Attention-Guided Fusion of EfficientNet-B0 and Swin Transformer for Cervical Cancer Classification

Twisibile Mwalughali[1], Emmanuel C. OGU[2], Evason Karanja[3]
Department of Mathematical Sciences
Pan African University Institute for Basic Sciences Technology and Innovation (PAUSTI), Nairobi, Kenya[1]
Department of Information Technology, Babcock University, Ilishan-Remo, Ogun State, Nigeria[2]
School of Computing and Information Technology (SCIT),
Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya[3]

Abstract—**The interpretation of colposcopy images is a critical yet subjective component of cervical cancer screening. To enhance this process, we propose a novel hybrid deep learning framework for the classification of cervical lesions. Our model integrates EfficientNet-B0, adept at extracting localized hierarchical features, with a Swin-Tiny Transformer, which excels at modeling long-range dependencies and global context. Moving beyond basic fusion techniques, we introduce a novel cross-attention fusion mechanism, augmented with channel and spatial attention modules. This design selectively highlights the most discriminative inter-feature relationships while maintaining computational efficiency. Evaluated on the International Agency for Research on Cancer (IARC) colposcopy image dataset, our framework achieves an accuracy of 94.76%, significantly outperforming a concatenation-based fusion model (83.99%). This represents an absolute improvement of 10.77 percentage points and captures 67.3% of the residual performance margin toward perfect accuracy. The model also demonstrates robust performance across other metrics, including a precision of 94.68%, recall of 94.82%, F1-score of 94.74%, and a Cohen's Kappa of 89.48%. These results indicate that our approach can enhance both the accuracy and reliability of cervical cancer screening, offering valuable support for clinical decision-making.**

Keywords*—*Cervical cancer classification; deep learning model; colposcopy; cross-attention fusion; EfficientNet; Swin Transformer

I. Introduction

Cancer arises when cells in the body begin to grow uncontrollably, forming malignant tumors that can spread and affect various organs [1]. Cervical cancer specifically develops in the cervix, the narrow passage connecting the uterus to the vagina [2]. It mainly develops from long-lasting infections with certain high-risk types of human papillomavirus (HPV) [3], [4], particularly HPV-16 and HPV-18 [5]. While the immune system typically clears the virus naturally, persistent infections can lead to cellular changes that may eventually develop into cancer [4]. This disease represents a significant global health burden, particularly in developing countries where it ranks as the third deadliest cancer among women [6]. Approximately 85% of cervical cancer deaths occur in low- and middle-income countries [3], reflecting unequal access to vaccination, screening, and treatment services. The burden is overwhelmingly highest in sub-Saharan Africa, where countries like Eswatini, Zambia and Malawi report incidence rates up to 25 times greater than those in Western Asia, underscoring a critical need for targeted public health interventions [7]. In response, WHO has launched a global strategy aiming to achieve 90% HPV vaccination coverage, 70% screening participation and 90% treatment access by 2030 [8], [9].

Despite the availability of effective prevention strategies, cervical cancer continues to pose a significant global health challenge [6]. According to projections, the global burden of cervical cancer is expected to rise substantially over the coming decades due to population growth and aging. Estimates indicate that the number of cases will increase by 56.8% between 2022 and 2050, reaching nearly 1 million cases, while deaths are projected to rise by 80.7%, totaling approximately 630,000 by 2050 [7]. Patient prognosis varies significantly depending on the stage at which cervical cancer is diagnosed. When detected early, the five-year relative survival rate is approximately 91% [8]. However, survival rates decline sharply in advanced stages, with metastatic cases showing a five-year relative survival rate of about 19% . This big difference shows why it is so important to find and treat cervical cancer early, especially in places where people do not have easy access to healthcare and are often diagnosed late.

Getting an accurate and timely diagnosis is crucial to reduce mortality from cervical cancer [10]. The standard tiered approach utilizes Pap smears, HPV testing, and colposcopy [6]. Primary screening typically begins with a Pap smear to detect abnormal cells or HPV testing to identify high-risk viral strains. Positive results from these initial tests are followed by colposcopy, which provides a magnified visual examination of the cervix to locate and assess abnormalities and guide targeted biopsies for definitive diagnosis. Although Pap smears and HPV tests are effective screening tools, their utility is limited by cost, laboratory requirements, and the potential for false results [11]. Colposcopy is therefore a critical diagnostic follow-up that enables timely intervention [12]. However, its effectiveness is constrained by reliance on clinician expertise for interpreting visual findings, a challenge that is particularly pronounced in resource-limited settings [6]. This dependence on subjective interpretation and the scarcity of trained specialists underscore the need for assistive technologies, thereby motivating the integration of artificial intelligence (AI) to standardize and enhance colposcopic evaluation [6].

Recent advances in artificial intelligence, particularly deep learning models such as convolutional neural networks (CNNs) and vision transformers like the Swin Transformer, have emerged as powerful tools in medical imaging [13], [14]. These models excel at analyzing complex patterns with high accuracy and consistency, often detecting subtle tissue changes that are imperceptible to the human eye [15]. Such capabilities make deep learning a promising pathway for developing diagnostic tools that address current limitations in early detection and support the WHO's goal of eliminating cervical cancer as a public health problem [10]. To build on this progress, hybrid architectures that integrate CNNs and Transformer-based models are gaining recognition for their effectiveness in cancer classification [16]. CNNs specialize in local spatial feature extraction [17], while Transformers capture long-range dependencies [18], [19], and their combination enables a more holistic analysis of cervical images. Despite this potential, most fusion strategies still rely on limited approaches such as concatenation or element-wise addition, which fail to fully exploit the complementarity between the two architectures. Addition assumes features can be summed directly, but this often results in features with larger magnitudes overwhelming those with smaller values, erasing critical discriminative information [20]. Conversely, concatenation preserves all information by simply appending feature vectors. While this avoids the suppression issue, it naively treats all features as equally important and substantially increases dimensionality, leading to higher computational cost and reduced model generalization [20], [21].

To address these limitations, we propose a novel hierarchical attention-guided fusion mechanism that differs fundamentally from existing fusion strategies. Unlike prior approaches that use either simple concatenation or addition, our method introduces three complementary attention modules in a carefully designed sequence, each addressing specific weaknesses of conventional fusion. Channel attention adaptively adjusts the importance of each feature channel, ensuring that channels carrying diagnostically meaningful patterns receive appropriate emphasis [27]. Spatial attention highlights the most relevant anatomical regions within the feature channels, guiding the model to focus on clinically meaningful lesion areas rather than diffuse background regions [28]. Finally, the cornerstone of our method, bidirectional cross-attention enables context-aware interaction between the CNN and Transformer branches. Unlike naïve fusion techniques, cross-attention allows each branch to selectively query and integrate information from the other, resulting in a richer and semantically aligned joint representation. Collectively, this integrated architecture minimizes redundancy, aligns semantic representations, and substantially enhances the discriminative power of the hybrid model for cervical cancer diagnosis.

Our key contributions in this study are:

- We developed a hybrid attention-guided CNN–Transformer architecture by integrating EfficientNet-B0 and the Swin Transformer, resulting in a tailored model optimized for accurate cervical cancer classification from colposcopic images.
- We developed a multi-stage attention fusion mechanism that integrates channel attention, spatial attention, and bidirectional cross-attention to effectively refine and combine complementary local and global features, resulting in enhanced predictive performance.
- We validated the effectiveness of the proposed attention-guided fusion strategy through systematic comparison with a simple concatenation-based fusion approach.

II. Literature Review

A. Deep Learning

Deep learning is a branch of machine learning that uses artificial neural networks to automatically learn patterns and features directly from raw data eliminating the need for manual feature engineering [13]. Unlike traditional methods, deep learning excels in tasks like medical imaging by detecting intricate patterns through layered abstraction [40]. CNNs are particularly effective, using trainable filters to capture spatial features and pooling layers to improve invariance to scale, rotation, and translation. Their hierarchical structure enables learning from simple to complex patterns ideal for identifying subtle tissue abnormalities or cellular irregularities [17]. Advanced architectures such as ResNet [41], DenseNet [42], EfficientNet [29], further enhance diagnostic accuracy and efficiency by incorporating residual connections, dense connectivity and compound scaling.

1) EfficientNet: EfficientNet represents a significant advancement in convolutional neural network (CNN) architecture through its innovative compound scaling methodology. Unlike traditional approaches that scale network depth, width, or input resolution independently, EfficientNet employs a principled compound scaling technique that simultaneously and proportionally scales all three dimensions [29]. This strategy, governed by a set of optimally determined scaling coefficients, ensures a balanced increase in model capacity becoming deeper and wider while processing higher-resolution images without compromising computational efficiency. The resulting architecture hierarchically captures increasingly complex features while maintaining a compact and efficient design. When benchmarked against conventional CNNs such as ResNet, DenseNet, and Inception, EfficientNet demonstrates superior performance. For instance, EfficientNet-B7 achieves state-of-the-art performance, attaining 84.3% top-1 accuracy on ImageNet with approximately 66 million parameters [29]. This outperforms deeper networks like ResNet-152, which uses a comparable parameter count but yields lower accuracy, and also surpasses both Inception-v3 and DenseNet in accuracy while requiring fewer computational resources. These characteristics make EfficientNet particularly suitable for medical imaging applications, where high diagnostic precision must often be achieved under significant computational constraints. Among the EfficientNet family (B0–B7), EfficientNet-B0 was selected as the backbone feature extractor in this study due to its optimal balance of accuracy, parameter efficiency, and computational cost. With only 5.3 million parameters and 0.39 GFLOPs, EfficientNet-B0 is capable of extracting fine-grained local features from high-resolution colposcopy images without requiring excessive memory or training time. This makes it particularly suitable for medical imaging applications where computational resources are often limited and high diagnostic precision is essential.

EfficientNet has been widely adopted in medical imaging, with successful applications in breast cancer diagnosis [43], [44], lung cancer detection [45] and cervical cancer [46], [47]. Despite these promising results, EfficientNet shares the limitations of CNN-based models in capturing long-range dependencies, which motivates its integration with transformer-based architectures for enhanced performance in complex classification tasks.

2) Transformers: The Transformer architecture, introduced by Vaswani et al. in 2017 [35], is distinguished by its capacity to learn long-range contextual relationships within sequential data. This is achieved by segmenting input sequences into smaller units (tokens) augmented with positional encodings. The model employs a self-attention mechanism to derive feature representations, enabling every token to interact with all other tokens in the sequence. This design allows direct and parallelized modeling of relationships between all token pairs, regardless of their distance, while maintaining uniform processing across the sequence. In contrast to convolutional neural networks (CNNs), which are constrained by limited receptive fields [19], Transformers leverage self-attention to incorporate global contextual information from the entire input [18]. This capability is especially beneficial in tasks requiring holistic sequence understanding such as natural language processing and image classification laying the groundwork for their adaptation in visual recognition.

