TelsNet: temporal lesion network embedding in a transformer model to detect cervical cancer through colposcope images

Lalasa Mukku [a,1,*], Jyothi Thomas [a,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 24, 2023
Revised September 13, 2023
Accepted November 1, 2023
Available online November 30, 2023

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
Transformer architecture
Deep learning
Cervical cancer
Colposcopy
Lesions

ABSTRACT

Cervical cancer ranks as the fourth most prevalent malignancy among women globally. Timely identification and intervention in cases of cervical cancer hold the potential for achieving complete remission and cure. In this study, we built a deep learning model based on self-attention mechanism using transformer architecture to classify the cervix images to help in diagnosis of cervical cancer. We have used techniques like an enhanced multivariate gaussian mixture model optimized with mexican axolotl algorithm for segmenting the colposcope images prior to the Temporal Lesion Convolution Neural Network (TelsNet) classifying the images. TelsNet is a transformer-based neural network that uses temporal convolutional neural networks to identify cancerous regions in colposcope images. Our experiments show that TelsNet achieved an accuracy of 92.7%, with a sensitivity of 73.4% and a specificity of 82.1%. We compared the performance of our model with various state-of-the-art methods, and our results demonstrate that TelsNet outperformed the other methods. The findings have the potential to significantly simplify the process of detecting and accurately classifying cervical cancers at an early stage, leading to improved rates of remission and better overall outcomes for patients globally.

1. Introduction

Cervical cancer has become a major health hazard for women worldwide, with its high mortality and morbidity rates [1]. The majority of fatalities occur in underdeveloped and developing countries [2]. Unlike other types of cancers that are genetically triggered, the cause of cervical cancer is known to be human papillomavirus (HPV) [3]. Cervical cancer can be completely cured if identified in its early stages [4]. In 2018, the World `Health Organization (WHO) urged global countries to work towards eradicating cervical cancer [5]. Globalized uniform cervical cancer screening can be a potential step toward achieving this goal [6]. Cervical cancer can be nipped off in the bud altogether through systematic screening and swift intervention.

There are certain limitations of identifying cervical cancer manually. First and foremost, there is the problem of interobserver variability [7]. It means that same colposcope image examined by different clinicians will have varied diagnostic decisions given by each of them. The concordance of diagnosis for colposcope images is 65% [8]. Reducing interobserver variability in colposcopy diagnosis is essential for improving the reliability of cervical cancer screening and early detection [9], [10]. Standardized training,

the use of technology to aid the decision making of the clinicians can be a step towards achieving the same. In addition to the variability problem, in many regions, especially low-income and remote areas, there is a severe shortage of skilled healthcare professionals, including gynecologists and pathologists, who are trained to accurately interpret cervical cancer screening results [10]. This scarcity of expertise can lead to delayed diagnoses and inadequate follow-up care. Hence, computational support for the existing clinicians with modest experience and boost the diagnostic accuracy. Also, traditional diagnostic methods, like Pap smears and VIA, have limitations in terms of sensitivity and specificity. Therefore, choosing colposcope test as the deciding examination for cervical malignancy identification can benefit in multiple angles like reducing the cost burden associated with pap smear test, overcoming the impediment of limited training of the clinicians and most importantly reducing the turn around time of final diagnostic decision.

Artificial intelligence (AI) assisted cancer screening [11] has gained notable traction in the past two decades, and cervical cancer diagnosis has benefitted remarkably from AI solutions [12]. Several researchers have embodied deep-learning solutions for cervical cancer detection through medical imaging. Images range from pap smear, colposcope, magnetic resonance imaging (MRI), and computerized tomography (CT) [13]. Singh *et al.* [14] published a chronological review of the deep learning methods in cervical cancer screening. The outcome of the survey ascertains that deep learning CAD solutions are a bridge to developing automatic screening of cervical cancer. Colposcopy examination is a pivotal tool for cervical cancer screening that offers a greater degree of accuracy than the human papillomavirus (HPV) and Thin-Prep cytologic test (TCT) tests [15]. During the colposcopy examination, a 5% acetic acid solution is topically administered to the cervical region to accentuate cancerous characteristics [16]. Subsequently, a colposcope is utilized to capture detailed images of the cervix, where lesions become conspicuously visible within a few minutes following acetowhite application. In some cases, the cervix images are captured in a time series fashion with saline, acetic acid, and Lugol's iodine application. The classification of colposcopy images is primarily employed for diagnostic purposes, aiming to discern between benign lesions and low/high squamous intraepithelial lesions or cervical intraepithelial neoplasia (CIN) [17] or cervical intraepithelial neoplasia (CIN). Clinician's experience and expertise are the basis of diagnostic accuracy in traditional colposcope exams, a scarce resource in many low-income countries. There are insufficient experienced specialists to accommodate the number of patients who need screening. Parallelly, several researchers are investigating the use implementation 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. This research aims to create 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.

A transformer architecture is a neural network that was initially developed keeping in mind the natural language processing tasks (NLP). Nevertheless, it has worked well in the image classification problem sector. Integrating a transformer model in computer vision, which assesses the attention weights of specific local regions in an image, has the potential to enhance image classification tasks. This is because the model can direct its attention towards the most relevant areas of an image, allowing it to better capture subtle differences and fine-grained details that may be crucial for accurate classification. The transformer model is applied to an image classification task to calculate and assess the attention weights of the regions of interest in an image. In this approach, the image is divided into smaller regions, or "patches", and each patch is treated as a sequence of pixels. The transformer is then applied to these patch sequences to calculate attention weights that indicate the relative importance of each patch for the classification task.

A colposcope image frequently contains extraneous elements such as background noise and unwanted objects like vaginal walls and speculum [18]. 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. In this paper we use an enhanced gaussian mixture model (GMM) to precisely segment the cervix region of interest which can further be given as inputs to the subsequent classification module. Gaussian mixture model is a probabilistic model that represents the data as a

mixture of multiple Gaussian distributions. This is a statistical model used to describe the distribution of data in multiple dimensions. In the context of image segmentation, each pixel in the image is treated as a data point with multiple attributes. The key parameters of the gaussian mixture modelling are the eigen vector, covariance matrix, and mixture coefficients. Usually, these parameters are obtained with expectation-maximization optimization. However, in this study we used a nature-based metaheuristic optimizer for tuning the parameters of GMM. Hyperparameter tuning is a crucial step in the process of training machine learning models. It involves finding the best set of hyperparameters for a given model and dataset to optimize its performance. Hyperparameters are parameters that are not learned from the data but are set prior to training. In order to extract the optimal hyperparameters, a Mexican axolotl algorithm is used in this study. Metaheuristic optimization refers to a class of optimization algorithms that are used to find the best solution to a problem without guaranteeing optimality. The Mexican Axolotl Algorithm is one such optimization technique. Mexican axolotl works by an opposition-based learning strategy. Opposition-Based Learning (OBL) is a machine learning strategy or optimization technique that involves considering both the positive (conventional) and negative (opposite) solutions when searching for optimal solutions in a search space. The concept behind OBL is to use opposition or contrast to improve the exploration of the solution space and enhance the performance of optimization algorithms.

The evolution of lesions is dynamic between the saline, acetic acid, and iodine induced cervix images. To capture the same, the acetowhite lesion recognition is converted to a fine-grained visual classification problem. In order to extract a fine-grained feature of the image, attention weights for specific local regions are obtained by utilizing a transformer model. These attention weights, which denote the significance of various parts of the image, are computed by the transformer model during its analysis of the input dataset. By acquiring attention weights for designated regions of the image, we can pinpoint the areas that are the most pertinent to the feature we are attempting to extract. This enables us to identify the most crucial elements of the image and thus obtain a more precise and accurate fine-grained feature. Hence, we put together a CNN embedded in a transformer to extract highly accurate local lesion features that can solve the over-segmentation problem faced by previous models. The framework takes in the cervix images in small sections and generates latent features of the same. These latent features are integrated with information about where the lesion is located, thus making the features more informative. These features are sent as input to the proposed network, and subsequently, the attention-based model generates and learns the weights of lesion features. Based on the weight assignment, the lesion area is marked for model performance. As the last step, the features and attention weights are optimized by a metaheuristic loss model. 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.

