Computerized Medical Imaging and Graphics 97 (2022) 102052

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Multi-task network for automated analysis of high-resolution endomicroscopy images to detect cervical precancer and cancer

David Brenes [a,*], CJ Barberan [a], Brady Hunt [a,1], Sonia G. Parra [a,2], Mila P. Salcedo [b], Júlio C. Possati-Resende [c], Miriam L. Cremer [d], Philip E. Castle [e], José H.T.G. Fregnani [f], Mauricio Maza [g], Kathleen M. Schmeler [b], Richard Baraniuk [a], Rebecca Richards-Kortum [a]

[a] Rice University, Houston, TX 77005, USA
[b] University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[c] Barretos Cancer Hospital, Barretos, São Paulo, Brazil
[d] Cleveland Clinic, Cleveland, OH 44195, USA
[e] National Cancer Institute, Bethesda, MD 20814, USA
[f] A. C. Camargo Cancer Center, Liberdade, São Paulo, Brazil
[g] Basic Health International, San Savlador, El Salvador

ARTICLE INFO

Keywords:
Endomicroscopy
Cervical precancer
Multi-task learning
Point-of-care

ABSTRACT

Cervical cancer is a public health emergency in low- and middle-income countries where resource limitations hamper standard-of-care prevention strategies. The high-resolution endomicroscope (HRME) is a low-cost, point-of-care device with which care providers can image the nuclear morphology of cervical lesions. Here, we propose a deep learning framework to diagnose cervical intraepithelial neoplasia grade 2 or more severe from HRME images. The proposed multi-task convolutional neural network uses nuclear segmentation to learn a diagnostically relevant representation. Nuclear segmentation was trained via proxy labels to circumvent the need for expensive, manually annotated nuclear masks. A dataset of images from over 1600 patients was used to train, validate, and test our algorithm; data from 20% of patients were reserved for testing. An external evaluation set with images from 508 patients was used to further validate our findings. The proposed method consistently outperformed other state-of-the art architectures achieving a test per patient area under the receiver operating characteristic curve (AUC-ROC) of 0.87. Performance was comparable to expert colposcopy with a test sensitivity and specificity of 0.94 ($p = 0.3$) and 0.58 ($p = 1.0$), respectively. Patients with recurrent human papillomavirus (HPV) infections are at a higher risk of developing cervical cancer. Thus, we sought to incorporate HPV DNA test results as a feature to inform prediction. We found that incorporating patient HPV status improved test specificity to 0.71 at a sensitivity of 0.94.

1. Introduction

With around 300,000 annual deaths and more than 500,000 new cases each year, cervical cancer is the fourth most common cancer in women worldwide (Arbyn et al., 2020). The incidence of cervical cancer is particularly high in low- and middle-income countries (LMICs), and it is the most common cancer in women living in 42 low-income nations (Arbyn et al., 2020).

Human papillomavirus (HPV) vaccination as well as detection and treatment of cervical precancerous lesions are highly effective at preventing cervical cancer (Arbyn et al., 2020; Lei et al., 2020). Two cervical cancer screening methods are widely used: the Papanicolaou test (Pap test) and the HPV DNA test (Melnikow et al., 2018; William et al., 2018). While challenges remain, HPV DNA testing is becoming more

* Corresponding author.
E-mail addresses: drb12@rice.edu (D. Brenes), cj.barberan@rice.edu (C. Barberan), brady.hunt@dartmouth.edu (B. Hunt), sonia.parra@hsc.utah.edu (S.G. Parra), mpsalcedo@mdanderson.org (M.P. Salcedo), julio.possati@uol.com.br (J.C. Possati-Resende), mcremer@basichealth.org (M.L. Cremer), philip.castle@nih.gov (P.E. Castle), mdfregnani@terra.com.br (J.H.T.G. Fregnani), mmaza@basichealth.org (M. Maza), kschmele@mdanderson.org (K.M. Schmeler), richb@rice.edu (R. Baraniuk), rkortum@rice.edu (R. Richards-Kortum).
[1] Present address: Dartmouth College, Hanover, NH 03755 USA
[2] Present address: University of Utah Health, UT, 84132

https://doi.org/10.1016/j.compmedimag.2022.102052

Available online 26 February 2022

widely available in LMICs (Sankaranarayanan, 2014; Villa, 2012). Patients who screen positive are referred to colposcopy for follow-up diagnosis. The colposcopist uses a low magnification microscope called a colposcope to image the cervix following application of acetic acid and, in some cases, Lugol's iodine, and, when pathology is available, takes biopsies of clinically suspicious lesions (Chase et al., 2009; Olatunbosun et al., 1991). Biopsies are sent to a pathologist for diagnosis. Cervical cancer precursors, also referred to as cervical intraepithelial neoplasia (CIN), are graded as CIN 1, CIN 2, or CIN 3, in order of increasing severity. Patients may also receive a diagnosis of adenocarcinoma in situ (AIS), which indicates that there were cellular abnormalities in the columnar tissue. Patients with CIN 2, CIN 3, and AIS are at a higher risk of developing cervical cancer and in accordance with the World Health Organization guidelines should receive treatment, whereas patients with CIN 1 lesions do not require treatment (Organization et al., 2014). Therefore, clinicians aim to distinguish between <CIN 2 (normal or CIN 1) and CIN 2 + (CIN 2, CIN 3, AIS or cancer) lesions.

The three-visit process for screening, diagnosis, and treatment of cervical precancers has helped reduce the incidence and mortality of cervical cancer in high-income countries. However, it has been difficult to implement such strategies in low- and middle-income settings due to a scarcity of trained professionals (Wilson et al., 2018; Mwaka et al., 2013), lack of affordable equipment, and high rates of loss to follow-up (Hunt et al., 2018; Muñoz et al., 2008). There is a need for low-cost technologies that can be used to diagnose cervical cancer and its precursors at the point-of-care.

Researchers have evaluated whether optical imaging devices coupled with automated algorithms could serve as low-cost alternatives to diagnose cervical precancer in vivo; such tools could bypass the need for a trained colposcopist and pathologist (Thekkek and Richards-Kortum, 2008; Gordon et al., 2019; Asiedu et al., 2018). Several computer-aided diagnostic (CAD) systems have been developed to automate interpretation of colposcopic images, and algorithms that leverage advances in deep learning have improved diagnostic performance (Hu et al., 2019; Yuan et al., 2020; Li et al., 2020). Hu et al. (2019) was one of first to explore deep learning-based computer aided diagnostic systems for cervical cancer and precancer diagnosis using biopsy correlated cervicography data. They successfully trained a Faster R-CNN model to detect the presence of CIN 2 + lesions with a sensitivity and specificity of 100% and 57.5% respectively on a validation set of 192 HPV positive patients (Hu et al., 2019). Since then, several studies using colposcopy data to train deep learning CAD systems have been reported (Yuan et al., 2020; Li et al., 2020; Zhang et al., 2020; Saini et al., 2020; Cho et al., 2020).

Other strategies include high-resolution microendoscope (HRME) imaging to visualize subcellular tissue features; previous algorithms developed to analyze HRME images use morphologic features such as the nuclear-to-cytoplasm area ratio or the number of abnormally shaped or sized nuclei. Recently, Hunt et al. (2021) conducted a prospective evaluation of 1901 images from colposcopically abnormal lesions in Barretos, Brazil. They achieved 89% sensitivity and 54% specificity compared to histopathologic diagnosis (Hunt et al., 2021).

