

Adaptive Cross-Attention Fusion of Spatial–Frequency Features with Hierarchical Transformers for Cervical Cancer Classification

N. Chamundeeswari¹ and *R. Ramachandran²

¹Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, Tamilnadu, India.

²Assistant Professor/Programmer, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India.

*Corresponding Author
R. Ramachandran

Article History

Received: 08.09.2025

Revised: 08.10.2025

Accepted: 15.10.2025

Published: 30.10.2025

Abstract:

Cervical cancer is a global health issue that requires precise and timely detection techniques to enable efficient clinical decision-making. In this paper, we introduce a new deep learning paradigm for classification of cervical cancer image into six classes based on spatial and frequency-domain features. The input images are first improved using preprocessing operations like CLAHE, normalization, resizing, augmentation, and balancing. In addition, Discrete Wavelet Transform (DWT) is used to obtain frequency sub-bands that supplement the information in the spatial domain. Next, a dual-stream convolutional neural network (CNN) is used for extracting structural and textural features that are combined through an adaptive cross-attention block with learnable fusion weights. This allows for dynamic representation of spatial-frequency dependencies. The combined representation is then subjected to further refinement via a hierarchical transformer encoder, which is intended to learn local cellular patterns and global tissue-level dependencies. Lastly, a dense classification head with dropout predicts the stage of cervical cancer. The model is trained with an 80:20 train-test split, optimized with AdamW, and a composite loss function consisting of cross-entropy, focal loss, and attention consistency loss. Experimental outcomes prove that the proposed approach performs better accuracy, precision, recall, F1-score, and AUC than traditional baselines. Also, interpretability is guaranteed through Grad-CAM visualization for CNN streams, providing improved clinical explainability. This framework offers a robust and interpretable method for automatic cervical cancer diagnosis.

Keywords: Cervical Cancer Classification, Dual-Stream CNN, Discrete Wavelet Transform (DWT), Adaptive Cross-Attention Fusion, Hierarchical Transformer Encoder, Explainable Deep Learning, Attention Consistency Loss, Spatial-Frequency Feature Fusion.

INTRODUCTION

Cervical cancer continues to be among the most prevalent causes of cancer-related deaths in women globally, especially in low- and middle-income nations where access to frequent screening and diagnostic facilities is not readily available. The World Health Organization estimates that over half a million new cervical cancer cases and over 300,000 cervical cancer deaths occur every year, with rates being much higher in resource-limited areas. Early identification and proper classification of cervical precancerous and cancerous lesions are significant in minimizing disease burden and enhancing survival. Conventional cytology-based screening strategies, including Pap smears and colposcopy, have been largely practiced; these methods are, however, subject to subjectivity, inter-observer variations, and reliance on experienced pathologists [1]. To transcend these issues, computational pathology and artificial intelligence (AI)-based diagnostic systems are being studied more and more for automated cervical cancer classification.

Over the past few years, medical image classification tasks have been performed with impressive results using deep learning-based image analysis methods. Convolutional Neural Networks (CNNs) have, in particular, been used with great success for cervical

abnormality detection from colposcopic and histopathology images [2][3]. But current CNN-based approaches mainly deal with spatial image details, which usually constrain them from detecting subtle frequency and textural patterns that are discriminative in histopathology. Current work emphasized the significance of combining frequency and spatial features for enhancing classification performance in medical imaging [4]. In addition, although CNNs are best at local feature extraction, they tend to be poor in modeling global contextual dependencies, which are important to differentiate between morphologically similar cervical tissue subtypes.

While Transformer-based models have made a recent surge in popularity in medical imaging for their capacity to model long-range dependencies, the majority of current works make use of plain CNNs or single-stream Transformers, missing out on the synergistic nature of interacting spatial and frequency-domain representations. In addition, interpretability and explainability are still significant gaps in research on cervical cancer classification, since clinicians require not just valid predictions but also a clear understanding of the decision-making process [5]. Some explainable AI methods have been pursued to that extent, i.e., attention-based colposcopic image classifiers [6], but they do not generalize across heterogeneous patient

populations and datasets since they are based on limited feature modalities.

Spurred by these deficiencies, the current research puts forward a new dual-stream approach that utilizes both frequency and spatial information for cervical cancer classification. While the spatial stream is intended to capture structural and morphological information, the frequency stream operates on Discrete Wavelet Transform (DWT)-derived sub-bands to yield fine-grained texture and frequency features [7]. To properly integrate these disparate representations, we propose an Adaptive Cross-Attention Fusion (ACAF) mechanism, where cross-attention adaptively aligns frequency and spatial features, and trainable fusion weights modulate their relative influences based on the input. This is coupled with a Hierarchical Transformer Encoder, which encodes both global and local dependencies in the fused feature space, thus improving contextual comprehension [8].

The key contributions of this research are summarized as follows:

- Dual-stream feature representation – A new spatial–frequency dual-stream CNN model is proposed, whereby improved images undergo CNN-S (spatial) and DWT-based sub-bands undergo CNN-F (frequency) for all-around feature representation.
- Adaptive cross-attention fusion – ACAF is proposed to simultaneously align and fuse the spatial and frequency features, discovering interdependencies, and adaptively weighing their significances.
- Hierarchical transformer encoding – Fused features are encoded using a hierarchical multi-head self-attention transformer to represent both fine-grained and global contextual dependencies.
- Robust classification head with hybrid loss – A hybrid loss function involving cross-entropy, focal loss, and attention consistency loss is used, optimized with AdamW and weight decay along with sophisticated training techniques.
- Exhaustive examination and explainability – The developed model is systematically tested on a six-class cervical cancer dataset (80% training, 20% testing), and performance is indicated using accuracy, precision, recall, F1-score, and AUC. Visualization methods such as Grad-CAM and attention map analysis allow for the explainability of the model's outputs.

Through the combination of spatial–frequency dual-stream learning with hierarchical transformer encoding and adaptive cross-attention, this work overcomes the limitations of current cervical cancer classification approaches. The model improves not only predictive accuracy but also clinical trust through interpretability, towards AI-assisted precision diagnosis in cervical oncology.

2. Review of Literature

Umay Yadav et al. [9] introduce an intelligent machine learning-based cervical cancer detection system utilizing traditional algorithms. The work outlines the usability of decision trees, SVM, and random forests in medical image analysis. Even though their performance is competitive for small datasets, the absence of deep feature representation constrains the model to be scaled to larger and more diversified cervical image datasets. Rashmi Ashtagi et al. [10] take a machine learning pipeline to predict cervical cancer, with a focus on traditional predictive modeling techniques with minimal dependency on deep learning. They conclude that even though traditional methods provide computational complexity, they are beaten by deep learning and Transformer-based models in accuracy and generalization. He et al. [11] investigate texture and morphological features using SVM classifiers for the classification of histopathology images. Their contribution highlights the importance of manually crafted features in medical image processing. Even though deep learning methods have mostly surpassed classical techniques, this work emphasizes that classical techniques hold significance in resource-poor settings where deep networks cannot be employed. Abinaya and Sivakumar [12] introduce a 3D CNN fused with a Vision Transformer for cervical cancer classification. Their method takes advantage of volumetric feature extraction utilizing the 3D CNN and contextual learning improvement with the Transformer. The work shows enhanced classification accuracy over traditional CNN-based models, especially in the ability to extract local and global dependencies within Pap smear datasets. Nonetheless, the computational expense of employing 3D CNNs will restrict real-time application in clinical environments.

Mohammed et al. [13] examine SWIN-Transformers combined with CNNs for the diagnosis of early cervical cancer. According to their findings, there are dramatic improvements in feature extraction, particularly in addressing histopathological variability between samples. The Transformer backbone enables improved global context modeling over CNN-only designs. However, the research highlights the imperative to have high-volume datasets to utilize the maximum generalization capability of Transformer-based networks. Xue Feng et al. [14] suggest an advanced Vision Transformer model for cervical pre-cancer classification. It is an improvement over typical ViTs in that it includes attention optimisation techniques to more adeptly capture cell morphology and tissue-level variation. The outcomes indicate large improvements in classification accuracy. While robust, the method is computationally expensive, which could limit its deployment in low-resource clinical settings. Zhang et al. [15] describe a ConvNeXt-based method for high-accuracy classification of cervical precancerous lesions. Their method greatly improves feature learning over

conventional CNNs, particularly in dealing with intricate morphological patterns. The work shows the increasing power of next-generation CNN models for medical imaging applications. Shurong Niu et al. [16] propose a dual-module hybrid feature fusion model for lesion detection and classification in cervical cytology images. Their approach blends hand-designed features and deep learning representations to achieve strong performance at lesion severity discrimination. The combined approach emphasizes the value of combining complementary modalities. While encouraging, the method might necessitate lengthy preprocessing procedures, which may limit its ability to scale on various datasets.

The TelsNet model [17] proposes a Temporal Lesion Network embedding within a Transformer model for cervical cancer detection. The methodology specifically addresses temporal lesion development by incorporating temporal relationships into the classification process. The model outperforms others in lesion progression tracking, paving the way for a new direction in the incorporation of temporal data in cervical cancer research. But the dependency on sequential image sets might limit its use in standard cytology or histopathology analysis. Sreelatha and Shivashetty [18] introduce a deep ensemble learning architecture with uncertainty-guided prediction ranking for Pap smear image classification. It uses an ensemble of multiple deep learning models to enhance robustness and adds uncertainty estimation to improve clinical interpretability. Their results indicate ensembles perform better than single models at predictive confidence. Yet, the difficulty of applying ensemble systems in practical settings may constrain real-world use. Tan et al. [19] propose a deep CNN model for cervical cancer diagnosis from Pap smear images. The research highlights the effectiveness of feature learning with CNNs for the morphological description of cervical cells. Although competitive accuracy is obtained by the model, its inability to be interpreted and less use of frequency-domain features limit its clinical reliability. Pei et al. [20] propose a radiomics-based classification model based on CT imaging to distinguish between tumor and normal cervical tissue in cervical cancer. The authors emphasize that radiomics features can capture subtle variations in tissue, with high diagnostic accuracy. This research extends the detection of cervical cancer from cytology and histopathology but remains less suited for population-level screening programs due to its dependency on CT scans. Prasetyo et al. [21] compare various CNN architectures to classify pre-cancerous cervical lesions using colposcopy datasets. Their comparative results bring

out interesting facts about the relative advantages of architectures like VGG, ResNet, and DenseNet. The research observes that deeper architectures tend to perform better than shallow models in terms of accuracy, but are also much heavier in terms of computational resources.