Artificial intelligence (AI) is playing an increasingly important role in medical image processing. AI algorithms can be used to automatically analyze and interpret medical images, which can help clinicians with diagnostic decision making [19]. The application of AI in gynecological cancer research [20] has gained traction in the past couple of decades. Diagnosis of endometriosis [21] vaginal cancer [22], ovarian abnormalities [23], uterine cancer [24], cervical cancer [25] and vulvar cancer [26] have all been benefited by machine learning and deep learning models. A significant amount of research is aimed at segmenting and classifying colposcope images [27]. Fan *et al.* [28] 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. Yan *et al.* [29] designed a BFCNN, a bilinear fuse convolutional neural network for the segmentation and classification of cervigrams. Yuzhen Cao [15] developed a multiscale feature fusion classification network to classify cervical transformation zone and reported an accuracy of 88.49% with 90.12% sensitivity. Asiedu *et al.* [30] used machine learning methods of using boundary boxes to extract ROI and classify the region through support vector machines.

Park *et al.* [31] used anatomical maps with texture and color 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.* [32] 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.* [33] 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.* [34] created a computer-generated diagnostic program based on an AW opacity index, which yielded a diagnosis with 84% specificity and 88% sensitivity. Authors of [35] 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. 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 states of information, like sequences of cervix examination images.

As an extension to the above approach, Li *et al.* [36] built a convolutional network with graph and edge features (E-GCN) and noted a 78.33% accuracy from using time series image features. Perkins *et al.* [17] contradicted these findings by fusing 17-time series colposcope images. The study reported no meaningful increase in accuracy after analyzing the performance of 17 fused images. It provides a scope to ponder over the speculation of the possibility of adding non-image information to meaningfully increase classification accuracy. Peng *et al.* [28] analyzed multimodal feature changes by building a multistate convolution neural network with an extension of genetic algorithm optimization technique. They declared 86.3% accuracy. Parallelly, Yinuo Fan *et al.* [37] built a multimodal fusion colposcopic convolutional neural network (CMF-CNN) that made use of Squeeze-and-Excitation fusion to combine to achieve 92.70% accuracy. 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.* [36] approached the time series imaging problem by building a graph convolutional network with edge features (E-GCN) to fuse sequential images of the cervigrams (images captured at 60s, 90s, 120s, 150s) and achieved an accuracy of 78.33. A bird's eye view of the results seen 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.

This paper proposes a transformer model (TelsNet) embedded with 3D CNN to extract the latent features and their weights corresponding to the acetowhite lesion region. In addition to that, the paper presents a preprocessing technique unique to colposcope images. The contributions of this paper are as follows :

- A preprocessing mechanism using a Gaussian mixture model is proposed to segment the whole image to remove the noise and artifacts in the cervix image.
- A novel specular reflection removal model is proposed through bi-dimensional histogram decomposition and Laplacian transformation.
- This is the first experiment using a transformer model that is embedded with 3-D CNN to extract the latent features and their weights corresponding to the acetowhite lesion region.
- A nature inspired meta-heuristic optimization algorithm is proposed to enhance the performance of the model to reach convergence.
- The proposed model is evaluated on a sequential cervix image dataset obtained from international archives for research in cancer (IARC)

The remainder of the paper is structured as follows: Section 2 presents the methodology of TelsNet transformer architecture. Section 3 discusses the results and comparative analysis of the performance of the model. Section 4 concludes the study.

2. Method

The schematic architecture of the current study is given in Fig. 1. The architecture of the proposed model contains a 3-dimensional CNN embedded transformer module and a metaheuristic optimization module for the GMM based segmentation.

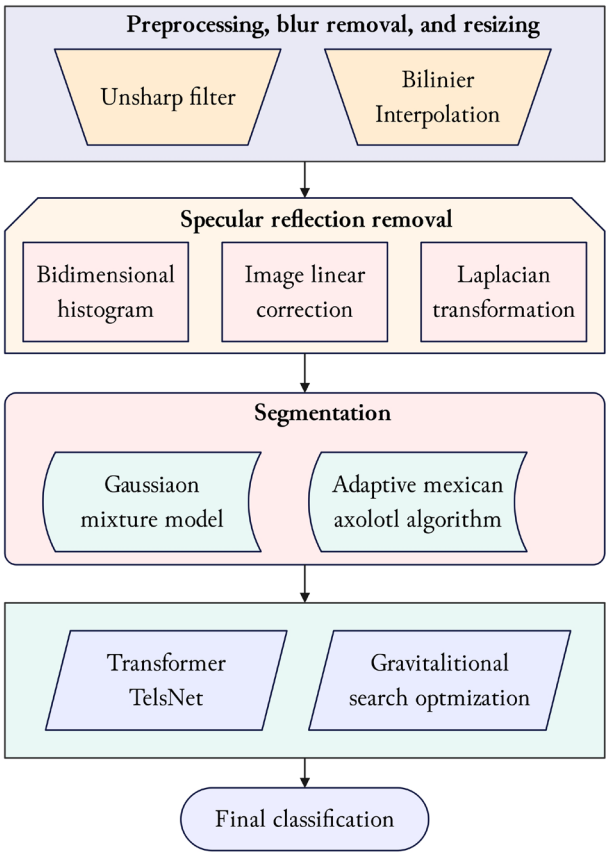


Fig. 1. Schematic architecture of the proposed model

2.1. Preprocessing

Preprocessing an image before inputting it into a deep learning model is a routine practice. Medical images typically have issues like contrast enhancement, blur, medical artifact involvement etc. In the current model, we address two problems: unblur and resizing of the image.

2.1.1. Unsharp filter for blur removal

The colposcope images are typically blurred to an extent due to the bodily movements of the patient. In order to remove the blur, a noise removing machine learning filter called unsharp filter is employed. It is an image enhancement technique frequently employed to sharpen the images. The filter works by creating and subtracting a blurred version of the original image from the original image. This is accomplished by applying a Gaussian blur kernel [38] to the original image through convolution, which has the effect of diminishing the high-frequency details present in the image. Once the blurred image is subtracted from the original image using equation in Fig. 2, a scaling factor is used to add the high pass image to the original image resulting in a non-blurred high-quality image without any loss of information. Fig. 3 shows the blurred image and the resulting image after applying the unsharp mask.

Input: Input image X, Standard deviation of gaussian kernel (σ Sigma), and Scaling factor 'k'

(output)

Step 1: Generate a gaussian kernel of size $(2\omega + 1)^*(2\omega + 1)$ for pixel (i, j), $Y(i, j) = \sum_{\mu=-\omega}^{\omega} \sum_{\mu=-\omega}^{\omega} X(i + \mu, j + v)G(\mu, v)$

Step 2: Perform convolution operation on the image with the gaussian kernel, $f(x, y) * G(x, y) = \sum_i \sum_j f(x, j)g(x - i, y - j)$

Step 3: Obtain the high pass image by subtracting the obtained blurred image from the original image using $A_G(x'y) = A(x'y) - A_B(x'y)$

Step 4: Multiply the image with scaling factor 'K'

Step 5: Add the scaled image to original image to obtain the clearer version of the image without blur.

