ISSN 2442-6571

CMT-CNN: colposcopic multimodal temporal hybrid deep learning model to detect cervical intraepithelial neoplasia

Lalasa Mukku [a,1,*], Jyothi Thomas [b,2]

[a] CHRIST(Deemed to be University), Kengeri, Bangalore -560074, India
[1] m.lalasa@res.christuniversity.in; [2] j.thomas@christuniversity.in
* corresponding author

ARTICLE INFO

Article history
Received June 28, 2023
Revised September 12, 2023
Accepted March 9, 2024
Available online May 31, 2024

Selected paper from The 2023 6th International Symposium on Advanced Intelligent Informatics (SAIN'23), Yogyakarta (Virtually), September 21, 2023, http://sain.ijain.org/2023/. Peer-reviewed by SAIN'23 Scientific Committee and Editorial Team of IJAIN journal.

Keywords
Cervical cancer
EfficientNet
Attention mechanism
Deep learning
Specular reflections

ABSTRACT

Cervical cancer poses a significant threat to women's health in developing countries, necessitating effective early detection methods. In this study, we introduce the Colposcopic Multimodal Temporal Convolution Neural Network (CMT-CNN), a novel model designed for classifying cervical intraepithelial neoplasia by leveraging sequential colposcope images and integrating extracted features with clinical data. Our approach incorporates Mask R-CNN for precise cervix region segmentation and deploys the EfficientNet B7 architecture to extract features from saline, iodine, and acetic acid images. The fusion of clinical data at the decision level, coupled with Atrous Spatial Pyramid Pooling-based classification, yields remarkable results: an accuracy of 92.31%, precision of 90.19%, recall of 89.63%, and an F-1 score of 90.72. This achievement not only establishes the superiority of the CMT-CNN model over baselines but also paves the way for future research endeavours aiming to harness heterogeneous data types in the development of deep learning models for cervical cancer screening. The implications of this work are profound, offering a potent tool for early cervical cancer detection that combines multimodal data and clinical insights, potentially saving countless lives.

1. Introduction

Cervical cancer represents a substantial health concern for women worldwide. 604,000 new cases were recorded in the year 2020, and among them, 342,000 deaths were reported [1]. Most of these cases and deaths, around 90%, happened in low incomes countries. In India, cervical cancer ranks as the third most prevalent malignancy, exhibiting an incidence rate of 18.3%, which translates to 123,907 reported cases, and stands as the second foremost contributor to mortality, with a fatality rate of 9.1%, according to data from GLOBOCAN 2020 [2].

Unlike other forms of cancers which are genetically triggered, the causation factor of cervical cancer is known to be a virus called human papillomavirus (HPV) [3], [4]. Cervical cancer can be completely cured if identified in its early stages [5]. Cervical cancer can be nipped off in the bud altogether through systematic screening and swift intervention. The World Health Organization (WHO) urged global countries to work towards the eradication of cervical cancer [6]. Globalized uniform cervical cancer screening can be a potential step toward achieving this goal [7].

Artificial intelligence (AI) [8] assisted cancer screening [9], [10] has gained notable traction in the past two decades, and cervical cancer diagnosis has benefitted remarkably from AI solutions [11]–[13]. Several researchers have embodied deep learning solutions for cervical cancer detection through medical

https://doi.org/10.26555/ijain.v10i2.1527

imaging [14]. Cervical cancer diagnostic images range from pap smear, colposcope, magnetic resonance imaging (MRI) to computerized tomography (CT) [15]. Singh et al. [16] published a chronological review of the deep learning solutions in screening of cervical cancer. The outcome of the survey ascertains that deep learning CAD solutions are a bridge to developing automatic screening of cervical cancer.

Colposcopy imaging is considered the gold standard for identifying cervical. Colposcopy examination is a pivotal tool for cervical cancer screening that offers a greater degree of accuracy than the Thin-Prep cytologic test (TCT) tests and human papillomavirus (HPV) tests [17]. During a colposcopy test, 5% acetic acid is smeared on the cervix to highlight the cancerous features[18]. A colposcope is used to capture the cervix image, which would have the lesions appearing distinctly within a few minutes of acetowhite treatment. In some cases, the cervix images are captured in a time series fashion with acetic acid, saline, and Lugol's iodine application. Colposcopy image classification for diagnosis is generally done to differentiate benign and low/high squamous intraepithelial lesions [19] or cervical intraepithelial neoplasia (CIN). The CIN level (Fig. 1) is used as the class label in this classification study. Clinicians experience and expertise is the basis of diagnostic accuracy in traditional colposcope exam. It is a scarce resource in several low to middle income countries. There is an insufficient number of experienced specialists to accommodate the number of patients who need screening. Parallelly, several researchers are investigating the use of deep learning to distinguish between cervical lesions seen in colposcopy images to help with patients triaging in clinical settings and improve clinicians' diagnostic accuracy. The objective of this research is to develop a novel technique to handle multi stage cervix images and patient data to provide classification support to expert clinicians to enable efficient diagnosis of cervical cancer.

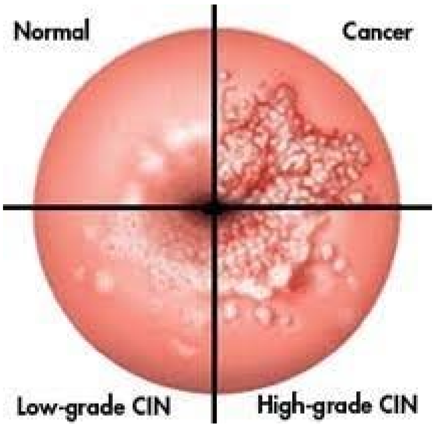

Fig. 1. Cervix image representing the features in various levels of malignancy

A well-structured literature review is essential as it provides a foundation of existing knowledge, contextualizes the research, and identifies gaps that this study aims to address.

A significant amount of research is aimed at segmenting and classifying colposcope images [20], [21]. Fan et al. [22] used a Mask R-CNN to segment the cervix area of interest, encoded the input images through EfficientNet B3 architecture, and attained 92.7% accuracy with 0.9856 AUC in classifying cancer. Yan et al. [23] designed a BFCNN, a bilinear fuse convolutional neural network for the segmentation and classification of cervigrams. Yuzhen Cao [24] developed a multiscale feature fusion classification network design for classifying the cervical transformation zone, which demonstrated an accuracy of 88.49% with 90.12% sensitivity. Asiedu et al. [25] used machine learning methods of using boundary boxes to extract ROI and classify the region through support vector machines. Despite their satisfactory performance, these models suffer from methodological fallibility of using a single acetic acid image as input. In order to overcome the said drawback, the input of the model could be enhanced to harbor multiple types of information like clinical reports, manual analysis findings, video sequences of target organ examination etc.

