses on key feature areas even when common endoscopic image noise, such as endoscopy, is present in the sample content. As demonstrated below, the model can effectively eliminate specular reflection noise by learning numerous features. Even without the use of the specular reflection removal algorithm, the model remains relatively resistant to interference.

C. Ablation Studies

In this section, we conducted ablation experiments to confirm the significance of our preprocessing approach and the model's components.

1) Effect of Key Components: This experiment demonstrates the validity of MFViT and multi-scale modules (MSFE and MSFF). The results are shown in Table VII. When the multi-scale modules (MSFE and MSFF) are removed but the MFViT module is kept, the accuracy is 87.5%, precision is 88.04%, and F1 score is 85.86%. When the MFViT module is removed but the multi-scale modules (MSFE and MSFF) are retained,

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Params	Accuracy	Precision	Sensitivity	Specificity	F1 Score
Alexnet	61.10M	79.42%	71.40%	80.00%	78.82%	75.46%
Vgg16	138.36M	79.25%	71.16%	83.40%	76.20%	76.80%
MobilenetV2	3.50M	83.23%	83.03%	84.82%	82.67%	83.92%
Resnet50	25.56M	83.68%	90.39%	78.35%	91.81%	83.94%
ShufflenetV2	1.37M	84.21%	93.49%	78.39%	94.28%	85.28%
Densenet121	7.9M	84.28%	89.27%	80.16%	90.33%	84.47%
ViT	86.33M	85.59%	91.58%	83.66%	92.29%	87.44%
Ours	**3.3M**	**89.20%**	**92.30%**	**88.16%**	**91.92%**	**90.18%**

TABLE VI
CLASSIFICATION TIME FOR SINGLE CASE SAMPLES

Model	Milisecond(average)
ViT	220
AlexNet	172
ResNet50	176
MobileNetV2	176
EfficientNetV2	180
ShufflentNet	856
RegNet	186
Ours	**172**

TABLE VII
ABLETION STUDIES OF MFEM-CIN

Module	1	2	3
MFViT	✓		✓
MSFE		✓	✓
MSFF		✓	✓
Accuracy	87.50	86.34	89.20
Precision(%)	88.04	88.93	92.30
F1 Score(%)	85.86	86.22	90.18
Param(M)	2.3	2.7	3.3

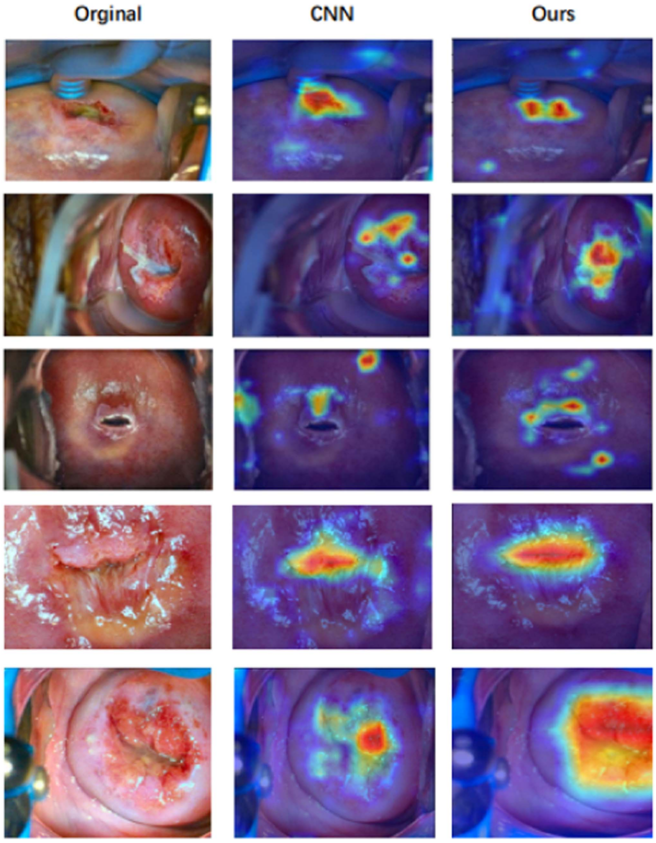

Fig. 8. Heat map of CNN with our model.

TABLE VIII
ABLETION STUDIES OF MFEM-CIN

Model	unenhanced	enhanced
AlexNet	77.31%	79.42%
ResNet50	81.47%	83.68%
MobileNetV2	80.96%	83.23%
DenseNet	81.29%	84.28%
ViT	83.26%	85.59%
Ours	**86.52%**	**89.20%**

the accuracy is 86.34%, precision is 88.93%, and F1 score is 86.22%. The comparison in Table V shows that the performance on these metrics was higher than that of the comparison model. The experimental results indicate that integrating these two modules improved the metrics compared to the first two groups of experiments. This confirms the necessity of correlating local features with global features and multi-scale feature extraction in our diagnosis of colposcopic images.