In this paper, we present, to our knowledge, the first application of deep learning to detect cervical precancer and cancer from HRME images. We describe a compact multi-task convolutional neural network (CNN) architecture that first performs the auxiliary task of nuclear segmentation to inform representation learning for HRME image classification to identify the presence of CIN 2 + lesions. Inclusion of an auxiliary task can support representation learning for the main classification task by drawing attention to relevant features or preventing overfitting (Caruana, 1997; Thung et al., 2017; Yang et al., 2017; Gao et al., 2020; Liu et al., 2019). The network also incorporates patient HPV status as an additional clinical attribute to inform prediction. To train, validate, and test our methods, we used data from two large diagnostic studies of HRME imaging conducted in rural Brazil (Hunt et al., 2018; 2021). To validate generalizability of the trained model, we tested its performance on an independent screening study of HRME imaging conducted in El Salvador by a separate group of clinicians (Parra et al., 2021).

We demonstrate that our method trained from random initialization outperforms classification based on morphologic features as well as state-of-the-art deep learning architectures trained either from a random initialization or pretrained on ImageNet (Deng et al., 2009). We also show that our method retains a high performance relative to other deep learning benchmarks and morphologic algorithms when training data are reduced. Results from the independent validation set show that our method performs on par with expert colposcopy.

2. Materials and methods

2.1. High-resolution microendoscope

The HRME is a low-cost, fiber optic fluorescence microscope that is used to image nuclear morphology in vivo (Hunt et al., 2018; Grant et al., 2015; Pierce et al., 2012). Following topical application of 0.01% w/v proflavine (Pantano et al., 2018), the fiber optic probe is placed in gentle contact with the cervix to collect HRME images. Images are collected at 12 fps; the field of view is 790 μm and the lateral resolution is 4.4 μm (Quang et al., 2016).

2.2. Data acquisition and partitioning

Data were assembled from three clinical studies designed to compare the accuracy of HRME to colposcopy using histopathology as the gold standard. The first two studies enrolled screen-positive women in Barretos, Brazil; in these populations the prevalence of histologically detected CIN 2 + was high (35% and 29%, respectively) (Hunt et al., 2018; 2021). The third study enrolled women undergoing cervical screening in San Salvador, El Salvador; in this population the prevalence of CIN 2 + was relatively low (1.5%) (Parra et al., 2021).

In the two high-prevalence studies in Brazil, study procedures took place during a single patient visit. Pap test or HPV DNA (cobas 4800 HPV test) screen-positive participants underwent colposcopy with 5% acetic acid and Lugol's iodine. The colposcopic impression of any abnormal areas (low-grade, high-grade, or suspected cancer) was recorded. Following colposcopy, 0.01% proflavine was applied and HRME images were acquired from areas noted as abnormal by colposcopy and from each quadrant with no lesions. In each quadrant without a lesion, HRME images were acquired from a randomly selected colposcopically normal site at the squamocolumnar junction. When the squamocolumnar junction was not present, a random biopsy from an apparently normal area was acquired at the clinician's discretion. The clinician acquired a single HRME image per site of interest. The HRME result (morphologic abnormality score and classification), colposcopic impression of tissue type (squamous, columnar, or metaplasia), and colposcopic impression (normal, low-grade, high-grade, or suspected cancer) for each site were recorded. Sites identified as abnormal by colposcopy and/or by HRME were biopsied. In cases where a lesion spanned multiple quadrants, the area within the lesion with the most severe colposcopic impression was biopsied. If no abnormal sites were identified by either method, then a single biopsy was taken from a clinically normal site imaged by the HRME (Hunt et al., 2018; 2021).

In the El Salvador study, procedures took place during two patient visits. During Visit 1 (screening visit) HPV DNA testing (QIAGEN careHPV) and VIA screening were performed. Screen-positive women and 10% of women who screened negative (HPV- and VIA-), were invited back for a second visit (Visit 2: triage and diagnostic visit). During Visit 2, VIA was repeated followed by colposcopy with acetic acid and Lugol's iodine. Any abnormalities detected by either VIA or colposcopy were noted along with the clinical impression (low-grade, high-grade, or suspected cancer). 0.01% proflavine was applied and

Table 1
Number of patients and number of HRME images in the training, validation, test, and external evaluation sets stratified by histopathologic diagnosis. Number of HRME images are in parentheses.

Histopathology	Training	Validation	Test	External Evaluation
Negative	547 (689)	191 (238)	193 (237)	274 (437)
CIN1	125 (158)	42 (54)	44 (53)	206 (414)
CIN2	69 (104)	24 (37)	22 (35)	6 (18)
CIN3	203 (280)	71 (103)	70 (99)	21 (37)
AIS	–	–	–	1 (1)
Invasive Carcinoma	15 (23)	5 (9)	6 (11)	–
Total	959 (1254)	333 (441)	335 (435)	508 (907)

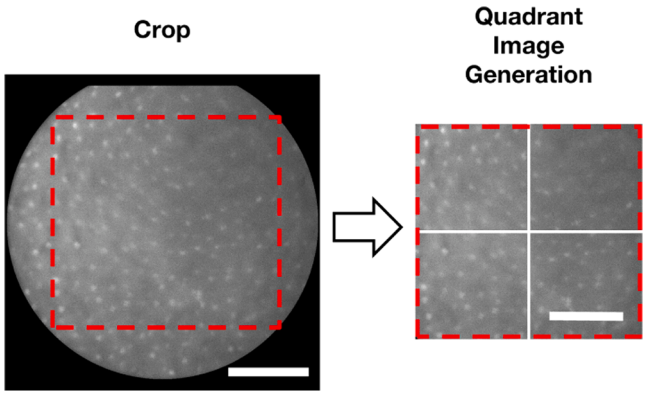

Fig. 1. Pre-processing of high-resolution microendoscopy (HRME) images before the Multi-Task Network's classification. The original HRME image is cropped and divided into four quadrants. The Multi-Task Network is then used to calculate the probability that each quadrant corresponds to CIN2 + . Scale bar is 200 μm. (View in color.).

HRME images were obtained of each abnormality noted during VIA and/or colposcopy along with one normal area of the cervix. All abnormalities detected by colposcopy were biopsied. The colposcopically apparent normal area was biopsied if abnormal by HRME. If there were no abnormalities during the clinical exam, then an HRME image was taken of each quadrant of the cervix and the worst scoring area by HRME was biopsied, regardless of whether the score was abnormal or normal (Parra et al., 2021). The clinician acquired a single HRME image per site of interest. Data were divided into training, validation, and test partitions in a 3:1:1 ratio stratified by patient histopathology. All data from a given patient were assigned to the same partition. The number of imaged sites and corresponding patients included in the training, validation, test, and external evaluation sets can be found in Table 1; results are stratified by histologic diagnosis.

Studies in Brazil were approved by the Barretos Cancer Hospital Ethics Research Committee, the Brazilian National Ethics Research Commission/CONEP (CAAE: 37774314.3.0000.5437, 61743416.1.0000.5437) and the Institutional Review Boards of Rice University (ID#: 653693, 2017–293) and The University of Texas MD Anderson Cancer Center (ID#: 2015–0442, 2017–0096). The study in El Salvador was approved by the Comité Nacional de Ética de la Investigación en Salud (National Ethics Committee of Health Research, ID#: CNEIS/005/2015) in El Salvador and the institutional review boards at The University of Texas MD Anderson Cancer Center (ID#: 2015–0620), Cleveland Clinic (ID#: 15–1162), and Rice University (ID#: 2017–347). Written informed consent was provided by all patients.

2.3. Morphologic analysis

HRME images were analyzed using a prospective morphologic image analysis algorithm. The algorithm segments nuclei within an automatically determined region of interest (ROI) that excludes areas of low- and high-intensity; classifies each segmented nucleus as normal or abnormal based on pre-defined area and eccentricity thresholds; and then reports the number of abnormal nuclei per unit area (Grant et al., 2015). The number of abnormal nuclei per unit area is used to classify each imaged site as normal or abnormal, using a pre-set threshold (Hunt et al., 2018).