However, the approach of using more than one input datatype (data modality) along with temporally acquired images is low in practice. Park et al. [26] used anatomical maps with texture and colour to identify cancerous regions, then employed k-means clustering to divide these regions into sub-regions. Using a CRF classifier, they amalgamated the categorization results of surrounding areas. in a probabilistic way and finally determined the overall classification results with the help of KNN and LDA integration, thus enabling automatic recognition of normal, CIN, and SCC (squamous cells of the cervix). Xu et al. [27] carried out a study in which they took three pyramid features (PLBP, PLAB, and PHOG) and manually extracted them, then compared seven traditional classifiers and one convolutional neural network (CNN). The cancer classification was then completed, and it was found that CNN was more effective than the standard machine learning classifiers. Chen et al. [28] tested a multimodal deep fusion technique called MultiFuseNet to classify cervical dysplasia. They proposed Multimodal Fusion Learning for Cervical Dysplasia Diagnosis for feature fusion of image modality with metadata and reported an accuracy of 87.4% with 86.1% specificity and 88.6% sensitivity. Li et al. [29] created a computer-generated diagnostic program based on an AW opacity index, which yielded a diagnosis with 84% specificity and 88% sensitivity. Authors of [30] developed a diagnostic image analysis system based on acetowhite lesion-based statistical features and evaluated its diagnostic accuracy. The reported sensitivity and specificity were 79% and 88%, respectively. Li et al. [31]constructed a convolutional network incorporating graph and edge features, denoted as E-GCN, and reported an accuracy of 78.33% when leveraging time series image characteristics. Perkins et al. [19]offered a contrasting perspective by integrating 17 time series colposcope images, their investigation revealing an absence of a significant improvement in accuracy upon analyzing the amalgamation of these 17 images. This finding prompts contemplation regarding the potential efficacy of incorporating non-image data to meaningfully enhance classification accuracy. Peng et al. [22] scrutinized the alterations in multimodal features through the development of a multi-state convolutional neural network employing a genetic algorithm approach, culminating in an impressive accuracy rate of 86.3%. Concurrently, Yinuo Fan et al. [32] devised a multimodal fusion colposcopic convolutional neural network (CMF-CNN), leveraging Squeeze-and-Excitation (SE) fusion techniques, yielding an outstanding accuracy rate of 92.70%. The above two multimodal approaches have used image and clinical data. However, they have the limitation of using a single acetic acid image as input. Adding meaningful information from the cervix image via saline and Lugol's iodine solution application is the way forward to assert superior, interpretable, and dependable results. Li et al. [31] approached the time series imaging problem by building a graph convolutional network(E-GCN) with edge features to fuse sequential images of the cervigrams (images captured at 60’s, 90’s, 120’s, 150’s) and achieved an accuracy of 78.33. A bird's eye view of the results presented in this section comes down to a scattered version of accuracies. One explanation to account for the varied results is that the strength of a deep learning model is dependent on the dataset size and quality. Higher accuracies with low sensitivity and specificity may represent the overfitting that could have occurred. In the same manner, the lesser accuracy with consistent specificity and sensitivity indicates the robustness of the trained model. In this article, we propose a colposcopic multimodal temporal CMT-CNN deep attention module that trains on heterogenous time series cervix image data in combination with clinical findings to classify cervical intra epithelial neoplasia.

The "Colposcopic Multimodal Temporal Convolution Neural Network (CMT-CNN)" represents a crucial component of our study. This innovative model was specifically designed to address the unique challenges associated with cervical cancer detection using colposcopic imagery. CMT-CNN utilizes a combination of temporal convolutional neural networks and multimodal data integration techniques to enhance the accuracy of cervical cancer diagnosis.

The key contributions of the paper are as follows

- A novel specular reflection removal method is proposed to remove the specular reflections from surface of cervix image.
- Mask Region convolution neural network is employed for segmenting the cervix region of interest from the colposcope image.
- The visual features are extracted using an EfficientNet architecture which are further fused with the corresponding clinical data vectors at decision level.

- Finally, a deep neural network with Atrous special pyramid-based hybrid classification segregates the cancer cases

To provide a more detailed understanding of CMT-CNN's functionality, we will elaborate on its key components, including the architecture, data preprocessing methods, and the rationale behind its multimodal approach in the methodology section. Additionally, we will highlight how CMT-CNN differs from existing techniques and the advantages it offers in terms of improved diagnostic accuracy and early detection in discussion section. By incorporating this information, we aim to ensure that readers have a comprehensive understanding of the innovative CMT-CNN model and its role in our research.

The paper is structured as follows: Section 2 explains the methods and materials, Section 3 presents the results and discusses them, and Section 4 concludes the study

2. Method

The schematic architecture of the current study is given in Fig. 2. The proposed CMT-CNN performs the fusion learning of acetic acid, saline, and iodine images and clinical data, then makes a CIN classification. The approach was broken down into four main steps: preprocessing and specular reflection removal; segmentation of the cervical region; image feature extraction, combining multiple features from clinical data; and finally, the classification of the features. In the first step, the specular reflections are identified, removed, and inpainted. In the second step, the cervix region of interest is segmented to be extracted from the colposcope images using Mask R-CNN. Subsequently, the images were resized and normalized to a uniform level of 512x512. Next, networks designed to encode image features were chosen to extract the features separately. For this purpose, EfficientNet B7 [33] architecture was chosen as a feature encoding network. Following that step, a colposcopic multimodal temporal (CMT model) network for feature fusion, including an atrous spatial pyramid pooling (ASPP) [34] block, Conv block, squeeze and excitation block [35] is introduced, which compresses and fuses the multitype image features of the input. The clinical data is fused with the image features using the One-Hot encoding process, making it a preprocessed multimodal dataset ready for classification. In the last step, the clinical and image features are concatenated and classified using dropout, FS, and swish layers.

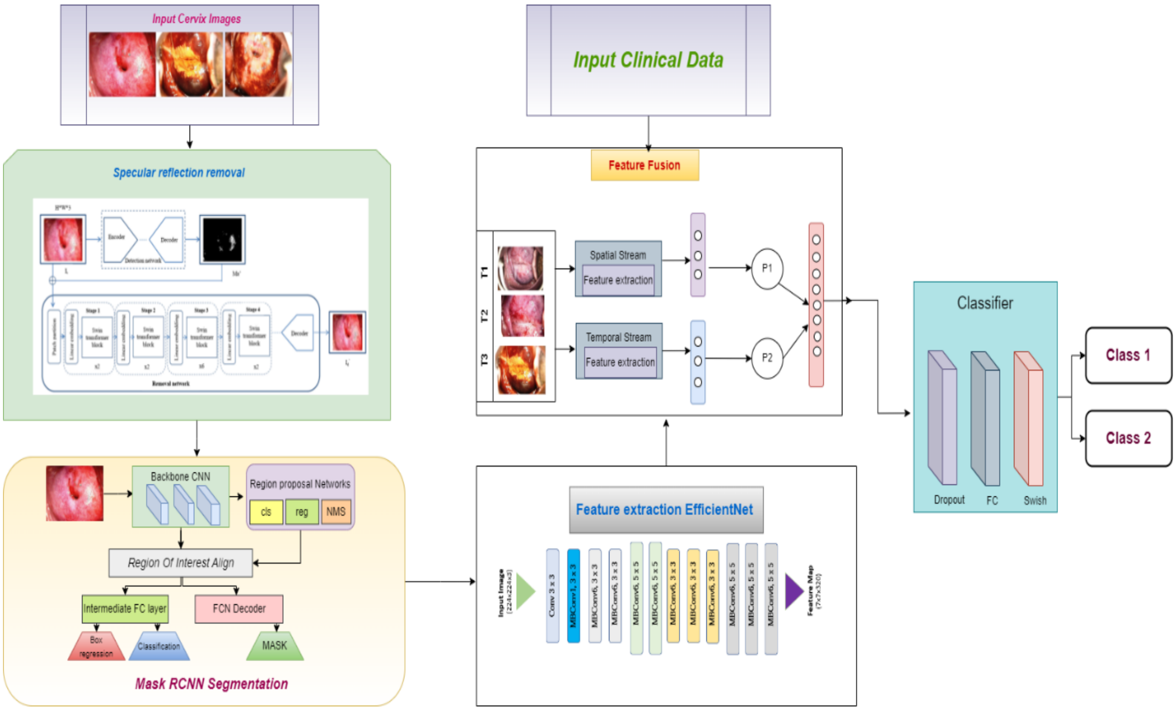

Fig. 2. Schematic architecture of the CMT-CNN

2.1. Preprocessing and specular reflection

Before beginning with the methodology, the entire dataset is preprocessed to remove specular reflections (SR). SR are high-intensity bright spots on the cervix caused by the moisture on the cervix reflecting the light from the colposcope. These spots are identical to the acetowhite lesions. AW lesions are the key features for the identification of the presence of cervical cancer. Since specular reflections (SR) have the same morphological appearance as acetowhite lesions, the diagnosis will be hindered by SR. In order to overcome the said drawback, it is essential to remove SR before classifying the cervigram. Over the last couple of decades, several researchers have proposed SR removal techniques using machine and deep learning methods.