The Vision Transformer (ViT) represents a significant shift in image classification methodology by adapting the Transformer architecture for visual data [48]. Rather than using convolutional operations to induce spatial hierarchies, ViT partitions an image into fixed-size non-overlapping patches, treating each as an individual token. These patch-level tokens are processed by a standard Transformer encoder, which models relationships across the entire image [49]. A typical configuration, ViT-Base, consists of 12 encoder layers that apply multi-head self-attention to extract discriminative features. The final stages include layer normalization which aids training stability at a computational cost and a classification token that aggregates global image representations for prediction.

Building on ViT, the Swin Transformer introduces a hierarchical representation learning strategy [30]. Images are divided into non overlapping patches, which are progressively merged into larger blocks, enabling the model to capture features at multiple scales from local details to global context. A key innovation is the use of shifted windows between layers, which promotes cross-window interaction and enlarges the effective receptive field without resorting to global self-attention. This design enhances flexibility across resolutions and complexities, making Swin Transformer particularly suited for vision tasks requiring fine-grained spatial and contextual understanding. Transformers have shown promising results in medical image analysis, with successful applications in breast cancer classification [50], [51], pulmonary nodule detection [52] and cervical image analysis [53], [54]. These studies underscore the effectiveness of transformer-based models in capturing complex spatial dependencies and enhancing diagnostic accuracy across diverse clinical tasks.

The Swin-Tiny variant was chosen to construct a hybrid backbone with EfficientNet-B0. Its compact architecture provides a compelling balance of representational power and computational efficiency, enabling the effective processing of high-resolution colposcopy images. This lightweight design, which avoids the excessive demands of larger models like ViT-B or Swin-Base, ensures a practical and efficient integration for our diagnostic task.

B. CNN-Transformer Hybrid Models in Medical Imaging

Recent studies have demonstrated the promise of CNN–Transformer hybrids in cervical cancer diagnosis. For example, Mohammed et al. [22] introduced a hybrid model for colposcopy image classification that fused EfficientNet-B0 with a Swin-Tiny Transformer, achieving an AUC of 0.96 and an accuracy of 87.5%. Despite these promising results, their fusion method weighted all features equally, highlighting the need for attention-based mechanisms to better prioritize diagnostically relevant patterns. Similarly, Wang et al. [23] proposed a hybrid architecture combining a 3D convolutional neural network (CNN) and a transformer for predicting N-staging and survival outcomes in non-small cell lung cancer (NSCLC). Evaluated on the NSCLC Radiogenomics and NSCLC-Radiomics datasets, their framework leveraged the CNN to extract local spatial features from volumetric images while the transformer captured global contextual dependencies. Features from both modules were fused via simple concatenation prior to classification. Although the model achieved competitive accuracies of 0.805, 0.828, and 0.819 on the training, validation, and test sets, respectively, the reliance on direct feature concatenation may limit effective cross-modal interaction between CNN and transformer representations. This highlights the need for more sophisticated fusion strategies that can dynamically weigh and integrate complementary features for improved representation learning.

Chen et al. [24] proposed the MFEM-CIN framework, a hybrid CNN–Transformer model for colposcopic image analysis that employs multi-scale feature extraction (MSFE) and multi-scale feature fusion (MSFF) to integrate local and global features. While the multi-scale fusion aims to enrich semantic representation by allowing shallow and deep features to interact, the study does not critically evaluate how effectively the fusion mechanism prioritizes diagnostically relevant patterns. The MSFF approach, though conceptually sound, relies on heuristic fusion without attention-guided weighting, which may limit the model's ability to emphasize clinically significant regions over less informative ones. Consequently, while the framework demonstrates high overall accuracy, the fusion strategy itself may not fully exploit the complementary strengths of CNN and Transformer features, suggesting room for improvement through more advanced, attention-based fusion techniques.

Other notable work includes the research of Liu et al. [25], who developed the CVM-Cervix framework for Pap smear classification. In this approach, a CNN module extracted local cellular details, while a Vision Transformer captured global spatial dependencies within the images. Moving beyond simple fusion strategies, the authors employed a multilayer perceptron (MLP) to integrate these complementary feature sets, achieving an accuracy of 91.72%.

In another study, Abinaya and Sivak [55] proposed a deep learning-based cervical classification system integrating 3D convolutional neural networks (3D CNN) with Vision Transformer (ViT) modules. The 3D CNN extracted spatiotemporal features from cervical images, while multiple ViT models captured higher-level global representations. A 3D feature pyramid network (FPN) was used for feature concatenation, and a 3D squeeze-and-excitation (SE) block reweighted the fused features, enhancing the discriminative power of diagnostically relevant information. Classification was then performed using a kernel extreme learning machine (KELM) with a radial basis kernel. This fusion strategy, particularly the SE-based feature reweighting, contributed to strong performance, achieving an accuracy of 98.6% and demonstrating its potential as an effective diagnostic support tool for cancer detection.

In a domain beyond human healthcare, Dula et al. [26] investigated the integration of CNNs and Transformers for muzzle-based cattle identification. The study proposed a novel Multi-Head Attention Feature Fusion (MHAFF) mechanism, which effectively combines CNN and Transformer features while preserving their complementary strengths. Experimental evaluations on two public cattle datasets yielded outstanding accuracy rates of 99.88% and 99.52%, surpassing both conventional fusion methods and existing identification systems.

While advanced attention-based fusion strategies, such as the Multi-Head Attention Feature Fusion (MHAFF) used in cattle identification [26], have achieved near-perfect accuracy by effectively balancing CNN and Transformer features, this level of sophistication remains largely untapped in cervical cancer diagnosis. Most current approaches still employ uniform or simplistic fusion schemes that cannot adaptively highlight subtle lesion characteristics. This gap highlights a significant opportunity to develop attention-guided frameworks capable of selectively prioritizing clinically relevant features, thereby enhancing the accuracy of cervical cancer classification.

III. Methodology

This section presents the proposed cervical cancer classification model, illustrated in Fig. 1. The model performs binary classification (normal vs. abnormal) by fusing EfficientNet-B0 and Swin-Tiny through attention-guided mechanisms. Both models accept inputs of 224 × 224 pixels, facilitating seamless alignment and fusion of features within the hybrid backbone. The complementary strengths of these architectures form a robust foundation for feature extraction. To further refine the representations, channel and spatial attention modules (SAM) were applied separately to each branch, enhancing informative channels and highlighting salient regions within their respective domains. Optimizing each branch independently ensures that the subsequent cross-attention mechanism receives more distinct and powerful features, enabling a more effective and informative fusion process.

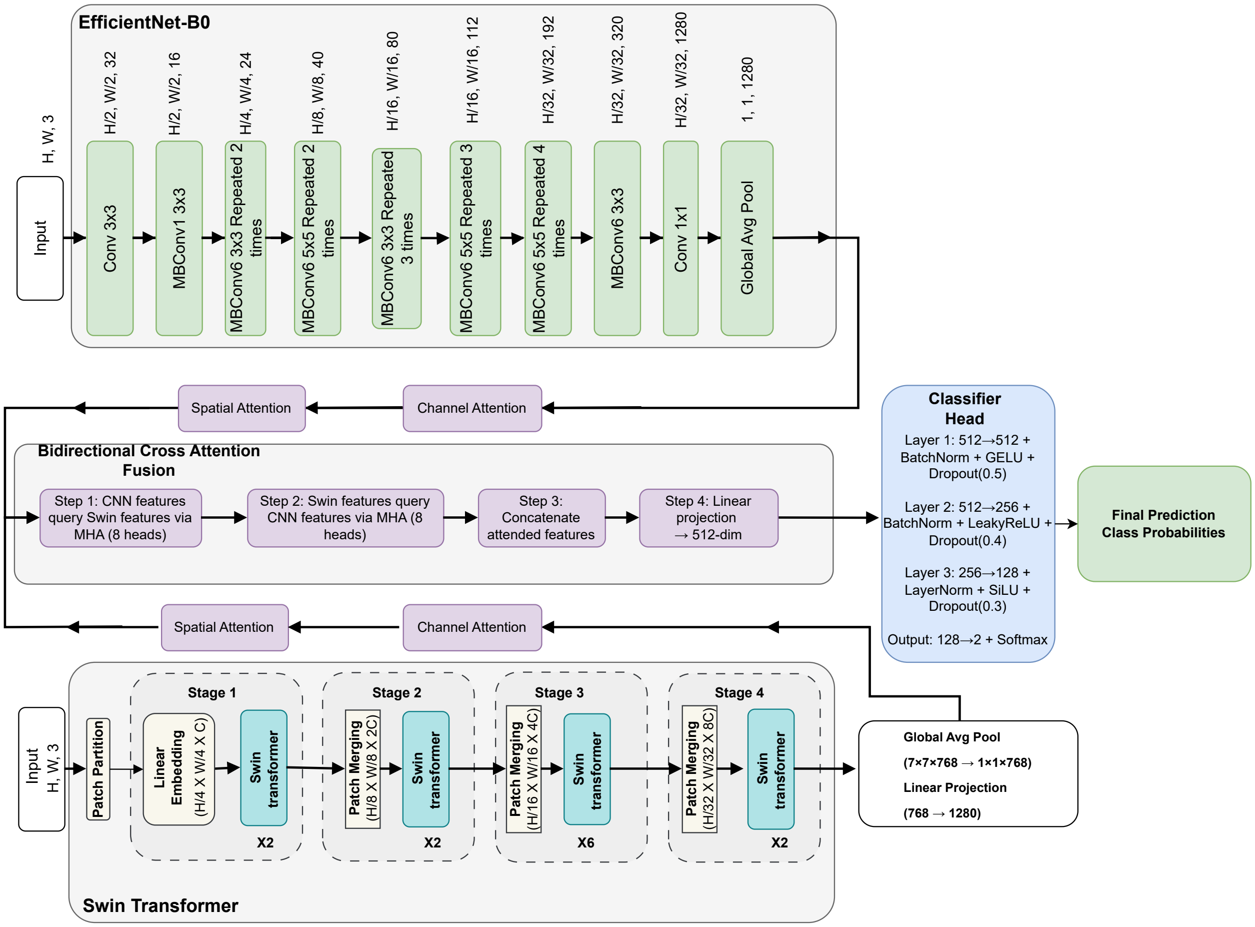
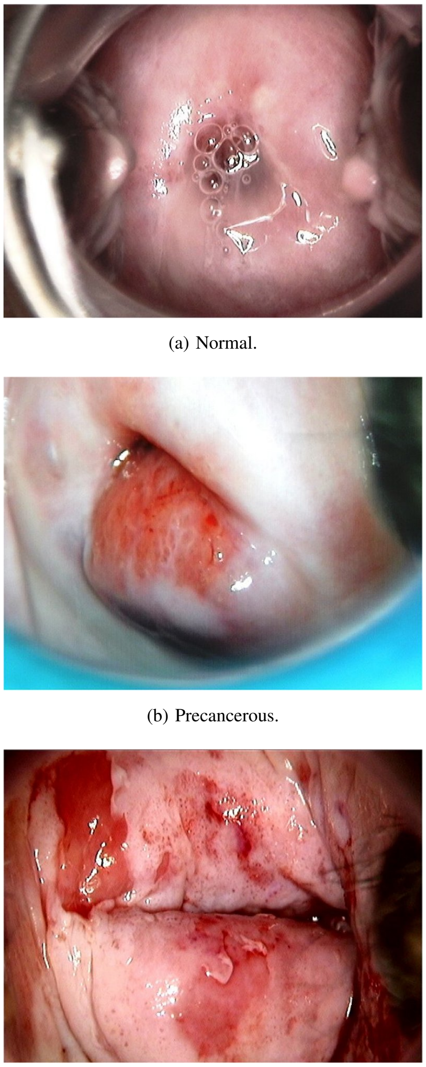