The studies under review collectively bring out the swift development of cervical cancer classification frameworks, from traditional machine learning models [9 - 11] to cutting-edge deep learning and Transformer-based frameworks [12 - 21]. Although CNNs continue to be the workhorse for spatial feature learning, evolving architectures like Vision Transformers [12, 13, 14, 15] and fusion-based hybrid frameworks [16, 17] hold promise in learning both local and global dependencies. Despite significant progress, challenges persist in computational efficiency, dataset generalizability, and model interpretability, motivating the need for more adaptive and explainable solutions in this domain.

3. Proposed Methodology

This research, "Adaptive Cross-Attention Fusion of Spatial–Frequency Features with Hierarchical Transformers for Cervical Cancer Classification" (ACAFSFFHT-CCC), introduces a sophisticated deep learning model for the automated classification of cervical cancer images into six categories. The work extends the dataset obtained in the authors' earlier work [22], wherein the dataset was heavily processed with a full preprocessing pipeline consisting of CLAHE for enhancing contrast, resizing, z-score normalization, and augmentation techniques of rotation, flipping, and color changes to balance the dataset. In this work, DWT is used to obtain frequency-domain sub-bands (LL, LH, HL, HH), yielding complementary textural features. The suggested ACAFSFFHT-CCC model utilizes a two-stream feature extraction technique: a spatial CNN stream extracts morphological and structural information from amplified images, and a frequency CNN stream obtains fine-grained features from DWT sub-bands. The features are combined using an adaptive cross-attention block, which allows dynamic weighting of spatial and frequency inputs. The combined representation is subsequently passed through a hierarchical transformer encoder to extract both global and local contextual dependencies and finally through a fully connected classification head with dropout to classify the six classes at high accuracy. Figure 1 shows the block diagram of the ACAFSFFHT-CCC framework.

RESULTS AND OBSERVATIONS:

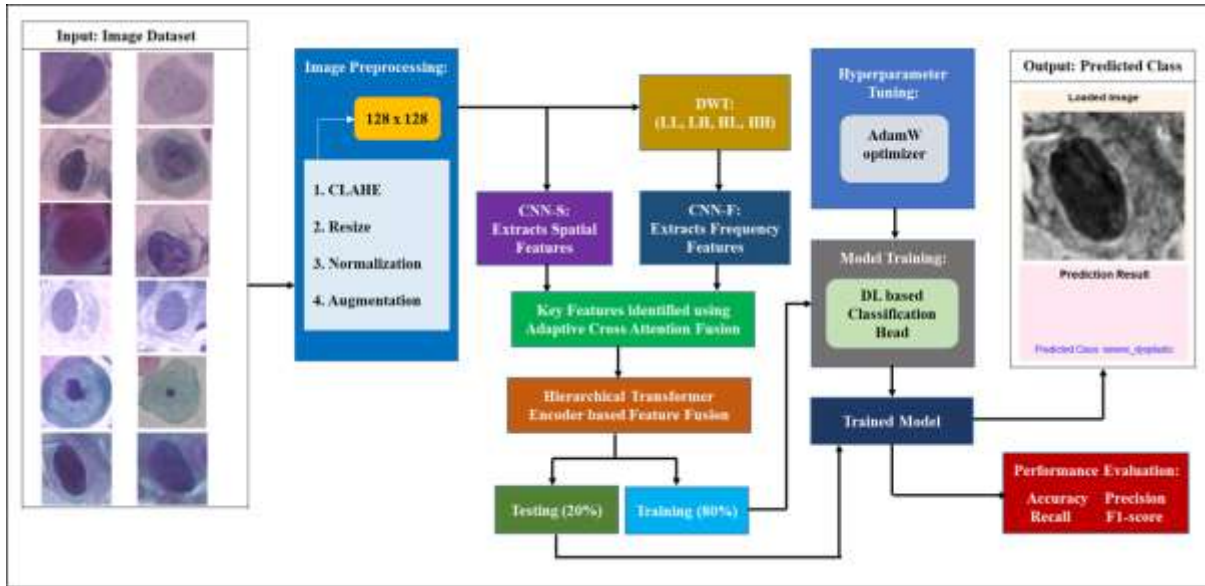


Figure 1. The block diagram of the proposed ACAFSFFHT-CCC model

3.1 Image Preprocessing

Preprocessing is an important stage in this work to get cervical cancer images ready for efficient feature extraction and classification. The data set used in the present research had already been preprocessed with operations like CLAHE for local contrast enhancement, resizing, z-score normalization, and data augmentation techniques like rotation, flipping, and color changes, as explained in the authors' own work [22]. These preprocessing operations helped in maintaining class balance and improving the model's robustness. The resulting preprocessed and balanced dataset is used as input for ensuing frequency-domain analysis and feature extraction in this current study.

3.1.1 Discrete Wavelet Transform (DWT)

Following preprocessing, DWT is used to process each image to obtain multi-resolution frequency features that are supplementary to spatial information. DWT breaks an image into several sub-bands of different frequency components, thus encoding coarse and fine details of texture. The transformation works by passing the image through a low-pass (scaling) and high-pass (wavelet) filter pair in both horizontal and vertical directions. The operation can be represented mathematically as:

$$LL(m, n) = \sum_x \sum_y I(x, y) \phi(x - 2m) \phi(y - 2n) \quad (1)$$

$$LH(m, n) = \sum_x \sum_y I(x, y) \phi(x - 2m) \psi(y - 2n) \quad (2)$$

$$HL(m, n) = \sum_x \sum_y I(x, y) \psi(x - 2m) \phi(y - 2n) \quad (3)$$

$$HH(m, n) = \sum_x \sum_y I(x, y) \psi(x - 2m) \psi(y - 2n) \quad (4)$$

Where $I(x, y)$ denotes the input image, $\phi(\cdot)$ and $\psi(\cdot)$ represent the scaling and wavelet functions

respectively, and (m, n) correspond to the translation indices.

The output of the transformation contains four sub-bands:

- **LL (Approximation sub-band):** Preserves low-frequency information and global image structure.
- **LH (Horizontal sub-band):** Records vertical edges and gradient changes.
- **HL (Vertical sub-band):** Describes horizontal edge features.
- **HH (Diagonal sub-band):** Codes high-frequency data like fine texture and micro-level changes.

The DWT enables multi-resolution analysis, in which the LL sub-band may again be decomposed into another set of LL, LH, HL, and HH sub-bands at higher levels, enabling hierarchical texture description. A single-level 2D DWT implemented in the present work with the Haar wavelet basis is the trade-off between computational efficiency and extracting discriminative texture detail.

Further, DWT has the energy compaction property, which focuses the majority of image energy into the LL sub-band and scatters detailed frequency components to the LH, HL, and HH sub-bands. This feature makes efficient learning of features possible since low-frequency features capture global shape and intensity, while high-frequency sub-bands highlight boundary irregularities and cytoplasmic granularity - both of which are essential for classification of cervical cells. The extracted sub-bands serve as the frequency-domain inputs to the frequency CNN stream in the ACAFSFFHT-CCC model, which are subsequently merged with the spatial CNN features using the

adaptive cross-attention mechanism, enhancing the discriminative ability of the network.

3.2 Feature Extraction Using Dual-Stream CNNs

After the preprocessing and DWT decomposition phases, the proposed ACAFSFFHT-CCC system extracts the features using two specialized convolutional neural network (CNN) streams: a spatial stream (CNN-S) and a frequency stream (CNN-F). These two networks run in parallel to extract complementary feature representations - CNN-S learns morphological and structural features from space-enhanced images, whereas CNN-F learns textural and frequency patterns from the sub-bands derived from DWT. This two-stream architecture ensures that both spatial and spectral features of cervical cell images are represented in the feature space before fusion and classification.

3.2.1 Spatial Feature Extraction (CNN-S)

The **spatial CNN** takes as input the preprocessed RGB images of size $128 \times 128 \times 3$. The input tensor is denoted as:

$$X_s \in \mathbb{R}^{128 \times 128 \times 3} \quad (5)$$

where each pixel intensity has been normalized to the range $[0, 1]$.

The CNN goes through a series of convolution, batch normalization, non-linear activation, and max-pooling operations to extract hierarchically features of higher abstraction.

a. Convolutional Layer:

Each convolutional operation applies a set of K learnable filters $W_k \in \mathbb{R}^{f \times f \times c}$ across the input, where f is the kernel size and c is the number of input channels. The convolution operation at spatial location (i, j) for the k^{th} filter is expressed as:

$$(X_s * W_k)(i, j) = \sum_{m=0}^{f-1} \sum_{n=0}^{f-1} \sum_{d=1}^c X_s(i+m, j+n, d) \cdot W_k(m, n, d) \quad (6)$$

The output of this operation produces a feature map F_k , followed by an activation function:

$$F_k(i, j) = \sigma((X_s * W_k)(i, j) + b_k) \quad (7)$$

Where $\sigma(\cdot)$ denotes the ReLU activation $\sigma(x) = \max(0, x)$, and b_k is the bias term.