2.4. Deep learning benchmarks

To benchmark the performance of the proposed method, we trained, validated, and tested state-of-the-art, off-the-shelf CNN architectures including ResNet18, ResNet34, ResNet50 (He et al., 2015), ResNext50_32×4d (Xie et al., 2017), InceptionV3 (Szegedy et al., 2016), Wide ResNet50_2 (Zagoruyko and Komodakis, 2016), SqueezeNet1_1 (Iandola et al., 2016), EfficientNet-B1 (Tan and Le, 2019), and VGG16bn (Simonyan and Zisserman, 2015) models. Two different initializations were used, Xavier initialization (Glorot and Bengio, 2010) and transfer learning from ImageNet. Models initialized through transfer learning were trained by full optimization (all weights in the model were learnable), or by fine-tuning (only weights in the last fully connected layers were learnable). Five models were independently trained for each architecture and initialization-training strategy pair. Original HRME images were center cropped, as shown in Fig. 1, and resized to the network's input size. For networks initialized through transfer learning, input images were normalized with the mean and standard deviation of ImageNet. Otherwise, input images were normalized with the mean and standard deviation of the training set. All networks were trained to classify HRME images as either < CIN 2 or CIN 2 + until the training AUC-ROC reached one, and the model with the highest validation AUC-ROC was selected for testing. Standard data augmentation techniques such as rotation, flipping, and random cropping were applied, and grid search was used for hyperparameter parameter tuning.

2.5. Proposed method: Multi-task learning with learning via proxy labels

The proposed Multi-Task Network (MTN), shown in Fig. 2, performs two tasks - nuclear segmentation and classification - and is based on Y-Net architecture proposed by Mehta et al. (2018a). The segmentation component of the network has an encoder-decoder structure. In the encoder, efficient spatial pyramid (ESP) modules handle the primary feature extraction operations. The decoder receives the encoder's final feature representation and uses upsampling and pyramid spatial pooling (PSP) modules to construct a nuclear mask with same spatial resolution as the input. Concatenating skip connections from the encoder to the decoder enables information sharing between the two. The MTN's diagnostic component is built on more ESP modules that culminate on an average global pooling module followed by two fully connected layers. Downsampling operations are handled by a single convolution, which reduces spatial resolution of the feature maps by half. Upsampling is performed by bilinear interpolation. Batch normalization and ReLU activation are applied after each downsampling operation, upsampling operation, ESP and PSP module. In the next sections, we describe the modules that compose the MTN and the learning via proxy labels.

2.5.1. Efficient spatial pyramid module

The ESP module consists of two components: an initial point-wise convolution followed by a spatial pyramid of dilated convolutions (Mehta et al., 2018b). Point-wise convolution applies a 1 by 1 kernel to the input, reducing the number of channels and downstream computations. The spatial pyramid of dilated convolutions takes in the output of the point-wise convolution and applies four parallel convolutional filters, each with a 3 by 3 kernel applied at different dilation factors. This

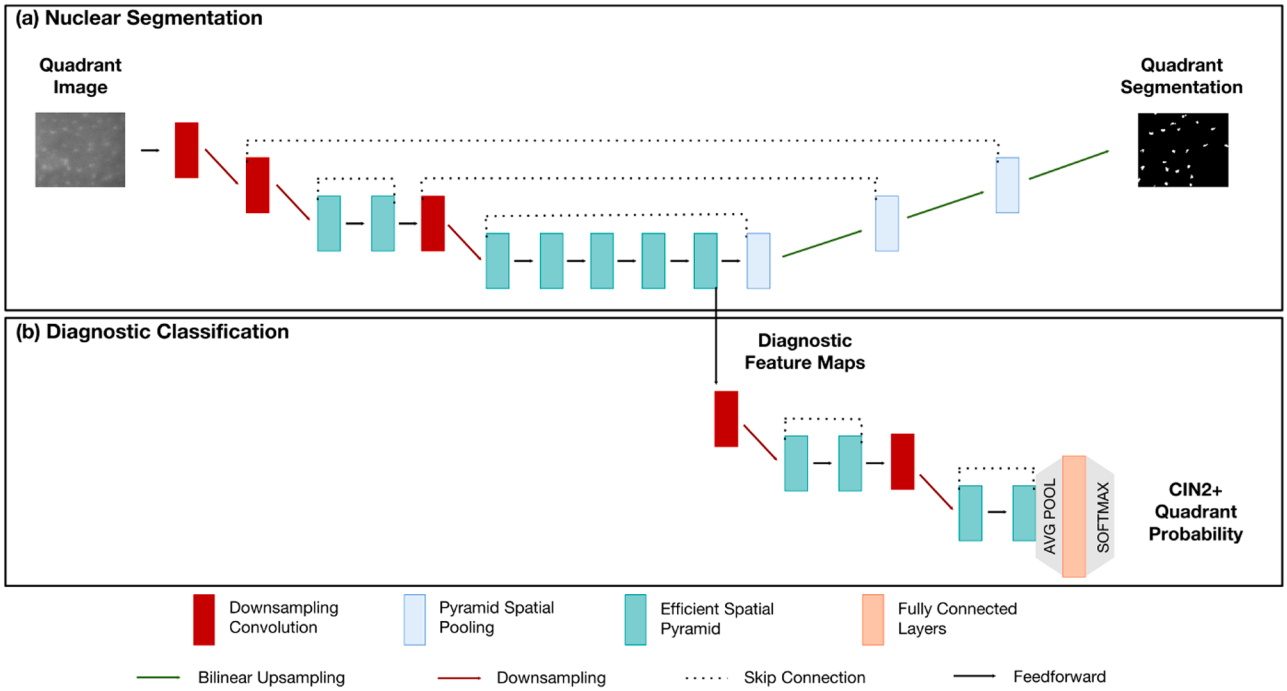

Fig. 2. The Multi-Task Network architecture consists of two main components: (a) Nuclear segmentation component and (b) Diagnostic classification component. In the nuclear segmentation component, the network generates a binary mask corresponding to cervical cell nuclei in the quadrant image. In the diagnostic classification component, diagnostic feature maps from the nuclear segmentation component are used to calculate the probability that each quadrant image corresponds to CIN2 +. (View in color.).

dilation increases the network's receptive field, incorporating multi-scale information without increasing the number of parameters (Yu and Koltun, 2015). Outputs from each dilated convolution are hierarchically summed, beginning with the lower levels, and the summations are concatenated. A residual connection between the input and output of the ESP module aids in gradient stability. In addition, a series of consecutive ESP modules has a concatenating skip connection joining the input and output feature maps.

2.5.2. Pyramid spatial pooling module

The PSP module extends the network's receptive field by integrating multi-scale features (Zhao et al., 2017). The PSP module uses average pooling operations at four different levels. The output of each pooling operation undergoes a point-wise convolution that compresses the number of channels and bilinear upsampling to match their spatial resolution to that of the input. The input and output feature maps of each pooling operating are concatenated.

2.5.3. Learning via proxy labels and two stage training

The MTN is trained in two stages, starting with the segmentation component. As shown in Fig. 1, cropped HRME images were divided into four non-overlapping image quadrants and passed through the network separately. The nuclear segmentation resulting from the morphologic algorithm was used to supervise the nuclear segmentation performed with the MTN. This stage was trained using stochastic gradient descent for three hundred epochs with a cosine annealing scheduled learning rate with restarts every 30 epochs and batch size of 10 (Loshchilov and Hutter, 2016). The model with the highest validation mean intersection over union (mIOU) was selected to initialize training stage two.