EfficientNet-B0
H, W, 3
Input
H/2, W/2, 32
Conv 3x3
H/2, W/2, 16
MBConv1 3x3
H/4, W/4, 24
MBConv6 3x3 Repeated 2 times
H/8, W/8, 40
MBConv6 5x5 Repeated 2 times
H/16, W/16, 80
MBConv6 3x3 Repeated 3 times
H/16, W/16, 112
MBConv6 5x5 Repeated 3 times
H/32, W/32, 192
MBConv6 5x5 Repeated 4 times
H/32, W/32, 320
MBConv6 3x3
H/32, W/32, 1280
Conv 1x1
1, 1, 1280
Global Avg Pool
Spatial Attention
Channel Attention
Bidirectional Cross Attention Fusion
Step 1: CNN features query Swin features via MHA (8 heads)
Step 2: Swin features query CNN features via MHA (8 heads)
Step 3: Concatenate attended features
Step 4: Linear projection → 512-dim
Classifier Head
Layer 1: 512→512 + BatchNorm + GELU + Dropout(0.5)
Layer 2: 512→256 + BatchNorm + LeakyReLU + Dropout(0.4)
Layer 3: 256→128 + LayerNorm + SiLU + Dropout(0.3)
Output: 128→2 + Softmax
Final Prediction Class Probabilities
Spatial Attention
Channel Attention
Input H, W, 3
Patch Partition
Stage 1
Linear Embedding (H/4 X W/4 X C)
Swin transformer
X2
Stage 2
Patch Merging (H/8 X W/8 X 2C)
Swin transformer
X2
Stage 3
Patch Merging (H/16 X W/16 X 4C)
Swin transformer
X6
Stage 4
Patch Merging (H/32 X W/32 X 8C)
Swin transformer
X2
Global Avg Pool
(7×7×768 → 1×1×768)
Linear Projection
(768 → 1280)
Swin Transformer

Fig. 1. Proposed dual-branch architecture combining EfficientNet-B0 and Swin Transformer. The model incorporates attention-guided feature fusion to enhance classification performance.

A. Datasets

This study utilized two datasets. The first dataset was obtained upon request from the Colposcopy Image Bank of the International Agency for Research on Cancer (IARC)[31]. The original collection included 913 images from 200 patients. After excluding 15 images from three cases with inconclusive clinical assessments, 898 images from 197 cases were retained. These were classified into three categories based on histopathological metadata: normal tissue (93 cases), precancerous lesions (78 cases), and invasive cancer (26 cases). Representative examples from each category are shown in Fig. 2.



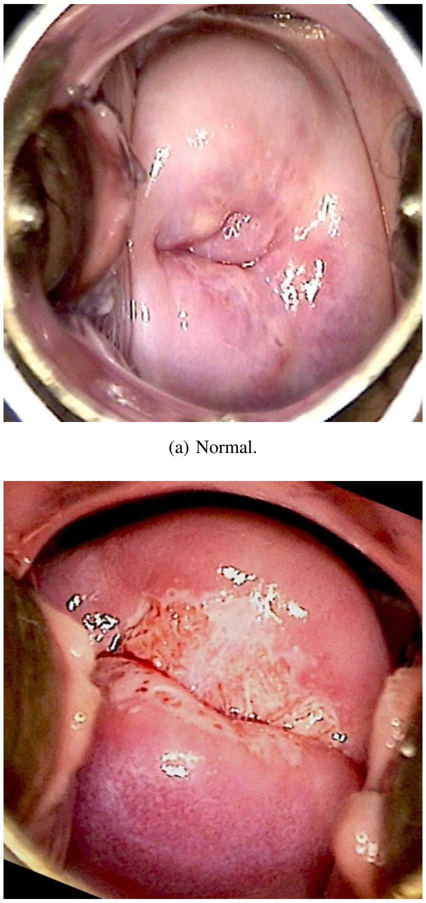
(a) Normal.

(b) Precancerous.

(c) Cancer.

Fig. 2. Colposcopy image samples from the IARC dataset.

The second dataset was sourced from the publicly available Colposcopy Images collection on Kaggle, curated by Siti Nursyafiqah [32]. This dataset contains 2617 images with binary labels: Normal (1445 images) and Abnormal (1172 images). Sample images from this collection are also shown in Fig. 3.



(a) Normal.

(b) Abnormal.

Fig. 3. Representative colposcopy image samples from the Kaggle dataset: Normal and abnormal.

The IARC dataset was reserved exclusively for testing to evaluate the model's generalizability to a new, clinically validated source. A binary classification task was created by merging the precancerous and cancer categories into a single Abnormal class, and the test set was balanced at the case level to avoid bias from patients contributing many images. Iodine-stained images were removed from the dataset to ensure consistency in image appearance. In total, the test set comprised 317 Normal images from 90 cases and 370 Abnormal images from 90 cases (including all 26 cancer cases and 64 precancerous cases).

The Kaggle dataset, which originally contained 1445 Normal and 1172 Abnormal images, was used for training and validation. To ensure consistency with the IARC distribution, a validation set of 317 Normal and 370 Abnormal images was first selected directly from the original Kaggle dataset prior to augmentation. The remaining 1128 Normal and 802 Abnormal images were then augmented fourfold using controlled transformations: brightness adjustment (factors between 0.5× and 1.5×), random zooming up to 0.6×, and random cropping to simulate variability in image framing.All augmented images were stored separately to ensure traceability, producing an augmented pool of 4512 Normal and 3208 Abnormal images. From this augmented pool, 2536 Normal and 2960 Abnormal images were allocated to the training set. Together with the 317/370 validation images and the 317/370 IARC test images, this produced an exact 80:10:10 split across training, validation, and testing, yielding a per class total of 3170 Normal and 3700 Abnormal images for the experiment(see Fig. 4 and Table I). To address the class imbalance, class weights were applied in the weighted categorical cross-entropy loss function, giving higher importance to the underrepresented Normal class and reducing bias toward the majority Abnormal class.

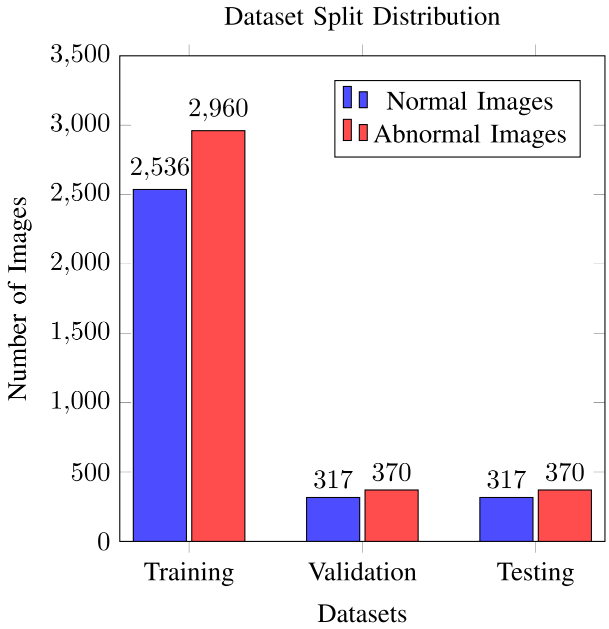


Fig. 4. Distribution of normal and abnormal images across training, validation, and testing sets.

TABLE I. NUMERICAL DISTRIBUTION OF NORMAL AND ABNORMAL IMAGES ACROSS TRAINING, VALIDATION, AND TESTING SETS.

Dataset Split	Normal Images	Abnormal Images
Training	2536	2960
Validation	317	370
Testing	317	370
Total	3170	3700

B. Data Preprocessing

All images in the dataset were processed through a standardized preprocessing pipeline to ensure consistency, improve model convergence, and enhance generalization. The pipeline consisted of the following steps:

Resizing: All images were resized to 224 × 224 pixels to match the input requirements of the ImageNet-pretrained architectures used in this study, specifically EfficientNet-B0 and Swin Transformer-Tiny [29], [30]. This standardization ensures consistent spatial dimensions across samples and facilitates seamless feature fusion between model branches.

Normalization: Pixel values were first scaled from the original 0–255 range to [0,1] by dividing by 255. They were then normalized to the range [-1,1] by subtracting 0.5 and dividing by 0.5 for each color channel. This centers the data distribution, stabilizes gradient-based learning, accelerates convergence, and aligns input statistics with those used during pretraining of the backbone models [13].

Data Augmentation: To increase the diversity of the training data and mitigate overfitting [33], several augmentation techniques were applied during training. These included brightness adjustment (random scaling of pixel intensity between 0.5× and 1.5×), random zooming (scaling factor up to 0.6× of the original image size), and random cropping to simulate variability in framing and viewpoint.

C. Feature Extraction

Feature extraction is the process of transforming complex, high-dimensional raw data, such as images or text into a compact set of informative numerical attributes known as a feature vector [36], [37]. In this study, feature extraction was performed using two parallel networks: EfficientNet-B0 and Swin Transformer (Tiny configuration with patch size 4 and window size 7), both initialized with ImageNet-pretrained weights. To adapt the networks to the colposcopy image classification task, the original classification heads were discarded and the remaining network parameters were fine-tuned during training. This design allows the models to retain their pretrained representational capacity while adjusting to domain-specific visual characteristics present in colposcopy data, forming a stable and efficient foundation for the subsequent stages of the proposed framework [34].