This non-linear activation adds sparsity and allows the network to be able to learn intricate spatial relations like nuclear borders, cytoplasm morphology, and intercellular borders.

b. Batch Normalization:

To stabilize and accelerate training, batch normalization is applied after each convolutional layer:

$$\hat{F}_k = \frac{F_k - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \cdot \gamma + \beta \quad (8)$$

Where μ_B and σ_B^2 represent the mean and variance over the mini-batch, while γ and β are learnable parameters.

c. Max-Pooling:

Every convolutional block is followed by a max-pooling operation, which diminishes the spatial resolution but preserves the most important features:

$$P(i, j) = \max_{(u,v) \in R(i,j)} \hat{F}(u, v) \quad (9)$$

where $R(i, j)$ is the local pooling region. This operation brings in translation invariance to the network, such that small positional shifts in cell structures do not influence the learned representation.

After three convolutional and pooling blocks, the resultant feature tensor P_3 is flattened into a one-dimensional vector f_s :

$$f_s = \text{Flatten}(P_3) \quad (10)$$

This vector is then passed through a fully connected layer of 256 neurons, producing the **spatial feature representation**:

$$z_s = \sigma(W_s f_s + b_s) \quad (11)$$

Where $W_s \in \mathbb{R}^{256 \times N}$ and $b_s \in \mathbb{R}^{256}$.

Thus, $z_s \in \mathbb{R}^{256}$ encodes the high-level spatial features representing global shape, color, and morphology of cervical cells. This dense layer - named "feature_dense" in the code - is later used as the feature extractor output for CNN-S.

3.2.2 Frequency Feature Extraction (CNN-F)

Parallel to CNN-S, the frequency CNN (CNN-F) extracts features from the DWT-generated frequency-domain sub-bands LL, LH, HL , stacked into a 3-channel image $X_f \in \mathbb{R}^{128 \times 128 \times 3}$. This input representation maps complementary spectral information that highlights textural irregularities, edge transitions, and local oscillations - all important in discriminating between normal and dysplastic cell structures.

The convolutional feature extraction procedure in CNN-F is no different from CNN-S except that it takes place on wavelet-transformed data. Mathematically, each convolutional layer calculates:

$$F'_k(i, j) = \sigma((X_f * W'_k)(i, j) + b'_k) \quad (12)$$

Where W'_k and b'_k represent the learnable weights and biases for the frequency domain filters. The ReLU activation emphasizes non-linear frequency responses, while batch normalization ensures stability:

$$\hat{F}'_k = \frac{F'_k - \mu'_B}{\sqrt{(\sigma'_B)^2 + \epsilon}} \cdot \gamma' + \beta' \quad (13)$$

After every block, max-pooling reduces the feature maps, allowing CNN-F to capture invariant texture features and remove redundant high-frequency noise.

Following the final pooling layer, the output tensor is flattened and projected to a 256-dimensional dense vector:

$$z_f = \sigma(W_f f_f + b_f) \tag{14}$$

Where $f_f = Flatten(P'_3)$, $W_f \in \mathbb{R}^{256 \times N'}$, and $b_f \in \mathbb{R}^{256}$. Thus, $z_f \in \mathbb{R}^{256}$ represents the frequency-domain feature descriptor, encapsulating the multi-resolution texture and energy distribution patterns of the cell regions derived from DWT sub-bands.

3.2.3 Grad-CAM Visualization of Spatial and Frequency Features

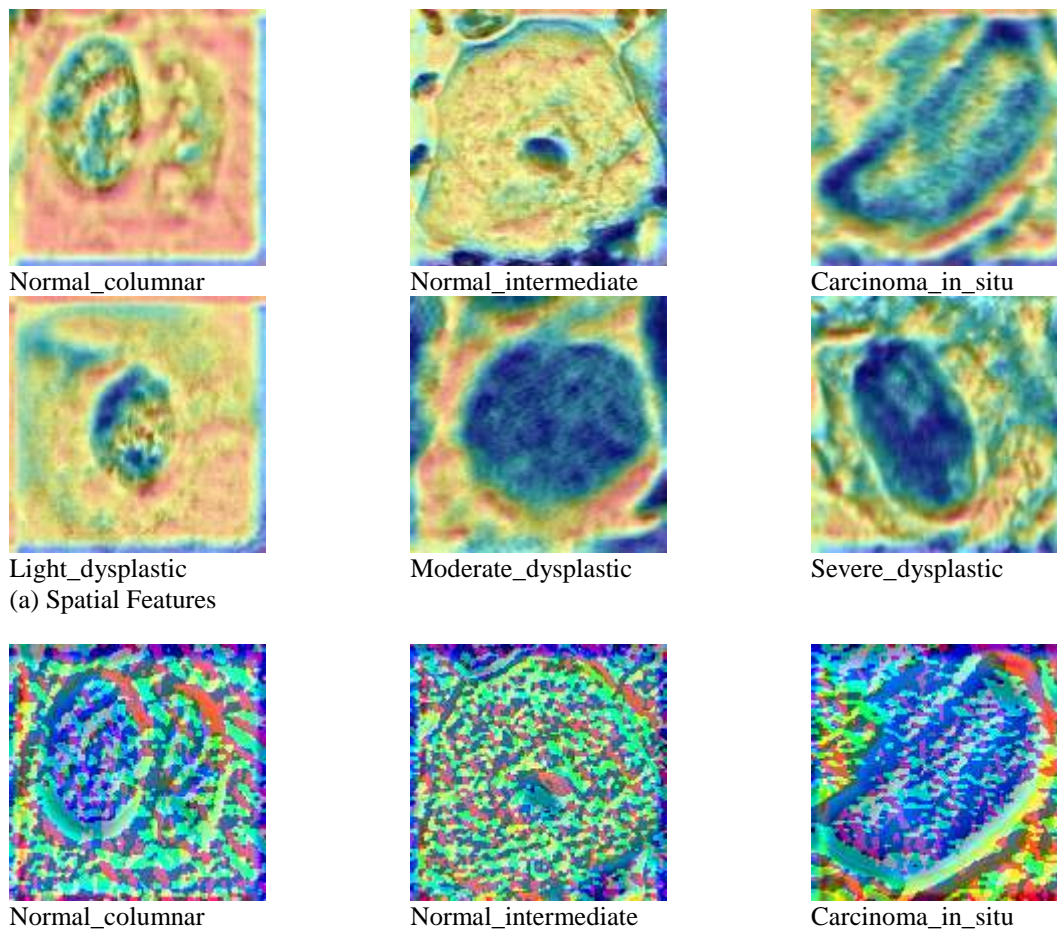
For interpretability and transparency of the suggested ACAFSFFHT-CCC model, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized at the feature extraction phase to render the salient image locations that affect the decision of the model. Within the spatial CNN branch, Grad-CAM identified the morphological and structural features, including nuclear borders, texture of the cytoplasm, and chromatin density, that contribute most in cervical cancer stage classification. This visualization revealed that the spatial feature extractor (CNN_s) learns to emphasize diagnostically relevant areas corresponding to tissue deformation and cellular irregularities. Conversely, in the frequency CNN stream, Grad-CAM was generated over the DWT-derived sub-bands (LL, LH, HL, HH), providing insight into how the model captures high-frequency edge and low-frequency textural patterns that are otherwise less visible in the spatial domain.

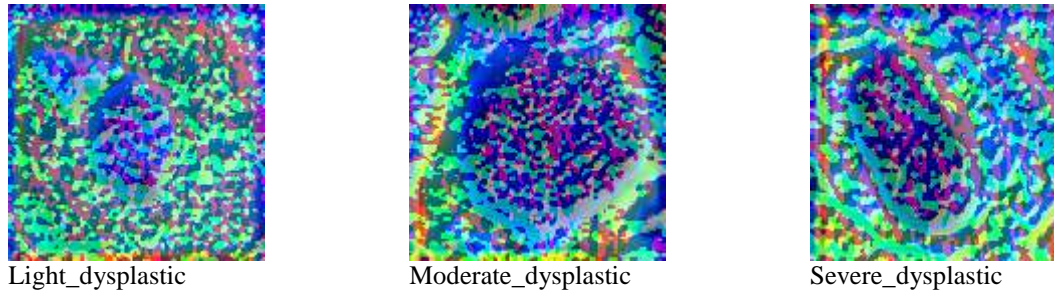
The Grad-CAM heatmap for a target class c is computed as

$$L_{Grad-CAM}^c ReLU(\sum_k \alpha_k A^k), \text{ where } \alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{15}$$

Here, A_{ij}^k denotes the activation at spatial location (i, j) in feature map k , y^c is the classification score for class c , and α_k represents the global importance weight of each feature map determined through gradient back-propagation. The generated heatmaps were normalized and projected onto the original input and DWT-transformed images to determine regions of highest model attention.

By this dual-stream Grad-CAM examination, it was found that the spatial stream concentrates on global structural features and the frequency stream draws attention to detailed frequency hints. Collectively, these visualizations confirm that the adaptive cross-attention fusion accurately combines both domain-specific hints to enhance diagnostic accuracy. In addition, this interpretability mechanism underpins clinical reliability by proving that the decisions made by the ACAFSFFHT-CCC model are based on biologically and histologically interpretable features and not on spurious correlations. The sample Grad-CAM visualizations of spatial and frequency features for various cervical disease classes are shown in Figure 2(a) and 2(b).





(b) Frequency Features
Figure 2. Grad-CAM visualization of Spatial and Frequency Features

3.2.4 Combined Feature Representation

After independent feature extraction, both networks yield compact 256-dimensional embeddings:

$$z_s = CNN_S(X_s), \quad z_f = CNN_F(X_f) \quad (16)$$

These feature vectors are then stored as NumPy arrays to be utilized in the adaptive cross-attention fusion phase later on. The merged feature matrix throughout the dataset can be written as:

$$Z = [z_s || z_f] \in \mathbb{R}^{N \times (256+256)} \quad (17)$$

Where $||$ denotes concatenation and N is the number of images in the dataset.