The fundamental principle of specular reflection areas involves the reflection of light from a smooth, shiny surface. This type of reflection produces a clear and sharp image of the light source as opposed to diffuse reflection, which produces a more scattered and diffused image. The angle at which the light strikes the surface, as well as the angle at which it is reflected, plays a pivotal role in understanding the characteristics of the reflected light. Additionally, the smoothness and shininess of the surface can affect the clarity and sharpness of the reflected image. (Fig. 3).

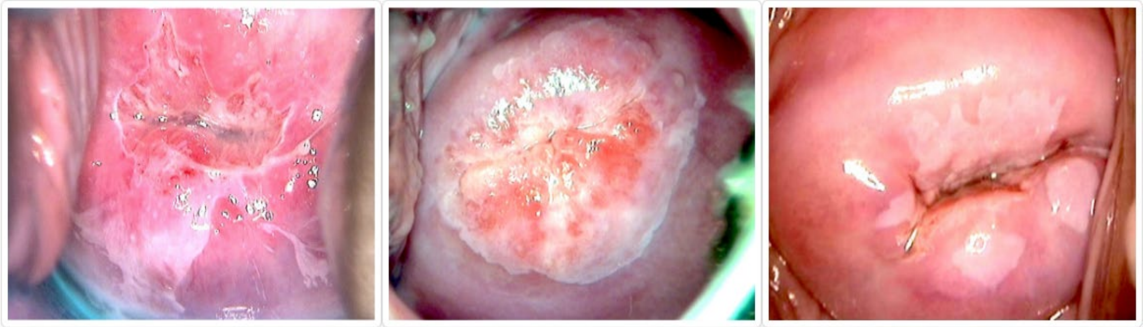

Fig. 3. Specular reflections on the cervix surface

Specular reflections obstruct the efficient analysis of cancerous changes in surface regions. For instance, [6] has explored the role of SR in confusing the endoscope procedure. SR removal has two phases. The first is to locate the specular region and remove the SR pixels. The second is to paint these areas back to their original morphology. During the phase of detection, generally, the image is projected into a diverse color space to facilitate further processing of the region of interest (ROI). For instance, the image formats used are RGB, grey-level [36], HSV, HSI [37], and a threshold value to identify the SR. Subsequently, the removed pixels are replaced with inpainting to preserve the image morphology.

2.1.1. Specular reflection identification

A specular reflection is a type of reflection in which the reflected light rays are at an angle to each other. In other words, the reflection is in the opposite direction as the incident light. In a bi-dimensional histogram, specular reflection refers to the symmetrical nature of the histogram when it is reflected along the x-axis or the y-axis. This means that the shape of the histogram remains the same after it is reflected, and the relative frequencies of the data points are preserved. Therefore, a bi-directional histogram decomposition is used to detect specular reflections whose formula is given in equations (1&2)

$$m = \frac{1}{3}(b + g + r) \tag{1}$$

where '*m*' stands for pixel intensity.

$$s = \begin{Bmatrix} \frac{1}{2}\left(2r - g - b \ = \ \frac{3}{2}(r - m)\right) \ in\ the\ event\ (b + r) > 2g \\ \frac{1}{2}\left(r + g - b \ = \ \frac{3}{2}(m - b)\right) \ in\ the\ event\ (b + r) \geq 2g \end{Bmatrix} \tag{2}$$

$$\begin{Bmatrix} m_p \geq \frac{1}{2}\, m_{max} \\ S_p \leq \frac{1}{3} s_{max} \end{Bmatrix}$$

Here, 's' denotes saturation, and (r, g, b) = (red, green, blue). Two important threshold values (mmax, Smax) determine the specular reflection pixels through a bi-dimensional histogram. Two independent criteria that must be met for a pixel to be considered as SR are given in equation (2).

2.1.2. Specular reflection removal

Image linear correction is a simple and effective way usually employed to enhance image quality and is often used as a preprocessing step for more advanced image analysis techniques. It involves applying a linear transformation to the pixel values of the image in order to stretch or compress the range of intensity values. There are several different techniques that can be used for linear image correction. The pixel replacement should be executed in a manner that ensures the preservation of the essential information contained within the cervix image. Routinely, the SR pixels are replaced with the mean of pixels surrounding the pixel that needs to be replaced.

2.1.3. Inpainting of deleted specular pixels

The Laplacian equation is a partial differential equation that describes the behavior of a two-dimensional surface. The Laplacian equation can be used in image repainting, a technique used to restore damaged images. In this context, the Laplacian equation can be used to identify SR in the image, which can then be used to repaint the SR areas. In order to apply the Laplacian equation in image repainting, the image is first convolved with a Laplacian kernel to enhance the edges and boundaries. The repainting is then performed in the areas of the image that have SR, using the enhanced edges and boundaries as a guide. The final step is to smooth the repainted areas and blend them with the rest of the image, to produce a seamless and natural-looking result. The equation for Laplace transformation is given in equation (3)

$$F(s) = \int_0^{\infty} f(t)e^{-st}t' \tag{3}$$

2.2. Cervix region of interest (ROI) extraction

A colposcope image frequently contains extraneous elements such as background noise and unwanted objects like vaginal walls and speculum [38]. The cervix region must be precisely cropped for subsequent efficient classification. The previous research on cervix ROI extraction is broadly classified into machine learning and deep learning methods. ML methods like gaussian mixture modelling [39], K means clustering [40], etc., were used extensively. As for deep learning methods, Deeplab V3 [41], Mask R CNN [42], and Faster R CNN are prominently employed. In this experiment, we adopted a supervised Mask R-CNN model because of its superior semantic object segmentation ability.

A Mask Regional Convolutional Neural Network (Mask R-CNN) is a cutting-edge deep learning architecture specifically designed for instance partition and object detection tasks in computer vision. In our study, we leveraged Mask R-CNN as a pivotal component for cervix region segmentation within colposcopic images. Mask R-CNN is a CNN architecture for object detection and instance segmentation [43]. It draws out the Faster R-CNN object detection model by incorporating a new branch for object mask prediction along with the pre-existing one for bounding box recognition [42]. This allows the model to identify not only what objects are present in an image but also where they are located and what their precise shape is. This property makes Mask R-CNN an invaluable tool for tasks and image segmentation and object counting within the stream of medical imaging [43].