EfficientNet-B0 uses compound scaling to balance network depth, width, and input resolution, allowing it to extract hierarchical features efficiently while keeping computational costs low [29]. The input image $\mathbf{X} \in \mathbb{R}^{3\times224\times224}$ is processed through a sequence of convolutional layers and mobile inverted residual blocks, which progressively capture features from low-level spatial details, such as edges and textures, to high-level semantic patterns relevant for classification. After these layers, a global average pooling operation is applied, which compresses each feature map into a single value by averaging over its spatial dimensions. This produces a 1280-dimensional feature vector, where each dimension represents a summary of a distinct feature detected by the network. This vector is then refined through channel and spatial attention modules,followed by bidirectional cross-attention with the Swin Transformer features, making it suitable for integration with them for subsequent processing.

The Swin Transformer (Tiny variant, patch size 4×4, window size 7×7) processes the image as a sequence of non-overlapping patches [30]. For an input of size 224×224, the image is partitioned into $\frac{224}{4} \times \frac{224}{4} = 3136$ patches. After propagation through the transformer blocks, the final patch embeddings are averaged to produce a 768-dimensional feature vector. To ensure dimensional compatibility with the EfficientNet branch, a linear projection is applied as shown in Eq. (1):

$$\mathbf{f}_{\text{swin}} = W^{\top} \cdot \text{Swin-T}(\mathbf{X}) + b, \quad \mathbf{f}_{\text{swin}} \in \mathbb{R}^{1280}, \tag{1}$$

where, $W \in \mathbb{R}^{768\times1280}$ and $b \in \mathbb{R}^{1280}$ denote the learnable projection matrix and bias term, respectively.

D. Channel Attention Module

A Channel Attention Module, inspired by the Squeeze-and-Excitation network [27], is integrated to adaptively recalibrate channel-wise feature responses by modeling interdependencies between channels. Given an input feature map $\mathbf{f} \in \mathbb{R}^{B\times C\times H\times W}$, a global average pooling operation is first applied to generate a compact channel descriptor $\mathbf{z} \in \mathbb{R}^{B\times C}$. The squeeze operation aggregates spatial information into a single channel-wise statistic, formulated as Eq. (2):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_c(i,j), \tag{2}$$

where, z_c represents the aggregated response of the c-th channel. Next, an 'excitation' step is performed to capture inter-channel dependencies through a bottleneck structure comprising two fully connected layers. This transformation is expressed in Eq. (3):

$$\mathbf{a} = \sigma(\mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \mathbf{z})), \tag{3}$$

where, $\mathbf{W}_1 \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times C/r}$ are learnable parameters, $r = 16$ is the reduction ratio, δ denotes the ReLU activation, and σ is the sigmoid function. The resulting activation vector $\mathbf{a}$ serves as channel attention weights. Finally, these weights are applied to the original feature map via channel-wise multiplication, as shown in Eq. (4):

$$\tilde{\mathbf{f}} = \mathbf{a} \otimes \mathbf{f}, \tag{4}$$

where, $\otimes$ denotes element-wise channel scaling. This mechanism strengthens informative channels while suppressing less relevant ones, thereby improving the discriminative power of the learned representation.

E. Spatial Attention Module (SAM)

To complement the channel-wise focus, a SAM is incorporated to direct the model's focus towards the most informative spatial regions within the feature maps [28]. This module generates a spatial attention map that highlights where an informative part is located, regardless of channel-specific responses. Given an input feature map $\mathbf{f} \in \mathbb{R}^{B\times C\times H\times W}$, we first aggregate channel information by applying both average pooling and max pooling operations along the channel axis, as shown in Eq. (5) and Eq. (6), respectively:

$$f_{\text{avg}} = \frac{1}{C} \sum_{c=1}^{C} f_c, \quad f_{\text{avg}} \in \mathbb{R}^{B\times 1\times H\times W} \tag{5}$$

$$f_{\max} = \max_{c=1,\dots,C} f_c, \quad f_{\max} \in \mathbb{R}^{B\times 1\times H\times W} \tag{6}$$

These maps are then concatenated along the channel dimension to form a composite feature descriptor:

$$f_{\text{cat}} = [f_{\text{avg}}; f_{\max}], \quad f_{\text{cat}} \in \mathbb{R}^{B\times 2\times H\times W} \tag{7}$$

where, $[;]$ denotes concatenation.

A convolutional layer with a 7 × 7 kernel is applied to this concatenated descriptor in Eq. 7 to produce a preliminary spatial map. This is followed by a sigmoid activation function σ to generate the final spatial attention map $\mathbf{a_s}$ in Eq. (8), with values constrained between 0 and 1:

$$\mathbf{a_s} = \sigma\left(\text{Conv}^{7\times 7}(f_{\text{cat}})\right), \quad \mathbf{a_s} \in \mathbb{R}^{B\times 1\times H\times W} \tag{8}$$

The output of the module is obtained by multiplying this attention map in Eq. (8) across every channel of the original input feature map, as shown in Eq. (9):

$$\tilde{f} = \mathbf{a_s} \otimes \mathbf{f} \tag{9}$$

where, $\otimes$ denotes element-wise multiplication. This mechanism emphasizes feature responses in salient regions while suppressing information from irrelevant areas, thereby enhancing the model's spatial discriminative capabilities.

F. Bidirectional Cross-Attention Mechanism

To facilitate synergistic interaction between the feature representations from the EfficientNet-B0 and Swin Transformer branches, a bidirectional cross-attention mechanism with multi-head attention (MHA) was implemented. This allows each branch to dynamically attend to and assimilate relevant features from the other, fostering a more cohesive and discriminative fused representation.

Let $\mathbf{F}_e \in \mathbb{R}^{B\times L\times C}$ and $\mathbf{F}_s \in \mathbb{R}^{B\times N\times D}$ denote the reshaped feature tensors from the EfficientNet and Swin Transformer branches, respectively, where $L = H \times W$ represents the spatial token length, N is the number of Swin tokens, and C, D are their corresponding feature dimensions. The bidirectional cross-attention is formalized through two symmetric attention blocks:

1) EfficientNet to Swin Transformer (E → S) direction: The Swin Transformer features act as the Query, seeking information from the EfficientNet branch's Key and Value [35]:

$$Q_s = F_s W_s^Q, \qquad Q_s \in \mathbb{R}^{B\times N\times d_k}, \tag{10a}$$
$$K_e = F_e W_e^K, \qquad K_e \in \mathbb{R}^{B\times L\times d_k}, \tag{10b}$$
$$V_e = F_e W_e^V, \qquad V_e \in \mathbb{R}^{B\times L\times d_v}. \tag{10c}$$

2) Swin Transformer to EfficientNet (S → E) direction: The EfficientNet features act as the Query, incorporating contextual information from the Swin branch:

$$Q_e = F_e W_e^Q, \qquad Q_e \in \mathbb{R}^{B\times L\times d_k}, \tag{11a}$$
$$K_s = F_s W_s^K, \qquad K_s \in \mathbb{R}^{B\times N\times d_k}, \tag{11b}$$
$$V_s = F_s W_s^V, \qquad V_s \in \mathbb{R}^{B\times N\times d_v}. \tag{11c}$$

where, $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ in Eq. (10) and Eq. (11) are learnable projection weights.

The Multi-Head Attention (MHA) mechanism with $h = 8$ heads is applied independently in each direction. For the E $\rightarrow$ S direction, the projected Query, Key, and Value tensors are first split into h distinct heads along the feature dimension. The operations for the i-th head are then defined as shown in the Eq. (12) :

$$Q_s^i = Q_s W_{Q_s}^i, \qquad Q_s^i \in \mathbb{R}^{B\times N\times d_h}, \tag{12a}$$

$$K_e^i = K_e W_{K_e}^i, \qquad K_e^i \in \mathbb{R}^{B\times L\times d_h}, \tag{12b}$$

$$V_e^i = V_e W_{V_e}^i, \qquad V_e^i \in \mathbb{R}^{B\times L\times d_h}. \tag{12c}$$

where, $d_h = d_k/h = d_v/h$ is the dimension per head, and $\mathbf{W}_{Q_s}^i, \mathbf{W}_{K_e}^i, \mathbf{W}_{V_e}^i$ are the projection weights for the i-th head. For each head, the scaled dot-product attention is computed independently. This involves calculating a compatibility score between each Query and Key vector, which determines how much focus to place on different parts of the input [see Eq. (13)]:

$$\text{Attention}(Q_s^i, K_e^i, V_e^i) = \text{softmax}\left(\frac{Q_s^i (K_e^i)^\top}{\sqrt{d_h}}\right) V_e^i. \tag{13}$$

The softmax function is applied row-wise to generate a probability distribution (attention weights) over the L spatial locations for each of the N tokens. The output for each head is consequently a weighted sum of the Value vectors, with weights determined by the Query-Key compatibility. The outputs of all h heads are then concatenated along the feature dimension, as shown in the Eq. (14):

$$\begin{aligned} H_s &= [\text{Head}_s^1; \text{Head}_s^2; \ldots; \text{Head}_s^h], \\ H_s &\in \mathbb{R}^{B\times N\times(h\cdot d_h)} = \mathbb{R}^{B\times N\times d_v}. \end{aligned} \tag{14}$$

Finally, the concatenated output is linearly projected back to the original dimension D using a learned output matrix $\mathbf{W}_O \in \mathbb{R}^{d_v\times D}$, [see Eq. (15)]:

$$\text{MHA}(Q_s, K_e, V_e) = H_s W_O. \tag{15}$$

This multi-head design allows the model to jointly attend to information from different representation subspaces at different positions, effectively capturing a diverse set of complementary features. The final refined feature for the Swin branch is obtained via a residual connection and layer normalization, as shown in Eq. (16):

$$\tilde{F}_s = \text{LayerNorm}(F_s + \text{MHA}(Q_s, K_e, V_e)). \tag{16}$$

An identical process is applied in the S $\rightarrow$ E direction to obtain $\tilde{\mathbf{F}}_e$. The enriched feature representations, $\tilde{\mathbf{F}}_s$ and $\tilde{\mathbf{F}}_e$, produced by the bidirectional cross-attention module are fused via global average pooling followed by concatenation, yielding a unified feature vector $\mathbf{z} \in \mathbb{R}^{B\times 512}$. This vector is subsequently processed by a Multi-Layer Perceptron (MLP) classifier to transform the high-level features into a probability distribution over the C target classes, where C = 2 in our case.