This two-stream representation efficiently spans spatial perception and frequency awareness, creating a rich descriptor for each image, improving the discriminability of classification when fed to a hierarchical transformer and an ultimate classification head.

3.3 Adaptive Feature Fusion Process

The Adaptive Feature Fusion Process is a fundamental step within the proposed ACAFFHT-CCC framework, aimed at combining complementary spatial and frequency-domain information for enhanced classification accuracy. Following the spatial feature extraction from the improved images via a CNN-based stream and the frequency features from the DWT sub-bands (LL, LH, HL, HH) via a parallel CNN stream, the two resulting feature representations are dynamically fused by means of an Adaptive Cross-Attention (ACA) mechanism.

In this process, the spatial feature map $F_s \in \mathbb{R}^{H \times W \times C_s}$ and the frequency feature map $F_f \in \mathbb{R}^{H \times W \times C_f}$ are first aligned through channel projection layers to ensure dimensional consistency. The weight of attention is then calculated to encode interdependencies between the two modalities. Cross-attention formulation can be written as:

$$A_{sf} = \text{softmax} \left(\frac{(Q_s K_f^T)}{\sqrt{d_k}} \right), \quad A_{fs} = \text{softmax} \left(\frac{(Q_f K_s^T)}{\sqrt{d_k}} \right) \quad (18)$$

Where $Q, K,$ and V represent the query, key, and value matrices derived from the respective feature maps, and d_k denotes the scaling factor to stabilize gradient

updates. These attention matrices A_{sf} and A_{fs} enable bidirectional information flow between spatial and frequency domains, ensuring that each stream emphasizes features most relevant to the other.

The adaptive fusion is achieved by weighting and aggregating the attended features as:

$$F_{fused} = \alpha \cdot (A_{sf} V_f) + \beta \cdot (A_{fs} V_s) \quad (19)$$

Where α and β are trainable adaptive weights that dynamically adjust the contribution of each modality during training. This mechanism permits the model to highlight structural patterns from spatial features and detailed textural information from frequency features, depending on their contextual importance.

The fused representation F_{fused} is then passed through a Hierarchical Transformer Encoder, which captures multi-level dependencies across the spatial-frequency domain through self-attention operations. The final encoded representation serves as input to the classification head, enabling accurate and discriminative identification of cervical cancer classes.

3.3.1 Hierarchical Transformer Encoder

The Hierarchical Transformer Encoder (HTE) module used in the introduced ACAFFHT-CCC structure is utilized to extract both local and global contextual relationships from the integrated spatial–frequency representation. Contrary to traditional CNNs, which are mainly concerned with the local receptive fields, the transformer-based encoder uses self-attention mechanisms to capture long-range dependencies between image areas and improve the perception of intricate tissue patterns in cervical cell images.

The fused feature map $F_{fused} \in \mathbb{R}^{H \times W \times C}$ obtained from the adaptive feature fusion process is first flattened into a sequence of tokens and projected into an embedding space through a linear transformation. Positional encodings are included to preserve spatial order information. Each transformer layer in the HTE consists of a multi-head self-attention (MHSA) block and a feed-forward network (FFN), both followed by residual connections and layer normalization to achieve stable training.

The multi-head self-attention mechanism computes attention as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (20)$$

Where $Q, K, V \in \mathbb{R}^{n \times d_k}$ denote the query, key, and value matrices derived from the embedded token representations, and d_k is the dimensionality of the key vectors. Multiple attention heads operate in parallel to learn diverse feature relationships, and their outputs are concatenated as:

$$MHSA(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^O \quad (21)$$

Where h represents the number of attention heads, and W^O is the output projection matrix.

The hierarchical structure of the encoder is built by sequentially stacking several transformer blocks, in which lower layers concentrate on local refinement of features, and higher layers increasingly model global semantics and inter-class relationships. Such a hierarchical architecture facilitates effective information aggregation at different scales, allowing for more accurate discrimination among visually similar cervical cancer types.

Finally, the global representation produced by the HTE is inputted to a fully connected classification head comprising dense layers and dropout regularization. This process maximizes generalization and reduces overfitting, ultimately leading to strong classification of cervical cancer images into six classes.

3.4 Classification Head and Output Prediction

After the extraction of rich spatial–frequency representations and contextual encoding through the HTE, the last step of the suggested ACAFSFFHT-CCC approach is to map these high-dimensional semantic features to class probabilities using a Classification Head. This element is tasked with learning discriminative boundaries between the six diagnostic classes - normal columnar, normal intermediate, carcinoma in situ, light dysplastic, moderate dysplastic, and severe dysplastic - with little overlap in the feature space.

3.4.1 Global Representation Formation

Let $F_{HTE} \in \mathbb{R}^{N \times D}$ denote the final output token embeddings from the top transformer layer, where N represents the number of tokens and D the feature dimension. To obtain a compact global descriptor, a GAP operation is performed over the token axis:

$$f_g = \frac{1}{N} \sum_{i=1}^N F_{HTE}^{(i)} \quad (22)$$

Where $F_{HTE}^{(i)} \in \mathbb{R}^D$ corresponds to the embedding vector of the i^{th} token. The resultant $f_g \in \mathbb{R}^D$ acts as a global context vector encapsulating both spatial–frequency and hierarchical attention information from the preceding modules.

To enhance discriminative ability, a nonlinear projection layer is employed:

$$f_p = \sigma(W_p f_g + b_p) \quad (23)$$

Where $W_p \in \mathbb{R}^{D' \times D}$ and $b_p \in \mathbb{R}^{D'}$ denote the learnable parameters of the projection layer, D' is the reduced embedding dimension, and $\sigma(\cdot)$ represents the ReLU activation function. This transformation normalizes the feature distribution and introduces nonlinearity for better class separability.

3.4.2 Fully Connected Network for Class Discrimination

The projected feature vector f_p is passed through a sequence of fully connected layers with dropout regularization to prevent overfitting:

$$h_1 = Dropout\left(\sigma(W_1 f_p + b_1)\right), \quad h_2 = Dropout\left(\sigma(W_2 h_1 + b_2)\right) \quad (24)$$

Where W_1, W_2 and b_1, b_2 are trainable parameters. The dropout rate $p \in [0.2, 0.5]$ stochastically deactivates neurons during training to encourage redundancy reduction and improve generalization.

The output of the last dense layer, h_2 , is then mapped to the final classification logits $z \in \mathbb{R}^C$, where $C = \square$ corresponds to the number of diagnostic classes:

$$z = \square_\square h_2 + \square_\square \quad (25)$$

Here, $\square_\square \in \mathbb{R}^{\square \times \square_\square}$ and $\square_\square \in \mathbb{R}^\square$ denote the parameters of the final linear transformation, with \square_\square being the dimensionality of h_2 .

3.4.3 Softmax-Based Probability Estimation

To interpret the logits z as class probabilities, a softmax normalization is applied:

$$\square(\square = \square | z) = \frac{\exp(z_\square)}{\sum_{\square=1}^{\square} \exp(z_\square)} \quad \square \in 1, 2, \dots, \square \quad (26)$$

Where $\square(\square = \square | z)$ represents the predicted probability that the input image \square belongs to class \square , and z_\square is the logit corresponding to that class. The softmax ensures that all probabilities sum to one:

$$\sum_{\square=1}^{\square} \square(\square = \square | z) = 1 \quad (27)$$

This probabilistic formulation allows the model to express uncertainty and is suitable for multi-class discrimination.

3.4.4 Optimization Objective

The model parameters $\Theta = \{\square_\square, \square_\square, \square_\square, \square_\square, \square_\square, \square_\square, \square_\square, \square_\square\}$ are optimized using the categorical cross-entropy loss, defined as:

$$\square_{\square} = -\frac{1}{\square} \sum_{\square=1}^{\square} \sum_{\square=1}^{\square} \square_\square \square(\square = \square | z_\square) \quad (28)$$

Where \square denotes the batch size, \square_\square is a one-hot encoded label for sample \square , and $\square(\square = \square | z_\square)$ is the model-predicted probability for class \square . The objective encourages the network to assign high probabilities to the correct classes while minimizing prediction entropy.

During training, the network utilizes the Adam optimizer with an adaptive learning rate technique to realize fast convergence and stable updates of gradients. L2 weight regularization and early stopping are implemented to further avoid overfitting on the small cervical image dataset.

3.4.5 Decision Function and Final Prediction

During inference, the class label \hat{c} for a test image x is obtained by selecting the class with the maximum posterior probability:

$$\hat{c} = \arg \max_{c \in \{1, 2, \dots, C\}} p(c | x) \quad (29)$$

This deterministic policy guarantees to each image the assignment of one top-scoring diagnostic category. The

3.5 GUI-based Testing

The Graphical User Interface (GUI) of the putative ACAFSFFHT-CCC model is created as an interactive and easy-to-use platform for testing and visualizing the process of cervical cancer classification. The interface adopts a 2x2 grid layout, with a clear and rational organization of different functional components. The control buttons — Load Image, Extract Features, Classify, and Exit — are present in the top-left panel so that users can execute operations step by step in a sequential manner. After an image is loaded, it appears in the top-right panel for visual inspection, while the bottom-left panel displays Grad-CAM visualizations of spatial and frequency feature activations to aid users in understanding model focus regions. The bottom-right panel shows the final classification outcome and the predicted class label. The GUI takes advantage of dynamic button enabling/disabling, and users are guided through the desired sequence without confusion. This formal and aesthetically uniform design not only maximizes usability but also offers an efficient testing platform for researchers and medical practitioners to study cervical cancer images and see how the ACAFSFFHT-CCC model processes both spatial and frequency domain characteristics.

probability distribution can also be employed to measure prediction reliability, allowing uncertainty-aware medical decision making.