In stage two, the diagnostic and segmentation components were optimized together. The diagnostic branch was appended to the architecture, and weights learned in stage one were used to initialize the architecture. This second stage was trained via stochastic gradient descent for 300 epochs with an annealing cosine scheduled learning rate with restarts every 20 epochs with a batch size of 5. Hyperparameters were optimized via grid search. In both stages, 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.

2.5.4. Ablation of nuclear segmentation task

To evaluate the contribution of the nuclear segmentation component to the diagnostic performance of the MTN, an ablation study was conducted where the nuclear segmentation decoder of the MTN was removed. This ablated structure, consisting of the segmentation encoder and the diagnostic branch, was initialized using Xavier initialization and trained to classify HRME quadrants as <CIN 2 or CIN 2 +. The network was optimized via stochastic gradient descent for 300 epochs with an annealing cosine scheduled learning rate with restarts every 20 epochs with a batch size of 5. Hyperparameters were optimized via grid search. Input data augmentation techniques such as rotation, flipping, and random cropping were applied.

2.5.5. Incorporating patient HPV status

One-hot encoding was applied to convert results of the HPV DNA test into a numeric value (high-risk HPV positive: 1, high-risk HPV negative: 0). This attribute was appended to the feature vector of the MTN after average pooling and passed on to the fully connected layers for classification. Patients were excluded if HPV DNA test results were not available or were indeterminant, resulting in exclusions of 101, 34, and 24 patients from the training, validation, and test sets respectively. HPV DNA test results were available for all patients in the external evaluation set. The prevalence of high-risk HPV in the test and evaluation sets was 65% and 36%, respectively. Patient HPV results were also incorporated on the second training stage. The performance of the MTN with and without incorporating patient HPV DNA test results was compared for the test and external evaluation sets.

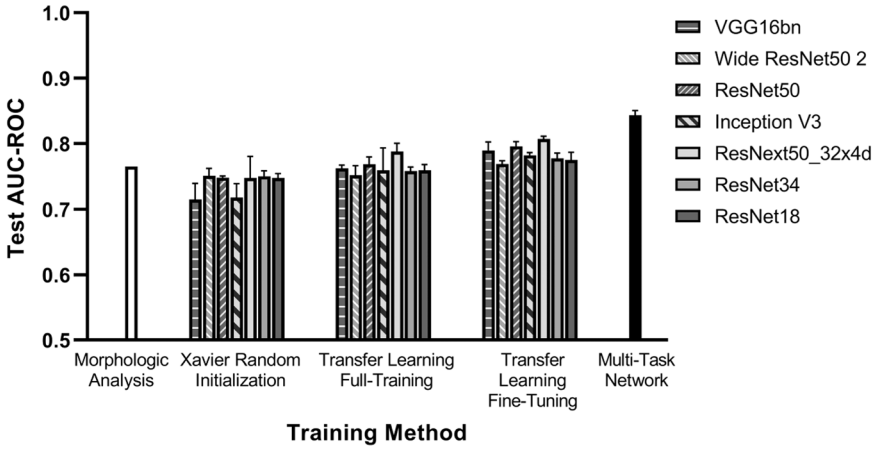

Fig. 3. Area under the per site ROC curve for algorithms applied to the test set, including morphologic analysis, the deep learning benchmark architectures using three different initialization-training strategies, and for the Multi-Task Network. The Multi-Task Network outperforms all the benchmarks, including morphologic analysis.

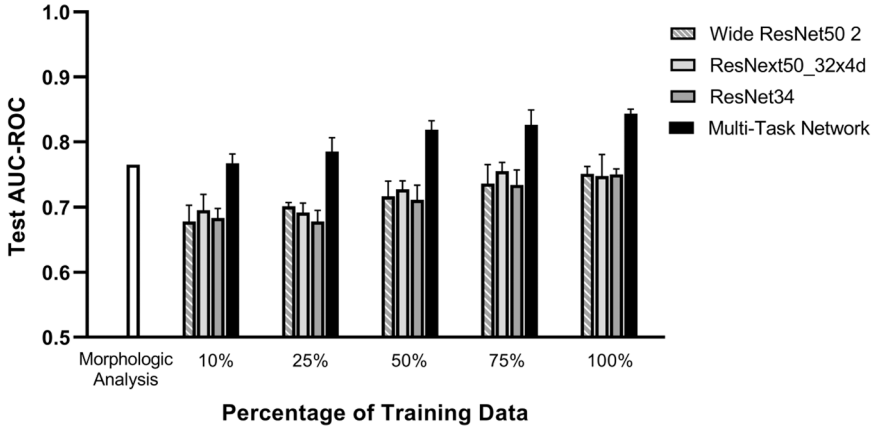

Fig. 4. Area under the per site ROC curve for algorithms applied to the test set stratified by amount of training data; algorithms include the Multi-Task Network and the top three performing architectures from the Xavier random initialization group. The Multi-Task Network consistently outperforms the deep learning benchmarks and morphologic analysis, even as the number of patient images in the training set is reduced from 100% to 25%.

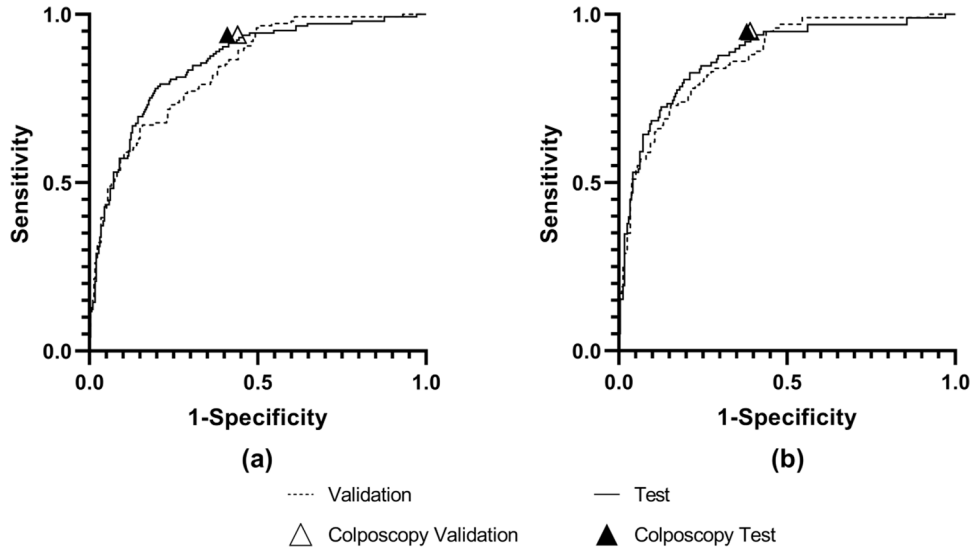

Fig. 5. ROC curves for the Multi-Task Network for HRME images in the validation and test sets when data are analyzed per site (a) or per patient (b) using histopathologic diagnosis as the gold standard. The performance of colposcopy is also shown for the validation and test datasets (triangle marker). The Multi-Task Network achieves similar sensitivity and specificity to that of colposcopy.