There is a backbone network of convolution neural network. This backbone network extracts hierarchical features from the input image, preserving spatial information at different scales. Mask R-CNN includes an region proposal network (RPN), which scans the feature maps generated by the backbone network to propose potential object regions. These proposals are ranked based on their likelihood of containing an object. After obtaining the region proposals, Region-of-Interest Align is used to extract fixed-size feature maps from the backbone network's output for each proposed region. This step ensures that each region has a consistent size, making it suitable for further processing. Subsequently a mask prediction functionality is implemented. This is the unique aspect of Mask R-CNN. For each proposed region, it predicts a binary mask that specifies which pixels belong to the object of interest and which do not. This is achieved through a convolutional head that operates on the RoI feature maps.

Mask regional convolution neural networks can attain greater accuracy in segmentation while taking the same amount of time due to its implementation of ROI Align technology, as compared to Faster R-CNN. After being processed at the ROI align layer, convolution block layer, linear layer, and RPN layer, the cervical region is extracted and highlighted with the help of a 'bounding box'. The information on the location and size of cervical region detected was recorded by recording the x- coordinates and y-coordinates, along with the height and width of the upper origin of the rectangular box drawn (Fig. 4). It was then extracted and resized.

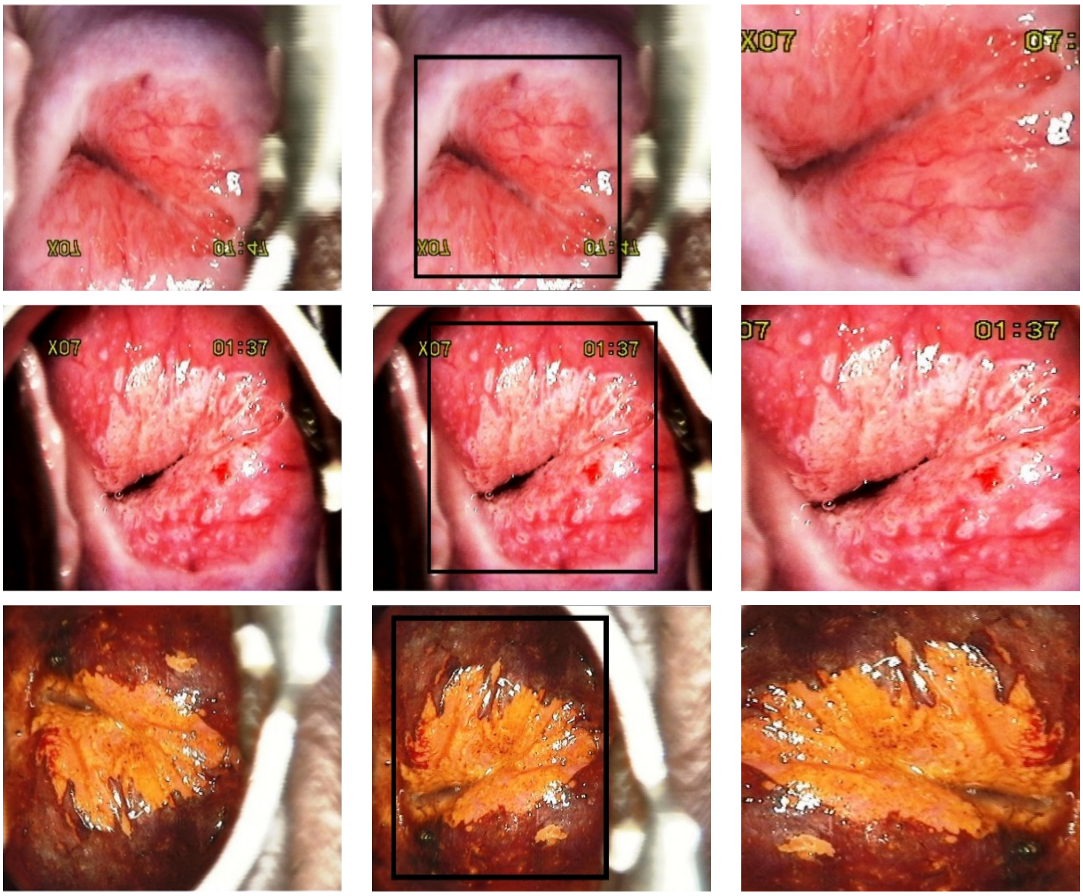

Fig. 4. a) Initial image b) ROI marked c) Resized

2.3. Feature extraction

EfficientNet is a CNN architecture that has been designed to maximize both the accuracy and efficiency of a model. One way it achieves this is by using A scaling method that uniformly modifies the dimensions of network layers based on the magnitude of the input data. This allows the model to maintain a high level of accuracy even when working with small or large input images. Another way EfficientNet improves efficiency is by using a novel compound scaling method that scales the network in a more structured way, allowing it to achieve better performance with fewer parameters.

EfficientNet introduces a concept called "compound scaling," which involves systematically scaling the model in three dimensions: width (channel count), depth (layer count), and resolution (input image size). The goal is to find the right balance between these dimensions to maximize accuracy. To scale the depth of the network, EfficientNet repeats building blocks (combinations of convolutional and pooling layers) multiple times. The number of repetitions is determined by a parameter called the "depth multiplier." Increasing the depth allows the network to capture more complex hierarchical features. Width scaling involves increasing the number of channels (also referred to as the filter count or width multiplier) in each layer of the network. This parameter controls the network's capacity and helps it learn richer feature representations. In addition to that EfficientNet scales the input image resolution, allowing the model to process higher-resolution images. EfficientNet frequently utilizes depth-wise separable convolutions, a technique that dissects the conventional convolution process into two distinct layers: point-wise convolution and depth-wise convolution. Squeeze-and-Excitation (SE) blocks are implemented to enhance feature representations. SE blocks dynamically recalibrate feature responses on a per-channel basis., focusing on more informative channels and suppressing less relevant ones.

The encoding part of the EfficientNet model was used for feature extraction from the saline, acetic acid, and iodine temporal sequential images. The model used inverted bottleneck convolution (MBconv) as a backbone network. This MBConv [44] made use of a separable deep convolution function for separating channels from regions, thus reducing the parameters needed. The SE attention mechanism introduced by Hu et al. in [45] SENet has been leveraged in this experiment to expand the sensing region. Further, based on each channel's importance, the parameters were allotted. The length of the sensing field is increased by dilating the convolution. Thereby enabling information to penetrate the neural network, thus assuring generalization and dependable accuracy. The use of the EfficientNet model significantly lowered the parameter count while retaining doubly harvested efficiency and accuracy.

The layers of the network were kept at 30, among which the first layer is BachNorm (BN), followed by 26 MBconv layers, after which a single convolution layer (whose equation is given by (4)) is placed. Finally, another BN layer is put in place. The conv layer is of size 3 x 3 with a growth rate of 6. Once the EfficientNet model had completed the training process, its parameters were locked in place and could no longer be adjusted.

$$a(m,n) * \; b(m,n) = \sum_{z=-\propto}^{\propto} \sum_{z_2=-\propto}^{\propto} a(z_1,z_2) \cdot b(m-z_1,n-z_2) \tag{4}$$

2.3.1. Multimodal temporal feature fusion

The time series images of saline, acetic acid, and iodine are individually processed to extract the relevant features (Fig. 5).

