Multiscale Optical Imaging Fusion for Cervical Precancer Diagnosis: Integrating Widefield Colposcopy and High-Resolution Endomicroscopy

David Brenes, Mila P. Salcedo, Jackson B. Coole, Yajur Maker, Alex Kortum, Richard A. Schwarz, Jennifer Carns, Imran S. Vohra, Júlio C. Possati-Resende, Márcio Antoniazzi, Bruno de Oliveira Fonseca, Karen C. Borba Souza, Iara V. Vidigal Santana, Flávia Fazzio Barbin, Regis Kreitchmann, Nirmala Ramanujam, Kathleen M. Schmeler, and Rebecca Richards-Kortum

Abstract*—*Objective: **Early detection and treatment of cervical precancers can prevent disease progression. However, in low-resource communities with a high incidence of cervical cancer, high equipment costs and a shortage of specialists hinder preventative strategies. This manuscript presents a low-cost multiscale in vivo optical imaging system coupled with a computer-aided diagnostic system that could enable accurate, real-time diagnosis of high-grade cervical precancers.** ***Methods:*** **The system combines portable colposcopy and high-resolution endomicroscopy (HRME) to acquire spatially registered widefield and microscopy videos. A multiscale imaging fusion network (MSFN) was developed to identify cervical intraepithelial neoplasia grade 2 or more severe (CIN 2+). The MSFN automatically identifies and segments the ectocervix and lesions from colposcopy images, extracts nuclear morphology features from HRME videos, and integrates the colposcopy and HRME information.** ***Results:*** **With a threshold value set to achieve sensitivity equal to clinical impression (0.98 [p = 1.0]), the MSFN achieved a significantly higher specificity than clinical impression (0.75 vs. 0.43, p = 0.000006).** ***Conclusion:*** **Our findings show that multiscale optical imaging of the cervix allows the highly sensitive and specific detection of high-grade precancers.** ***Significance:*** **The multiscale imaging system and MSFN could facilitate the accurate, real-time diagnosis of cervical precancers in low-resource settings.**

***Index Terms*—Optical imaging, cervical cancer, multimodality fusion, machine learning, deep learning.**

I. Introduction

CERVICAL cancer is the fourth leading cause of cancer death in women worldwide and is estimated to be responsible for over 340000 deaths annually [1]. Although cervical cancer is preventable through the implementation of known strategies for prevention, screening, and early intervention, the resources and well-trained personnel to implement those strategies are not available in many low- and middle-income countries (LMICs) [2]. As a result, cervical cancer remains the leading cause of cancer death in many countries in sub-Saharan Africa, Melanesia, South America, and Southeast Asia [1].

Cervical cancer is a human papillomavirus (HPV) associated cancer, making HPV vaccination the primary prevention strategy [3]. However, HPV vaccination is not widely accessible around the world and must be administered at a young age (9 to 13 years of age) to be the most effective [4], [5]. For the adult population at risk, the World Health Organization (WHO) recommends secondary prevention strategies [1]. The WHO recommends two strategies: screen-and-treat or screen-triage-treat. In both scenarios, HPV DNA detection is the preferred screening test. Recommended triage tests include partial HPV genotyping, colposcopy with or without biopsy, visual inspection with acetic acid (VIA), or cytology. Among triage tests, colposcopy with biopsies taken from abnormal areas is preferred since a biopsy allows for a confirmatory histopathologic diagnosis [1]. The WHO recommends the treatment of patients with histologically confirmed high-grade precancers (cervical intraepithelial neoplasia grade 2 or more severe [CIN 2+]). The decision to triage and the selection of a triage test are contingent on the availability of equipment and trained physicians. In low-resource settings,

Manuscript received 31 October 2023; revised 3 February 2024; accepted 10 March 2024. Date of publication 20 March 2024; date of current version 22 August 2024. This work was supported by the National Cancer Institute of the National Institutes of Health under Grant R01CA251911 and Grant U01CA269192. *(Corresponding author: Rebecca Richards-Kortum.)*

David Brenes, Jackson B. Coole, Yajur Maker, Alex Kortum, Richard A. Schwarz, Jennifer Carns, and Imran S. Vohra are with the Rice University, USA.

Mila P. Salcedo and Kathleen M. Schmeler are with The University of Texas MD Anderson Cancer Center, USA.

Júlio C. Possati-Resende, Márcio Antoniazzi, Bruno de Oliveira Fonseca, Karen C. Borba Souza, Iara V. Vidigal Santana, and Flávia Fazzio Barbin are with the Barretos Cancer Hospital, São Paulo, Brazil.

Regis Kreitchmann is with the Federal University of Health Sciences of Porto Alegre, Porto Alegre, Brazil.

Nirmala Ramanujam is with the Duke University, USA, and also with the Calla Health Foundation, USA.

Rebecca Richards-Kortum is with Rice University, Houston, TX 77005 USA (e-mail: rkortum@rice.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2024.3379898, provided by the authors.