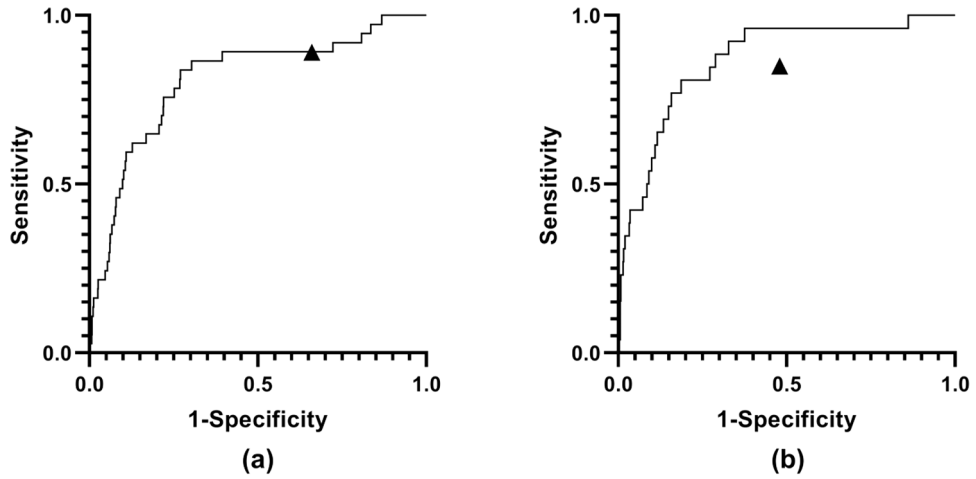

Fig. 6. ROC curves for the Multi-Task Network for HRME images in the external evaluation set when data are analyzed per site (a) or per patient (b) using histopathologic diagnosis as the gold standard. The performance of colposcopy is also shown (triangle marker).

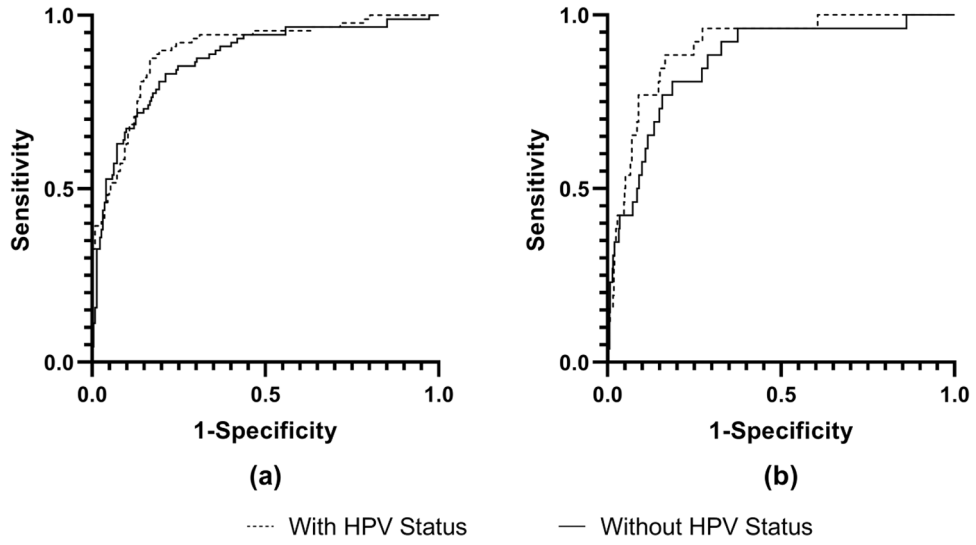

Fig. 7. Per patient ROC curves for the test (a) and external evaluation (b) sets with and without incorporating patient high-risk HPV DNA test results as a feature.

2.6. Training set size reduction

To test performance of the MTN when training data are limited, we reduced the amount of training data to 10%, 25%, 50%, and 75% of its original size. These reductions were applied at the patient level to simulate a smaller-scale study where sample diversity may be limited. Sampling was stratified by pathology to retain the same disease prevalence in each reduced set. The MTN and benchmarks were trained using these reduced sets in replicates of five, where each replicate contained a different set of patients sampled from the training data.

2.7. Visualizations

After the networks were trained, we used guided-backpropagation to visualize which input features were prioritized by the network (Springenberg et al., 2014). We compared results of guided-backpropagation of the MTN to that of the best-performing off-the-shelf CNN.

To visualize the impact tissue-type has on the feature space of the MTN, we generate a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the penultimate fully connected layer of the MTN (der Maaten and Hinton, 2008). We implemented t-SNE with a principal component analysis initialization and perplexity of 50. K-means clustering was used to demarcate image groups. t-SNE plots were constructed for the validation and the test set.

2.8. Diagnostic evaluation metrics

The diagnostic performance of morphologic analysis and the MTN were evaluated by constructing the receiver operating characteristic (ROC) curves, computing the area under the ROC curve (AUC-ROC), and the sensitivity and specificity at relevant operating points. Statistical significance testing for differences in sensitivity and specificity between colposcopy and the MTN was performed using McNemar's test (McNemar, 1947). The Y-Net's operating point was determined by selecting a threshold at which its per site sensitivity matches that of the colposcopic impression for the validation set. This selection criterion ensures that the instrument's performance was on par with expert colposcopy with regards to sensitivity, encouraging a low number of false negatives. Since patients can have multiple suspicious lesions, results are given both at a per site and per patient level. Clinical decision making is done based on the worst histopathology result across all lesions.

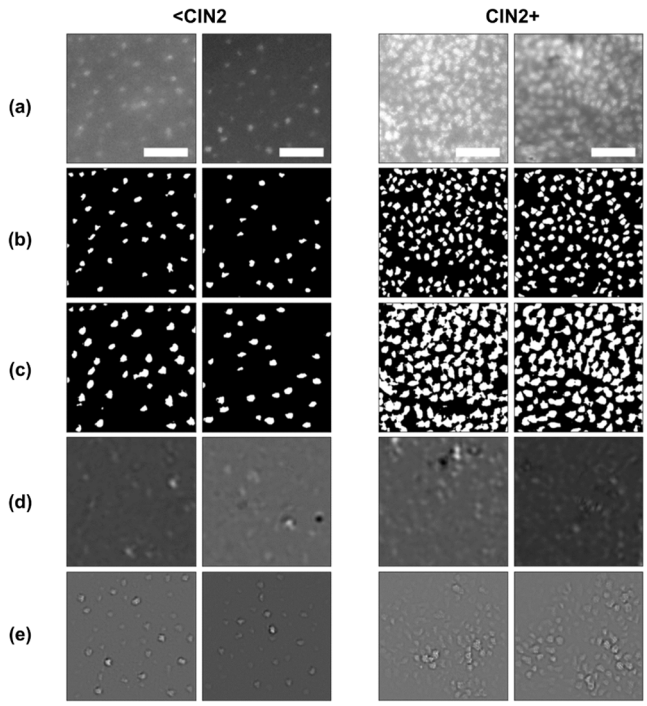

Fig. 8. Representative test set image quadrants from sites diagnosed as <CIN2 (left column) and CIN2 + (right column) that were correctly classified by the Multi-Task Network. Scale bar is 100 μm. (a) Original images; (b) Proxy label of nuclear segmentation; (c) Binary mask resulting from Multi-Task Network nuclear segmentation; (d) Guided backpropagation using the best-performing deep learning benchmark (ResNext50_32×4d transfer learned, fine-tuned model); and (e) Guided-backpropagation using the Multi-Task Network model. Multi-Task Network guided-backpropagation more consistently emphasizes pixels associated with nuclei than the best-performing deep learning benchmark.

3. Results

As shown in Fig. 3, the MTN outperformed all CNN benchmarks and the morphologic algorithm when applied to the test set. The per site AUC-ROC for the MTN was 0.85. In contrast, the best-performing CNN benchmark architecture within the same initialization-training strategy was a ResNext50_32×4d model with a per site AUC-ROC of 0.77. Using transfer learning to initialize the networks improved the performance of the CNN benchmarks. The best-performing CNN benchmark that underwent transfer learning initialization was a ResNext50_32×4d model from the Transfer Learning Fine-Tuning group, with a per site AUC-ROC of 0.82. The per site AUC-ROC of the ablated MTN without the nuclear segmentation decoder was 0.76.