Fig. 2. Algorithm for unsharp mask

Original blurred images

Unblurred images

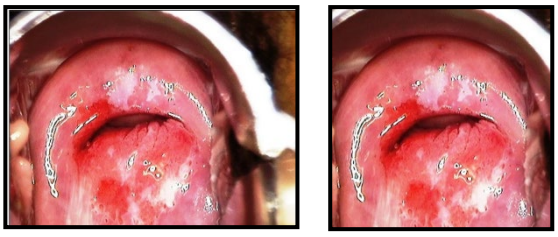
Fig. 3. Unsharp mask blur removal

2.1.2. Bilinear interpolation for resizing

The images are resized to a uniform size of 512 x 512 using this technique. The bilinear interpolation algorithm calculates the location of each new pixel by dividing the location of the original pixel by the scaling factor using:

$$F(x, y) = (1 - x)(1 - y)P_{00} + (1 - y)P_{10} + (1 - x)P_{01} + xyP_{11} \quad (1)$$

For example, if the original image is being scaled down by a factor of 2, the new image will have half as many pixels in each dimension. Fig. 4 demonstrates the resizing of the cervix images after using bilinear interpolation



Original image Cropped image

Fig. 4. Images after resizing to 512 x 512 uniform size using bilinear interpolation

2.2. Specular reflection removal

Acetowhite lesions (AW) are abnormalities that appear white or pale in color when viewed under post-acetic acid application to the uterine cervix. In the context of cervix images, acetowhite lesions may be indicative of precancerous changes in the tissue [39]. 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 [40].

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 critical role in determining 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. 5)

Fig. 5. Specular reflections on the cervix surface

Specular reflections obstruct the efficient analysis of cancerous changes of surface regions [40]. 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. In the detection phase, Generally, the image is transformed into a different color space to facilitate further processing of the region of interest (ROI). For instance, the image formats used are RGB [30], grey-level [41], HSV, HSI [42] and a threshold value to identify the SR. Subsequently, the removed pixels are replaced with inpainting to preserve the image morphology.

2.2.1. 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 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 equation (2).

$$m = \frac{1}{3}(b + g + r) \tag{2}$$

where 'm' stands for pixel intensity.

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

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

Here, '*s*' denotes saturation, and (*r, g, b*) = (red, green, blue). Two important threshold values (m_{max}, S_{max}) 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 (3)

2.2.2. Specular reflection removal

Image linear correction is a simple and effective way usually employed to improve the quality of an image 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 pixels must be replaced in such a way that the information of the cervix image is preserved. Routinely, the SR pixels are replaced with the mean of pixels surrounding the pixel that needs to be replaced.

2.2.3. Inpainting of deleted specular pixels

The Laplacian equation is a partial differential equation that describes the behaviour 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 (4).

$$F(s) = \int_0^\infty f(t)e^{-st}t' \tag{4}$$

2.3. Segmentation Using Enhanced Gaussian Mixture Model

A colposcope image typically includes the surrounding organ and medical equipment interferences in addition to the cervix region. In order to efficiently implement the transformer deep learning model, the image needs to be segmented to rid the extraneous noise. Hence, we have used a multivariate gaussian mixture model enhanced with an adaptive mexican axolotl algorithm. Colposcope image show as Fig. 6.

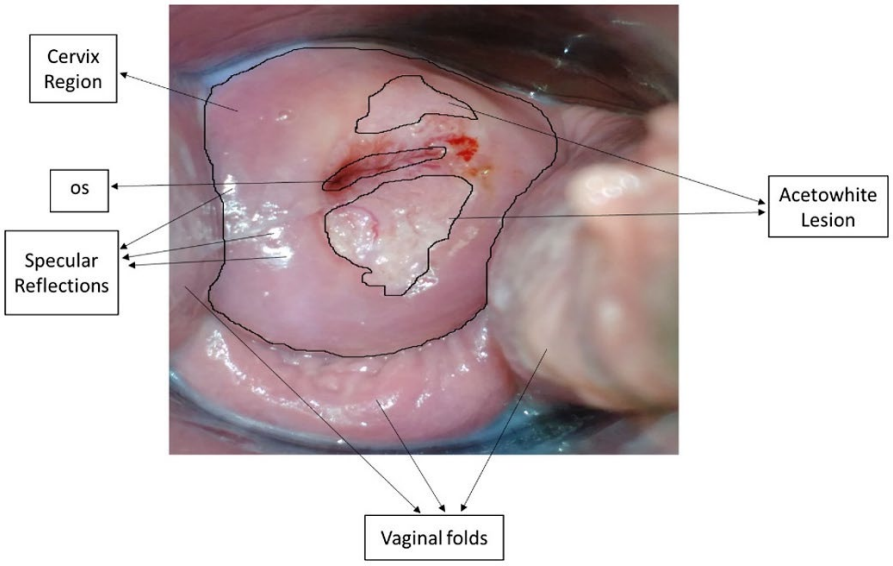


Fig. 6. Colposcope image

Multivariate gaussian distribution is an extension of the univariate model that can fit vectors, which are the pixels in this case. It is a probabilistic model that represents the data as a mixture of multiple Gaussian distributions. This is a statistical model used to describe the distribution of data in multiple

dimensions. In the context of image segmentation, each pixel in the image is treated as a data point with multiple attributes. X is an input vector with 'd' values. The distribution is parameterized by mean μ (a length' d' vector) and a covariance matrix Σ (d x d matrix). Subsequently, the equation of the probability density function is given by:

$$P(x;\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (5)$$

Where:

$\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}$ is a constant that ensures the integral value is 1.

(x- μ): row vector; Σ^{-1}: d x d matrix ; $(x-\mu)^T$ is a column vector; μ: weighted mean, which is a length 'd' row vector; Σ: variance, a d x d matrix ; |Σ|: matrix determinant.

The updated mean $\hat{\mu}$ is the imperial average of the pixel vectors in X. It is calculated by $\hat{\mu} = \frac{1}{m}\sum_j x^j$. The covariance matrix Σ is also the empirical average of the vector product calculated by $\hat{\Sigma} = \frac{1}{m}\sum_j(\underline{x}^{(j)} - \underline{\hat{\mu}})^T(\underline{x}^{(j)} - \underline{\hat{\mu}})$. The product of (x- μ), Σ^{-1}, $(x-\mu)^T$ gives a scalar number which is the probability or likelihood of the value 'x' belonging in the cluster k. These parameters were optimized using a metaheuristic algorithm called Mexican axolotl. The segmented cervix image is presented in Fig. 7.

Fig. 7. Preprocessed and segmented cervix image

2.4. Adaptive mexican axolotl optimization algorithm

The hyperparameters of the Gaussian mixture described above are the mean vector value, mixture coefficient value, and the covariance matrix's eigenvector. These values are computed using a nature-based metaheuristic algorithm, Mexican axolotl [43]. A meta-heuristic optimization algorithm is an algorithm that searches for near-optimal solutions to a given problem by using a combination of heuristic techniques. These algorithms are useful for solving complex problems where an exact solution is not practical or possible. Meta heuristics are not problem-specific but rather provide a framework to generate solutions to any given problem. Mexican axolotl works by an opposition-based learning strategy. The algorithm of Mexican axolotl is given below:

Here;

$\hat{\pi}_k$ →Mixture coefficient

$\hat{\mu}_k$ → Mean vector

$\hat{\Sigma}_k$ → Covariant matrix of k^{th} component

Step 1: initializing the solution: In the Mexican axolotl algorithm, the initial position of each particle in the population also greatly influences the evaluation of the population. The initial solution is created from the parameters: normalized mixture coefficient, d^{th} dimension of the mean vector, and the d^{th} Eigenvalue. Initially, the values are assigned randomly. The population of this algorithm is defined as follows:

$$P = \{A_1, A_2, \ldots, A_{np}\} \quad A_n \in P, 1 \leq n \leq np \tag{6}$$

Np represents the population size, and 'A' is the axolotl (solution), which is described as such:

$$A_k = \{\rho_k \ \mu_{k1} \ldots \mu_{kD} \ \lambda_{k1} \ldots \lambda_{kD}\} \tag{7}$$

Here, ρ_k represents the mixture coefficient, μ_{kd} which is yet un-normalized, represents the mean vector of d^{th} dimension, and the d^{th} Eigenvalue is represented by λ_{kd}. The ρ_k (mixture coefficient) value is selected between the range [0,1], the value of μ_{kd} is chosen to be between $\lfloor x_{\min,d}, x_{\max,d} \rfloor$, and the value of λ_{kd} (eigenvector) is selected between $[0, \eta]$, where η is the sample largest Eigenvalue from the covariance matrix is determined from the image. $x_{\min,d}$ and $x_{\max,d}$ denote the minimum and maximum values of this image.