Digital Object Identifier 10.1109/TBME.2024.3379898

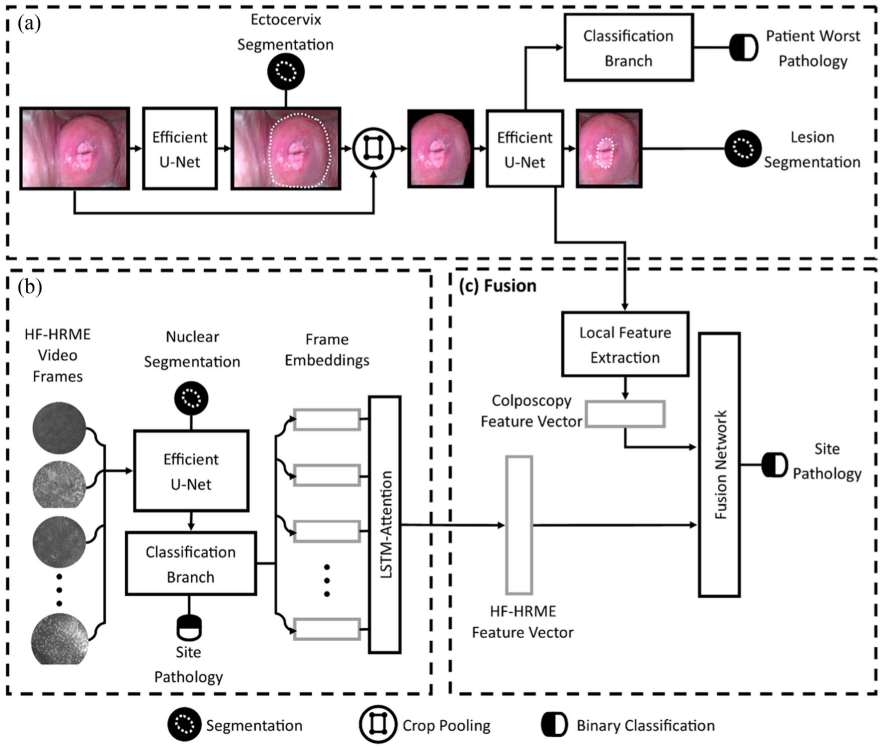

Fig. 1. Multiscale network is composed of three components: (a) colposcopy, (b) HRME, and (c) fusion modules. The colposcopy module processes the colposcopy image generating an ectocervix mask, a lesion mask, and a prediction of whether the patient has CIN 2+. The HRME module processes a sequence of HF-HRME frames using a nuclear segmentation and single image classifier to predict whether the site contains CIN 2+. The embeddings in the last fully connected layer of the single image classifier are passed into an LSTM network with global attention that aggregates the information in the sequence to perform video classification. The fusion module combines colposcopy and HRME features extracted from the two modules to predict whether a site contains CIN 2+. Boxes with a grey outline represent features, while boxes with a back outline represent modules. See Supplement A for more detail on the modules. HF-HRME, high frame rate fiber optic microendoscope; LSTM, long short-term memory network; CIN 2+, cervical intraepithelial grade 2 or more severe. HRME, high-resolution microendoscope; CIN 2+, cervical intraepithelial grade 2 or more severe; LSTM, long short-term memory network.

high equipment costs, lack of supplies, and lack of qualified specialists make it challenging to implement many of these triage tests. Furthermore, screen-triage-treat strategies that require multiple visits can result in the loss of patients to follow-up in settings where access to care is limited. Thus, low-resource settings often rely on screen-and-treat strategies, which may lead to high overtreatment rates.

Optical imaging systems combined with computer-aided diagnostic software could be used in low-resource settings to triage patients with an abnormal screening test. Previously proposed optical imaging systems for cervical precancer detection include widefield imaging systems that image the entire cervix at sub-mm spatial resolution or high-resolution imaging systems that provide sub-cellular resolution from sub-mm fields of view. Portable low-cost colposcopes, such as the EVA (MobileODT, Tel Aviv-Yafo, Israel) or the Pocket Colposcope (Calla Health Technologies, Durham, NC, USA), are examples of widefield imaging systems [6], [7]. These systems capture images of the entire surface of the cervix with lateral resolutions as low as 28 um. Several computer-aided diagnostic systems have been built on these platforms for lesion detection and pathology prediction [7], [8], [9]. High-resolution systems, such as the High-Resolution Microendoscope (HRME), capture subcellular tissue features and have a lateral resolution of $\sim$4 um [10]. Automated image analysis algorithms have been developed to analyze widefield and microscopy images to detect the presence of histologically confirmed CIN 2+. For example, a recent study showed that the HRME coupled with a deep learning model performs comparably to expert clinical impression with a sensitivity of 0.94 ($p = 0.3$) and specificity of 0.58 ($p = 1.0$) in CIN 2+ diagnosis [11]. Yet, the specificity of these systems remains low due to many false positive predictions arising from benign confounding morphologies in polyps, inflammation, or columnar tissue [12].

To date, most computer-aided diagnostic systems proposed for cervical precancer detection only consider a single optical imaging modality. However, combining the multiscale data obtained from widefield and high-resolution imaging could help improve the detection accuracy of CIN 2+. Here, we present the first implementation of a computer-aided diagnostic system that employs multiscale in vivo imaging data to detect cervical precancers. The optical imaging system, which combines portable colposcopy and high-resolution endomicroscopy, was used to acquire a large dataset of spatially registered widefield and microscopy videos from patients referred for colposcopy [13]. This study developed a multiscale fusion deep learning model to analyze the registered images and predict which areas of the cervix contained CIN 2+ lesions; results were compared

to the gold standard of histology. At a sensitivity equal to clinical impression (0.98 [p = 1.0]), the model achieved a significantly higher specificity than clinical impression (0.75 vs. 0.43, p = 0.000006).

II. Materials and Methods

A. Multiscale Imaging System

A multiscale optical system was used to acquire co-registered widefield and high-resolution images of cervical tissue simultaneously; the system has previously been described in detail [13]. Briefly, widefield images are acquired with a Pocket Colposcope, a low-cost, portable, low-magnification system with a 3.5-5.0 cm field of view, 3.5-4.5 cm working distance, and 20 um lateral resolution. High-resolution images are acquired with a high frame rate fiber optic microendoscope (HF-HRME) that captures images of proflavine-stained nuclei from epithelial tissue [14]. The HF-HRME has a 790 um field of view and a 4 um lateral resolution.

B. Study Participants

Patients were recruited from the cervical cancer prevention program at Barretos Cancer Hospital (BCH) in Barretos, Brazil, based on the following inclusion criteria: (a) were scheduled to undergo colposcopy due to a history of dysplasia or an abnormal cervical cancer screening result (abnormal cytology [atypical squamous cells of undetermined significance or more severe] or positive for high-risk HPV [cobas 4800 HPV]), (b) were older than 25 years of age, (c) had a negative pregnancy test, and (d) had the ability to provide written informed consent. Women were excluded from the study if they: (a) had undergone a hysterectomy with removal of the cervix, (b) had an allergy to proflavine or acriflavine, or (c) were breastfeeding at the time of enrollment.

The study was approved by the BCH Research Committee, the Brazilian National Ethics Research Commission/CONEP (CAAE: 38969820.9.1001.5437) and the Institutional Review Board of Rice University (ID#2020-342) and The University of Texas MD Anderson Cancer Center (ID#2021-0356). All participants provided written informed consent. The protocol was registered with ClinicalTrials.gov (NCT05078528). All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000.