G. Multi-Layer Perceptron (MLP) Classifier

The fused feature vector $z \in \mathbb{R}^{512}$, obtained from the bidirectional cross-attention module, is processed by a Multi-Layer Perceptron (MLP) classifier. This network transforms the input features into a probability distribution over the target classes. The classifier comprises three hidden layers, progressively reducing the dimensionality from 512 to 128, followed by a linear output layer mapping to the number of output classes C. The dimensionality reduction follows the sequence: $512 \rightarrow 256 \rightarrow 128 \rightarrow C$.

The input to the first layer is the fused feature vector $h^{(0)} = z$. The layer applies a linear transformation using a weight matrix $W^{(1)} \in \mathbb{R}^{512\times 512}$ and bias vector $b^{(1)} \in \mathbb{R}^{512}$, as shown in Eq. (17). Batch normalization is applied to the linear output, followed by a GELU activation and dropout. The resulting output, hereafter denoted as $h^{(1)}$, serves as the input to the next layer.

$$z^{(1)} = h^{(0)} W^{(1)} + b^{(1)} \tag{17}$$

The second layer receives $h^{(1)} \in \mathbb{R}^{512}$ and projects it to 256 dimensions using $W^{(2)} \in \mathbb{R}^{512\times 256}$ and $b^{(2)} \in \mathbb{R}^{256}$, as shown in the Eq. (18).The linear output is subsequently normalized through batch normalization, followed by the application of a LeakyReLU activation function with a negative slope and dropout, producing $h^{(2)}$, which serves as the input to the third layer.

$$z^{(2)} = h^{(1)} W^{(2)} + b^{(2)} \tag{18}$$

The third hidden layer reduces the dimensionality of $h^{(2)}$ from 256 to 128 using the weight matrix $W^{(3)} \in \mathbb{R}^{256\times 128}$ and bias vector $b^{(3)} \in \mathbb{R}^{128}$, as shown in Eq. (19). The linear output is subsequently normalized through layer normalization, followed by the application of a SiLU (Swish) activation function and dropout, yielding the final hidden representation $h^{(3)}$.

$$z^{(3)} = h^{(2)} W^{(3)} + b^{(3)} \tag{19}$$

The output layer linearly projects $h^{(3)} \in \mathbb{R}^{128}$ to C dimensions using $W^{(4)} \in \mathbb{R}^{128\times C}$ and $b^{(4)} \in \mathbb{R}^{C}$, producing the logits u, as shown in Eq. (20). These logits are then converted into a probability distribution over the output classes using the softmax function [Eq. (21)]. Finally, the predicted class $\hat{c}$ is obtained by selecting the class with the highest probability, as defined in Eq. (22).

$$u = h^{(3)} W^{(4)} + b^{(4)} \tag{20}$$

$$\hat{y} = \text{softmax}(u) \tag{21}$$

$$\hat{c} = \arg\max_j \hat{y}_j \tag{22}$$

IV. Experimental Setup and Evaluations

A. Computing Environment

All experiments were conducted in the Google Colab environment equipped with an NVIDIA GPU, leveraging the CUDA architecture for accelerated deep learning computations. The software environment was built on Python 3.10.12 using the PyTorch deep learning framework. Key libraries included torchvision for image preprocessing and data loading, timm (Torch Image Models) for access to pre-trained backbone networks, and scikit-learn for comprehensive evaluation metrics.

B. Training Configuration and Hyperparameters

For both the baseline concatenation model and the proposed attention-guided fusion model, identical training configurations were employed to ensure a fair comparison. The training hyperparameters used in this study are summarized in Table II. The models were trained for 50 epochs with a batch size of 8 using the AdamW optimizer, a learning rate of 5×10^{-5}, and a weight decay of 0.01. These hyperparameters were selected based on prior transformer-based literature, preliminary experimental validation, and computational constraints associated with training dual-branch architectures. Model performance was evaluated using 5-fold stratified cross-validation to ensure robustness and reduce variance, and the reported metrics correspond to the aggregated performance across all folds.

AdamW was chosen because its decoupled weight decay improves generalization and helps mitigate overfitting, particularly in transformer-based networks [39]. The learning rate was determined through pilot experiments that demonstrated stable convergence during fine-tuning of pre-trained models, while a weight decay of 0.01 provided effective regularization without limiting model capacity. The batch size of 8 was necessitated by GPU memory limitations; however, smaller batch sizes have also been shown to promote better generalization in medical imaging tasks with moderately sized datasets. Training for 50 epochs ensured adequate convergence, as validation performance plateaued beyond this point during preliminary experiments.

The hidden layers utilized GELU, LeakyReLU, and SiLU activation functions to exploit their complementary strengths: GELU provides smooth non-linearity suitable for transformer components, LeakyReLU prevents the dying neuron problem in convolutional layers, and SiLU introduces self-gating behavior that enhances feature representation. A weighted categorical cross-entropy loss function was applied to address class imbalance within the dataset. Maintaining this consistent training configuration across all experiments ensures that any observed performance differences can be attributed to architectural innovations rather than variations in optimization strategy.

C. Metrics for Evaluating Performance

Model performance was assessed using standard classification metrics. For binary classification, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The following metrics were computed [see Eq. (23) to Eq. (27)]:

TABLE II. Training Hyperparameters used for Both the Baseline Concatenation Model and the Proposed Attention-Guided Fusion Model

Hyperparameter	Value
Epochs	50
Batch Size	8
Optimizer	AdamW
Learning Rate	5×10^{-5}
Weight Decay	0.01
Loss Function	Weighted Categorical Cross-Entropy
Activation Function (Hidden Layers)	GELU, LeakyReLU, SiLU
Activation Function (Output Layer)	Softmax

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

Accuracy measures the overall proportion of correctly classified samples out of all predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

Precision evaluates how many of the samples predicted as positive are actually positive, reflecting reliability in identifying abnormal cases.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

Recall (or sensitivity) measures how many actual positive samples are correctly identified, which is critical in medical diagnosis to minimize missed cases.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

The F1-Score balances precision and recall, providing a single measure of overall classification effectiveness, especially useful when dealing with imbalanced datasets.

$$\text{Cohen's Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{27}$$

Cohen's Kappa measures the agreement between predicted and actual classifications while accounting for agreement occurring by chance. It provides a more robust evaluation than accuracy alone, particularly in imbalanced datasets.

The observed agreement P_o is defined as Eq. (28):

$$P_o = \frac{TP + TN}{TP + TN + FP + FN} \tag{28}$$

The expected agreement P_e is computed as Eq. (29):

$$P_e = \frac{(TP+FP)(TP+FN)}{(TP+TN+FP+FN)^2} + \frac{(FN+TN)(FP+TN)}{(TP+TN+FP+FN)^2} \tag{29}$$

Kappa values range from -1 to 1, where $\kappa = 1$ indicates perfect agreement, $\kappa = 0$ indicates agreement equivalent to random chance, and $\kappa < 0$ indicates disagreement worse than chance. Higher Kappa values, therefore, indicate stronger and more reliable classification performance beyond random coincidence.

To complement these quantitative metrics, performance is further analyzed using visualization techniques such as confusion matrices, Receiver Operating Characteristic (ROC) curves, and learning curves, which provide deeper insight into class-wise behavior, discriminative capability and convergence characteristics of the models.

D. Experimental setup

To rigorously assess the effectiveness of the proposed attention-guided fusion strategy, we implemented two experimental configurations. The baseline model, which used simple concatenation to combine CNN and transformer representations, provided a reference for performance without attention mechanisms. The proposed model incorporated an attention-guided fusion strategy. This approach dynamically reweights features from both branches to emphasize diagnostically relevant patterns.

1) Concatenation-based fusion model (baseline): The concatenation-based fusion model was designed as a baseline to assess the contribution of attention mechanisms. In this configuration, feature representations extracted from EfficientNet-B0 and Swin Transformer–Tiny were directly concatenated along the feature dimension without applying any attention operation. Specifically, the 1280-dimensional feature vector produced by EfficientNet-B0 and the 768-dimensional feature vector obtained from Swin Transformer–Tiny were concatenated to form a 2048-dimensional fused representation. The resulting feature vector was processed by a Multi-Layer Perceptron (MLP) classifier consisting of three sequential hidden layers with 512, 256, and 128 units, respectively. Each hidden layer incorporated normalization and nonlinear activation functions, employing GELU, LeakyReLU, and SiLU activations in successive layers, along with dropout rates of 0.5, 0.4, and 0.3, to improve generalization and mitigate overfitting. The classifier was finalized with a linear output layer for binary classification. Model optimization was performed using the AdamW optimizer with a learning rate of5×10^{-5} and a weight decay of 0.01. This baseline model provides a strong yet simple fusion strategy, serving as a reference point for evaluating the performance gains achieved by the proposed attention-based fusion framework.

2) Attention-based fusion model (proposed model): As illustrated in the Methodology section, the proposed model introduces a multi-stage attention-based fusion framework to effectively integrate features extracted from EfficientNet-B0 and Swin Transformer–Tiny. Following independent feature extraction, the 768-dimensional output of the Swin Transformer–Tiny branch is projected into a 1280-dimensional feature space via a linear transformation to ensure dimensional compatibility with the EfficientNet-B0 branch. Prior to feature fusion, each branch is refined separately using sequential Channel Attention and Spatial Attention modules. The Channel Attention module employs adaptive average pooling followed by a sigmoid-activated MLP to compute channel-wise importance weights, while the Spatial Attention module emphasizes salient spatial regions through mean and max pooling operations followed by a one-dimensional convolution. The core fusion mechanism is implemented through a CrossAttnFusion module, in which features are projected into Query (Q), Key (K), and Value (V) vectors and cross-attention is computed in both directions. The attended outputs from both directions are concatenated using the torch.cat operation and passed through a linear layer with ReLU activation to form a unified representation. This fused representation is subsequently processed by a multi-layer classifier that progressively reduces the feature dimensionality from 512 to 256 and finally 128 units. Each stage incorporates normalization and nonlinear activation functions GELU, LeakyReLU, and SiLU along with dropout rates of 0.5, 0.4, and 0.3, respectively, to enhance robustness and generalization. The final classification layer maps the 128-dimensional representation to two output units, corresponding to the normal and abnormal classes. Training was conducted using the Weighted Categorical Cross-Entropy loss function, implemented via PyTorch's `nn.CrossEntropyLoss` with class weights to account for dataset imbalance. The model was optimized using the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.01. During inference, class predictions were obtained using `argmax(1)`, while class probabilities for ROC analysis and other evaluation metrics were computed using `torch.softmax(outputs, 1)`.

This experimental design isolates the effect of the attention mechanism, allowing direct attribution of any performance differences to the fusion strategy rather than extraneous factors.