Step 2: Opposite solution generation: Subsequently, for every solution initialized, alternative solutions are created. The opposite solution A'_i can be deducted as:

$$A'_i = x + y - A_i \tag{8}$$

where $A_i \in [x, y]$ is a real number.

Step 3: Fitness calculation: once the solution is initialized, for every solution, the fitness is calculated. The maximum likelihood (MLE) function is taken as the function to calculate fitness. The maximum likelihood function enhances the resulting segmentation accuracy. The fitness is given in equation (9)

$$Fitness = max\left(\sum_{i=1}^{N} log\left\{\sum_{k=1}^{K} \hat{\pi}k.fN\left(X_i|\hat{\pi}k, \widehat{\Sigma_k}\right)\right\}\right) \tag{9}$$

As stated in (8), the best parameter value is chosen as the one with the highest fitness score. If this is not the case, then the solution is adjusted in the subsequent step.

Step 4: Updating the solution space using AMAO: The three steps that make up this algorithm are transitioning, injury and reviving, reproducing, and sorting.

The most well-adapted male axolotl, denoted as, is determined by the fitness of the solution and the transition parameter within the range of 0 to 1. This male axolotl changes the coloration of its body parts in accordance with the value of,

$$X_{nm} \leftarrow X_{nm} + (X_{best,m} - X_{nm}) \bullet y \tag{10}$$

Similarly, female axolotls are identified as progressing from larvae to adults to highly adapted females through the use of equation 10. The female axolotl is represented by Y_n.

$$Y_{nm} \leftarrow Y_{nm} + (Y_{best,m} - X_{nm}) \bullet y \tag{11}$$

A number rand between 0 and 1 is chosen randomly to decide which individuals to pick for random transition. Additionally, the inverse probabilities of transition for female and male axolotls are calculated by

$$pX_n = \frac{F(X_n)}{\Sigma F(X_n)} \tag{12}$$

$$pY_n = \frac{F(Y_n)}{\Sigma F(Y_n)} \quad (13)$$

$F(Y_n)$ and $F(X_n)$ represent the fitness quotient in female and male axolotls.

The random transition of individuals will occur as (13) and (14). This situation holds true under the condition of rand being less than the inverse probability value.

$$X_{nm} \leftarrow min_m + (max_m - min_m) \times rand_m \quad (14)$$

$$Y_{nm} \leftarrow min_m + (max_m - min_m) \times rand_m \quad (15)$$

Injury & Restoration: When walking on the water, axolotls can be at risk for accidents and injuries. This risk has been taken into account during the healing and rehabilitation stages.

$$P_{nm} \leftarrow min_m + (max_m - min_m) \times rand_m \quad (16)$$

Reproduction and Assortment: For female axolotls, a male is selected through a process of competition in order to produce offspring. The male axolotl will deposit sperm, which the female will then coat and place into the sperm to create an egg containing genetic material from both parents. This pair will generate two eggs. Afterward, the female will lay the eggs and watch for them to hatch. The hatchlings then compete with their parents for survival using a fitness function. If the young axolotls are more adapted compared to their parents, they are able to take their place.

Step 5: Termination: The aforementioned steps are iterated until achieving the optimal solution or the initial value. Alternatively, if this condition is not met, the algorithm will be concluded. The chosen value is then applied to the encryption process

2.5. TelsNet embedding

A dataset containing one saline, acetic acid, and iodine images pertaining to each patient is inputted into the model, and the latent features are developed. The original image is divided into smaller patches of size $s \times s$, and then these patches are flattened to create a sequence of images denoted by x_p. The rate of change between the sequential images is not highly significant. In the event of attempting to analyze an image sequence by projecting it into latent features using a linear or convolutional layer, we may miss out on the spatial relationships between adjacent image patches. This is because these layers typically process individual patches in isolation without considering their context. This could result in the latent features being incomplete and failing to capture the full meaning and details of the original image sequence. The proposed TelsNet has the ability to consider both the temporal and spatial dimensions of the sequential image frames when performing convolution operations.

The temporal dimension refers to the time-based aspect of the series of images, while the spatial dimension refers to the visual aspects of the individual images. By taking both of these dimensions into account, the TelsNet model can analyze how the visual elements of the images change over time, and how these changes relate to the lesion evolution over time. We take the key acetic acid frame use the TelsNet model to analyze the feature dynamics between the image patch and its adjacent patches, thus projecting the latent features into vector subspace. The model is given by

$$I_{p\,x\,q} = F_{TelsNet}\left(\sum X_{p\times m\times n}\,\omega + b\right) \quad (17)$$

$P = 3$ (image sequence), q is the 3-dimensional feature generated by the network, N, and M are the width and length of the images (512x512), b and ω are parameters of the network.

2.6. Transformer embedding

Multi-head self-attention technique is used to identify relationships between different elements in a sequence of data. Multi-head self-attention builds by incorporating multiple attention mechanisms, or "heads," that operate in parallel to evaluate the different elements. Each attention head processes the

input sequence independently, generating its own set of attention weights. These weights are then combined to capture complex relationships between the different elements in the sequence. In the current problem, each image patch is assigned a weight based on its importance, as determined by the attention mechanism. Subsequently, the network classifies the attention weights. The current network is built using a transformer encoder to solve the cervix image classification and recognition problem.

Due to the mild changes in epithelium, the changes in acetowhite lesions are sometimes very minimal, making it hard to capture the change. However, the remaining cervix region does not change and remains the same, thus directing the multi-head self-attention transformer to target the dynamically evolved patch areas. When training models with attention mechanisms, some image patches may be assigned a low weight, indicating that they are less relevant to the task at hand. To help the model converge more efficiently and learn the attention weights of each patch more accurately, these low-weight patches are excluded from the training process. By focusing only on the most important patches during training, the model is able to learn the attention weights more effectively and achieve better overall performance. This technique, known as dropout, improves the efficiency and accuracy of the model by prioritizing the patches that are most relevant to the task at hand. The architecture of the proposed model is shown in Fig. 8.

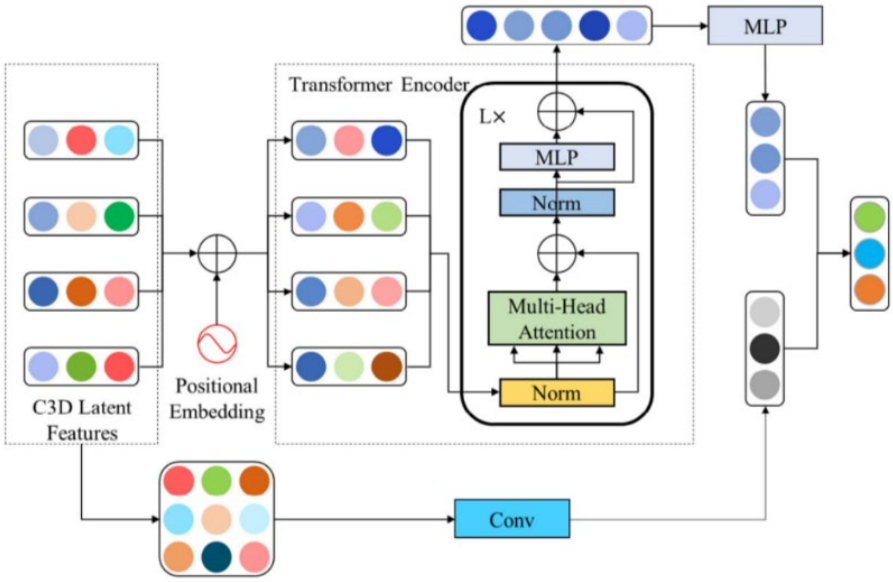


Fig. 8. Encoder of Transformer architecture used in TelsNet. The output of the model, embedding space is fused with positional embedding into the transformer encoder containing MLP blocks for output classes

Working of the TelsNet:

- Image sequencing: TelsNet takes a sequence of cervix images captured over time as input. These images include saline, acetic acid, and iodine images, which accentuate different features of the cervical region.
- Patch level analysis: The original image is divided into smaller patches or regions. These patches are then flattened to create a sequence of images, denoted as xp. These patches represent local areas within each image
- Transformer model: It calculates and assesses attention weights for the regions of interest within the images.
- Attention mechanism: The transformer model uses a multi-head self-attention mechanism to identify relationships between different elements in the sequence of data. Each image patch is

assigned a weight based on its importance, as determined by the attention mechanism. These weights represent the significance of various parts of the image.