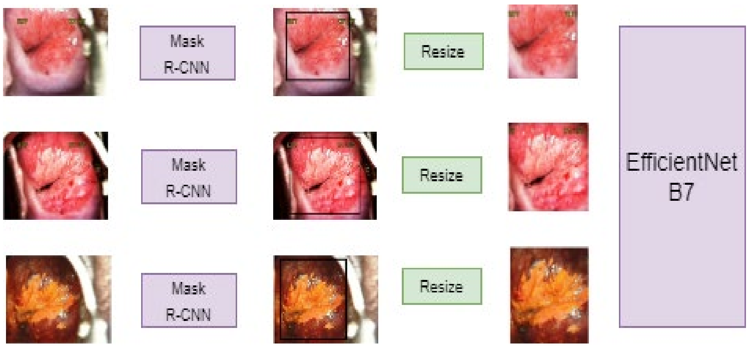

Fig. 5. Proposed architecture of CMT-CNN (a)

The proposed model integrates these image features with selected clinical features to classify the malignancy. The dimensionality of the temporal image is reduced using the fusion module given in Fig. 6.

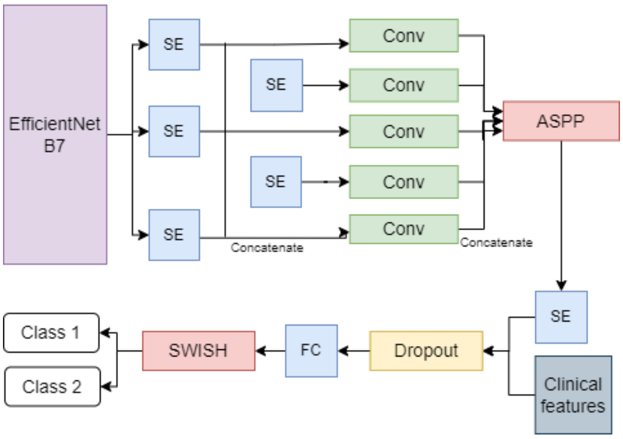

Fig. 6. Proposed architecture of CMT-CNN (b)

2.3.2. Meta data module

In this study we propose a hybrid attention module to integrate information from three colposcope images take in time series with the meta data of that case. Initially the image data has the features extracted through an EfiicientNet B7 model, whose output is a tensor. Subsequently the tensor is passed through a feature compression layer added at the end of the EfiicientNet, so as to convert the tensor into a vector. Parallelly, the metadata features are extracted through a random forest model. The output of the RF is a vector. In the next stage, the self-attention module assigns weights to the image and meta features separately. In addition to that, the hybrid attention module uses the Value from meta data to select weights for Queary and Key of the image vectors and vice versa. This enables the model to systematically select the important features with the help of complementary information. The final feature vector obtained from the hybrid attention module is classified into classes "normal", "CIN1" and "CIN2" groups.

2.4. Transfer learning

In this paper, transfer learning [46] methods are employed to accelerate the network's convergence and improve its ability to preserve the characteristics of various designs sourced from various models. Make sure to keep the originality of each model's pattern and ensure that the features of each design remain distinct. Initially, using EfficientNet B7, which is an ImageNet pretrained model, the saline, acetic acid, and iodine applied images were trained, leading to an accelerated convergence rate for the network. The feature extraction and fusion networks were not trained separately. Instead, they were combined into a single network so that the loss would not converge owing to the complexity of mastering the features. The feature encoding layers of EfficientNet-B3 were frozen, and the outputs of the feature maps were incorporated as input in a multi-modal feature fusion classification system for image classification. Frozen parameters would ensure that none of the layers are updated during training, thereby preserving the benefits of the features extracted of iodine stained images and acetic acid images, ensuring there is no intercession between them, and avoiding any negative effect on results. Post the optimal features acquisition, they were then fed into the successive colposcopic multimodal temporal classification framework (CMT-CNN) for the combination of two datatypes used in this experiment (Fig. 7).

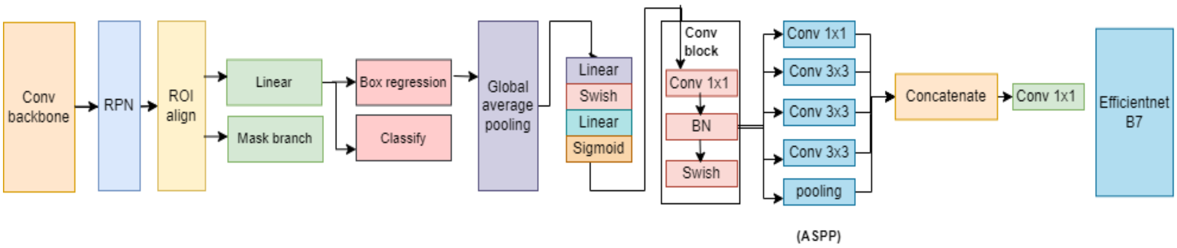

Fig. 7. Proposed architecture of CMT-CNN(c)

3. Results and Discussion

This section presents the details of the experimental environment, dataset information, and evaluation metrics for assessing the performance of the model

3.1. Experimental setup

The model was implemented with a Python Google colab environment, utilizing an Intel Xeon W-2233-based workstation equipped with 32 GB of RAM, a 1 TB HDD, and a 256 GB SSD processor. The computational power was further augmented by an NVIDIA Quadro P5000 GPU with 32 GB of dedicated memory. Essential libraries such as OpenCV and Matplotlib, among others, were employed to facilitate the model development and analysis.

3.2. Dataset

The dataset contained 906 uterine cervix images of saline, acetic acid, and Lugol's iodine stages. The target classes are Normal, CIN 1 (cervical intraepithelial neoplasia), and CIN 2. Each case contained three cervix images captured by a colposcope in the time intervals 0 seconds, 60 seconds, and 120 seconds. In addition to that, clinical data pertaining to age, HPV test result, CIN grade, observations, proposed course of treatment etc., corresponding to images is available for each case. The data was split in an 80:20 ratio for training the model and testing it. The IARC Cervical Cancer Image Bank curates images sourced from diverse clinical environments. Images gathered retrospectively are eligible for inclusion solely when they conform to rigorous criteria concerning the image collection procedure and quality. A panel of experts meticulously assesses these images prior to their incorporation into the image bank. The process of image collection is a collaborative effort involving colposcopists and adheres to standardized formats. Moreover, access to the image bank is subject to stringent verification protocols, overseen either by duly regulated research teams or by commercial AI developers.

3.3. Evaluation criteria

We use the metrics accuracy, recall, precision and F-1 score to evaluate the performance of the model. Equations (5-8) encompass the mathematical formulations for the evaluation metrics acquired. Here, TP refers to true positive, TN true negative, FP false positive, FN false negative.

$$Accuracy = \frac{TP1+TN1}{TP1+FP1+TN1+FN1} \tag{5}$$

$$Recall = \frac{TP1}{TP1+FN1} \tag{6}$$

$$Precision = \frac{TP1}{TP1+FP1} \tag{7}$$

$$F1-Score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \tag{8}$$

3.4. Segmentation results

Visual noise from the vaginal wall, speculum, and cotton swabs can disrupt the diagnosis in colposcopy images. To increase the accuracy of the diagnosis, preprocessing methods have been applied to extract the region of interest without extraneous noise. Region-based convolution neural network was implemented to isolate the cervical area of cervigrams, followed by resizing the ROI into a predetermined ratio of 512 x 512. The uniform size of the image makes it easy to feed them to the CMT- CNN framework. The average precision of Mask R-CNN was noted to be 93.12 %.