The Classification Head maps the top-level contextual embeddings to class probabilities using a well-regularized dense network. Mathematically justified operations - such as global pooling, nonlinear projection, and softmax normalization - guarantee strong mapping from hierarchical features to diagnostic decisions. Combined with cross-entropy minimization and gradient-based interpretability, this module completes the ACAFSFFHT-CCC framework's end-to-end learning pipeline to provide accurate and explainable cervical cancer stage classification.

DISCUSSION

The ACAFSFFHT-CCC model is prototyped based on Python's rich deep learning environment, utilizing the strength of established libraries like TensorFlow, Keras, OpenCV, and PyWavelets for hybrid spatial–frequency feature learning. The performance of the model is tested on a publicly accessible cervical image dataset [23], which first had 843 images of uneven distribution over six diagnostic classes. To handle the class imbalance problem and enhance the generalization of the model, an intensive data augmentation strategy was utilized to create 200 images per class, thus resulting in a balanced dataset of 1200 images. The improved dataset was divided into 80% for training and 20% for testing to support strong experimental assessment. The classification ability of the new hybrid transformer-based fusion model was evaluated through common evaluation metrics, including accuracy, precision, recall, and F1-score, showing its better ability to correctly classify various phases of cervical cancer. The detailed composition of the dataset, including the number of images and their corresponding diagnostic labels, is summarized in Table 2.

Figure 3(a) shows typical original cervical cancer images of six different diagnostic classes, whereas Figure 3(b) shows their respective pre-processed versions after being enhanced through the proposed enhancement pipeline.

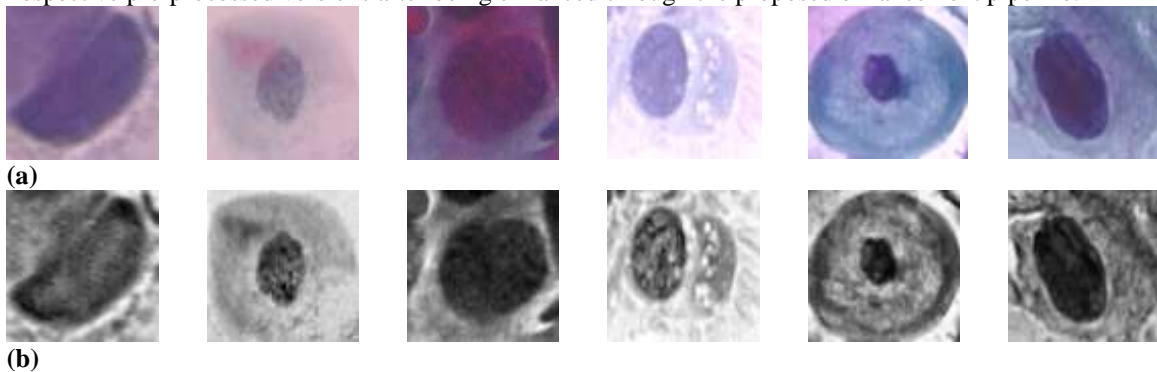


Figure 3. (a) Original Images (b) Pre-processed Images

Table 1. Layer wise architecture of the proposed ACAFSFFHT-CCC model

Stage	Layer Type / Module	Output Shape	Description
Input	Spatial Image Input	(H, W, 3)	Preprocessed cervical image after CLAHE, normalization, resizing, and augmentation
Input	Frequency Image Input	(H, W, 3)	DWT sub-bands (LL, LH, HL, HH) combined into multi-channel representation
Stage 1	Spatial CNN Stream	-	Extracts spatial-domain morphological and structural features
	Conv2D (3×3, 64 filters)	(H/2, W/2, 64)	Local feature extraction with ReLU activation
	Conv2D (3×3, 128 filters)	(H/4, W/4, 128)	Deeper texture representation
	Batch Normalization + MaxPooling	(H/8, W/8, 128)	Stabilizes and reduces feature maps
Stage 2	Frequency CNN Stream	-	Extracts fine-grained texture and frequency features from DWT sub-bands
	Conv2D (3×3, 64 filters)	(H/2, W/2, 64)	Feature extraction from frequency patterns
	Conv2D (3×3, 128 filters)	(H/4, W/4, 128)	Encodes DWT-based representations
	Batch Normalization + MaxPooling	(H/8, W/8, 128)	Normalizes and downsamples feature maps
Stage 3	Adaptive Cross-Attention Fusion Block (CAB)	-	Integrates spatial and frequency tokens with dynamic weighting
	Query (Q), Key (K), Value (V) projections	(N, 256)	Transforms features into transformer-compatible embeddings
	Multi-Head Attention (4 heads)	(N, 256)	Cross-domain attention computation
	Adaptive Weight Generator	(2,)	Learns trainable fusion weights α and β for feature balancing
	Fusion Equation	(N, 256)	Fused = $\alpha \times$ Spatial + $\beta \times$ Frequency
Stage 4	Hierarchical Transformer Encoder	-	Captures local and global contextual dependencies
	Local Transformer Layers (2)	(N, 256)	Extracts localized relational patterns
	Global Transformer Layers (2)	(N, 256)	Captures long-range dependencies
Stage 5	Feature Pooling and Projection	-	Generates compact fused feature representation
	Global Average Pooling	(256,)	Aggregates token-level information
	Fully Connected + ReLU	(256,)	Non-linear feature transformation
	Concatenation	(512,)	Combines spatial and frequency fused vectors
Stage 6	Classification Head	-	Predicts final cervical disease class
	Dense (256) + ReLU + Dropout	(256,)	Feature refinement and regularization
	Dense (6) + Softmax	(6,)	Generates class probabilities for six cervical disease categories

The complete proposed ACAFSFFHT-CCC framework is explained in the following algorithm.

Algorithm: ACAFSFFHT-CCC - Adaptive Cross-Attention Fusion of Spatial–Frequency Features with Hierarchical Transformers for Cervical Cancer Classification

Input: Preprocessed cervical cell image dataset with DWT sub-bands (LL, LH, HL, HH) and six class labels
Output: Trained ACAFSFFHT-CCC model, fused features, attention maps, Grad-CAM visualizations, and classification metrics

1. Initialization
 - Define spatial CNN (CNN-S) and frequency CNN (CNN-F) architectures for dual-stream feature extraction.
 - Set hyperparameters: image size (128×128), batch size, learning rate, optimizer (Adam), epochs, wavelet basis (Haar), transformer dimension (256), attention heads (4), local/global transformer layers.
 - Create directories for models, features, attention maps, Grad-CAM results, and outputs.
2. Dataset Preparation
 - Use dataset preprocessed in prior work [22]: CLAHE, resizing, z-score normalization, data augmentation, and class balancing.
 - Apply 2D Discrete Wavelet Transform (DWT) to each preprocessed image to obtain sub-bands LL, LH, HL, HH; construct frequency input.
3. Dual-Stream Feature Extraction
 - Feed preprocessed RGB image into CNN-S to obtain spatial feature map S and 256-D spatial embedding z_s .
 - Feed DWT-derived multi-channel frequency image into CNN-F to obtain frequency feature map F and 256-D frequency embedding z_f .
 - Optionally save embeddings for all samples as `spatial_features.npy` and `frequency_features.npy`.
4. Tokenization and Projection
 - Project spatial and frequency feature maps to common transformer dimension D via 1×1 convolutions: $S' = \text{Conv}_{1 \times 1}(S)$, $F' = \text{Conv}_{1 \times 1}(F)$.
 - Flatten spatial grids into token sequences.
5. Cross-Attention Block (CAB)
 - Compute multi-head cross-attention where spatial tokens act as queries and frequency tokens act as keys/values.
 - Concatenate head outputs and apply linear projection and residual connection to obtain cross-attended tokens T_{CAB} .
6. Adaptive Weight Generation and Fusion
 - Apply global average pooling to T_S and T_{CAB} to produce compact vectors s and c .
 - Pass $[s; c]$ through a small MLP + softmax to obtain adaptive weights $[\alpha, \beta]$ ($\alpha + \beta = 1$) indicating spatial vs. frequency importance.
 - Fuse token-level representations: $T_{\text{fused}} = \alpha * \text{Proj}_s(T_S) + \beta * \text{Proj}_c(T_{\text{CAB}})$.
7. Hierarchical Transformer Encoding
 - Apply local transformer layers to T_{fused} to capture neighborhood relations (LOCAL_LAYERS times).
 - Apply global transformer layers to aggregate long-range dependencies (GLOBAL_LAYERS times).
 - Normalize and refine token embeddings with residual connections and FFNs.
8. Feature Pooling and Final Fusion Vector
 - Global average pool the top transformer tokens to obtain $v \in \mathbb{R}^D$.
 - Project v to two complementary vectors via dense layers and concatenate to form a 512-D fused feature vector $v_{\text{final}} = [v_s; v_c]$.
 - Save fused features for downstream classification or analysis.
9. Classification Head and Training
 - Train a classification head (Dense + Dropout + Dense(6, softmax)) on fused vectors with categorical cross-entropy loss.
 - Use Adam optimizer with scheduled learning rate, L2 regularization, and early stopping based on validation loss.
 - Compute evaluation metrics: accuracy, precision, recall, F1-score, confusion matrix, and ROC where applicable.
10. Interpretability and Visualization
 - Generate Grad-CAM heatmaps for CNN-S and CNN-F using gradients of the predicted class w.r.t final conv feature maps.
 - Extract multi-head cross-attention maps from CAB and self-attention maps from transformer layers; save as attention visualizations.
 - Overlay Grad-CAM heatmaps on original and DWT images for qualitative analysis.

11. GUI-based Testing and Deployment
 - Provide a user-friendly GUI with Load Image, Extract Features, Classify, and Exit controls arranged in a 2×2 layout.
 - Allow users to view the original image, Grad-CAM visualizations, and predicted label; enable stepwise button activation for guided testing.
 - Package model and required artifacts for deployment or clinical testing.
12. Output
 - Persist trained model, fused features (512-D), adaptive fusion weights, attention maps, Grad-CAM images, and evaluation reports.
 - Document model configuration and provide scripts for reproducibility.