- Temporal convolution: In parallel with the transformer-based model, TelsNet uses temporal convolutional neural networks (CNNs) to analyze how the visual elements of the images change over time.
- Integration: The features obtained from both the transformer-based model and the temporal CNNs are integrated to provide a comprehensive understanding of the images' spatial and temporal aspects.
- Lesion recognition: The model utilizes the weights of the lesion features based on their importance. This weight assignment helps in marking the lesion area for model performance.
- Metaheuristic Optimization: To optimize the model's features and attention weights, a metaheuristic loss model, possibly the "gravitational search algorithm" mentioned earlier, is used to ensure convergence and improve the model's performance.

Feature vector f_{pxq} generated by TelsNet is given as input to the encoder. A classification token $f_{c,q}$ is generated and fused with f_{pxq}. In order to enable a transformer-based model to take positional information into account during training, a learnable positional embedding called f_{pos} is associated with each image patch. This embedding is then combined with the patch data and fed into the transformer encoder model. The process of combining the positional embedding with the patch data is referred to as weighted fusion. This step involves weighing the importance of the patch data and the positional embedding relative to each other based on the requirements of the task being performed. Fig. 9 demonstrates the layout of the layers in the model.

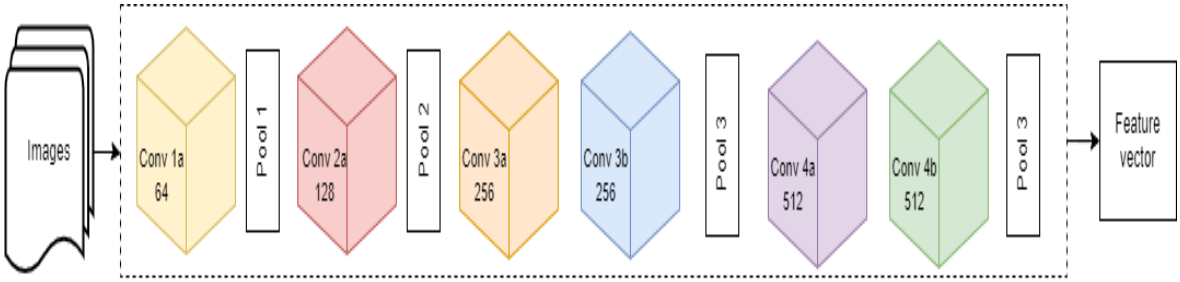


Fig. 9. Embedding architecture of TelsNet, the images are split into acetowhite, saline, and iodine patches and are projected into embedding space. The architecture contains six convolution and four pooling layers

The resulting new feature is a representation of both the image patch and its position in the sequence. By incorporating positional information in this way, the model is able to better understand the relationships between different patches in the sequence and achieve greater accuracy and effectiveness in its analysis. The feature equation is given by:

$$f_{(p-1)xq} = [f_{c,q}, f_{1,q}, f_{2,q}, .. f_{p,q}] + f_{pos} \quad (18)$$

Here, number of patches is denoted by p, $f_{c,q}$ is a class token, $f_{1,q}$ is the embedded patch. projection, f_{pos} denotes the position embedding. The class token $f_{c,q}$ is used to learn the attention weight of the lesion projection.

To enable a transformer model to encode the position of image lesions, it needs to know the order relationship of each patch in the sequence. The TelsNet embedding can provide this information because it captures both the temporal and spatial dimensions of the image patches. By using the TelsNet embedding to extract features that encode the relationships between each image lesion patch and its neighboring patches, we can ensure that the transformer model receives accurate positional information. This allows the model to better understand the context and structure of the image sequence, leading to

more accurate and effective analysis. The positional embedding f_{pos} is inputted into the transformer for conducting a weighted fusion with the new features. Lastly, the feature vectors that are generated by the transformer encoder with features produced by TelsNet are fused for cervix lesion identification.

$$y' = SoftMax(F_{LR}([f_{att}, f_{conv}])) \quad (19)$$

$$f_{att} = F_{MLP}(f_L) \quad (20)$$

$$f_i = F_{MLP,i}(F_{MSA,i}(f_{i-2})) \quad (21)$$

$$f_{conv} = F_{conv}(f_{pxq}) \quad (22)$$

Here, FLR stands for a fully connected layer, f_{att} denotes the attention layer, f_0 is the input feature, the encoder is denoted by $f_{(p-1)xq}$ and f_{conv} is the convolution layer.

The loss given by the metaheuristic GSA is 0.106, which is considered to be satisfactory and better than the traditional loss functions used in traditional networks. The value given by the GSA implicates that the network model is optimized with desired convergence.

The gravitational search algorithm (GSA) [44] is a type of optimization algorithm that uses Newton's gravitational law as its basis. In this algorithm, each agent is treated as an object, and gravitational forces act on them, causing all objects to move towards those with heavier masses, which represent the optimal solution in the search space [45]. The position of agents, which corresponds to a potential solution to the problem being solved, is updated repeatedly until a termination condition is met.

3. Results and Discussion

This section presents a comprehensive analysis of the results achieved and their significance in diagnosing cervical cancer through segmentation and classification of colposcope images.

3.1. Dataset

The dataset is downloaded from the International Agency for Research on Cancer (IARC) as a part of the study. Sequential colposcope images of 200 patients were collected after the application of saline, 5% acetic acid, and lugol's iodine in that order. A total of 916 images were collected, owing to the repeated duplicates of pictures of the cervix. These images are divided in 80-20 ratio for training the model and testing it. This is a standard practice in machine learning models.

The images are augmented by using transformation techniques like flip, rotate, etc., to increase the volume of the dataset, which is a crucial aspect for attaining higher and more reliable accuracy. Fig. 10 shows the transformation operations on a single image.

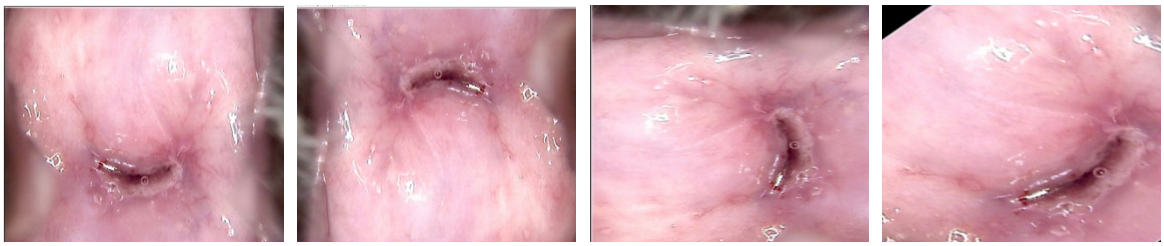


Fig. 10. Cervix images flipped, rotated, and transformed during augmentation

3.2. Experimental environment

The current segmentation framework is implemented on Google Colab in Python, with the system having 16GB of RAM with an Intel Core i7 processor and the Windows 10 operating system with 256 SSD. The efficiency of the recommended method is evaluated using a range of popular performance metrics.