3.5. Classification results

Using the correct type of neural network structure will help to identify important characteristics in images and boost the precision of classification. The same image feature encoding networks were compared against each other by using them to train, verify and test the saline, acetic, and iodine image datasets, respectively. The EfficientNet B7 module has performed with an accuracy of 92.31%, precision of 90.19%, recall of 89.63, and F-1 score of 90.72. The convergence of the CMT-CNN model was monitored by recording the losses from training and validation at each epoch (Fig. 8).

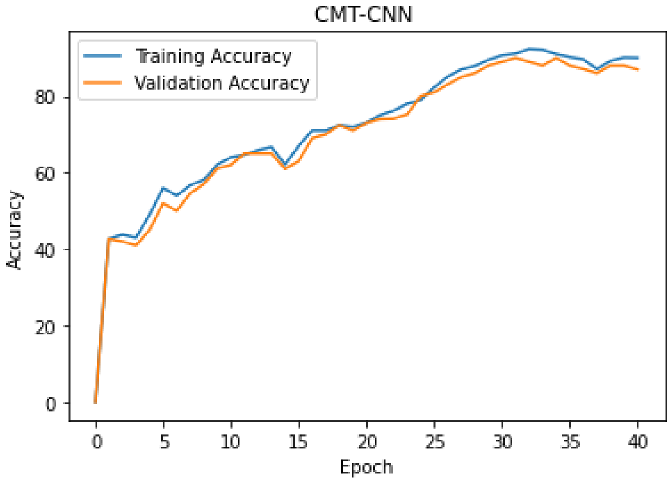

Fig. 8. Training and validation accuracy of CMT-CNN

Table 1 gives a comparison of the current framework's results with respect to state-of-the-art models. The results demonstrate that the CMT CNN model's results are superior to the previous reported works.

Table 1. Comparative analysis of the state of art models with proposed method

Study	Model	Accuracy(%)	Dataset
[47]	Colponet	81.35	Acetic acid image
[23]	BF-CNN	85.50	Acetic acid image and iodine image
[20]	CMF- CNN	92.70	Acetic acid, iodine image and text data.
[48]	ResNet 50	86.80	Acetic acid and text data
[31]	ResNet 50 + EGCN	78.33	5 acetic acid images taken in time series
[49]	MobileNet V	83.33	1 iodine image and 5 acetic acid images
[50]	ResNet 50	84.10	1 iodine image, 1 acetic acid image, 1 text data
Proposed	CMT CNN	92.31	1 saline image, 1 acetic acid image, 1 iodine image with clinical findings.

Table 1 presents a comparative analysis of the state-of-the-art models that attempted to classify cervical intraepithelial neoplasia. To contextualize the achievements of the CMT-CNN model, it is essential to contrast its efficiency with reference to other available models in cervical cancer classification tasks. Table 1 presents a summary of accuracy rates achieved by various models on different datasets.

The CMT-CNN model attained an accuracy of 92.31%.using a dataset consisting of a saline image, an acetic acid image, an iodine image, and clinical findings. This remarkable accuracy underscores the model's effectiveness in classifying cervical lesions. In comparison, the Colponet model [50] achieved an accuracy of 81.35% using only acetic acid images. While this model shows promise, it falls short of the CMT-CNN's performance. The BF-CNN model [24], which utilized acetic acid and iodine images, achieved an accuracy of 85.50%. Although an improvement over Colponet, it still lags behind the CMT-CNN.

The CMF-CNN model [21] stands out with an impressive accuracy of 92.70%, which is on par with the CMT-CNN. However, it's important to note that CMF-CNN utilized text data in addition to iodine images and acetic acid images. This highlights the potential advantages of incorporating diverse data types in cervical cancer classification. The ResNet 50 model , using acetic acid and text data, achieved an accuracy of 86.80%, while the ResNet 50 + EGCN model [33], which utilized a time series of five acetic acid images, achieved an accuracy of 78.33%. MobileNet V , which combined one iodine image with five acetic acid images, achieved an accuracy of 83.33%. Finally, the ResNet 50 model , using one iodine image, one acetic acid image, and text data, achieved an accuracy of 84.10%.

It is evident that a combination of 1 saline image, 1 acetic acid image, 1 iodine image with clinical findings is the input that brought about the best accuracy.

The findings of this study have significant implications for cervical cancer screening and the broader field of medical imaging. Firstly, the CMT-CNN model's accuracy in classifying cervical lesions indicates its potential as a valuable tool for healthcare providers. It has the capacity to assist in the early identification of cervical cancer, enabling timely intervention and improving patient outcomes. Moreover, the incorporation of diverse data types, including images and clinical findings, suggests a new direction in medical image analysis. The success of the CMF-CNN model, which used text data alongside images, underscores the importance of holistic patient information in diagnosis. This methodology holds promise for extension to additional medical imaging applications, enhancing the accuracy of disease detection and patient care.

Additionally, the CMT-CNN model's performance highlights the utility of temporal information in medical imaging. By considering sequential images taken over time, the model can capture dynamic changes in cervical lesions, further improving diagnostic accuracy. This temporal approach may find applications in other areas of medical imaging, where monitoring disease progression is crucial.

3.6. Discussion

In the field of colposcopy, the integration of clinical evidence with colposcopic images has long been the standard practice for cervical cancer diagnosis. While computer-aided diagnosis (CAD) algorithms based on colposcopic images have provided valuable support to medical practitioners, they have not reached a level of comprehensiveness where they can replace the expertise of clinicians entirely. In recognition of this, our current study has attempted to bridge this gap by developing the Colposcopic Multimodal Temporal Convolution Neural Network (CMT-CNN) framework. This innovative approach aims to enhance the precision of cervical lesion diagnosis by amalgamating multimodal clinical data with temporal information derived from the colposcopic image data. Atrous special pyramid-based hybrid classification" is a critical component of our proposed Colposcopic Multimodal Temporal Convolution Neural Network (CMT-CNN) architecture. "Atrous" pertains to atrous convolution, also recognized as dilated convolution, a technique that expands the convolution layer’s receptive field while minimizing the escalation in the number of parameters. Atrous convolutions empower the network to glean insights from a wider context, an advantageous trait, especially in image classification tasks.

The incorporation of multimodal clinical data represents a significant advancement, as it recognizes the intricate interplay of various patient-specific factors in the diagnostic process. By encompassing variables such as age, HPV test results, CIN grade, and histopathology findings, our model strives to provide a more holistic understanding of the health status of the patient. This integration not only contributes to the accuracy of the diagnosis but also aligns with the comprehensive approach adopted by medical practitioners.

Furthermore, the utilization of temporal information extracted from colposcopic images adds another layer of sophistication to our framework. This temporal dimension captures dynamic changes in cervical lesions over time, which can be invaluable in distinguishing benign from malignant lesions. It also allows for the identification of subtle alterations that may elude conventional static image analysis.

By integrating clinical data, our model gains access to valuable information that medical practitioners typically consider when making a diagnosis. For instance, the age of the patient can be a critical factor in assessing the risk of cervical cancer. Additionally, the CIN grade is a source of critical information on the severity of cervical lesions. The fusion of clinical data allows our model to make more precise and personalized diagnoses. It enables the model to consider individual patient characteristics, which can result in tailored and accurate predictions. For example, the presence of high-risk HPV in a younger patient might indicate a different level of concern compared to the same finding in an older patient. Our CMT-CNN framework uses both image features and clinical data to arrive at its final classification. This holistic approach ensures that the model's predictions are not solely reliant on visual cues from colposcopic images. Instead, it combines these cues with patient-specific information, mirroring the decision-making process of medical professionals.