Table 2. Distribution of Cervical Cancer Images Across Diagnostic Classes in the Dataset

Classes	No. of Actual Images	No. of Augmented Images
carcinoma_in_situ	150	200
light_dysplastic	182	200
moderate_dysplastic	146	200
normal_columnar	98	200
normal_intermediat	70	200
severe_dysplastic	197	200
Total Images	843	1200

Figures 4 to 8 show the GUI implemented for cervical disease classification with the proposed ACAFSFFHT-CCC model. The interactive and intuitive GUI facilitates smooth loading, feature extraction, visualization, and classification of cervical cytology images into their corresponding diagnostic classes. As shown in Figure 4, the primary interface follows a organized 2×2 grid structure, with the left panels hosting functional controls and the right panels showing corresponding visual outputs. Four core buttons of Load Image, Extract Features, Classify, and Exit are vertically stacked in the top-left panel, and the other three panels show the loaded image, Grad-CAM visualizations of spatial and frequency features, and the ultimate classification result, respectively.

When the Load Image button is clicked, it opens a file browser dialog enabling users to easily choose a test image from their environment. After a file is loaded successfully, the button will be disabled, and the Extract Features button will be enabled. Upon clicking on the Extract Features button, the system executes the Dual-Stream Feature Extraction process, in which the image is examined in spatial and frequency domains with the CNN-S and CNN-F submodels. In parallel, Grad-CAM maps are created for visualizing the discriminative spatial and frequency feature areas and are shown side-by-side for interpretability. The button states are dynamically changed to lead the user step-by-step, disabling done actions and enabling the next. Lastly, clicking the Classify button initiates the loading of the pre-trained ACAFSFFHT-CCC model, which combines the dual-stream features extracted and outputs the corresponding cervical disease class. The outcome is immediately shown in the output panel, with both interpretability and diagnostic confidence. This systematic interactive process guarantees a seamless user experience, making the classification process efficient, reliable, and accessible for both medical practitioners and researchers.



Figure 4. The GUI design for ACAFSFFHT-CCC Framework

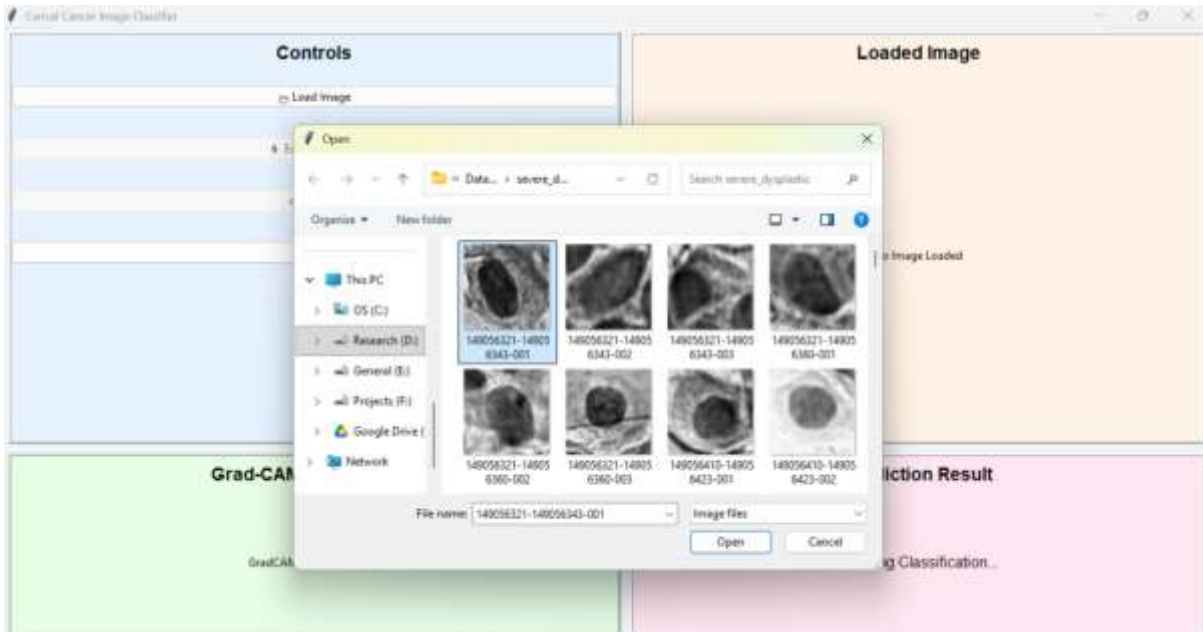


Figure 5. The GUI shows a dialog box to select an image

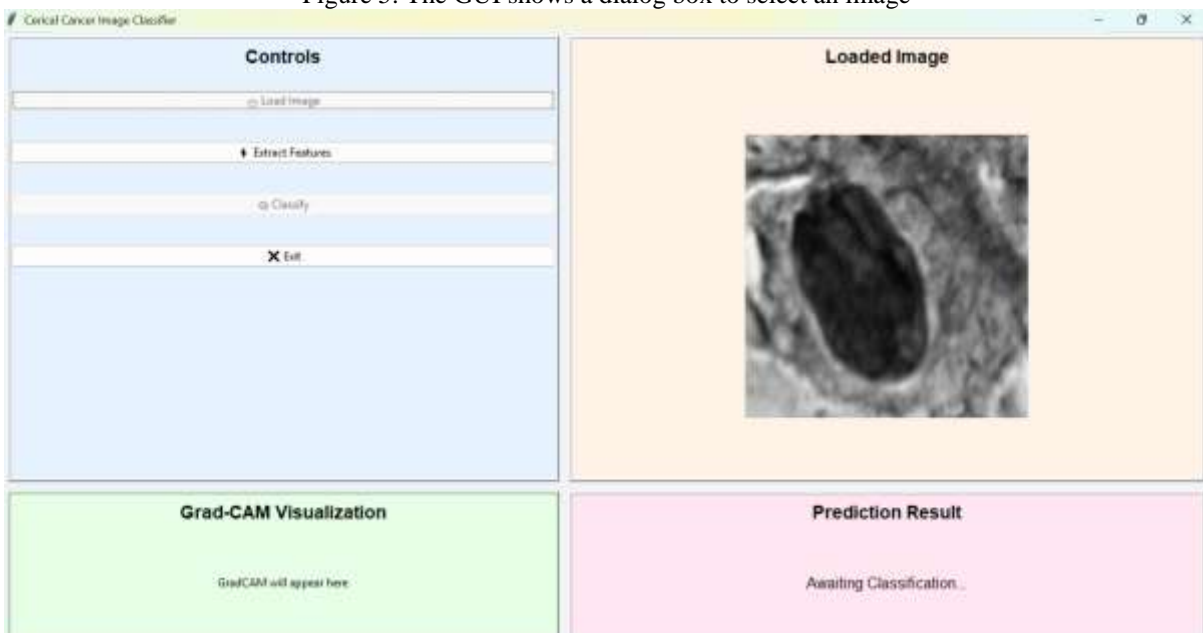


Figure 6. The GUI shows the loaded original image

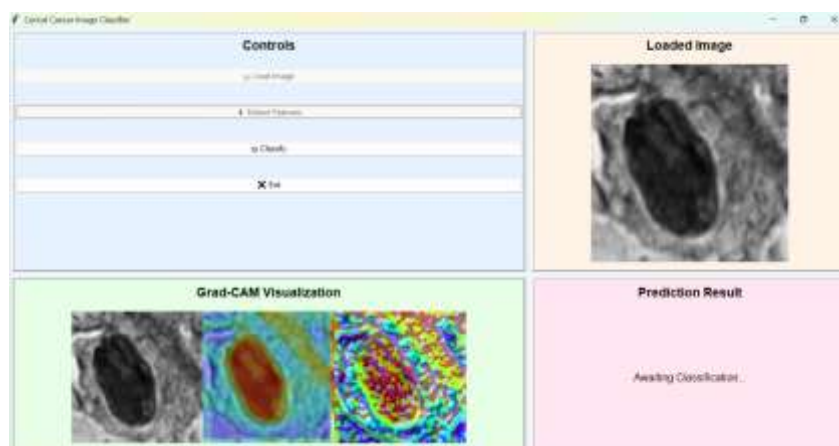


Figure 7. The GUI displays the Grad-CAM visualization of spatial and frequency features