V. Results and Discussion

A. Quantitative results

As shown in Table III, the proposed model demonstrates consistent and substantial improvements over baseline model across all evaluation metrics. Classification accuracy increases by 10.77 percentage points, indicating a marked enhancement in overall predictive performance. Similarly, the F1-score improves by 10.76 percentage points, reflecting more balanced classification across classes. Notably, Cohen's Kappa exhibits a substantial increase of 21.45 percentage points, highlighting a significantly stronger agreement beyond chance and confirming the robustness of the proposed attention-guided fusion strategy.

B. Qualitative Results

The confusion matrices presented in Fig. 5 and Fig. 6 provide a detailed class-wise evaluation of the models on the test dataset.The results demonstrate a clear performance improvement when using the attention-guided fusion model compared to the simple concatenation model. In the simple

TABLE III. Performance Comparison on the Test Dataset

Metric	Baseline Model (%)	Proposed Model (%)	Absolute Improvement(%)
Accuracy	83.99	94.76	+10.77
Precision	84.09	94.68	+10.59
Recall	84.28	94.82	+10.54
F1-score	83.98	94.74	+10.76
Cohen's Kappa	68.03	89.48	+21.45

concatenation approach, although the model achieved a reasonable number of correct classifications (298 abnormal and 279 normal), it produced a relatively high number of false negatives (72 abnormal cases misclassified as normal), which is particularly concerning in a medical diagnostic context where missed abnormal cases can delay treatment. In contrast, the attention-guided model substantially reduced false negatives from 72 to 22 and false positives from 38 to 14, while simultaneously increasing correct abnormal predictions (from 298 to 348) and correct normal predictions (from 279 to 303).These improvements indicate that the attention mechanism enhances feature discrimination by prioritizing more informative representations within the hybrid architecture, thereby improving overall classification reliability.

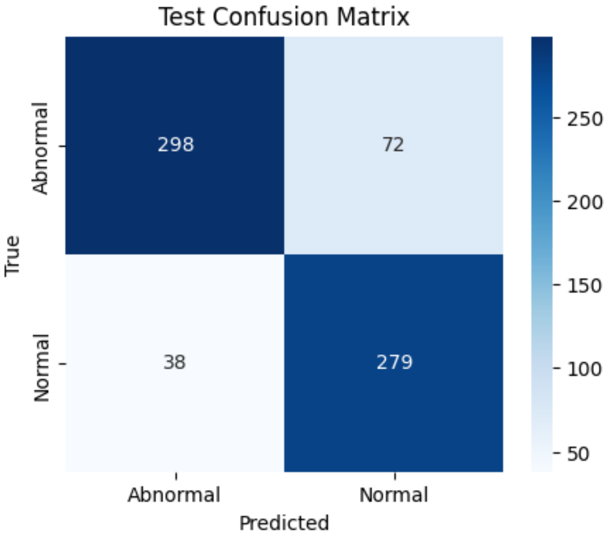


Fig. 5. Confusion matrix for the concatenation-based fusion model on the test dataset.

This finding is strongly corroborated by the Receiver Operating Characteristic (ROC) curves in Fig. 7 and Fig. 8. While the baseline model achieved a respectable Area Under the Curve (AUC) of 0.90 for both classes, the proposed model attained a near-perfect AUC of 0.99. This significant increase in AUC, consistent with the sharp reduction in confusion matrix errors, underscores that the attention mechanism not only boosts overall accuracy but also substantially improves the model's confidence and precision in separating the two classes.

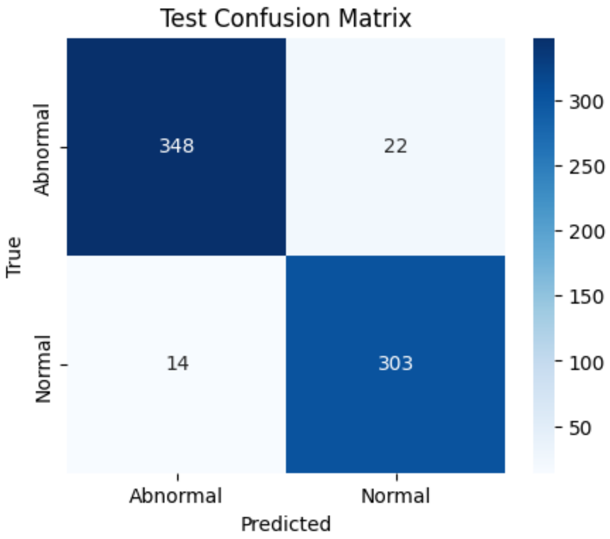


Fig. 6. Confusion matrix for the proposed attention-based fusion model on the test dataset.

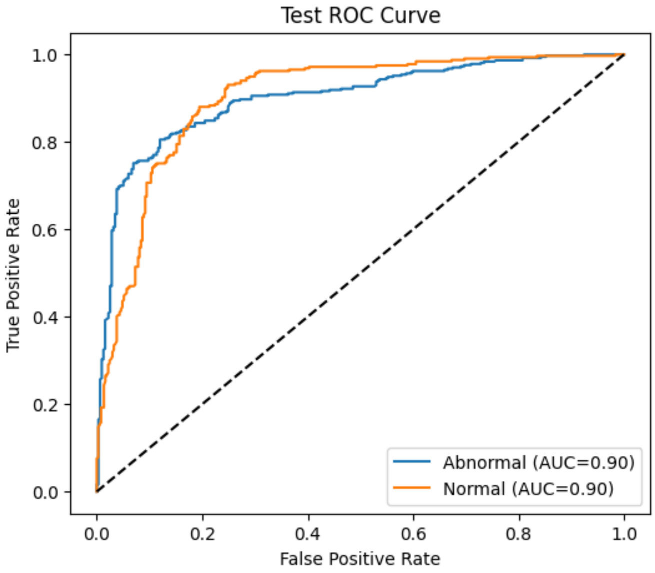


Fig. 7. ROC curve for the concatenation-based fusion model on the test dataset.

C. Learning Curve Analysis

The learning dynamics of the two approaches, illustrated in Fig. 9 and Fig. 10, provide additional insight into their behavior during training. The concatenation-based fusion model exhibits slower convergence, plateauing around epoch 30, and shows variability in generalization during intermediate epochs, which may indicate that the model struggles to locate the most critical regions in the images. By contrast, the proposed model stabilizes much earlier, between epochs 10 and 15, with closely aligned training and validation accuracies, reflecting faster and more consistent learning. This efficiency in convergence, combined with the improved predictive performance, underscores the effectiveness of the attention-guided fusion strategy in

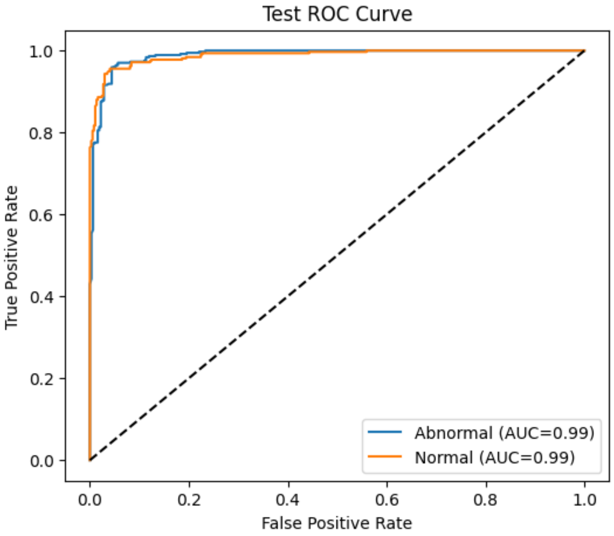


Fig. 8. ROC curve for the proposed attention-based fusion model on the test dataset.

prioritizing clinically relevant features while reducing training time. Collectively, these results highlight the potential of the proposed model as a robust and reliable tool for automated colposcopy image analysis. By improving classification accuracy, reducing misclassifications, and demonstrating stable learning, the model supports safer and more efficient cervical cancer screening workflows, complementing existing diagnostic procedures while minimizing the risk of missed diagnoses.

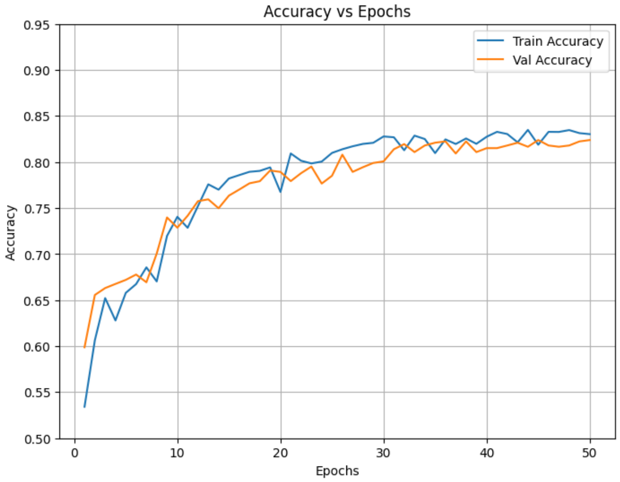


Fig. 9. Training and validation accuracy curve for the baseline model.

D. Model Classification Decisions

The Grad-CAM [38] visualizations in Fig. 11 illustrate that the model effectively highlights the image regions most influential to its classification decisions. The figure presents four representative samples (two normal and two abnormal), with the top row showing the original images and predicted confidence scores, and the bottom row displaying the corresponding Grad-CAM heatmaps. The model achieves confidence scores of 99.1% and 99.6% for normal samples and 88.5% and 99.1% for abnormal samples, indicating strong certainty in distinguishing healthy from diseased tissue.

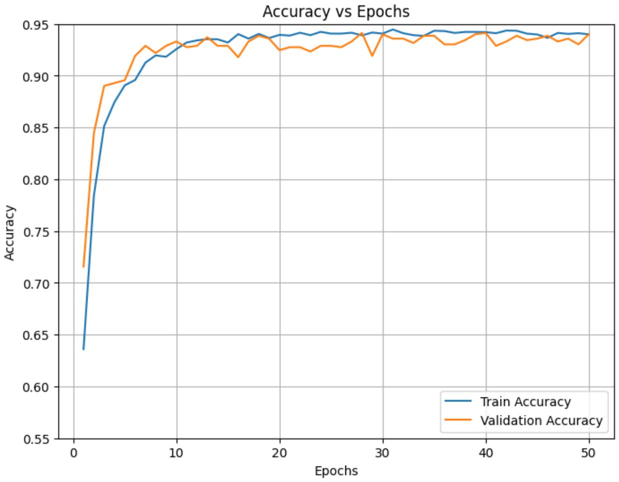


Fig. 10. Training and validation accuracy curve for the proposed attention-based fusion model.