3.3. Evaluation metrics

To evaluate the proposed framework, the following components are used: accuracy, sensitivity, specificity, Jaccard Index and dice coefficient.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (23)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (24)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad (25)$$

$$Jaccardindex = \frac{|S_g^1 \cap S_t^1|}{|S_g^1 \cup S_t^1|} = \frac{TP}{TP+FP+FN} \quad (26)$$

$$DiceCoefficien = \frac{2|S_g^1 \cap S_t^1|}{|S_g^1|+|S_t^1|} = \frac{2TP}{2TP+FP+FN} \quad (27)$$

Where S is the region of overlap. If S_g intersection S_t is empty, then $J(S_t,S_g) = 0$

Accuracy measures the overall correctness of a diagnostic model's predictions, reflecting the proportion of correctly classified cases (both positive and negative) out of the total cases evaluated. High accuracy provides confidence in the model's ability to make correct predictions. This is crucial for medical practitioners when making treatment decisions based on the model's output. Incorrect diagnoses can lead to inappropriate treatments or delayed interventions, potentially endangering patient safety. High accuracy reduces the risk of misdiagnosis and its associated consequences. On the other hand, sensitivity, also known as recall, measures the model's ability to correctly identify true positive cases among all actual positive cases. In cervical cancer, sensitivity is crucial for early detection. It ensures that the model can identify cases of cancer (or its markers) even at the earliest stages, when intervention is most effective. High sensitivity reduces the likelihood of false negatives, which occur when the model fails to detect actual cases of cancer. A single false negative diagnosis carries a substantial burden, particularly in the context of cancer. The ramifications of a missed cancer diagnosis are profound, as individuals may unknowingly continue their daily lives while the disease continues to progress within their bodies. From both moral and ethical standpoints, diagnostic support systems must prioritize the attainment of the highest possible sensitivity.

3.4. Results

This section discusses the results attained for segmentation and classification

3.4.1. Segmentation Results

Comparative analysis of the enhanced gaussian mixture model for segmentation is carried out with two state-of-the-art clustering methods, k-means, and univariate gaussian mixture model. It is noted that the proposed method consistently outperformed the baseline methods. Fig. 11 shows the segmentation outline, while Table 1 and Table 2 display the segmentation results.

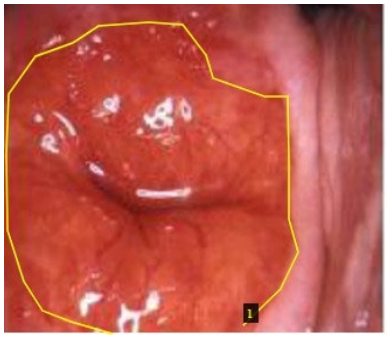
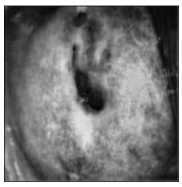
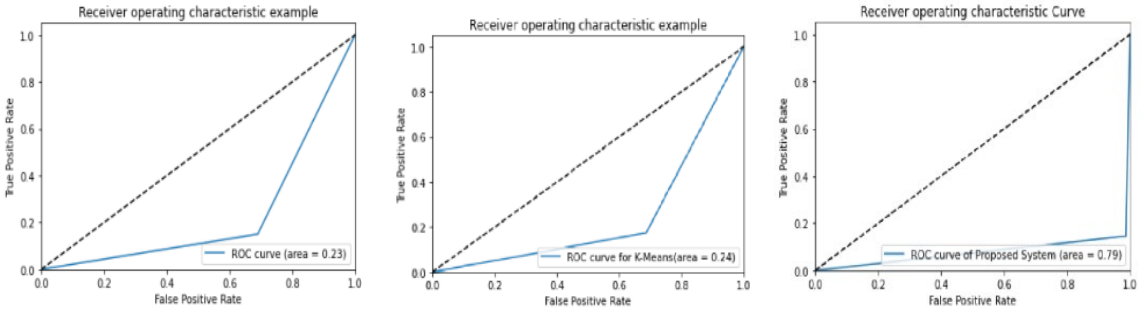


Fig. 11. Segmentation process

Table 1. Demonstrated

Cervix image	Metric	K-means	GMM	EGMM (Proposed)
	Accuracy	0.611	74.80	0.961
	Loss	0.631	0.621	0.173
	Specificity	0.321	0.580	0.916
	Sensitivity	0.841	0.7652	0.979
	Jaccard index	0.389	0.421	0.869
	Dice score	0.834	0.571	0.939

As demonstrated in Table 1, in comparison with K- means and GMM models, the traditional GMM achieved a better accuracy of 74.80%. However, it has been further enhanced as EGMM is improved with its parameters optimized through AMAO. The current model, EGMM, demonstrated an accuracy of 96.1%. Parallelly, the proposed framework displayed a sensitivity of 97.9%, specificity of 91.6%, dice score of 0.939, Jaccard index of 0.869, and menial loss of 0.173. The results suggest that the methodology proposed in this study produced better outcomes compared to the existing state-of-the-art approaches. This can be attributed to the utilization of the AMAO algorithm for optimizing parameters with the EGMM. Fig. 12 displays the ROC curve.



(a) (b) (c)

Fig. 12. ROC curve for above discussed (a) GMM, (b) K-means, and (c) EGMM

3.5. Comparative analysis

The proposed model is compared with four state-of-the-art pre trained IMAGENET models, namely, AlexNet, VGG16, ResNet50 using transfer learning techniques. The final layer is frozen to three classes to depict 'Normal', 'pre-cancer', and 'cancer' class labels. The details of the pre-trained models are given below. Additionally, the model is evaluated with respect to other proposed models to affirm the results achieved are the best.

3.5.1. AlexNet

AlexNet is a convolutional neural network architecture that has gained popularity due to its success in the ImageNet Large Scale Visual Recognition Challenge. It utilizes key building blocks such as max pooling, convolutions, and dense layers to extract features from input images. The architecture of the model is comprised of a total of eight layers, with each set of learnable parameters consisting of five convolutional layers that incorporate both fully connected and max pooling layers. Additionally, the model includes two normalizing layers and one softmax layer. Each layer in the architecture is composed of a convolutional layer paired with an activation function that utilizes the rectified linear unit (ReLU). The AlexNet model performed with an accuracy of 0.702. Fig 13 shows the accuracy plot of the AlexNet model.

3.5.2. ResNet 50

ResNet50 is a convolutional neural network comprising of 50 layers, among which there are 48 fully connected layers, additionally with a max pool layer and an average pool layer. Over top of that, it has the capability of performing floating-point calculations of more than 3.8 × 109. The ResNet50 model is

designed with a unique approach that utilizes convolutional filters of various sizes, addressing the challenge of degradation commonly found in deep CNN models. This approach has also contributed to faster training times. With 48 fully connected layers, a max pool layer, and an average pool layer, ResNet50 can process floating-point calculations with efficiency. Furthermore, the model's utilization of a limited number of filters translates to even quicker performance. ResNet architecture takes multiples of 32x32 dimensions of height, width and channel. ResNet model gave an accuracy score of 0.8462 which is the highest among the transfer learning models used in this study. The accuracy plot of ResNet is shown in Fig. 14.