C. Multiscale Imaging Session

Standard colposcopy was performed after applying 5% acetic acid to the cervix, and any clinical lesions were noted. Prior to multiscale imaging, 5% acetic acid solution was re-applied. The clinician inserted the Pocket Colposcope through the speculum and captured a representative widefield image of the cervix. Additional images were taken if the clinician deemed the representative image to be poor quality. Next, a 0.01% w/v proflavine solution was applied to the cervix, followed by the application of Lugol's iodine. Additional representative widefield images were collected. The proflavine solution was then reapplied in preparation for the HF-HRME imaging. The HF-HRME probe was gently placed in contact with the cervix and translated across any lesions identified during the initial colposcopic examination. The multiscale imaging system records simultaneous video from the Pocket Colposcope and the HF-HRME, allowing precise co-registration. Biopsies were acquired from lesion areas identified via standard-of-care colposcopy. Multiple biopsies were acquired if the colposcopist identified multiple lesions, one biopsy per lesion. If no lesions were identified, the probe was translated across a normal area of the cervix just outside the squamocolumnar junction, and a single biopsy was acquired from an area imaged by the HF-HRME. The biopsy samples were submitted for routine histologic analysis using standard diagnostic criteria; the final diagnosis was reported as benign, cervicitis, cervical intraepithelial neoplasia (CIN) 1, CIN 2, CIN 3, or cancer. The final diagnosis was reached by consensus of two expert pathologists; disagreements were resolved in a consensus session where the pathologists met to discuss and reevaluate the case. For analysis purposes, samples were grouped into two diagnostic categories: less than CIN 2 (<CIN 2) and CIN 2 or more severe (CIN 2+).

D. Annotation and Curation of Multiscale Imaging Data

After multiscale imaging, the clinician annotated the representative widefield image of the cervix to outline the ectocervix, the squamocolumnar junction, the os, any lesions, and the location of any biopsies. The clinical impression for each lesion was recorded as a low-grade intraepithelial lesion (LSIL), high-grade intraepithelial lesion (HSIL), or cancer. Video captured with the Pocket Colposcope was used to manually track the location of the HF-HRME probe on the cervix and determine the corresponding HF-HRME frames for each biopsy site.

All HF-HRME video segments obtained from regions of the tissue that were biopsied were reviewed for quality control. Manual quality control was first performed by three experts; video frames were judged to be of high quality if they met two criteria: 1) the frame was free from motion blur, and 2) more than 50% of the field of view was in focus. Sites for which experts agreed that no high-quality HF-HRME frames were captured were eliminated from further analysis. All frames were subject to an automated quality control algorithm that performed a rapid segmentation to identify the number of potential nuclei in the field of view. Frames containing less than a predetermined number of potential nuclei in the field of view were eliminated from further analysis.

E. Multiscale Imaging Fusion Network

A multiscale imaging fusion network (MSFN) was developed to analyze co-registered image data (Fig. 1); the MSFN is composed of separate colposcopy, HF-HRME, and fusion modules described below.

1) Colposcopy Module: The colposcopy module is designed to segment the ectocervix and any lesions as well as to predict the presence or absence of CIN 2+. The module utilizes two Efficient U-Net blocks that sequentially process the

colposcopy image. The Efficient U-Net has an encoder-decoder structure with concatenating skip connections and can perform semantic segmentation of color images, and is based on the segmentation portion of the Y-Net architecture proposed by Mehta et al. (2018) [15]. To minimize computations and reduce the model's memory footprint, the Efficient U-Net employs efficient spatial pyramid and pyramid spatial pooling blocks [15], [16].

The colposcopy module processes images sequentially in two steps. First, an Efficient U-Net segments the ectocervix. The resulting mask is used to isolate the cervix from the background through crop pooling and masking. Next, the cropped and masked image of the cervix is passed to a second Efficient U-Net, which segments lesions. Two possible lesion segmentation strategies were explored: segmentation of all colposcopically low-grade or more severe lesions (LSIL+) and segmentation of histologically confirmed CIN 2+ lesions. A classification branch (see Supplement A) was connected to the lowest embedding of the second Efficient U-Net and was trained to predict whether or not the cervix contained CIN 2+ using the highest-grade pathology result for the patient as ground truth.

2) HF-HRME Module: The HF-HRME module classifies whether both individual image frames and video segments were obtained from regions containing CIN 2+; this classification is aided by nuclear segmentation. The single image classifier is composed of one Efficient U-Net that performs nuclear segmentation and a classification branch that performs image classification (see Supplement A). The single image classifier was trained for CIN 2+ classification as described in [11] and was used without adjustment.

Video classification is enabled by a long short-term memory network (LSTM) with global attention (LSTM-Attention) [17]. Each HF-HRME frame was divided into quadrants that ran independently through the single image classifier. Quadrant embeddings in the last fully connected layer of the HF-HRME single image classification branch are combined using average pooling to form an HF-HRME frame embedding. These frame embeddings are passed into a unilateral LSTM with two layers that output a hidden state for each input. An attention component based on the general global attention mechanism proposed by Luong et al. (2015) is used (see Supplement A) [17]. The general global attention method generates a context vector, which is computed as the weighted sum of all previously generated hidden states modulated by attention weights. To compute the attention weights, previous hidden states are passed through a fully connected layer, and the output is multiplied by the current hidden state and passed through a softmax layer. To create the output of the LSTM-Attention, the context vector and current hidden state are first concatenated and passed through a fully connected layer. The resulting vector is then tanh-normalized. The output is further processed using a series of additional fully connected layers and a softmax layer, which predicts the likelihood that the video was obtained from a lesion containing CIN 2+.

The attention weights predicted by the LSTM-Attention module can be visualized with a one-dimensional heatmap and used to infer the contributions of individual frames to the classification of the HF-HRME sequence. These attention weights can also help illustrate how the model reacts to changes in image characteristics, such as blur and changes in the nuclear morphological structure. Several representative attention maps were selected and presented alongside the key HF-HRME frames from the sequence.

3) Fusion Module: The fusion model combines features from the colposcopy and HF-HRME modules to predict whether each biopsied site contains CIN 2+ (see Supplement A). HF-HRME features are extracted from the output of the LSTM-Attention network. Local colposcopy features are extracted from feature maps generated by the classifier of the lesion segmentation Efficient U-Net (see Supplement A). The feature maps are up-sampled to match the spatial resolution of the input colposcopy image and crop pooling is used to extract the features from a predefined area of interest surrounding the biopsy site. The feature maps are reduced to a vector using average pooling and passed through a fully connected layer that reduces the length of the colposcopy feature vector to equal that of the HRME feature vector. The colposcopy and HRME feature vectors are concatenated and passed through two additional fully connected layers, and a final softmax layer predicts the probability of CIN 2+ for the site.