However, it is essential to acknowledge the limitations and potential drawbacks in our experimental design. One notable limitation is the possibility of our approach failing to detect lesions in the cervical canal of patients with type 2 and type 3 transformation zones (TZ). Lack of exposure and visibility of these lesions on colposcopic images can impede their accurate diagnosis. This limitation underscores the need for complementary diagnostic methods and a multidisciplinary approach in challenging cases.

Another aspect that warrants consideration is the relatively limited size of our dataset. The dataset's size can impact the accuracy of feature extraction and classification, potentially limiting the model's ability to generalize to a broader population. Additionally, our study did not encompass an analysis of polyps and stenosis, which are important considerations in cervical health assessment. Future research endeavors should aim to address these limitations by expanding the dataset and exploring the incorporation of additional diagnostic factors.

4. Conclusion

Cervical cancer holds a high burden of malignancy and mortality in developing countries. It is caused by a sexually transmitted virus, HPV, left untreated for long durations, and this cancer does not have a genetic trigger, making it curable. Early diagnosis is the key to treatment and curing cancer completely. Clinicians do not limit their analysis to just the single cervix image when making a diagnosis but should also incorporate the saline image, iodine image, and other clinical data to make a more informed decision. This information can help them determine the best course of action for their patient. This study puts forth a colposcope multimodal temporal fusion model to integrate the cervix images taken in time series along with the clinical data of each case. This approach is complementary to the aggregation of information that doctors use. The CMT CNN framework aggregates all the information a clinician would have to consider to make a diagnosis. A Mask R-CNN was employed for the cervix region of interest extraction. A multimodal temporal fusion-based feature network is employed to integrate image and text features and subsequently classify them into three categories: normal, CIN1,2, and CIN2. This network includes components such as the Atrous Spatial Pyramid Pooling (ASPP), Squeeze-and-Excitation (SE) module, and Convolutional Blocks. The current model performed with a precision of 90.19%, accuracy of 92.31, F-1 score of 90.72, and recall of 89.63, which we consider satisfactory. A broader implication of our findings is that the integration of multimodal information could serve as the bridge that closes the gap between human and machine-based medical diagnosis. Overall, the method not only enables clinicians to make a more informed diagnosis but also serves as a direction to explore multi-stage inputs like adding the pap test images, HPV results etc., as inputs for better diagnostic performance.

Declarations

Author contribution. LM conducted the study and prepared the manuscript, JT supervised the experiment and corrected the manuscript

Funding statement. None.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] A. Znaor, A. Ryzhov, M. Corbex, M. Piñeros, and F. Bray, “Cervical cancer in the Newly Independent States of the former Soviet Union: incidence will remain high without action,” *Cancer Epidemiol.*, vol. 73, p. 2, 2021, doi: 10.1016/j.canep.2021.101944.

[2] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] A. B. Sravani, V. Ghate, and S. Lewis, "Human papillomavirus infection, cervical cancer and the less explored role of trace elements," *Biol. Trace Elem. Res.*, vol. 201, no. 3, pp. 1026–1050, 2023, doi: 10.1007/s12011-022-03226-2.

[4] W. Ahmed *et al.*, "Role of human Papillomavirus in various cancers: epidemiology, screening and prevention," *Mini Rev. Med. Chem.*, vol. 23, no. 10, pp. 1079–1089, 2023, doi: 10.2174/1389557523666230213140641.

[5] M. M. Kalbhor and S. V. Shinde, "Cervical cancer diagnosis using convolution neural network: feature learning and transfer learning approaches," *Soft Comput.*, pp. 1–11, Jul. 2023, doi: 10.1007/s00500-023-08969-1.

[6] M. Gultekin, P. T. Ramirez, N. Broutet, and R. Hutubessy, "World Health Organization call for action to eliminate cervical cancer globally," *Int. J. Gynecol. Cancer*, vol. 30, no. 4, pp. 426–427, Apr. 2020, doi: 10.1136/ijgc-2020-001285.

[7] A. Srinath, F. van Merode, S. V. Rao, and M. Pavlova, "Barriers to cervical cancer and breast cancer screening uptake in low- and middle-income countries: a systematic review," *Health Policy Plan.*, vol. 38, no. 4, pp. 509–527, Apr. 2023, doi: 10.1093/heapol/czac104.

[8] L. Mukku and J. Thomas, "A machine learning model to predict suicidal tendencies in students," *Asian J. Psychiatr.*, vol. 79, p. 103363, 2023, doi: 10.1016/j.ajp.2022.103363.

[9] X. Hou, G. Shen, L. Zhou, Y. Li, T. Wang, and X. Ma, "Artificial Intelligence in Cervical Cancer Screening and Diagnosis.," *Front. Oncol.*, vol. 12, p. 851367, 2022, doi: 10.3389/fonc.2022.851367.

[10] B. Hunter, S. Hindocha, and R. W. Lee, "The Role of Artificial Intelligence in Early Cancer Diagnosis," *Cancers (Basel).*, vol. 14, no. 6, p. 1524, Mar. 2022, doi: 10.3390/cancers14061524.

[11] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, "A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis," *Arch. Comput. Methods Eng.*, vol. 29, no. 4, pp. 2043–2070, Jun. 2022, doi: 10.1007/s11831-021-09648-w.

[12] L. Allahqoli *et al.*, "Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: a systematic review," *Diagnostics*, vol. 12, no. 11, p. 2771, 2022, doi: 10.3390/diagnostics12112771.

[13] C. Liu *et al.*, "Artificial intelligence in cervical cancer research and applications," *Acadlore Trans. AI Mach. Learn.*, vol. 2, no. 2, pp. 99–115, 2023, doi: 10.56578/ataiml020205.

[14] Y. Ming, X. Dong, J. Zhao, Z. Chen, H. Wang, and N. Wu, "Deep learning-based multimodal image analysis for cervical cancer detection," *Methods*, vol. 205, pp. 46–52, 2022, doi: 10.1016/j.ymeth.2022.05.004.

[15] C. Yang, L. Qin, Y. Xie, and J. Liao, "Deep learning in CT image segmentation of cervical cancer: a systematic review and meta-analysis," *Radiat. Oncol.*, vol. 17, no. 1, p. 175, Nov. 2022, doi: 10.1186/s13014-022-02148-6.

[16] Y. Singh, D. Srivastava, P. S. Chandranand, and D. S. Singh, "Algorithms for screening of Cervical Cancer: A chronological review," *Mach. Learn. arXiv*, p. 10, Nov. 2018. [Online]. Available at: https://arxiv.org/abs/1811.00849v1.

[17] M. Lalasa and J. Thomas, "A Review of Deep Learning Methods in Cervical Cancer Detection," in *International Conference on Soft Computing and Pattern Recognition*, 2022, pp. 624–633, doi: 10.1007/978-3-031-27524-1_60.

[18] X. Chen *et al.*, "Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions," *Cancer Med.*, vol. 12, no. 7, pp. 8690–8699, 2023, doi: 10.1002/cam4.5581.

[19] R. Perkins *et al.*, "Comparison of accuracy and reproducibility of colposcopic impression based on a single image versus a two-minute time series of colposcopic images," *Gynecol. Oncol.*, vol. 167, no. 1, pp. 89–95, Oct. 2022, doi: 10.1016/j.ygyno.2022.08.001.