The Multi-Task Network outperformed all CNN benchmarks and the morphologic algorithm even as the amount of training data was reduced. As illustrated in Fig. 4, reducing the amount of training data decreased the performance of the MTN. Nevertheless, the MTN consistently outperformed all CNN benchmarks and the morphologic algorithm even when only 25% of training data were used.

The per site validation and test ROC curves depicted in Fig. 5a for the best-performing MTN model show no generalization gap between the network's performance on the two sets. For the test set, the sensitivity of the MTN was 0.93 and specificity was 0.55 at the Q-point. There were no statistically significant differences in the sensitivity ($p = 0.2$) and specificity ($p = 1.0$) of the MTN and expert colposcopic impression. Similarly, in the per patient analysis shown in Fig. 5b, no generalization gap was observed between the validation and test sets and no significant differences were found in the sensitivity and specificity of the MTN and colposcopic impression. The MTN had a per-patient AUC-ROC of 0.87 and a sensitivity and specificity of 0.94 ($p = 0.3$) and 0.58 ($p = 1.0$), respectively. Table A. 1 provides a summary of the clinical performance of colposcopic impression, the morphologic analysis, and the MTN.

When applied to the external evaluation set, the MTN achieved a per site AUC-ROC of 0.81 as shown in Fig. 6a. While no significant difference in specificity was found between the MTN and colposcopic impression ($p = 1.0$), the MTN's sensitivity was significantly lower; the sensitivity of the MTN was 0.86 while that of colposcopic impression was 0.89 ($p < 0.0001$). However, in the per patient analysis depicted in Fig. 6b the MTN outperformed colposcopic impression. The MTN had a per patient AUC-ROC of 0.87 and a sensitivity and specificity of 0.96 ($p = 0.005$) and 0.59 ($p = 0.37$), respectively. For comparison, the per site and per patient AUC-ROCs of the best overall CNN benchmark (ResNext50_32×4d Transfer Learning Fine-Tuning) on the external evaluation set were 0.77 and 0.81, respectively.

Incorporating patient HPV DNA test results as a feature into the MTN increased the per patient AUC-ROC for both test and external evaluation sets. An increase in AUC-ROC of 3.4% (0.87–0.90) was observed for the test set, whereas the AUC-ROC increased by 4.6% (0.87–0.91) for the external evaluation set as shown in Fig. 7. This increase in AUC-ROC can

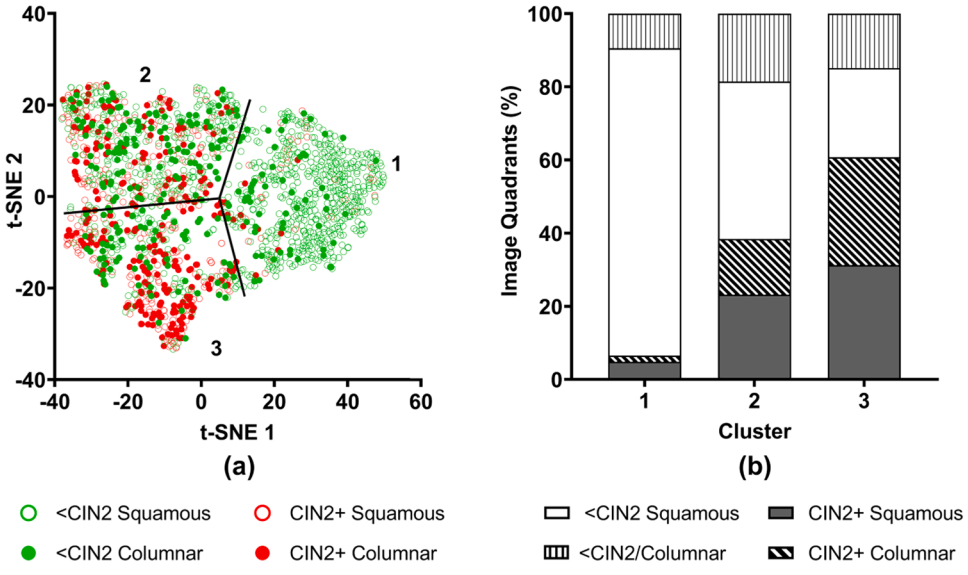

Fig. 9. t-SNE visualization of features from the penultimate fully connected layer of the Multi-Task Network for image quadrants in the test set (a). Each point corresponds to features from a single image quadrant, and is colored according to pathology and shaded according to tissue type. Image quadrant features cluster not only by pathology but also by colposcopic tissue type. K-means clustering ($k = 3$) produces: 1. A cluster of < CIN2, squamous tissue type sites; 2. A cluster with increased amounts of columnar tissue type and CIN2 + sites; and 3. A cluster of mostly columnar tissue with CIN2 + sites. Clusters were named by comparing their pathology and tissue-type distributions (b). This clustering highlights the confounding effect tissue type has on the Multi-Task Network's feature extraction for pathology prediction.

be attributed to an improvement in specificity at the high sensitivity operating points of the ROC. At the test set operating point corresponding to a sensitivity of 0.94, the specificity increased from 0.58 to 0.71.

Fig. 8 highlights differences in feature attention between the MTN and the best overall CNN benchmark (ResNext50_32×4d Transfer Learning Fine-Tuning) by visualizing the guided-backpropagations of a representative image set. The MTN guided-backpropagations had strong, well-defined activation in regions corresponding to cell nuclei. In contrast, the benchmark guided-backpropagations had sporadic activations in regions corresponding to cell nuclei as well as surrounding areas.

The test set TSNE plot in Fig. 9a illustrates the impact that tissue type has on the MTN's feature embedding. K-Means clustering ($k = 3$) yields three distinct clusters with an increasing percentage of both columnar tissue and CIN 2 + lesions (Fig. 9b): 1. A cluster containing 89% squamous tissue sites with 93% of sites < CIN 2; 2. A cluster containing 34% columnar tissue sites with 38% of sites containing CIN 2 +; and 3. A cluster of mostly columnar tissue sites (44%) with 61% of sites containing CIN 2 +. Fig. B1 exemplifies the visual differences between squamous and columnar/metaplasia images that could lead the MTN to generate different embeddings for each tissue type.

4. Discussion

Cervical precancer and cancer detection remains a challenge in low- and middle-income countries due to poor access to diagnosis, limited number of trained professionals, and lack of affordable equipment. To tackle this problem and provide an alternative to conventional biopsy and histopathology, we developed a deep learning-based CAD system to interpret high resolution images and detect precancer and cancer. The diagnostic performance of the MTN was on par with expert colposcopic impression for the test and external-evaluation sets in the per patient analysis. The MTN described here has several advantages compared to other deep learning architectures by leveraging nuclear segmentation as an auxiliary task for classification. As suggested by the guided-backpropagations, the segmentation task may have steered the MTN's attention to nuclei, favoring the extraction of nuclear morphology and the tissue organization information that aids in pathology prediction.

Moreover, incorporating HPV status as a diagnostic feature boosted the per patient specificity at high sensitivity operating points for the test and external-evaluation sets. High-risk HPV plays a critical role in the progression to cervical cancer. As high-risk HPV DNA testing becomes more available in LMICs (William et al., 2018) this clinical variable could become an integral input to CAD systems for automated cervical precancer diagnosis in low- and middle-income countries. This work defined the HPV status as a pooled result of high-risk HPV variants. However, future work may focus on the role of specific HPV variants, such as HPV 16 and HPV 18 which are known to be more oncogenic. Aside from high-risk HPV status, several works on colposcopy image classification have explored age and Pap test result among others as useful clinical variables to aid in patient diagnosis (Yuan et al., 2020; Xu et al., 2016). While this study does not study these variables, we acknowledge that they may also be of interest for future work.