4) Training and Validation: To train the MSFN, the multiscale imaging data collected in the study were partitioned temporally into training and validation sets; patients were ordered by enrollment date, with the first ∼50% of patients being used for training and the remaining patients for validation. The MSFN was trained end-to-end with all segmentation and classification targets optimized together. The single image classifier in the HF-HRME module was initialized with weights from Brenes et al. (2022) [11].

When training the MSFN, the colposcopy module was initialized using weights learned by training on data from a previous study conducted at BCH that enrolled 1600 participants (the CLARA study) who received abnormal cervical cytology results (atypical squamous cells of undermined significance or more severe) or positive high-risk HPV DNA test results (cobas 4800 HPV test) [12]. In the CLARA study, women underwent colposcopy, and widefield images were captured after the application of 5% acetic acid using two types of colposcopes: a standard colposcope (CP-M1255 colposcope, D.F. Vasconcellos, Brazil) or a mobile colposcope (EVA 3 Plus, MobileODT, Israel). Two expert colposcopists reviewed the data; a single representative image with the highest image quality was selected for each patient. Each reviewer annotated the image to denote the ectocervix, the squamocolumnar junction, and the os. For each image, the expert annotations were merged using union to create a final ground truth annotation.

The LSTM-Attention and fusion modules were initialized using Xavier random initialization. Hyperparameters were optimized via grid search. Input data augmentation techniques such as rotation, flipping, and random cropping were applied. All code was written in Python 3.6 using PyTorch 1.5.0. Experiments ran in a CUDA 10.2 enabled computer with two GeForce RTX 2080 Ti graphics processing units each with 12 GB VRAM.

F. Evaluation of Multiscale Fusion Network

The colposcopy and HF-HRME modules of the MSFN were evaluated independently on the multiscale dataset.

1) Colposcopy Module: The colposcopy module was first trained and evaluated with data from the CLARA study before being refined with the multiscale dataset. Data from the CLARA study were randomly partitioned at the patient level into training, validation, and test sets in a 3:1:1 split, stratified by the patient's worst histopathologic diagnosis. During training, the CLARA validation set was used to monitor performance. Lesion masks generated by the colposcopy module were compared to the clinical expert's annotations using mean intersection over union (mIOU). Performance was reported on the test set and stratified by the patient's worst histopathologic diagnosis. The ability of the colposcopy module to segment lesions was compared to that of a simple U-Net. The diagnostic performance of the colposcopy module was evaluated by comparing its prediction of whether the patient had CIN 2+ compared to the gold standard of histology using the test set area under the receiver operating curve (AUC); results from the colposcopy module were compared to that of five off-the-shelf deep learning models pretrained on ImageNet. When training the off-the-shelf models, all parameters in the networks were trainable. Each network's last fully connected layer was replaced to enable binary classification, and their weights were initialized using Xavier random initialization.

The colposcopy module was also refined with training data from the multiscale dataset, and its performance was evaluated on the corresponding multiscale validation set. The colposcopy module was initialized with weights learned from the CLARA study data. When training on the multiscale dataset, all parameters in the module were trainable. The ability to segment lesions was measured using mIOU stratified by histopathology, and the ability to predict whether the patient had CIN 2+ was measured by calculating the AUC.

2) HF-HRME Module: The ability of the HF-HRME module to predict whether a site contained CIN 2+ was evaluated with and without the LSTM-Attention block. When the LSTM-Attention block was excluded, the single image classifier scores were averaged to generate a sequence score. The HF-HRME module without the LSTM-Attention was evaluated in two states: (a) with weights from Brenes et al. (2022) and (b) after refinement with the multiscale dataset [11]. The HF-HRME module with the LSTM-Attention block was evaluated in two cases: (a) with fixed weights from Brenes et al. (2022) and a trainable LSTM and (b) with the entire network being trainable [11]. The performance of models was compared by calculating the per site AUC for the validation set and the specificity at a sensitivity matched to clinical impression. The specificities of the four HF-HRME module variants were compared to the specificity of clinical impression using McNemar's test for statistical significance [18].

G. Late Fusion Strategies

In addition to the proposed MSFN, late fusion models were also explored. In late fusion, output probabilities from the colposcopy and HF-HRME modules were combined using a weighted average to generate a new multiscale imaging score for the site. All late fusion models used scores from the best-performing HF-HRME module variant.

The HF-HRME scores were combined with one of three possible colposcopy module scores: (a) a score predicting whether the patient had CIN 2+, (b) a local lesion score calculated by averaging the probabilities of CIN 2+ within the biopsy site, or (c) a global lesion score, calculated by averaging the probabilities of CIN 2+ across the image. The performance of a colposcopy module that performed LSIL+ segmentation was evaluated for the three combinations of scores; similarly, the performance of a colposcopy module that performed segmentation of histologically confirmed CIN 2+ lesions was evaluated for the three score combinations.

TABLE I
NUMBER OF PATIENTS WITH COLPOSCOPY IMAGES IN THE TRAINING, VALIDATION, AND TEST SETS OF THE CLARA STUDY, STRATIFIED BY PATIENT-LEVEL HISTOPATHOLOGIC DIAGNOSIS[a]

Histopathology	Training	Validation	Test
Benign	457 (60%)	165 (63%)	156 (59%)
CIN 1	85 (11%)	27 (10%)	26 (10%)
CIN 2	46 (6%)	19 (7%)	17 (6%)
CIN 3	160 (21%)	50 (19%)	59 (22%)
Invasive Carcinoma	11 (1%)	2 (1%)	5 (2%)
Total	759 (60%)	263 (20%)	263 (20%)

[a]CIN 1, cervical intraepithelial grade 1; CIN 2, cervical intraepithelial grade 2; CIN 3, cervical intraepithelial grade 3.