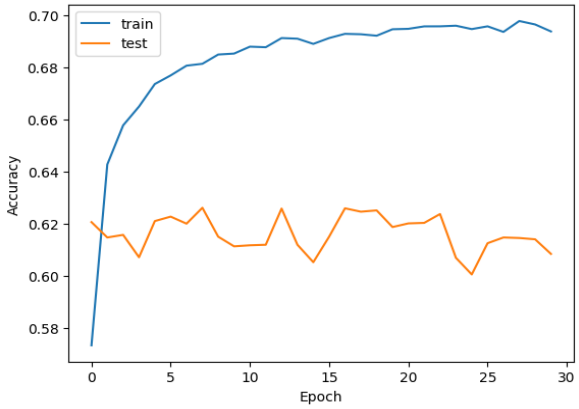


Fig. 13. Accuracy vs epoch plot for AlexNet

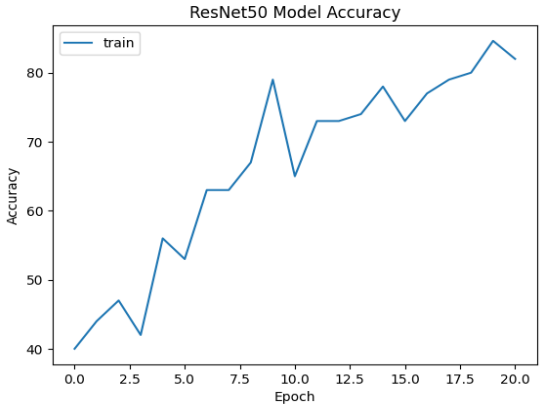


Fig. 14. Accuracy vs. epoch plot for ResNet

3.5.3. VGG 16

VGG16 is a convolutional neural network architecture that has gained popularity due to its deep structure and use of small 3x3 convolutional filters. The model comprises 16 layers, consisting of 13 convolutional layers and 3 fully connected layers. In addition to that, VGG16 incorporates max pooling layers and dropout regularization to mitigate the risk of overfitting. For classification tasks, the final layer of the model is often a SoftMax layer. VGG16 has proven to be effective in a variety of computer vision tasks, such as image classification, object detection, and segmentation. VGG 16 model has achieved an accuracy score of 0.8174 (Fig. 15), making it second best among the transfer learning models used in the current study.

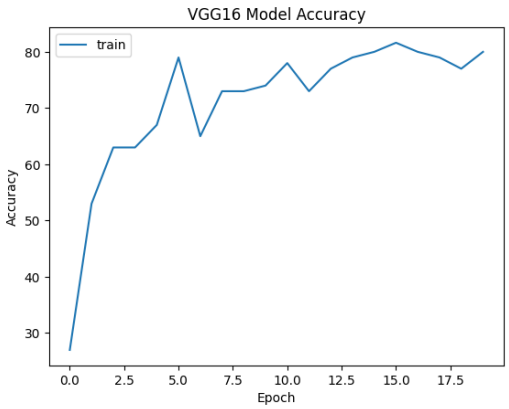


Fig. 15. Accuracy vs. epoch plot for VGG16

3.5.4. Comparative analysis with state-of-the-art models

In the context of cancer diagnosis, the impact of a false negative surpasses that of a false positive. The aforementioned findings unmistakably demonstrate that our proposed technique produced superior results than previous techniques. Table 2 shows the results in terms of accuracy, sensitivity, and specificity.

Table 2. Comparative analysis of TelsNet with ImageNet models

Model	Accuracy (%)	Sensitivity	Specificity
AlexNet	70.02	-	-
ResNet	84.62	-	-
VGG16	81.74	-	-
Fuzzy clustering segmentation [46]	91.2	78.2	-
Unimodal deep learning model [47]	76.3	-	-
K means with deep learning [28]	86.3	84.1	89.8
BFCNN and factorized bilinear pooling [29]	85.5	74.6	-
Multiscale feature fusion classification network [15]	88.49	-	-
Proposed TelsNet	92.7	73.4	82.1

TelsNet model attained 92.7% accuracy with 73.4% sensitivity and 82.1 specificity. Fig. 16 demonstrates the model's behavior over the training and validation accuracy. TelsNet's successful performance and high diagnostic accuracy hold immense promise for its integration into clinical practice. With its ability to accurately identify cervical cancer-related lesions in colposcope images, TelsNet can serve as a valuable decision support tool for clinicians. This integration would streamline and enhance the diagnostic process, ensuring that patients receive timely and accurate assessments of their cervical health.

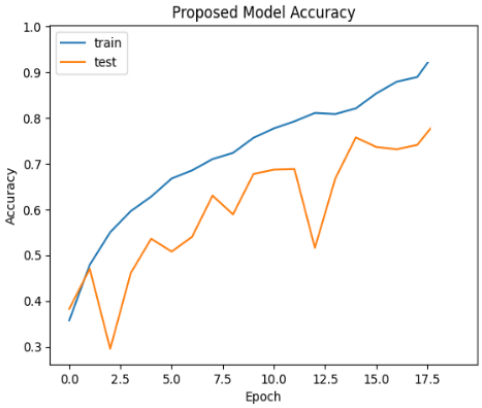


Fig. 16. Performance of the proposed TelsNet

TelsNet's potential influence on cervical cancer diagnosis extends beyond improved consistency. Its high accuracy, as demonstrated in our experiments, indicates that it has the potential to detect cervical cancer at an early stage with remarkable sensitivity and specificity. This implies that TelsNet can contribute significantly to the early identification of cervical malignancies, enabling timely interventions and potentially life-saving treatments.

Moreover, TelsNet's efficiency in processing colposcope images and generating rapid diagnostic assessments can reduce the turnaround time for results. Quicker diagnosis means that patients can receive follow-up care and treatment promptly, further enhancing their chances of positive outcomes. TelsNet can serve as a valuable tool in the hands of clinicians, offering them a second opinion and aiding in more accurate decision-making. While the ultimate diagnostic decision will still be made by a healthcare professional, TelsNet's support can enhance their confidence in the diagnosis and provide additional insights into lesion characteristics.

3.6. Discussion

Cervical cancer remains a significant global health concern, particularly in underdeveloped and developing countries, where high mortality and morbidity rates persist. Early intervention is pivotal for complete remission and cure, and global uniform cervical cancer screening is a step toward achieving this goal. However, manual diagnostic methods have several limitations that hinder early detection and accuracy. In this study, we developed a deep learning framework, incorporating advanced deep learning techniques, to enhance the diagnosis of cervical cancer using colposcope images.

The use of deep learning techniques, including transformer-based models and GMM segmentation with hyperparameter tuning, holds significant implications for cervical cancer diagnosis. This innovative model has the potential to revolutionize cervical cancer diagnosis by reducing interobserver variability, enabling early detection, expediting the diagnostic process, and enhancing the overall quality of cervical health assessments. With further validation and integration into clinical practice, TelsNet can make a substantial contribution to cervical cancer prevention and patient care.

By leveraging transformer-based models, we achieve improved accuracy and sensitivity in feature extraction, allowing for more precise lesion identification. This special focus on segmentation was given based on the limitations pointed out by the previous models. These models reviewed in introduction have pointed out that the deep learning approaches so far have overlooked the independent segmentation process and proceeded with automatic segmentation. In order to bridge this gap a novel segmentation was developed. The use of the Mexican Axolotl Algorithm for hyperparameter tuning enhances the optimization of GMM parameters, further improving segmentation accuracy. This innovative model has the potential to revolutionize cervical cancer diagnosis by reducing interobserver variability, enabling early detection, expediting the diagnostic process, and enhancing the overall quality of cervical health assessments. With further validation and integration into clinical practice, TelsNet can make a substantial contribution to cervical cancer prevention and patient care.

While our approach demonstrates promise in enhancing cervical cancer diagnosis, it is not without limitations. Although the sensitivity of the segmentation step is 97.9% the sensitivity of classification is modest compared to the same. As a cancer diagnosis support, we aim to mitigate the burdensome false negatives. As future work we plan to increase the sensitivity of the framework while preserving the accuracy and specificity. Another noteworthy shortcoming of the model is that the classifier heavily relies on the segmentation module. In real world point of care low cost colposcopes do not capture ideal cervix images [30]. This leaves scope for exploring possible solutions to solve the problem of automated segmentation processes.