2) Effect of Preprocessing Methods: In the previous section on data preprocessing, we introduced the image enhancement algorithm and the highlight removal algorithm used in this process. To demonstrate the practicality and necessity of these two parts of preprocessing, we designed an experiment to compare the classification performance of the model before and after preprocessing. The results are shown in Table VIII. The experiment demonstrated that preprocessing the predesigned data improved the model's feature learning on cervical images, resulting in better performance. However, it should be noted that the multi-channel model used in the study had over 1000 patients, while our data sampling only contains data from over 4000 patients. The team of professional physicians achieved a diagnostic accuracy of 81.00%. Therefore, it can be concluded that the automated diagnostic method proposed in this paper is superior to the professional physicians.

D. Comparison With Other Methods

Table IX shows a comparison between the results of our experiment and those of previous studies. Although other studies have different types of classification, our classification of CIN type LSIL/HSIL is more clinically relevant. The following works can be compared with ours in this regard. Our proposed model has a better accuracy rate of 89.2%, and our sample size is also

TABLE IX
COMPARISON WITH OTHER RELATED METHODS

Method	Year	Data sets	Accracy
Bae *et al.* [29]	2018	240 colposcopic images with acetic acid	80.80%
Asiedu *et al.* [30]	2018	1112 acetic acid laboratory colposcopy images, consisting of 345 positive (CINN2/3/cancer) and 767 negative (CIN1/Normal) cases	80.00%
Li *et al.* [9]	2020	7668 time-lapse colposcopic images divided into (Normal/LSIL+)	78.33%
Saini *et al.* [11]	2020	800 colposcopic images provided by NCI in Type1 (Normal/CIN1) Type2 (CIN1, CIN3 and CIN4)	81.33%
Peng *et al.* [31]	2021	300 acetic acid laboratory colposcopy images (75 each of Normal, CIN1, CIN2, CIN3) divided into HSIL/LSIL	86.30%
Ours	2023	4081 acetate charts HSIL/LSIL classification	89.20%

comparable to that of Li et al., who had a sample size of over 7000.

VI. Experimental Result and Analysis

Computer-aided automated diagnosis is crucial for the widespread availability of medical treatment in economically underdeveloped areas. This paper presents pre-processing methods, including image enhancement and specular reflection, to accelerate model training speed and improve model classification accuracy based on the characteristics of colposcopic images. The proposed MFEM-CIN model is based on CNN and Transformer. The model's design is based on the concept of combining a Vision Transformer with a CNN to achieve a global perceptual field and capture long-range dependencies while avoiding the drawbacks of ViT's large parameter count. The combination of MSFE and MSFF enable the model to extract and fuse multi-scale features, taking into account the interaction between shallow important features and high-level semantics. The classification performance of MFEM-CIN was validated using basic CNN and ViT models on the vagoscopic dataset. The experiments demonstrated that our model outperformed previous studies in terms of classification accuracy on highly complex datasets. Additionally, our model was shown to be highly focused on key features when faced with highly noisy samples, illustrating its strong learning ability.

In summary, this paper's research can assist physicians in making auxiliary clinical judgments and improving medical care, which is of great importance in reducing the incidence of cervical cancer in poor areas. However, the study has some limitations. Firstly, the pretreatment method is not a system with an automated diagnostic framework, so more efforts will be made to access it in the future. Secondly, some patients may have intracervical lesions that are not directly visible during colposcopy. To address this issue, we need to integrate cervical cytology and HPV results into the dataset. Additionally, we are working to collect colposcopy datasets from different regions and populations, along with other medical imaging datasets, to further validate the model's generalizability.

Human and Animal Rights

None

Conflict of Interest

The authors affirm that this study was carried out without any commercial or financial associations that would be interpreted as indicative of potential conflicts of interest.

Author Contributions

FL and PC conceived the study; FL, JZ, and BW participated in the methodology design; FL and PC carried it out and drafted the manuscript. All authors revised the manuscript critically. All authors read and approved the final manuscript.

Acknowledgment

The authors would like to express our gratitude to Wanan Medical College in Wuhu, China, for providing the colposcopic images used in this research.

References

[1] D. J. Li, J. Shi, J. Jin, N. Y. Du, and Y. T. He, "Epidemiological trend of cervical cancer," *Zhonghua zhong liu za zhi [Chin. J. Oncol.]*, vol. 43, no. 9, pp. 912–916, Sep. 2021, doi: 10.3760/cma.j.cn112152-20190904-00573.

[2] O. Ginsburg, F. Bray, M. P. Coleman, V. Vanderpuye, and L. Conteh, "The global burden of women's cancers: A grand challenge in global health," *Lancet*, vol. 389, pp. 847–860, 2016.

[3] G. Purwoto, H. D. Dianika, A. Putra, S. Purbadi, and L. Nuranna, "Modified cervicography and visual inspection with acetic acid as an alternative screening method for cervical precancerous lesions," *J. Cancer Prevention*, vol. 22, no. 4, pp. 254–259, 2017.