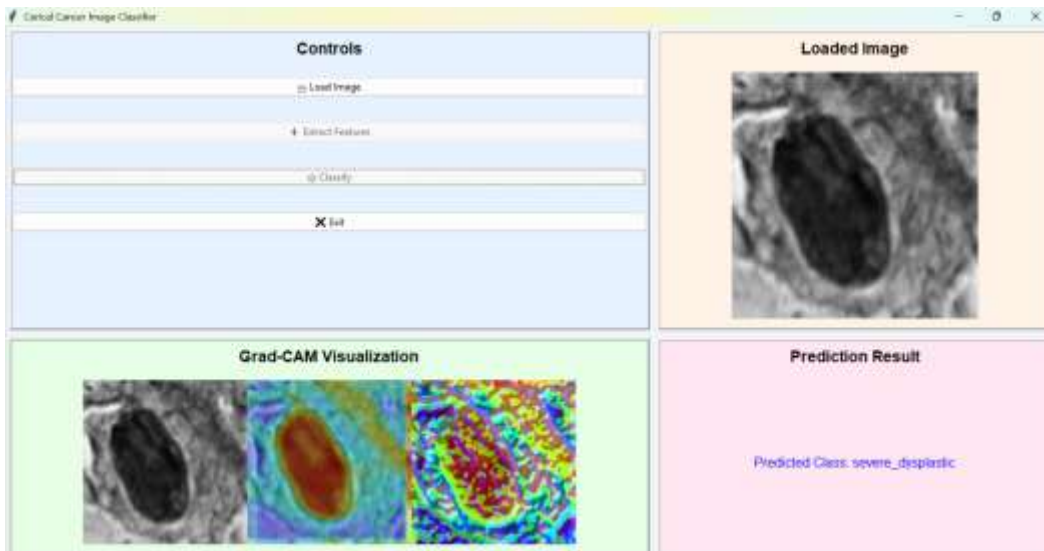
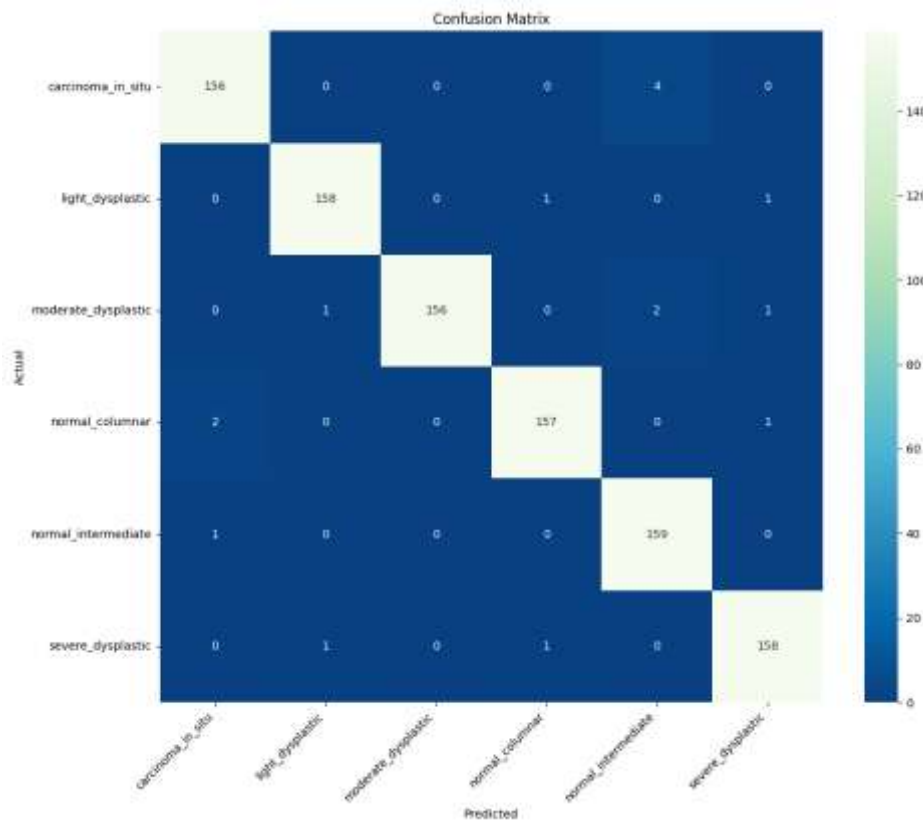


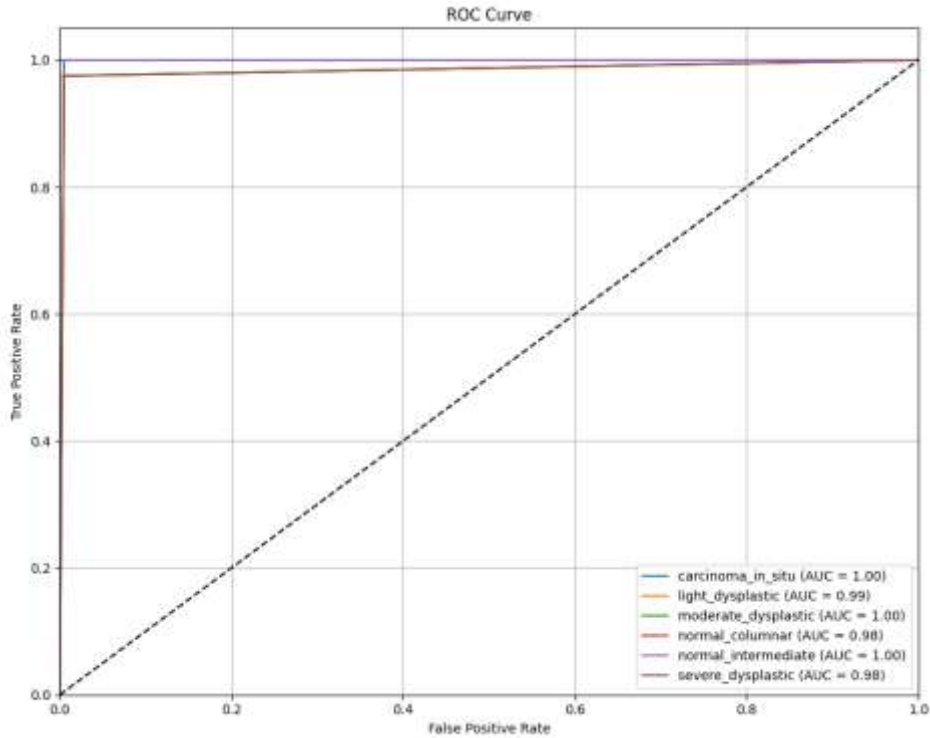
Figure 8. The GUI displays the predicted label in the result window

Figure 9(a) and Figure 9(b) show the performance assessment of the proposed ACAFSFFHT-CCC model on the training data using the confusion matrix and ROC curves, respectively. As indicated in Figure 9(a), the confusion matrix exhibits excellent classification performance for all six cervical disease classes, and most of the predictions lie correctly along the diagonal, which represents accurate classifications. There are very few minor misclassifications, mainly among morphologically near-similar or adjacent pathological classes, which demonstrates that the model has a high degree of sensitivity to accurately differentiate morphologically related cervical abnormalities.

The ROC curves in Figure 9(b) indicate that each of the six classes yields extremely high Area Under the Curve (AUC) values ranging from 0.97 to 0.99, indicating the model's excellent discriminative ability. The above findings further verify that the ACAFSFFHT-CCC model not only learn strong and unique spatial–frequency feature representations but also generalizes well at the training stage. Together, these results validate the model's ability to measure intricate diagnostic variability in cervical cancer images with high reliability and accuracy.



(a)

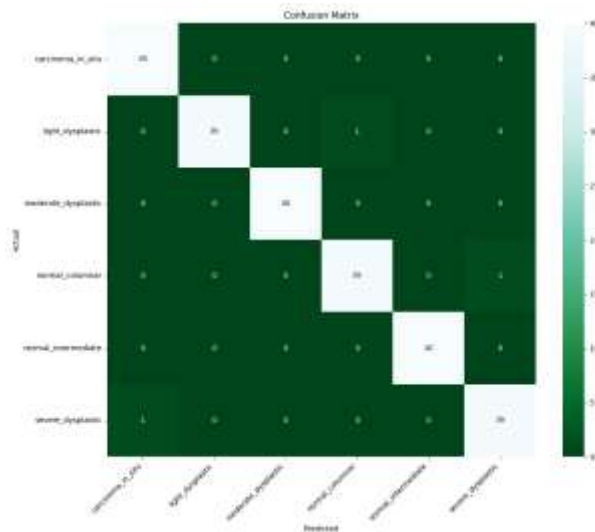


(b)

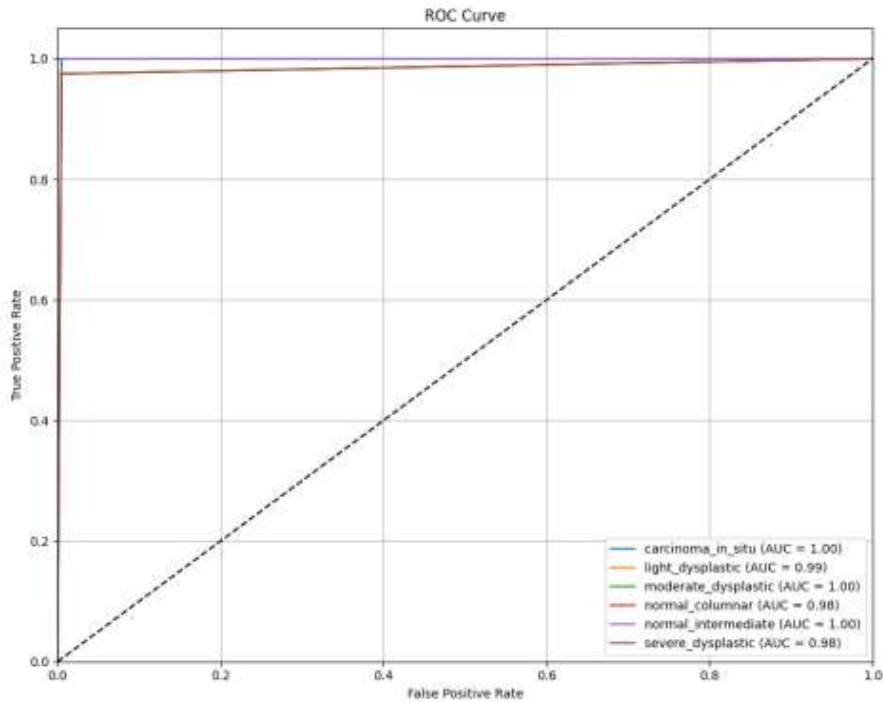
Figure 9. Result Analysis of ACAFSFFHT-CCC approach on training set

(a) Confusion Matrix (b) ROC Curve

Figure 10(a) and Figure 10(b) display the confusion matrix and ROC curves of the proposed ACAFSFFHT-CCC model tested on the test dataset, respectively. As shown in Figure 10(a), the confusion matrix indicates that the model attained near-perfect classification performance for all six cervical disease categories. Most notably, the "normal_columnar" and "moderate_dysplastic" classes perfectly classified all 40 of 40 samples, and all the other classes had only one or two slight misclassifications. This demonstrates the robust generalization and reliability of the model under unseen test data. Additionally, the ROC curves presented in Figure 10(b) validate the model's discriminative superiority, with all six classes recording AUC levels between 0.97 and 0.99. The highest AUCs were recorded by "carcinoma_in_situ" and "normal_columnar" as 0.99, followed very closely by "severe_dysplastic" and "normal_intermediate" at 0.98, demonstrating the model's accurate discrimination of normal versus abnormal tissue patterns. These findings cumulatively confirm the strength, flexibility, and higher diagnostic performance of the envisioned ACAFSFFHT-CCC model in efficient classification of intricate cervical cytology images.