H. Evaluation Metrics

The sensitivity and specificity of clinical impression in detecting histologically proven CIN 2+ were computed using a clinical impression of LSIL+ as a positive colposcopy result. The ability of the MSFN to predict whether a lesion contained CIN 2+ was measured using the area under the AUC for the validation set; we noted the specificity of the MSFN when the sensitivity was matched to that of clinical impression.

III. RESULTS

A. Data Collection and Curation

Of the 1600 women enrolled in the CLARA study, 1285 had recorded colposcopy images. Their colposcopy images were annotated. The prevalence of CIN 2+ among participants was 29%. A detailed breakdown of patient histology and data partition is shown in Table I.

Data from 286 participants in the multiscale imaging study were retrieved and reviewed for quality control; multiscale images from the 283 participants passing quality control were included in the multiscale imaging dataset. Of the 283 patients, 256 had one biopsy, and 27 had two biopsies. The prevalence of CIN 2+ among these patients was 32% (see Supplement B). Biopsies were acquired from 310 sites imaged with the HF-HRME. A breakdown of histologic diagnosis by site and the data partition is shown in Table II. See Supplement B for examples of the data collected.

TABLE II
NUMBER OF BIOPSIED SITES IN THE TRAINING AND VALIDATION SETS OF THE MULTISCALE IMAGING STUDY, STRATIFIED BY SITE-LEVEL HISTOPATHOLOGIC DIAGNOSIS[a]

Histopathology	Training	Validation
Benign	11 (6.9%)	8 (5.3%)
Cervicitis	91 (57.2%)	86 (57.0%)
CIN 1	7 (4.4%)	5 (3.3%)
CIN 2	9 (5.7%)	6 (4.0%)
CIN 3	41 (25.8%)	46 (30.5%)
Total	159 (100%)	151 (100%)

[a]CIN 1, cervical intraepithelial grade 1; CIN 2, cervical intraepithelial grade 2; CIN 3, cervical intraepithelial grade 3.

TABLE III
MEAN INTERSECTION OVER UNION (MIOU) BETWEEN THE GROUND TRUTH SEGMENTATIONS AND THE LESION SEGMENTATIONS PRODUCED BY THE COLPOSCOPY MODULE, EVALUATED ON THE MULTISCALE DATASET VALIDATION SET AND STRATIFIED BY HISTOPATHOLOGIC DIAGNOSIS[a]

Histopathology	mIOU
Benign	0.67
CIN 1	0.60
CIN 2	0.67
CIN 3	0.68

[a]mIOU, mean intersection over union; CIN 1, cervical intraepithelial grade 1; CIN 2, cervical intraepithelial grade 2; CIN 3, cervical intraepithelial grade 3.

B. Colposcopy Module Evaluation

The colposcopy module, trained and evaluated on the CLARA study to segment LSIL+ lesions and predict patient pathology, obtained an mIOU of 0.70, with an mIOU of 0.69 and 0.73 for histologically confirmed CIN 2 and CIN 3 lesions. The colposcopy module outperformed a simple U-Net model for lesion segmentation for all pathologic diagnoses except cancer, where both models had a mIOU of 0.84 (see Supplement C). The colposcopy module could predict the presence of CIN 2+ from colposcopy images with an AUC of 0.82, and was superior to all tested baseline deep learning models (see Supplement C). The second-best model, a large EfficientNet B7, had an AUC of 0.77. After refinement with the multiscale dataset, the colposcopy module achieved an overall mIOU of 0.67, with an mIOU of 0.67 and 0.68 for histologically confirmed CIN 2 and CIN 3 lesions, respectively (Table III). The refined module could predict the presence of CIN 2+ with an AUC of 0.80.

C. HF-HRME Module Evaluation

Fig. 2 compares the performance of the four HF-HRME module variants evaluated on the multiscale dataset validation set using ROC curves (Fig. 2(a)) and the operating sensitivities and specificities (Fig. 2(b)) to that of clinical impression. Expert clinical impression achieved a sensitivity of 0.98 and a specificity of 0.43 compared to the gold standard of histologic diagnosis. The single image classifier with fixed weight from Brenes et al. (2022) achieved an AUC of 0.844; at the same sensitivity of clinical impression, the specificity of this HF-HRME module was 0.54 [11]. When trained with the multiscale dataset, the single image classifier AUC improved to 0.894, with a specificity of 0.51. Adding the LSTM to the single image classifier improved the specificity of the module, both when the single image classifier was fixed with weights from Brenes et al. (2022) and when it was trained with the multiscale dataset [11]. The best-performing HF-HRME module variant measured by specificity was the trainable single image classifier with a trainable LSTM, which achieved a specificity of 0.66 ($p = 0.0009$). The scatter plot of the predicted probability of CIN 2+ for each site stratified by histopathologic diagnosis (Fig. 2(c)) suggests that adding the LSTM increases specificity by reducing the number of false positive results associated with cervicitis.

Fig. 3 shows the attention maps generated by the best-performing variant of the HF-HRME module for three sites, along with representative images. The attention map and HF-HRME frames (Fig. 3(a)–(d)) shown in the top row are from a site with CIN 1. Attention is high for the frames shown in Fig. 3(b) and Fig. 3(d), which have similar nuclear densities; attention is low for the frames shown in Fig. 3(c), which capture fewer nuclei. The middle row shows an attention map and HF-HRME frames (Fig. 3(e)–(h)) from a site with CIN 3. In the initial portion of the video, nuclei are sparse, small, and regularly shaped; attention is low in this region (Fig. 3f). At approximately frame 170, there is a sharp transition, showing increased nuclear size, dens and pleomorphism (Fig. 3(g)); attention is high throughout this region. Attention decreases again around frame 470 with motion blur (Fig. 3(h)). The bottom row shows an attention map and frames (Fig. 3(i)–(l)) from a site with CIN 3. Early portions of the video show debris regions (Fig. 3(j)), and attention is low in this region. Attention increases in regions of higher quality (Fig. 3(k)), then decreases in regions where a significant portion of the frame is not in focus (Fig. 3(l)).