In the heatmaps, red regions denote areas of highest importance, yellow indicate moderate importance, and blue correspond to regions with minimal influence. The strongest activations are consistently concentrated on irregular tissue patterns, including lesion margins, atypical epithelial structures, and abnormal coloration. This focused localization suggests that the model bases its predictions on clinically relevant pathological features rather than background artifacts.

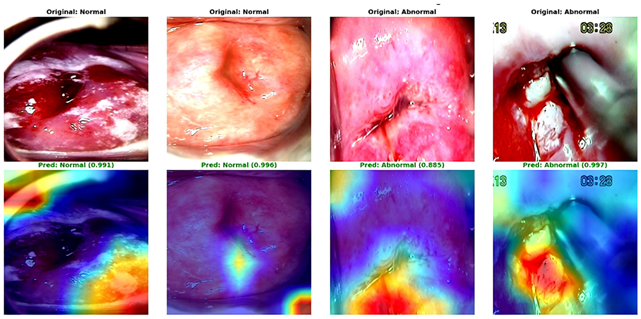


Fig. 11. Grad-CAM visualizations highlighting regions of high and low importance in colposcopy images.

E. Discussion

The attention mechanism fundamentally changes how the hybrid model integrates information from its two component architectures. Rather than simply combining features, it learns to prioritize different types of visual information depending on the input. This selective weighting explains the observed error reduction: false negatives dropped from 72 to 22, and false positives fell from 38 to 14. This pattern is particularly significant in screening contexts, where missed abnormalities carry greater clinical risk than over-calling normal tissue.

The training dynamics further demonstrate the value of attention. The baseline model required more epochs to plateau, converging around epoch 30, whereas the proposed model stabilized near epoch 10 with closely aligned training and validation curves. This efficiency suggests that attention focuses learning on informative regions from the outset.

The near-perfect AUC of 0.99 reflects how cleanly the model separates classes in feature space. This indicates that the internal representations of normal and abnormal cases form two distinct clusters with minimal overlap, enabling confident predictions across any decision threshold. The attention mechanism achieves this separation by ensuring that only relevant features such as lesion margins, epithelial changes, and vascular patterns contribute to the final representation, while effectively suppressing background noise.

Despite the promising performance, several limitations should be acknowledged. First, the model was trained and evaluated on a relatively small dataset, which may limit its generalizability to larger or more diverse populations. Second, variations in imaging devices, staining protocols, or patient demographics could affect performance when applied to external cohorts. Third, the proposed hybrid model is computationally demanding due to the large size of the EfficientNet and Swin Transformer branches, which constrained the training batch size to 8. Future work should focus on validating the model on larger, more diverse datasets, optimizing computational resource usage, and integrating complementary modalities, such as histopathology images and clinical metadata, to further enhance robustness and predictive performance. These findings demonstrate that the proposed attention-based fusion model outperforms the baseline in accuracy and reliability, providing a clinically meaningful approach for precise and efficient cervical cancer screening.

F. Comparative Analysis with Prior Studies

The proposed model was comparatively evaluated against several representative CNN–Transformer–based cervical image analysis frameworks reported in the literature. The CVM-Cervix model [25] demonstrated the benefit of integrating CNN-extracted cellular features with Vision Transformer–captured global contextual representations; Feature fusion was performed using a multilayer perceptron (MLP), which enables joint feature learning but lacks explicit attention mechanisms to selectively emphasize diagnostically salient regions. Similarly, the hybrid EfficientNet-B0–Swin-Tiny framework [22] highlighted the potential of CNN–Transformer integration but employed a simple concatenation-based fusion mechanism, achieving an accuracy of 87.5% after 30 training epochs. The MFEM-CIN framework [24] further contributed toward lightweight, multi-scale colposcopic image analysis; nevertheless, its multi-scale feature fusion (MSFF) strategy was primarily based on concatenation and linear projection, which may constrain the modeling of complex inter-feature dependencies. Beyond cervical imaging, a related hybrid approach combining a ResNet module with a transformer was proposed in [23] for non-small cell lung cancer (NSCLC) analysis. While the ResNet branch captured local volumetric features and the transformer modeled long-range contextual information, the use of simple feature concatenation resulted in a maximum testing accuracy of 81.9%, underscoring the limitations of uniform feature aggregation. In contrast, the baseline model implemented in this study, which also employed concatenation-based fusion, achieved an accuracy of 83.99%, aligning with the performance trends observed in prior concatenation-driven hybrid architectures. Notably, the proposed attention-guided fusion model substantially outperformed both the baseline and existing methods, achieving an accuracy of 94.76%. This improvement highlights the effectiveness of attention mechanisms in selectively emphasizing diagnostically relevant regions and suppressing redundant features during fusion (see Table IV).

TABLE IV. Comparison between the Proposed Model and Related Studies

Study	Model	Image type	Fusion Mechanism	Accuracy (%)
[25]	CNN + Transformer	Pap smear images	Multilayer Perceptron (MLP)	91.72
[22]	EfficientNet-B0 + Swin-Tiny	Colposcopy images	Concatenation	87.50
[24]	Multi-scale CNN + Transformer	Colposcopy images	Concatenation	89.20
[23]	ResNet + Transformer	NSCLC CT images	Concatenation	81.90
Our baseline	EfficientNet-B0 + Swin-T	Colposcopy images	Concatenation	83.99
Our proposed	**EfficientNet-B4 + Swin-T**	**Colposcopy images**	**Attention-guided fusion (Cross attention)**	**94.76**

G. Clinical Integration and Practical Implications

The proposed hybrid EfficientNet-B0 and Swin Transformer model with cross-attention fusion offers meaningful potential to support cervical cancer screening and diagnostic workflows. By processing colposcopy images in real-time, the model can assist clinicians during routine gynecological examinations, delivering rapid and data-informed interpretations. This capability helps mitigate the subjectivity and interobserver variability common in visual assessment, contributing to more standardized and reproducible evaluations across practitioners.

With its high predictive accuracy and robust generalization, the model serves as a reliable decision-support tool that complements established diagnostic techniques such as Pap smears, HPV testing, and biopsy. Although the current implementation is best suited for systems with GPU support, future optimizations may enable broader adoption even in resource-constrained clinical settings.

Importantly, the model produces interpretable attention maps that highlight key regions in the images, including lesion margins, abnormal epithelial pattern, and areas of unusual coloration. These visual explanations assist pathologists and radiologists in focusing on clinically relevant areas, prioritizing critical casesand making more confident, informed diagnostic decisions. Overall, the system is intended to support rather than replace expert judgment, improving both the speed and reliability of cervical cancer diagnosis in clinical practice.

VI. Conclusion

This study proposed a novel attention-guided hybrid framework for cervical cancer classification, integrating features from EfficientNet-B0 and Swin Transformer. By applying channel-wise and spatial attention prior to fusion, the model effectively captured both global contextual information and localized fine-grained details, addressing the limitations of simple concatenation. The framework achieved strong performance, with a test accuracy of 94.76%, precision of 94.68%, recall of 94.82%, F1-score of 94.74%, and a Cohen's Kappa of 89.48%. Attention mechanisms were crucial in prioritizing salient features, enhancing discriminative power, and reducing misclassifications, while the accelerated and stable convergence during training underscores the model's efficiency.

Although these results are promising, future work should focus on validating the framework on large-scale datasets from diverse clinical institutions to ensure robust generalizability. Incorporating complementary patient information, such as clinical history and histopathology reports, could further improve diagnostic accuracy and utility. Additionally, the flexible design of the framework presents an opportunity to extend its application to other gynecological cancers and imaging modalities, broadening its potential clinical impact.

Acknowledgment

We express our sincere gratitude to the Pan African University, Institute for Basic Sciences, Technology, and Innovation (PAUSTI), Nairobi, Kenya, and the African Union for providing the essential resources that made this research possible. Their support and encouragement were instrumental in enabling the successful completion of this work.

References

[1] S. K. Singh, R. Singh, and A. Goyal, "Semi-automatic segmentation of overlapping cells in pap smear image," in *Proceedings of the 4th International Conference on Computer Science (ICCS)*, Aug. 2018, pp. 161–165. https://doi.org/10.1109/ICCS.2018.00034

[2] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Applied Sciences*, vol. 1, no. 6, p. 641, 2019. doi: 10.1007/s42452-019-0645-7. [Online]. Available: https://doi.org/10.1007/s42452-019-0645-7

[3] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, and F. Bray, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," *The Lancet Global Health*, vol. 8, no. 2, pp. e191–e203, 2020. doi: 10.1016/S2214-109X(19)30482-6. [Online]. Available: https://doi.org/10.1016/S2214-109X(19)30482-6

[4] J. Jin, "HPV Infection and Cancer," *JAMA*, vol. 319, no. 10, p. 1058, 2018. doi: 10.1001/jama.2018.0687. [Online]. Available: https://doi.org/10.1001/jama.2018.0687

[5] M. Das, "WHO launches strategy to accelerate elimination of cervical cancer," *The Lancet Oncology*, vol. 22, no. 1, pp. 20–21, 2021. doi: 10.1016/S1470-2045(20)30729-4. [Online]. Available: https://doi.org/10.1016/S1470-2045(20)30729-4

[6] N. Youneszade, M. Marjani, and C. P. Pei, "Deep Learning in Cervical Cancer Diagnosis: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 11, pp. 6133–6149, 2023. https://doi.org/10.1109/ACCESS.2023.3235833

[7] J. Wu, Q. Jin, Y. Zhang, Y. Ji, J. Li, X. Liu, and Y. Huang, "Global Burden of Cervical Cancer: Current Estimates, Temporal Trend and Future Projections Based on the GLOBOCAN 2022," *Journal of the National Cancer Center*, 2025. doi: 10.1016/j.jncc.2024.11.006. [Online]. Available: https://doi.org/10.1016/j.jncc.2024.11.006

[8] P. N. Simelela, "WHO Global Strategy to Eliminate Cervical Cancer as a Public Health Problem: An Opportunity to Make It a Disease of the Past," *International Journal of Gynecology & Obstetrics*, vol. 152, pp. 1–3, 2021. doi: 10.1002/ijgo.13484. [Online]. Available: https://doi.org/10.1002/ijgo.13484

[9] K. T. Simms, J. Steinberg, M. Caruana, M. A. Smith, J. B. Lew, I. Soerjomataram, *et al.*, "Impact of scaled up human papillomavirus vaccination and cervical screening and the potential for global elimination of cervical cancer in 181 countries, 2020–99: a modelling study," *The Lancet Oncology*, vol. 20, no. 3, pp. 394–407, 2019. https://doi.org/10.1016/S1470-2045(18)30836-2