[4] L. T. Thomsen, S. K. Kjr, C. Munk, K. Frederiksen, D. Rnskov, and M. Waldstrm, "Clinical performance of human papillomavirus (HPV) testing versus cytology for cervical cancer screening: Results of a large Danish implementation study," *Clin. Epidemiol.*, vol. 12, pp. 203–213, 2020.

[5] D. Zhi, G. Li, and X. Xing, "Treatment of cervical intraepithelial neoplasia," *J. Practical Oncol.*, vol. 21, no. 005, pp. 478–480, 2007.

[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[7] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lui, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[9] E. Portelance, M. C. Frank, D. Jurafsky, A. Sordoni, and R. Laroche, "The emergence of the shape bias results from communicative efficiency," in *Proc. 25th Conf. Comput. Natural Lang. Learn.*, 2021, pp. 607–623.

[10] F. Shamshad et al., "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102802.

[11] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://arxiv.org/abs/2110.02178

[12] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 115–123.

[13] D. Y. Ma, J. H. Liu, J. Li, and Y. F. Zhou, "Cervical cancer detection in cervical smear images using deep pyramid inference with refinement and spatial-aware booster," *IET Image Process.*, vol. 14, no. 17, pp. 4717–4725, Dec. 2020, doi: 10.1049/iet-ipr.2020.0688.

[14] L. Zhang, L. Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao, "DeepPap: Deep convolutional networks for cervical cell classification," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1633–1643, Nov. 2017, doi: 10.1109/jbhi.2017.2705583.

[15] Y. Li et al., "Computer-aided cervical cancer diagnosis using time-lapsed colposcopic images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3403–3415, Nov. 2020.

[16] C. Buiu, V. R. Dnil, and C. N. Rdu, "MobileNetV2 ensemble for cervical precancerous lesions classification," *Processes*, vol. 8, no. 5, 2020, Art. no. 595.

[17] S. K. Saini, V. Bansal, R. Kaur, and M. Juneja, "ColpoNet for automated cervical cancer screening using colposcopy images," *Mach. Vis. Appl.*, vol. 31, no. 3, Mar. 2020, Art. no. 15, doi: 10.1007/s00138-020-01063-8.

[18] Z. Xue, S. Antani, L. R. Long, J. Jeronimo, and G. R. Thoma, "Comparative performance analysis of cervix ROI extraction and specular reflection removal algorithms for uterine cervix image analysis," *Proc. SPIE*, vol. 6512, pp. 1507–1515, 2007.

[19] Z. Yue, S. Ding, X. Li, S. Yang, and Y. Zhang, "Automatic acetowhite lesion segmentation via specular reflection removal and deep attention network," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3529–3540, Sep. 2021, doi: 10.1109/jbhi.2021.3064366.

[20] Y. Miyagi, K. Takehara, Y. Nagayasu, and T. Miyake, "Application of deep learning to the classification of uterine cervical squamous epithelial lesion from colposcopy images combined with HPV types," *Oncol. Lett.*, vol. 19, no. 2, pp. 1602–1610, 2020.

[21] Y. Li, Z. H. Liu, P. Xue, J. Chen, and Y. L. Qiao, "Grand: A large-scale dataset and benchmark for cervical intraepithelial neoplasia grading with fine-grained lesion description," *Med. Image Anal.*, vol. 70, no. 3, 2021, Art. no. 102006.

[22] H. Chen, Z. Yin, P. Zhang, and P. Liu, "SleepZzNet: Sleep stage classification using single-channel EEG based on CNN and transformer," *Int. J. Psychophysiol.*, vol. 168, 2021, Art. no. S168.

[23] Z. Hong et al., "Dual encoder network with transformer-CNN for multi-organ segmentation," *Med. Biol. Eng. Comput.*, vol. 61, pp. 661–671, 2023.

[24] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Trans. Consum. Electron.*, vol. 43, no. 1, pp. 1–8, Feb. 1997.

[25] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphic Gems IV*. San Diego, CA, USA: Academic Press Professional, 1994, pp. 474–485.

[26] Z. Yue, S. Ding, X. Li, S. Yang, and Y. Zhang, "Automatic acetowhite lesion segmentation via specular reflection removal and deep attention network," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3529–3540, Sep. 2021, doi: 10.1109/jbhi.2021.3064366.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[29] J. K. Bae et al., "Quantitative screening of cervical cancers for low-resource settings: Pilot study of smartphone-based endoscopic visual inspection after acetic acid using machine learning techniques," *JMIR mHealth uHealth*, vol. 8, no. 3, 2020, Art. no. e16467.

[30] M. N. Asiedu et al., "Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, Pocket Colposcope," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2306–2318, Aug. 2019.

[31] G. Peng, H. Dong, T. Liang, L. Li, and J. Liu, "Diagnosis of cervical precancerous lesions based on multimodal feature changes," *Comput. Biol. Med.*, vol. 130, 2021, Art. no. 104209.