D. MSFN Evaluation and Comparison to Other Fusion Strategies

Table IV compares the performance of all fusion strategies tested. Strategies 1-6 relied on late fusion, while the MSFN (strategy 7) relied on intermediate fusion. Fusion strategies 1-3 incorporated a colposcopy module trained to segment lesions with LSIL+, whereas fusion strategies 3-6 incorporated a colposcopy module trained to segment lesions with histologically confirmed CIN 2+. In general, when comparing similar scoring strategies, the AUC of late fusion models trained to segment lesions with CIN 2+ was higher than that of models trained with LSIL+ lesion segmentation. Of the three colposcopy scoring strategies tested, AUC was the highest for the global lesion score. Overall, the late fusion model with the highest AUC employed CIN 2+ lesion segmentation and the global lesion scoring strategy to achieve an AUC of 0.900. In comparison, strategy 7, the MSFN, achieved an AUC of 0.910.

Fig. 4(a) shows the probability of CIN 2+ predicted by the MSFN model for each site in the multiscale validation set, stratified by histopathologic diagnosis. Fig. 4(b) shows the receiver operating characteristic curve with an AUC of 0.910. At a sensitivity of 0.98 ($p = 1.0$), the MSFN achieved a specificity of 0.75

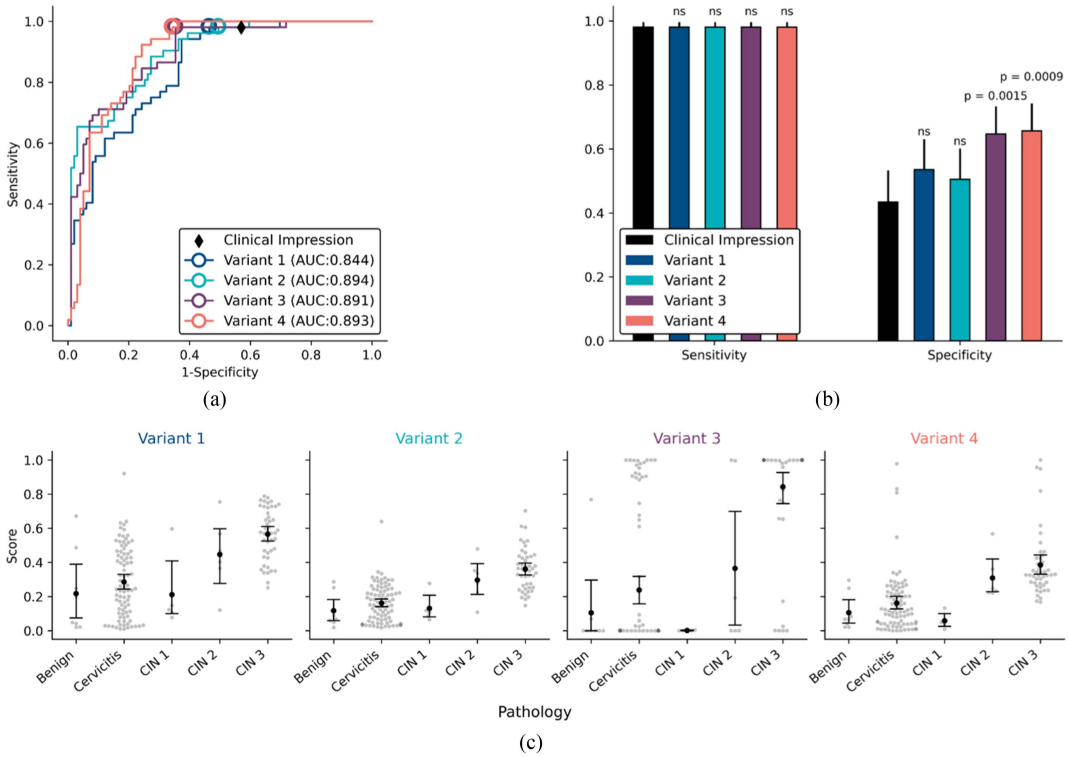

Fig. 2. Performance of four different HF-HRME module variants on the multiscale dataset validation set: Variant 1 – a fixed single image classifier with weights from Brenes et al. (2022), Variant 2 - a trainable single image classifier, Variant 3 - a fixed single image classifier with trainable LSTM, and Variant 4 - a trainable single image classifier with trainable LSTM [10]. (a) The ROC curves of the four variants. The diamond denotes the sensitivity and specificity of clinical impression, and the circles indicate the operating points with the same sensitivity as clinical impression. (b) Comparison of the sensitivities and specificities of clinical impression and the HF-HRME module variants at the operating point. (c) The predicted probability of CIN 2+ for each site, stratified by histologic diagnosis across all four variants. HF-HRME, high frame rate fiber optic microendoscope; LSTM, long short-term memory network; ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve; CIN 1, cervical intraepithelial grade 1; CIN 2, cervical intraepithelial grade 2; CIN 3, cervical intraepithelial grade 3; CIN 2+, cervical intraepithelial grade 2 or more severe.

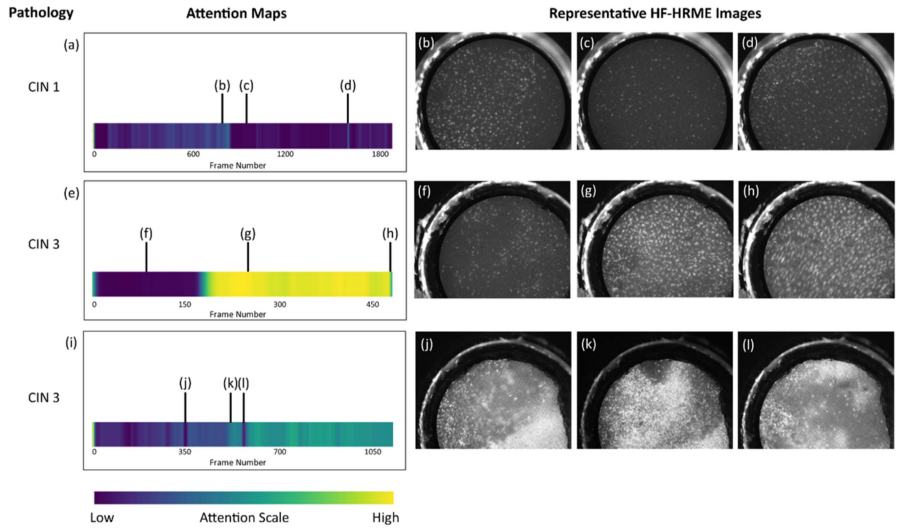

Fig. 3. Attention maps (a), (e), (i) generated by the best performing variant of the HF-HRME module for three sites, along with representative images. Data in the top row correspond to a site with CIN 1; attention is higher for frames (b) and (d) which show higher nuclear contrast and density than frame (c). Data in the middle row were acquired from a site with CIN 3; attention is initially low with frame (f) corresponding to a region with low nuclear density. Attention increases sharply at approximately frame 170, with frame (g) showing increased nuclear density. Attention decreases again near frame 470, with frame (h) showing evidence of motion blur. Data in the bottom row were acquired from a site with CIN 3; attention is low in regions where image quality is reduced due to debris (j) or poor probe contact (l) and higher in regions with high quality images with dense nuclei (k). HF-HRME, high frame rate fiber optic microendoscope; LSTM, long short-term memory network; CIN 1, cervical intraepithelial grade 1; CIN 3, cervical intraepithelial grade 3.