While the MTN was successful at diagnostic feature extraction, we observed that the network's performance to classify images of columnar sites was lower than that for squamous sites. Nuclei in benign columnar sites tend to be tightly packed and are often arranged in a glandular pattern, whereas nuclei in benign squamous sites are more evenly distributed. The t-SNE of the MTNs features suggests that the network may be sensitive to these morphometric differences. As a result, tissue type may act as a confounding variable limiting the model's performance. Future work should explore untangling tissue type prediction from pathology prediction to improve diagnostic performance.

The HRME does not operate as a standalone device and must rely on guidance from colposcopic impression. When colposcopists are unavailable this presents a significant limitation. Advancement in the field of colposcopy image analysis may aid in guiding appropriate placement of the HRME probe by localizing high-risk areas on the cervix.

Low-cost imaging technologies coupled with deep learning-based CAD systems could address the challenges faced by low- and middle-income countries in expanding and optimizing cervical cancer prevention programs. In this work, we show that the HRME powered by our deep learning based diagnostic system performs comparable to expert colposcopy for cervical precancer and cancer diagnosis.

Citation diversity

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are undercited relative to the number of papers in the field (Caplar et al., 2017; Dworkin et al., 2020; Maliniak et al., 2013; Dion et al., 2018). We recognize this bias and have worked diligently to ensure that we are referencing appropriate papers with fair gender and racial author inclusion.

CRediT authorship contribution statement

David Brenes: Methodology, Software, Investigation, Formal analysis, Writing – original draft. **CJ Barberan:** Validation, Investigation, Resources. **Brady Hunt:** Validation, Data curation. **Sonia G. Parra:** Data curation. **Mila P. Salcedo:** Data curation, Supervision. **Júlio C. Possati-Resende:** Investigation, Resources. **Miriam L. Cremer:** Project administration. **Philip E. Castle:** Supervision. **José H. T. G. Fregnani:** Project administration. **Mauricio Maza:** Project administration, Resources. **Kathleen M. Schmeler:** Conceptualization, Supervision, Funding acquisition. **Richard Baraniuk:** Conceptualization, Supervision, Funding acquisition. **Rebecca Richards-Kortum:** Conceptualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request through a data-sharing agreement that provides for: (a) a commitment to securing the data only for research purposes and not to identify any individual participant; (b) a commitment to securing the data using appropriate computer technology; and (c) a commitment to destroying or returning the data after analyses are completed.

Acknowledgements

This work was supported by National Cancer Institute: R01 CA251911, R01 CA186132, UH2/UH3 CA189910, P30 CA016672; NSF grants: CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014–18-12571, N00014–20-1–2534, N00014–18-1–2047, and MURI N00014–20-1–2787; AFOSR grant FA9550–18-1–0478; and a Vannevar Bush Faculty Fellowship.

Appendix A. . Overall per patient performance comparison

See Appendix Table A1.

Table A1
Summary of the per patient diagnostic performance of colposcopy impression, morphologic analysis, and the Multi-Task Network on the test set.

Method	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Colposcopy	0.95	0.62	0.50	0.97
Morphologic Analysis	0.91	0.60	0.48	0.94
Multi-Task Network	0.94	0.58	0.49	0.96

Appendix B. . Visual differences across tissue types

See Fig. B1.

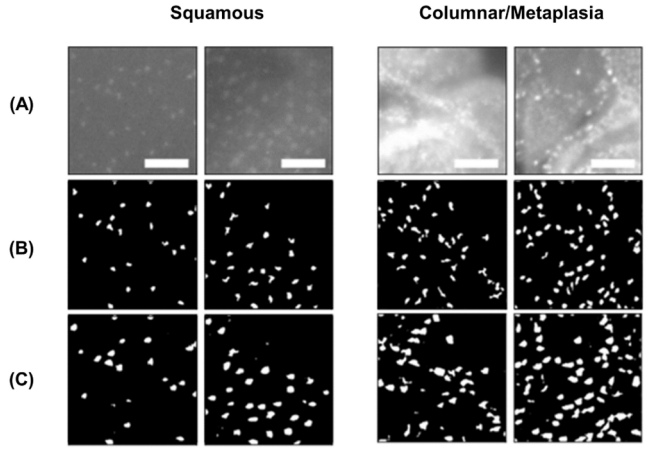

Fig. B1. Representative test set image patches from squamous (left column) and columnar (right column). (A) Original image, (B) Proxy label of nuclear segmentation, and (C) Multi-Task Network's nuclear segmentation prediction. High background, non-nuclear fluorescence in columnar images impedes nuclear segmentation resulting in less accurate nuclear masks as shown in columnar (B) and (C). Scale bar is 100 μm.

References

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., Bray, F., 2020. Estimates of incidence and mortality of cervical cancer in 410 2018: a worldwide analysis. Lancet Glob. Health 8, e191–e203.

Asiedu, M.N., Simhal, A., Chaudhary, U., Mueller, J.L., Lam, C.T., Schmitt, J.W., Venegas, G., Sapiro, G., Ramanujam, N., 2018. Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope. IEEE Trans. Biomed. Eng. 66, 2306–2318.

Caplar, N., Tacchella, S., Birrer, S., 2017. Quantitative evaluation of gender bias in astronomical publications from citation counts. Nat. Astron. 1, 141.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75.

Chase, D.M., Kalouyan, M., DiSaia, P.J., 2009. Colposcopy to evaluate abnormal cervical cytology in 2008. Am. J. Obstet. Gynecol. 200, 472–480.

Cho, B.J., Choi, Y.J., Lee, M.J., Kim, J.H., Son, G.H., Park, S.H., Kim, H.B., Joo, Y.J., Cho, H.Y., Kyung, M.S., et al., 2020. Classification of cervical neoplasms on colposcopic photography using deep learning. Sci. Rep. 10, 1–10.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database 2009 248 255.

der Maaten, L., Hinton, G., 2008. Visualizing data using T-sne. J. Mach. Learn. Res.

Dion, M.L., Sumner, J.L., Mitchell, S.M., 2018. Gendered citation patterns across political science and social science methodology fields. Political Anal. 26, 312–327.

Dworkin, J.D., Linn, K.A., Teich, E.G., Zurn, P., Shinohara, R.T., Bassett, D.S., 2020. The extent and drivers of gender imbalance in neuroscience reference lists. Nat. Neurosci. 23 (8), 918–926. https://doi.org/10.1101/2020.01.03.894378.

Gao, F., Yoon, H., Wu, T., Chu, X., 2020. A feature transfer enabled multi-task deep learning model on medical imaging. Expert Syst. Appl. 143, 112957.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings. pp. 249–256.

Gordon, S., Seadia, O., Levi, E., Landesman, I., 2019. A novel multimodal optical imaging device for cervical cancer screening and diagnosis. Proceedings, 108560L.

Grant, B.D., Schwarz, R.A., Quang, T., Schmeler, K.M., Richards-Kortum, R., 2015. High-resolution microendoscope for the detection of cervical neoplasia. In: Mobile Health Technologies. Springer, pp. 421–434.

He, K. , Zhang, X. , Ren, S. , Sun, J. , 2015. Deep residual Learn. Image Recognit. 2015.

Hu, L., Bell, D., Antani, S., Xue, Z., Yu, K., Horning, M.P., Gachuhi, N., Wilson, B., Jaiswal, M.S., Befano, B., et al., 2019. An observational study of deep learning and automated evaluation of cervical images for cancer screening. JNCI: J. Natl. Cancer Inst. 111, 923–932.