4. Conclusion

Cervical cancer ranks as the fourth most prevalent malignancy among women, presenting a substantial threat to women's health globally due to its elevated mortality rates. Automated colposcopy image analysis represents a pivotal stride towards the large-scale screening for cervical cancer. This paper introduces a pioneering approach for the early detection of cervical cancer by leveraging colposcope image analysis. Our proposed model, TelsNet, employs a transformer-based neural network combined with temporal lesion convolutional neural networks to identify cancerous regions within the images. To enhance model accuracy, we incorporated diverse preprocessing techniques, such as unsharp filters, bilinear interpolation, and bi-dimensional histogram equalization. We also used Gaussian mixture models to segment the images and identify regions of interest. Experimental results revealed that TelsNet achieved an exceptional accuracy of 92.7%, with a sensitivity of 73.4% and a specificity of 82.1%. Comparative evaluations against state-of-the-art methods showcased the superior performance of TelsNet. By automating colposcope image analysis, TelsNet has the potential to significantly enhance the efficiency and precision of cervical cancer screening, providing invaluable support to medical professionals

Declarations

Author contribution. LM conducted the study and prepared the manuscript; JT supervised the experiment and corrected the manuscript.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Ethical Statement: The authors bear responsibility for every facet of the work, ensuring that any queries concerning the accuracy or integrity of any portion of the work are thoroughly investigated and addressed.

Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

International Agency for Research on Cancer (IARC) dataset will be made available upon request for research and education purposes.

References

[1] P. Xue *et al.*, "Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis," *npj Digit. Med.*, vol. 5, no. 1, p. 19, Feb. 2022, doi: 10.1038/s41746-022-00559-z.

[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. C. Rodriguez *et al.*, "Longitudinal Study of Human Papillomavirus Persistence and Cervical Intraepithelial Neoplasia Grade 2/3: Critical Role of Duration of Infection," *JNCI J. Natl. Cancer Inst.*, vol. 102, no. 5, pp. 315–324, Mar. 2010, doi: 10.1093/jnci/djq001.

[4] X. Chao *et al.*, "Efficacy of different surgical approaches in the clinical and survival outcomes of patients with early-stage cervical cancer: protocol of a phase III multicentre randomised controlled trial in China," *BMJ Open*, vol. 9, no. 7, p. e029055, Jul. 2019, doi: 10.1136/bmjopen-2019-029055.

[5] 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.

[6] 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.

[7] A. Traverso *et al.*, "Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients," *Radiother. Oncol.*, vol. 143, pp. 88–94, Feb. 2020, doi: 10.1016/j.radonc.2019.08.008.

[8] F. A. Stuebs *et al.*, "Concordance Rate of Colposcopy in Detecting Cervical Intraepithelial Lesions," *Diagnostics*, vol. 12, no. 10, p. 2436, Oct. 2022, doi: 10.3390/diagnostics12102436.

[9] L. Pleș, J.-C. Radosa, R.-M. Sima, R. Chicea, O.-G. Olaru, and M.-O. Poenaru, "The Accuracy of Cytology, Colposcopy and Pathology in Evaluating Precancerous Cervical Lesions," *Diagnostics*, vol. 12, no. 8, p. 1947, Aug. 2022, doi: 10.3390/diagnostics12081947.

[10] C. Nakisige *et al.*, "Artificial intelligence and visual inspection in cervical cancer screening," *Int. J. Gynecol. Cancer*, vol. 33, no. 10, pp. 1515–1521, Oct. 2023, doi: 10.1136/ijgc-2023-004397.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Y. Cao *et al.*, "A deep learning-based method for cervical transformation zone classification in colposcopy images," *Technol. Heal. Care*, vol. 31, no. 2, pp. 527–538, Mar. 2023, doi: 10.3233/THC-220141.

[16] 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. Heal. Informatics*, vol. 25, no. 9, pp. 3529–3540, Sep. 2021, doi: 10.1109/JBHI.2021.3064366.

[17] 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.

[18] C. P. N. Khuong *et al.*, "Rapid and efficient characterization of cervical collagen orientation using linearly polarized colposcopic images," *J. Innov. Opt. Health Sci.*, vol. 16, no. 05, p. 16, Sep. 2023, doi: 10.1142/S1793545822410012.

[19] M. P. Recht *et al.*, "Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations," *Eur. Radiol.*, vol. 30, no. 6, pp. 3576–3584, Jun. 2020, doi: 10.1007/s00330-020-06672-5.

[20] D. P. Mysona *et al.*, "Applying Artificial Intelligence to Gynecologic Oncology: A Review," *Obstet. Gynecol. Surv.*, vol. 76, no. 5, pp. 292–301, May 2021, doi: 10.1097/OGX.0000000000000902.

[21] S. Guerriero *et al.*, "Artificial intelligence (AI) in the detection of rectosigmoid deep endometriosis," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 261, pp. 29–33, Jun. 2021, doi: 10.1016/j.ejogrb.2021.04.012.

[22] PDQ Adult Treatment Editorial Board, "Vaginal Cancer Treatment (PDQ®): Patient Version," *PDQ Cancer Information Summaries*, 2002. [Online]. Available at: https://www.cancer.gov/types/vaginal/patient/vaginal-treatment-pdq.

[23] M. Akazawa and K. Hashimoto, "Artificial Intelligence in Ovarian Cancer Diagnosis," *Anticancer Res.*, vol. 40, no. 8, pp. 4795–4800, Aug. 2020, doi: 10.21873/anticanres.14482.

[24] M. Toğaçar, "Detection of segmented uterine cancer images by Hotspot Detection method using deep learning models, Pigeon-Inspired Optimization, types-based dominant activation selection approaches," *Comput. Biol. Med.*, vol. 136, p. 104659, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104659.

[25] 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.

[26] S. M. Fragomeni *et al.*, "2022-RA-1299-ESGO How to predict preoperative risk of lymph node metastasis in vulvar cancer patients the Morphonode Predictive Model," in *Vaginal and vulvar cancer*, Oct. 2022, vol. 32, no. Suppl 2, p. A445.2-A446, doi: 10.1136/ijgc-2022-ESGO.961.

[27] B. Bai, P.-Z. Liu, Y.-Z. Du, and Y.-M. Luo, "Automatic segmentation of cervical region in colposcopic images using K-means," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 4, pp. 1077–1085, Dec. 2018, doi: 10.1007/s13246-018-0678-z.

[28] 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.

[29] 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.

[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, doi: 10.1109/TBME.2018.2887208.

[31] 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.

[32] 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.

[33] T. Chen *et al.*, "Multi-Modal Fusion Learning For Cervical Dysplasia Diagnosis," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019, vol. 2019-April, pp. 1505–1509, doi: 10.1109/ISBI.2019.8759303.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Y.-Q. Liu, X. Du, H.-L. Shen, and S.-J. Chen, "Estimating Generalized Gaussian Blur Kernels for Out-of-Focus Image Deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 829–843, Mar. 2021, doi: 10.1109/TCSVT.2020.2990623.

[39] 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.

[40] X. Wang *et al.*, "Integration of Global and Local Features for Specular Reflection Inpainting in Colposcopic Images," *J. Healthc. Eng.*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/5401308.

[41] 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.

[42] 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.

[43] Y. Villuendas-Rey, J. L. Velázquez-Rodríguez, M. D. Alanis-Tamez, M.-A. Moreno-Ibarra, and C. Yáñez-Márquez, "Mexican Axolotl Optimization: A Novel Bioinspired Heuristic," *Mathematics*, vol. 9, no. 7, p. 781, Apr. 2021, doi: 10.3390/math9070781.

[44] A. Hashemi, M. Bagher Dowlatshahi, and H. Nezamabadi-Pour, "Gravitational Search Algorithm," in *Handbook of AI-based Metaheuristics*, Boca Raton: CRC Press, 2021, pp. 119–150, doi: 10.1201/9781003162841-7.

[45] Y. Wang, S. Gao, Y. Yu, Z. Cai, and Z. Wang, "A gravitational search algorithm with hierarchy and distributed framework," *Knowledge-Based Syst.*, vol. 218, p. 106877, Apr. 2021, doi: 10.1016/j.knosys.2021.106877.

[46] J. Liu, L. Li, and L. Wang, "Acetowhite region segmentation in uterine cervix images using a registered ratio image," *Comput. Biol. Med.*, vol. 93, pp. 47–55, Feb. 2018, doi: 10.1016/j.compbiomed.2017.12.009.

[47] 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.