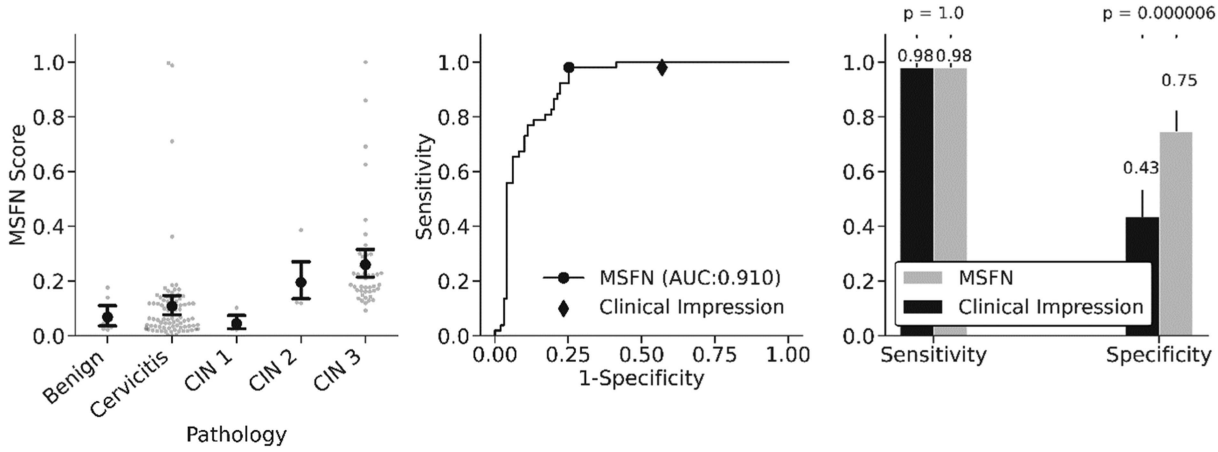

Fig. 4. Performance of the MSFN evaluated on the multiscale dataset validation set. (a) CIN 2+ probability predicted by the MSFN stratified by histopathologic diagnosis. The average CIN 2+ probability for sites with histologic diagnosis of benign, cervicitis, CIN 1, CIN 2, and CIN 3 were 0.07, 0.11, 0.03, 0.20, and 0.26 respectively. (b) The receiver operating characteristic curve with an AUC of 0.910 shows high specificities at high sensitivities. (c) MSFN outperformed colposcopic impression with a sensitivity of 0.98 ($p = 1.0$) and specificity of 0.75 ($p = 0.000006$). MSFN, multiscale imaging fusion network; long short-term memory network; AUC, area under the receiver operating characteristic curve; LSIL+, low-grade lesion or more severe; CIN 2+, cervical intraepithelial grade 2 or more severe.

TABLE IV
DIAGNOSTIC PERFORMANCE OF THE FUSION MODELS TO PREDICT WHETHER A SITE CONTAINS CIN 2+ USING THE MULTISCALE DATASET VALIDATION SET[a]

Fusion Model	Colposcopy		HF-HRME Score	Fusion Strategy	AUC	Sensitivity	Specificity
	Segmentation	Score					
1	LSIL+	Patient Pathology Prediction	Score of Trainable HF-HRME Module with LSTM	Late	0.873	0.98	0.60
2		Local Lesion			0.874	0.98	0.60
3		Global Lesion			0.897	0.98	0.71
4	CIN 2+	Patient Pathology Prediction			0.869	0.98	0.43
5		Local Lesion			0.859	0.98	0.68
6		Global Lesion			0.900	0.98	0.74
7 (MSFN)	CIN 2+	N/A	N/A	Intermediate	0.910	0.98	**0.75**

[a]HF-HRME, high frame rate fiber optic microendoscope; LSTM, long short-term memory network; AUC, area under the receiver operating characteristic curve; LSIL+, low-grade lesion or more severe; CIN 2+, cervical intraepithelial grade 2 or more severe; MSFN, multiscale imaging fusion network.

($p = 0.000006$), significantly higher than the 0.43 specificity of clinical impression (Fig. 4(c)).

IV. DISCUSSION

When combined with computer-aided diagnostic software, optical imaging systems hold great potential for detecting and diagnosing cervical precancers in low-resource settings. However, previous imaging systems based on colposcopy or HRME imaging have achieved high sensitivity but only modest specificity, which limits their utility as triage tests. This study demonstrates that combining colposcopy and HRME imaging can achieve high specificity and high sensitivity. The proposed MSFN model outperformed both colposcopy-only and HRME-only models. In addition, the MSFN surpassed colposcopic impression with a specificity of 0.75 ($p = 0.000006$) at an equal sensitivity of 0.98.

Results to optimize the model show that MSFN intermediate feature fusion is superior, outperforming late fusion strategies. Of the late fusion methods tested, combining the global lesion colposcopy score and the HF-HRME module score performed the best. The global lesion colposcopy score has a larger effective field of view than the local lesion score and MSFN, as it captures colposcopy features from the entire cervix instead of features from an area of interest surrounding the biopsy site. This extended field of view may factor in its high performance among

late fusion strategies. It is possible that allowing the MSFN to consider a larger area of colposcopy features could also improve its performance. However, at this time, the limited availability of registered HF-HRME images outside areas of interest precluded us from implementing multiscale fusion across larger areas.