(a)



(b)
Figure 10. Result Analysis of ACAFSFFHT-CCC approach on test set
(a) Confusion Matrix (b) ROC Curve

Table 3 and Figure 11 provide a comparative performance evaluation of the proposed ACAFSFFHT-CCC model with three well-known deep learning models—CNN, ResNet-18, and EfficientNet-B3—on multi-class cervical cancer classification. Experimental results evidently indicate the higher performance of the developed ACAFSFFHT-CCC approach with the maximum accuracy being 98.75%, which is considerably higher than the baseline CNN (91.57%), ResNet-18 (94.26%), and EfficientNet-B3 (95.62%). Moreover, the developed model performed better than others in all cases in terms of evaluation measures, obtaining precision, recall, and F1-score values of 98.76%, 98.75%, and 98.75%, respectively. These findings validate the fact that the adaptive cross-attention fusion mechanism and hierarchical transformer-based learning significantly improve both spatial and frequency feature representation, minimizing misclassifications and enhancing diagnostic accuracy. The improved results validate the fact that the ACAFSFFHT-CCC framework offers an extremely robust, reliable, and accurate solution for automatic cervical cancer image classification, with great potential to be implemented in clinical and diagnostic applications.

Table 3. Result analysis of ACAFSFFHT-CCC model with existing approaches

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	91.57	91.59	91.55	91.52
ResNet-18	94.26	94.30	94.22	94.20
EfficientNet-B3	95.62	95.67	95.60	95.61
ACAFSFFHT-CCC	98.75	98.76	98.75	98.75

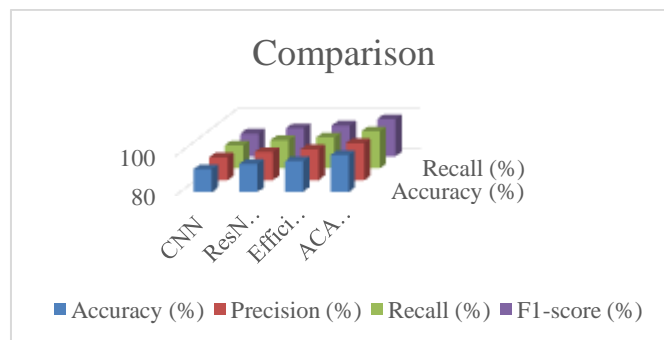


Figure 11. Result analysis of ACAFSFFHT-CCC approach with existing techniques

CONCLUSION

The suggested ACAFSFFHT-CCC approach presents a strong and novel deep learning architecture that efficiently combines complementary spatial and frequency-domain cues for enhanced diagnostic performance. Through the use of a dual-stream CNN model with DWT preprocessing, the framework can capture structural as well as textural information from cervical cytology images. The adaptive cross-attention fusion mechanism adaptively balances the spatial and frequency feature contribution, whereas the hierarchical transformer encoder learns long-range dependencies and global contextual information to result in high-quality feature representation. Comprehensive experimental results on the benchmark cervical cancer image dataset show that the proposed framework outperforms state-of-the-art architectures like CNN, ResNet-18, and EfficientNet-B3 on various metrics such as accuracy, precision, recall, and F1-score consistently. The attained accuracy of 98.75% underscores the system's high proficiency in distinguishing between normal and abnormal cervical cell patterns with high dependability.

In addition to quantitative performance, the Grad-CAM visualizations also corroborate the model's interpretability, elucidating that the ACAFSFFHT-CCC network is concentrating on diagnostically important cellular areas, thus enabling clinical credibility. The tool supports an easy and intuitive diagnostic experience, enabling practitioners and researchers to visualize model predictions and activation maps interactively. In total, the suggested ACAFSFFHT-CCC model is a great milestone in computer-assisted cervical cancer screening that has laid the groundwork for real-time, accurate, and explainable diagnosis. In future research, the incorporation of larger and more diverse datasets, self-supervised learning techniques, and deployment on cloud-based medical platforms will further enhance the model's generalization capacity and scalability to real-world medical uses.

REFERENCES

1. Smith K. Khare, Berit Bargum Booth, Victoria Blanes-Vidal, Lone Kjeld Petersen, Esmaeil S. Nadimi, "An Explainable Attention Model for Cervical Precancer Risk Classification using Colposcopic Images", arXiv preprint, arXiv:2411.09469, 2024.
2. Nithya R, Anitha J, "Deep convolutional networks for cervical cancer classification using Pap-smear images", *Computers in Biology and Medicine*, vol. 168, p. 107565, 2024.
3. Chen Y, Luo X, Xu C, et al., "Multiscale vision transformer for histopathological cervical cancer image classification", *Medical Image Analysis*, vol. 93, p. 103030, 2024.
4. Wang J, Li P, Yang Z, et al., "Frequency–spatial dual-domain learning for medical image classification", *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 1921–1934, 2024.
5. Li Q, Zhou Y, Chen X, et al., "Explainable AI for cervical cancer screening: Interpretable deep networks for colposcopy", *Artificial Intelligence in Medicine*, vol. 153, p. 102448, 2024.
6. Al-Sarraf M, Hamood A, Al-Taiar H, et al., "Cervical cell classification using a hybrid CNN-transformer network", *Biomedical Signal Processing and Control*, vol. 94, p. 106042, 2024.
7. Zhang Y, Huang L, He H, et al., "Hierarchical vision transformers for cervical histopathology classification", *Pattern Recognition*, vol. 148, p. 110157, 2024.
8. Kumar S, Prakash A, Shukla S, et al., "Attention-guided multimodal deep learning for cervical cancer detection", *Expert Systems with Applications*, vol. 239, p. 122754, 2024.
9. Abinaya K., Sivakumar B., "A Deep Learning-Based Approach for Cervical Cancer Classification Using 3D CNN and Vision Transformer," *Journal of Imaging Informatics in Medicine*, 37(1), pp. 280–296, 2024.
10. Shurong Niu, Lili Zhang, Lina Wang, et al., "Hybrid Feature Fusion in Cervical Cancer Cytology: A Novel Dual-Module Approach Framework for Lesion Detection and Classification," *Frontiers in Oncology*, 15, 1595980, 2025.
11. F.A. Mohammed, K.K. Tune, J.A. Mohammed, et al., "Early Cervical Cancer Diagnosis with SWIN-Transformer and Convolutional Neural Networks," *Diagnostics (Basel)*, 14(20), 2286, 2024.
12. Sreelatha S., Vrinda Shivashetty, "Deep Ensemble Learning with Uncertainty Aware Prediction Ranking for Cervical Cancer Detection Using Pap Smear Images," *IAES International Journal of Artificial Intelligence*, 14(2), pp. 1450–1460, 2025.
13. Umay Yadav, V.D. Bondre, S.V. Bondre, et al., "Intelligent Cervical Cancer Detection: Empowering Healthcare with Machine Learning Algorithms," *IAES International Journal of Artificial Intelligence*, 14(1), pp. 298–306, 2025.
14. Rashmi Ashtagi, Vaishali Rajput, Sonali Antad, et al., "Cervical Cancer Prediction Using Machine Learning," *Journal of Electrical Systems*, 20(1s), 2024.
15. TelsNet Model Authors, "TelsNet: Temporal Lesion Network Embedding in a Transformer Model to Detect Cervical Cancer," *International Journal of Advances in Intelligent Informatics (IJAIN)*, 9(3), 2024.
16. Xue Feng, Mohamed Ziad Altabel, Rahib H. Abiyev, "Enhancing Cervical Pre-Cancerous Classification Using Advanced Vision Transformer," *Diagnostics (Basel)*, 13(18), 2884, 2023.
17. Tan S. L., Selvachandran G., Ding W., Paramesran R., Kotecha K., "Cervical Cancer Classification From Pap Smear Images Using Deep

- Convolutional Neural Network Models”, *Interdisciplinary Sciences: Computational Life Sciences*, vol. 16, no. 1, 2023.
18. Pei J., Yu J., Ge P., Bao L., Pang H., Zhang H., “Constructing a Classification Model for Cervical Cancer Tumor Tissue and Normal Tissue Based on CT Radiomics”, *Technology in Cancer Research & Treatment*, 2024.
 19. Zhang Y., Huang L., He H., et al., “High Precision Cervical Precancerous Lesion Classification Method Based on ConvNeXt”, *Bioengineering*, vol. 10, no. 12, article 1424, 2023.
 20. He S., Xiao B., Wei H., Huang S., Chen T., “SVM classifier of cervical histopathology images based on texture and morphological features”, *The Journal of Healthcare Engineering*, vol. 2023, Article ID 220031, 2023.
 21. Prasetyo N., Nurmaini S., Rini D. P., “Comparison of CNN Architectures for Pre-Cancerous Cervical Lesion Classification Based on Colposcopy Images Using IARC and AnnoCerv Datasets”, *Jurnal Sisfokom*, vol. 14, no. 2, 2024.
 22. N. Chamundeeswari and R. Ramachandran, "A Lightweight Method for Cervical Cancer Classification Using Preprocessing Pipeline and Attention-Guided Shallow-CNN", *Utilitas Mathematica*, Vol. 122 (2), 2025.
 23. <http://mde-lab.aegean.gr/index.php/downloads>