[10] X. Hou, G. Shen, L. Zhou, Y. Li, T. Wang, and X. Ma, "Artificial intelligence in cervical cancer screening and diagnosis," *Frontiers in Oncology*, vol. 12, p. 851367, 2022. https://doi.org/10.3389/fonc.2022.851367

[11] V. Chandran, M. G. Sumithra, A. Karthick, T. George, M. Deivakani, B. Elakkiya, *et al.*, "Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images," *BioMed Research International*, vol. 2021, no. 1, p. 5584004, 2021. doi: 10.1155/2021/5584004. [Online]. Available: https://doi.org/10.1155/2021/5584004

[12] G. Purwoto, H. D. Dianika, A. Putra, S. Purbadi, and L. Nuranna, "Modified cervicography and visual inspection with acetic acid as an alternative screening method for cervical precancerous lesions," *Journal of Cancer Prevention*, vol. 22, no. 4, p. 254, 2017. https://doi.org/10.15430/JCP.2017.22.4.254

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539

[14] P. Sahoo, S. Saha, S. Mondal, M. Seera, S. K. Sharma, and M. Kumar, "Enhancing Computer-Aided Cervical Cancer Detection Using a Novel Fuzzy Rank-Based Fusion," *IEEE Access*, vol. 11, pp. 145281–145294, 2023. doi: 10.1109/ACCESS.2023.3346764. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3346764

[15] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A Survey of the Vision Transformers and their CNN-Transformer Based Variants," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 2917–2970, 2023. doi: 10.1007/s10462-023-10595-0. [Online]. Available: https://doi.org/10.1007/s10462-023-10595-0

[16] M. Kalbhor, S. Shinde, H. Joshi, and P. Wajire, "Pap smear-based cervical cancer detection using hybrid deep learning and performance evaluation," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 11, no. 5, pp. 1615–1624, 2023. https://doi.org/10.1080/21681163.2023.2219469

[17] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2021. doi: 10.1109/TNNLS.2021.3084827. [Online]. Available: https://doi.org/10.1109/TNNLS.2021.3084827

[18] I. Zimerman and L. Wolf, "On the Long Range Abilities of Transformers," *arXiv preprint arXiv:2311.16620*, 2023. doi: 10.48550/arXiv.2311.16620. [Online]. Available: https://doi.org/10.48550/arXiv.2311.16620

[19] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing Properties of Vision Transformers," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 23296–23308. doi: 10.48550/arXiv.2105.10497. [Online]. Available: https://doi.org/10.48550/arXiv.2105.10497

[20] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional Feature Fusion," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2021, pp. 3560–3569. doi: 10.1109/WACV48630.2021.00360. [Online]. Available: https://doi.org/10.1109/WACV48630.2021.00360

[21] J. Wei, S. Wang, and Q. Huang, "F³Net: Fusion, Feedback and Focus for Salient Object Detection," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12321–12328. doi: 10.1609/aaai.v34i07.6916. [Online]. Available: https://doi.org/10.1609/aaai.v34i07.6916

[22] F. A. Mohammed, K. K. Tune, J. A. Mohammed, T. A. Wassu, and S. Muhie, "Early Cervical Cancer Diagnosis with SWIN-Transformer and Convolutional Neural Networks," *Diagnostics*, vol. 14, no. 20, p. 2286, 2024. doi: 10.3390/diagnostics14202286. [Online]. Available: https://doi.org/10.3390/diagnostics14202286

[23] L. Wang, C. Zhang, and J. Li, "A Hybrid CNN-Transformer Model for Predicting N Staging and Survival in Non-Small Cell Lung Cancer Patients Based on CT-Scan," *Tomography*, vol. 10, no. 10, pp. 1676–1693, Oct. 2024. doi: 10.3390/tomography10100123. [Online]. Available: https://doi.org/10.3390/tomography10100123

[24] P. Chen, F. Liu, J. Zhang, and B. Wang, "MFEM-CIN: a lightweight architecture combining CNN and transformer for the classification of Pre-cancerous lesions of the cervix," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 216–225, 2024. doi: 10.1109/OJEMB.2024.3367243. [Online]. Available: https://doi.org/10.1109/OJEMB.2024.3367243

[25] W. Liu, C. Li, N. Xu, T. Jiang, M. M. Rahaman, H. Sun, and M. Grzegorzek, "CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multi-layer perceptron," *Pattern Recognition*, vol. 130, p. 108829, 2022. https://doi.org/10.1016/j.patcog.2022.108829.

[26] R. Dulal, L. Zheng, and M. A. Kabir, "MHAFF: Multihead Attention Feature Fusion of CNN and Transformer for Cattle Identification," *IEEE Transactions on AgriFood Electronics*, 2025. doi: 10.1109/TAFE.2025.3574708. [Online]. Available: https://doi.org/10.1109/TAFE.2025.3574708

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141. https://doi.org/10.48550/arXiv.1709.01507.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19. https://doi.org/10.48550/arXiv.1807.06521.

[29] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114. https://doi.org/10.48550/arXiv.1905.11946.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.

[31] International Agency for Research on Cancer, "Colposcopy Image Bank," 2024. [Online]. Available: https://screening.iarc.fr/cervicalimagebank.php. Accessed: May 17, 2025.

[32] S. Nursyafiqah, "Dataset: Colposcopy Images," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/sitinursyafiqahba/dataset-colposcopy-images.

[33] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019. https://doi.org/10.1186/s40537-019-0197-0.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017. https://doi.org/10.1145/3065386.

[35] A. Vaswani *et al.*, "Attention Is All You Need," arXiv preprint arXiv:1706.03762, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[36] G. Kumar and P. K. Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems," in *2014 Fourth International Conference on Advanced Computing & Communication Technologies (ACCT)*, February 2014, pp. 5–12, IEEE. doi: 10.1109/ACCT.2014.74.

[37] S. Khalid, T. Khalil, and S. Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning," in *2014 Science and Information Conference*, 2014, pp. 372–378, IEEE. https://doi.org/10.1109/SAI.2014.6918213. [Online]. Available:https://doi.org/10.1109/SAI.2014.6918213

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74

[39] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 6–9, 2019. [Online]. Available: https://dblp.org/rec/conf/iclr/LoshchilovH19.html

[40] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.48550/arXiv.1512.03385. [Online]. Available: https://doi.org/10.48550/arXiv.1512.03385

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017. doi: 10.48550/arXiv.1608.06993. [Online]. Available: https://doi.org/10.48550/arXiv.1608.06993

[43] O. S. Oyebanji, A. R. Apampa, P. I. Idoko, A. Babalola, O. M. Ijiga, O. Afolabi, and C. I. Michael, "Enhancing Breast Cancer Detection Accuracy through Transfer Learning: A Case Study Using EfficientNet," *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 01, pp. 285–318, 2024. doi: 10.30574/wjaets.2024.13.1.0415. [Online]. Available: https://doi.org/10.30574/wjaets.2024.13.1.0415

[44] M. Behzadpour, B. L. Ortiz, E. Azizi, and K. Wu, "Breast Tumor Classification Using EfficientNet Deep Learning Model," *arXiv preprint arXiv:2411.17870*, 2024. doi: 10.48550/arXiv.2411.17870. [Online]. Available: https://doi.org/10.48550/arXiv.2411.17870

[45] A. A. Nafea, M. S. Ibrahim, M. M. Shwaysh, K. Abdul-Kadhim, H. R. Almamoori, and M. M. Al-Ani, "A Deep Learning Algorithm for Lung Cancer Detection Using EfficientNet-B3," *Wasit Journal of Computer and Mathematics Science*, vol. 2, no. 4, pp. 68–76, 2023. doi: 10.31185/wjcms.209. [Online]. Available: https://doi.org/10.31185/wjcms.209

[46] S. Aouadi, T. Torfeh, O. Bouhali, S. A. Yoganathan, S. Paloor, S. Chandramouli, and N. Al-Hammadi, "Prediction of Cervix Cancer Stage and Grade from Diffusion Weighted Imaging Using EfficientNet," *Biomedical Physics & Engineering Express*, vol. 10, no. 4, p. 045042, 2024. doi: 10.1088/2057-1976/ad5207. [Online]. Available: https://doi.org/10.1088/2057-1976/ad5207

[47] A. Suphalakshmi, A. Ahilan, A. Jeyam, and M. Subramanian, "Cervical Cancer Classification Using EfficientNet and Fuzzy Extreme Learning Machine," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 5, pp. 6333–6342, 2022. doi: 10.3233/JIFS-220296. [Online]. Available: https://doi.org/10.3233/JIFS-220296

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. doi: 10.48550/arXiv.2010.11929. [Online]. Available: https://doi.org/10.48550/arXiv.2010.11929

[49] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022. doi: 10.1109/TPAMI.2022.3152247.

[50] J. Zhang, Z. Zhang, H. Liu, and S. Xu, "SaTransformer: Semantic-aware transformer for breast cancer classification and segmentation," *IET Image Processing*, vol. 17, no. 13, pp. 3789–3800, 2023. doi: 10.1049/ipr2.12897. [Online]. Available: https://doi.org/10.1049/ipr2.12897

[51] D. R. Nayak, "RDTNet: A residual deformable attention based transformer network for breast cancer classification," *Expert Systems with Applications*, vol. 249, p. 123569, 2024. doi: 10.1016/j.eswa.2024.123569. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.123569

[52] D. P. Yadav, B. Sharma, J. L. Webber, A. Mehbodniya, and S. Chauhan, "EDTNet: A spatial aware attention-based transformer for the pulmonary nodule segmentation," *PLOS ONE*, vol. 19, no. 11, p. e0311080, 2024. doi: 10.1371/journal.pone.0311080. [Online]. Available: https://doi.org/10.1371/journal.pone.0311080

[53] M. Darwish, M. Z. Altabel, and R. H. Abiyev, "Enhancing cervical precancerous classification using advanced vision transformer," *Diagnostics*, vol. 13, no. 18, p. 2884, 2023. doi: 10.3390/diagnostics13182884. [Online]. Available: https://doi.org/10.3390/diagnostics13182884

[54] N. Sharma, K. Gaurav, and T. K. R. Bollu, "CerviTransX: Explainable Transformer-Based Cervical Cancer Classification," in *Proceedings of the 2025 National Conference on Communications (NCC)*, pp. 1–6, Mar. 2025. doi: 10.1109/NCC63735.2025.10983448. [Online]. Available: https://doi.org/10.1109/NCC63735.2025.10983448

[55] K. Abinaya and B. Sivakumar, “A Deep Learning-Based Approach for Cervical Cancer Classification Using 3D CNN and Vision Transformer,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, p. 280, 2024. doi: 10.1007/s10278-023-00911-z. [Online]. Available: https://doi.org/10.1007/s10278-023-00911-z