[20] Y. Fan, H. Ma, Y. Fu, X. Liang, H. Yu, and Y. Liu, "Colposcopic multimodal fusion for the classification of cervical lesions.," *Phys. Med. Biol.*, vol. 67, no. 13, Jun. 2022, doi: 10.1088/1361-6560/ac73d4.

[21] J. Kim, C. M. Park, S. Y. Kim, and A. Cho, "Convolutional neural network-based classification of cervical intraepithelial neoplasias using colposcopic image segmentation for acetowhite epithelium," *Sci. Rep.*, vol. 12, no. 1, p. 17228, Oct. 2022, doi: 10.1038/s41598-022-21692-5.

[22] 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, p. 104209, Mar. 2021, doi: 10.1016/j.compbiomed.2021.104209.

[23] L. Yan *et al.*, "Multi-state colposcopy image fusion for cervical precancerous lesion diagnosis using BF-CNN," *Biomed. Signal Process. Control*, vol. 68, p. 102700, Jul. 2021, doi: 10.1016/j.bspc.2021.102700.

[24] Y. Cao *et al.*, "A deep learning-based method for cervical transformation zone classification in colposcopy images," *Technol. Heal. Care*, no. Preprint, pp. 1–12, 2022, doi: 10.3233/THC-220141.

[25] 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, doi: 10.1109/TBME.2018.2887208.

[26] S. Y. Park, D. Sargent, R. Lieberman, and U. Gustafsson, "Domain-Specific Image Analysis for Cervical Neoplasia Detection Based on Conditional Random Fields," *IEEE Trans. Med. Imaging*, vol. 30, no. 3, pp. 867–878, Mar. 2011, doi: 10.1109/TMI.2011.2106796.

[27] T. Xu *et al.*, "Multi-feature based benchmark for cervical dysplasia classification evaluation," *Pattern Recognit.*, vol. 63, pp. 468–475, Mar. 2017, doi: 10.1016/j.patcog.2016.09.027.

[28] T. Chen *et al.*, "Multi-Modal Fusion Learning For Cervical Dysplasia Diagnosis College of Computer Science and Technology Real Doctor AI Research Centre University of Notre Dame Department of Computer Science and Engineering Department of Gynecologic Oncology , Women ' s H," *2019 IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019)*, no. Isbi, pp. 1505–1509, 2019, doi: 10.1109/ISBI.2019.8759303.

[29] W. Li, S. Venkataraman, U. Gustafsson, J. C. Oyama, D. G. Ferris, and R. W. Lieberman, "Using acetowhite opacity index for detecting cervical intraepithelial neoplasia," *J. Biomed. Opt.*, vol. 14, no. 1, p. 014020, 2009, doi: 10.1117/1.3079810.

[30] S. Young Park *et al.*, "Automated image analysis of digital colposcopy for the detection of cervical neoplasia," *J. Biomed. Opt.*, vol. 13, no. 1, p. 014029, Jan. 2008, doi: 10.1117/1.2830654.

[31] Y. Li *et al.*, "Computer-Aided Cervical Cancer Diagnosis Using Time-Lapsed Colposcopic Images," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3403–3415, Nov. 2020, doi: 10.1109/TMI.2020.2994778.

[32] H. Yu *et al.*, "Segmentation of the cervical lesion region in colposcopic images based on deep learning," *Front. Oncol.*, vol. 12, p. 952847, Aug. 2022, doi: 10.3389/fonc.2022.952847.

[33] S. Aggarwal, A. K. Sahoo, C. Bansal, and P. K. Sarangi, "Image Classification using Deep Learning: A Comparative Study of VGG-16, InceptionV3 and EfficientNet B7 Models," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2023, pp. 1728–1732, doi: 10.1109/ICACITE57410.2023.10183255.

[34] Y. Qiu, Y. Liu, Y. Chen, J. Zhang, J. Zhu, and J. Xu, "A2SPPNet: attentive atrous spatial pyramid pooling network for salient object detection," *IEEE Trans. Multimed.*, 2022, doi: 10.1109/TMM.2022.3141933.

[35] M. Patacchiola, J. Bronskill, A. Shysheya, K. Hofmann, S. Nowozin, and R. Turner, "Contextual squeeze-and-excitation for efficient few-shot image classification," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36680–36692, 2022. [Online]. Available at: https://arxiv.org/abs/2206.09843.

[36] D.-F. Shen, J.-J. Guo, G.-S. Lin, and J.-Y. Lin, "Content-aware specular reflection suppression based on adaptive image inpainting and neural network for endoscopic images," *Comput. Methods Programs Biomed.*, vol. 192, p. 105414, Aug. 2020, doi: 10.1016/j.cmpb.2020.105414.

[37] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, and X. Zheng, "Unsupervised-Learning-Based Continuous Depth and Motion Estimation With Monocular Endoscopy for Virtual Reality Minimally Invasive Surgery," *IEEE Trans. Ind. Informatics*, vol. 17, no. 6, pp. 3920–3928, Jun. 2021, doi: 10.1109/TII.2020.3011067.

[38] C. P. N. Khuong *et al.*, "Rapid and efficient characterization of cervical collagen orientation using linearly polarized colposcopic images," *J. Innov. Opt. Health Sci.*, p. 2241001, 2022, doi: 10.1142/S1793545822410012.

[39] A. D. Magaraja *et al.*, "A Hybrid Linear Iterative Clustering and Bayes Classification-Based GrabCut Segmentation Scheme for Dynamic Detection of Cervical Cancer," *Applied Sciences (Switzerland)*, vol. 12, no. 20. p. 14, 2022, doi: 10.3390/app122010522.

[40] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, 2022, doi: 10.1016/j.ins.2022.11.139.

[41] Z. YANG, X. PENG, Q. ZHU, and Z. YIN, "Image segmentation algorithm with adaptive attention mechanism based on deeplab v3 plus," *J. Comput. Appl.*, vol. 42, no. 1, p. 230, 2022. [Online]. Available at: http://www.joca.cn/EN/10.11772/j.issn.1001-9081.2021010137.

[42] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: a survey," *Comput. Intell. Pattern Recognit.*, pp. 657–668, 2020, doi: 10.1007/978-981-13-9042-5_56.

[43] T. Cheng, X. Wang, L. Huang, and W. Liu, "Boundary-preserving mask r-cnn," in *European conference on computer vision*, 2020, pp. 660–676, doi: 10.1007/978-3-030-58568-6_39.

[44] J. Shang, K. Zhang, Z. Zhang, C. Li, and H. Liu, "A high-performance convolution block oriented accelerator for MBConv-Based CNNs," *Integration*, vol. 88, pp. 298–312, 2023, doi: 10.1016/j.vlsi.2022.10.012.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[46] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Computers in Biology and Medicine*, vol. 128. p. 39, 2021, doi: 10.1016/j.compbiomed.2020.104115.

[47] 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, p. 15, 2020, doi: 10.1007/s00138-020-01063-8.

[48] L. Liu *et al.*, "Computer-aided diagnostic system based on deep learning for classifying colposcopy images," *Ann. Transl. Med.*, vol. 9, no. 13, pp. 1045–1045, Jul. 2021, doi: 10.21037/atm-21-885.

[49] C. Buiu, V.-R. Dănăilă, and C. N. Răduţă, "MobileNetV2 ensemble for cervical precancerous lesions classification," *Processes*, vol. 8, no. 5, p. 595, 2020, doi: 10.3390/pr8050595.

[50] C. Yuan *et al.*, "The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images," *Sci. Rep.*, pp. 1–12, 2020, doi: 10.1038/s41598-020-68252-3.