In addition to the benefits of combining multiscale imaging data, the improved specificity of the MSFN can also be attributed to computationally increasing the effective field-of-view of the HF-HRME. Typically, microscopy modalities are limited by a relatively small field of view. For example, the HF-HRME has a field-of-view of only 790 um. Combining information across HF-HRME frames, the LSTM-Attention model expands the effective field-of-view of the HF-HRME. By processing a sequence of frames acquired across a larger tissue area, the algorithm can gather more contextual information and generate a more robust representation for predicting the probability of CIN 2+. The attention maps show that this contextualization also helps the network learn more abstract concepts like image quality, which could further improve the diagnostic performance.

Strengths of this study include that it was carried out in a middle-income country where the incidence of cervical cancer and the prevalence of HPV are both high, leading to a rich dataset that encompasses a broad range of pathologies. More than 190000 HF-HRME frames were recorded from biopsy sites and mapped to specific locations on colposcopy images. Colposcopy images were annotated with precise outlines of anatomical and clinical features by expert colposcopists. The histopathology results were derived from the consensus diagnosis of two expert pathologists, increasing their reliability. Limitations of this study include that the MSFN does not currently distinguish between CIN 2-3 and invasive carcinoma, which are pathologies that require different treatments. Furthermore, the analysis presented in this study relied on the manual correlation of HF-HRME and colposcopy data. Video captured with the Pocket Colposcope was used to track the location of the HF-HRME probe on the cervix and determine which HF-HRME frames were assigned to a site of interest during fusion. If the MSFN is to be deployed, the HF-HRME frames must be assigned to the inference area in real time. Therefore, successful deployment of the MSFN will require a real-time method to register the HF-HRME frames to their corresponding locations on the cervix, a non-trivial problem due to the handheld nature of the Pocket Colposcope and HF-HRME probe. Improved registration strategies could also enable more complex methods for integrating imaging information acquired across the cervix rather than at specific areas.

In the eventual deployment of the MSFN in a low-resource setting, the system will run on a commercially available laptop. The main building block of the MSFN is the Efficient U-Net, a computationally inexpensive network architecture. In previous works, we have successfully deployed the HF-HRME single image classifier, which uses one Efficient U-Net, for real-time image interpretation in low-resource settings, including Brazil and Mozambique [13], [19]. The model was deployed on a Surface Book 3 (Microsoft, Redmond, Washington, USA) and could predict up to 90 frames per second when plugged into power using the GPU. Compared to the single image classifier, there is little added computational complexity in the MSFN architecture. The colposcopy module may take longer to run since it employs two Efficient U-Nets in series (up to 3 seconds), but it will only run once per clinical case. In addition, while the fusion module must run each time an HF-HRME frame is collected, it only contributes a relatively small number of computations. Preliminary benchmarking tests have shown that the MSFN can infer up to 48 frames per second.

V. Conclusion

The multiscale imaging system powered by the MSFN offers a semi-automated diagnosis with performance that surpasses that of expert colposcopy while relying on low-cost multiscale imaging instrumentation. These attributes make the system accessible to low-resource settings where expert clinicians and financial resources are limited. Operating as a triage test, this system could reduce overtreatment in settings where screen-treat strategies are currently implemented, improving patient outcomes and reducing the clinical and financial burdens of unnecessary treatment. Overall, this work shows how multiscale data can be integrated with computer-aided diagnostic software to improve the diagnosis of cervical precancer.

Conflict of Interest

Nimmi Ramanujam is the co-founder of Calla Health Foundation and currently serves as a technical advisor.

Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to acknowledge Graziela de Macêdo Matsushita for her contribution to the study.

References

[1] World Health Organization and others, *WHO Guideline for Screening and Treatment of Cervical Pre-Cancer Lesions for Cervical Cancer Prevention*. Geneva, Switzerland: World Health Org., 2021.
[2] M. Bonjour et al., "Global estimates of expected and preventable cervical cancers among girls born between 2005 and 2014: A birth cohort analysis," *Lancet Public Heal.*, vol. 6, no. 7, pp. e510–e521, 2021.
[3] J. Lei et al., "HPV vaccination and the risk of invasive cervical cancer," *New England J. Med.*, vol. 383, no. 14, pp. 1340–1348, 2020.
[4] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
[5] World Health Organization and others, *Best Buys' and Other Recommended Interventions for the Prevention and Control of Noncommunicable Diseases*. Geneva, Switzerland: World Health Org., 2017.
[6] J. Mink and C. Peterson, "MobileODT: A case study of a novel approach to an mHealth-based model of sustainable impact," *Mhealth*, vol. 2, 2016, Art. no. 12.
[7] M. N. Asiedu et al., "Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2306–2318, Aug. 2019.
[8] L. Hu et al., "An observational study of deep learning and automated evaluation of cervical images for cancer screening," *JNCI J. Nat. Cancer Inst.*, vol. 111, no. 9, pp. 923–932, 2019.
[9] E. Skerrett et al., "Multicontrast pocket colposcopy cervical cancer diagnostic algorithm for referral populations," *BME Front.*, vol. 2022, 2022, Art. no. 9823184.

[10] B. D. Grant et al., "High-resolution microendoscope for the detection of cervical neoplasia," in *Mobile Health Technologies*. Berlin, Germany: Springer, 2015, pp. 421–434.
[11] D. Brenes et al., "Multi-Task network for automated analysis of high-resolution endomicroscopy images to detect cervical precancer and cancer," *Comput. Med. Imag. Graph.*, vol. 97, 2022, Art. no. 102052.
[12] B. Hunt et al., "Cervical lesion assessment using real-time microendoscopy image analysis in Brazil: The CLARA study," *Int. J. cancer*, vol. 149, pp. 431–441, 2021.
[13] J. B. Coole et al., "Development of a multimodal mobile colposcope for real-time cervical cancer detection," *Biomed. Opt. Exp.*, vol. 13, no. 10, pp. 5116–5130, 2022.
[14] B. Hunt et al., "High frame rate video mosaicking microendoscope to image large regions of intact tissue with subcellular resolution," *Biomed. Opt. Exp.*, vol. 12, no. 5, pp. 2800–2812, 2021.
[15] S. Mehta et al., "Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 893–901.
[16] S. Mehta et al., "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 552–568.
[17] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv1508.04025*.
[18] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
[19] J. B. Coole et al., "Multi-modal mobile colposcope for real-time cervical precancer detection: A pilot study in Mozambique," *Proc. SPIE*, vol. 12369, pp. 40–42, 2023.