Hunt, B., Fregnani, J.H.T.G., Schwarz, R.A., Pantano, N., Tesoni, S., PossatiResende, J.C., Antoniazzi, M., de Oliveira Fonseca, B., de Macêdo Matsushita, G., Scapulatempo-Neto, C., et al., 2018. Diagnosing cervical neoplasia in rural brazil using a mobile van equipped with in vivo microscopy: 24 a cluster-randomized community trial. Cancer Prev. Res. 11, 359–370.

Hunt, B., Fregnani, J.H.T.G., Brenes, D., Schwarz, R.A., Salcedo, M.P., PossatiResende, J. C., Antoniazzi, M., de Oliveira Fonseca, B., Santana, I.V.V., de Macêdo Matsushita, G., 2021. Cervical lesion assessment using real-time microendoscopy image analysis in brazil: the clara study. Int. J. Cancer.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K., 2016. Squeezenet: alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. ArXiv Prepr. ArXiv 1602, 07360.

Lei, J., Ploner, A., Elfström, K.M., Wang, J., Roth, A., Fang, F., Sundström, K., Dillner, J., Sparén, P., 2020. Hpv vaccination and the risk of invasive cervical cancer. N. Engl. J. Med. 383, 1340–1348.

Li, Y., Chen, J., Xue, P., Tang, C., Chang, J., Chu, C., Ma, K., Li, Q., Zheng, Y., Qiao, Y., 2020. Computer-aided cervical cancer diagnosis using timelapsed colposcopic images. IEEE Trans. Med. Imaging 39, 3403–3415.

Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Trans. Med. Imaging 39, 718–728.

Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. Arxiv preprint ArXiv 1608, 03983.

Maliniak, D., Powers, R., Walter, B.F., 2013. The gender citation gap in international relations. Int. Organ. 67, 889–922.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157.

Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H., 2018b. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in: Proceedings of the european conference on computer vision (ECCV), pp. 552–568.

Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L., 2018a. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. Arxiv 25, 893–901.

Melnikow, J., Henderson, J.T., Burda, B.U., Senger, C.A., Durbin, S., Weyrich, M.S., 2018. Screening for cervical cancer with high-risk human papillomavirus testing: updated evidence report and systematic review for the us preventive services task force. Jama 320, 687–705.

Muñoz, N., Franco, E.L., Herrero, R., Andrus, J.K., de Quadros, C., Goldie, S.J., Bosch, F. X., 2008. Recommendations for cervical cancer prevention in latin america and the caribbean. Vaccine 26, L96–L107.

Mwaka, A.D., Wabinga, H.R., Mayanja-Kizza, H., 2013. Mind the gaps: a qualitative study of perceptions of healthcare professionals on challenges and proposed remedies for cervical cancer help-seeking in post conflict northern uganda. BMC Fam. Pract. 14, 1–14.

Olatunbosun, O.A., Okonofua, F.E., Ayangade, S.O., 1991. Screening for cervical neoplasia in an african population: Simultaneous use of cytology and colposcopy. Int. J. Gynecol. Obstet. 36, 39–42.

Organization, W.H., et al., 2014. Who guidelines for treatment of cervical intraepithelial neoplasia 2–3 and adenocarcinoma in situ: Cryotherapy, large loop excision of the transformation zone, and cold knife conization.

Pantano, N., Hunt, B., Schwarz, R.A., Parra, S., Cherry, K., Possati-Resende, J.C., Longatto-Filho, A., Fregnani, J.H.T.G., Castle, P.E., Schmeler, K., et al., 2018. Is proflavine exposure associated with disease progression in women with cervical dysplasia? a brief report. Photochem. Photobiol. 94, 1308–1313.

Parra, S.G., López-Orellana, L.M., Molina Duque, A.R., Carns, J.L., Schwarz, 515 R.A., Smith, C.A., Ortiz Silvestre, M., Diaz Bazan, S., Escobar, P.A., Felix, 26 J.C., et al., 2021. Cervical cancer prevention in el salvador: a prospective evaluation of screening and triage strategies incorporating high-resolution microendoscopy to detect cervical precancer. Int. J. Cancer 148, 2571–2578.

Pierce, M.C., Guan, Y., Quinn, M.K., Zhang, X., Zhang, W.H., Qiao, Y.L., Castle, P., Richards-Kortum, R., 2012. A pilot study of low-cost, high-resolution microendoscopy as a tool for identifying women with cervical precancer. Cancer Prev. Res. 5, 1273–1279.

Quang, T., Schwarz, R.A., Dawsey, S.M., Tan, M.C., Patel, K., Yu, X., Wang, G., Zhang, F., Xu, H., Anandasabapathy, S., et al., 2016. A tablet-interfaced high-resolution microendoscope with automated image interpretation for realtime evaluation of esophageal squamous cell neoplasia. Gastrointest. Endosc. 84, 834–841.

Saini, S.K., Bansal, V., Kaur, R., Juneja, M., 2020. Colponet for automated cervical cancer screening using colposcopy images. Mach. Vis. Appl. 31, 1–15.

Sankaranarayanan, R., 2014. Screening for cancer in low-and middle-income countries. Ann. Glob. Health 80, 412–417.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional net535 works for large-scale image recognition. Tech. Rep.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. ArXiv Preprint ArXiv 1412, 6806.

Szegedy, C. , Vanhoucke, V. , Ioffe, S. , Shlens, J. , Wojna, Z. , 2016. Rethinking the inception architecture for computer vision. 2016 2818 2826.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning. PMLR 6105–6114.
Thekkek, N., Richards-Kortum, R., 2008. Optical imaging for cervical cancer detection: solutions for a continuing global problem. Nat. Rev. Cancer 8, 725–731.
Thung, K.H., Yap, P.T., Shen, D., 2017. Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 160–168.
Villa, L.L., 2012. Cervical cancer in latin america and the caribbean: the problem and the way to solutions. Cancer Epidemiology and Prevention. Biomarkers 21, 1409–1413.
William, W., Ware, A., Basaza-Ejiri, A.H., Obungoloch, J., 2018. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Comput. Methods Prog. Biomed. 164, 15–22.
Wilson, M.L., Fleming, K.A., Kuti, M.A., Looi, L.M., Lago, N., Ru, K., 2018. Access to pathology and laboratory medicine services: a crucial gap. Lancet 391, 1927–1938.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. 2017 1492 1500.
Xu, T. , Zhang, H. , Huang, X. , Zhang, S. , Metaxas, D.N. , 2016. Multimodal Deep Learn. Cerv. dysplasia Diagn. 2016 115 123.
Yang, X., Zeng, Z., Yeo, S.Y., Tan, C., Tey, H.L., Su, Y., 2017. A novel multi-task deep learning model for skin lesion segmentation and classification. ArXiv 1703, 01025.
Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. ArXiv 1511, 07122.
Yuan, C., Yao, Y., Cheng, B., Cheng, Y., Li, Y., Li, Y., Liu, X., Cheng, X., Xie, X., Wu, J., et al., 2020. The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. Sci. Rep. 10, 1–12.
Zagoruyko, S. , Komodakis, N. , 2016. Wide residual Netw. arXiv Prepr. arXiv 1605 2016 07146.
Zhang, T., Luo, Ym, Li, P., Liu, Pz, Du, Yz, Sun, P., Dong, B., Xue, H., 2020. Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images. Biomed. Signal Process. Control 55, 101566.
Zhao, H. , Shi, J. , Qi, X. , Wang, X. , Jia, J. , 2017. Pyramid Scene parsing Netw., : Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2017 2881 2890.