

Computer-Aided Diagnosis of Label-Free 3-D Optical Coherence Microscopy Images of Human Cervical Tissue

Yutao Ma, *Member, IEEE*, Tao Xu, Xiaolei Huang, *Member, IEEE*, Xiaofang Wang, Canyu Li, Jason Jerwick, Yuan Ning, Xianxu Zeng, Baojin Wang, Yihong Wang, Zhan Zhang, Xiaohan Zhang, and Chao Zhou, *Senior Member, IEEE*

Abstract— Objective: Ultrahigh-resolution optical coherence microscopy (OCM) has recently demonstrated its potential for accurate diagnosis of human cervical diseases. One major challenge for clinical adoption, however, is the steep learning curve clinicians need to overcome to interpret OCM images. Developing an intelligent technique for computer-aided diagnosis (CADx) to accurately interpret OCM images will facilitate clinical adoption of the technology and improve patient care. **Methods:** 497 high-resolution 3-D OCM volumes (600 cross-sectional images each) were collected from 159 *ex vivo* specimens of 92 female patients. OCM image features were extracted using a convolutional neural network (CNN) model, concatenated with patient information (e.g., age and HPV results), and classified using a support vector machine classifier. Ten-fold cross-validations were utilized to test the performance of the CADx method in a five-class classification task and a binary classification task. **Results:** An $88.3 \pm 4.9\%$ classification accuracy was achieved for five fine-grained classes of cervical tissue, namely normal, ectropion, low-grade and high-grade squamous intraepithelial lesions (LSIL and HSIL), and cancer. In the binary classification task (low-risk [normal, ectropion and LSIL] vs. high-risk [HSIL and cancer]), the CADx method achieved an area-under-the-curve (AUC) value of 0.959 with an $86.7 \pm 11.4\%$ sensitivity and $93.5 \pm 3.8\%$ specificity. **Conclusion:** The proposed deep-learning based CADx method outperformed four human experts. It was also able to identify morphological characteristics in OCM images that were consistent with histopathological interpretations. **Significance:** Label-free OCM imaging, combined with deep-learning based CADx methods, hold a great promise to be used in clinical settings for the effective screening and diagnosis of cervical diseases.

Index Terms—Cervical cancer, optical coherence tomography, optical coherence microscopy, deep learning, computer-aided diagnosis.

This work was supported in part by the National Basic Research Program of China (Grant No. 2014CB340404), start-up fund from Lehigh University, US National Science Foundation (Grant No. DBI-1455613), US National Institutes of Health (Grant Nos. K99/R00-EB010071 and R01-EB025209), Medical Science and Technology projects of China (Grant Nos. 201503117 and 161100311100), and Henan Province Medical Technology Challenger Program (SBGJ2018050). Yutao Ma and Tao Xu contributed equally to this work. Corresponding author: Chao Zhou, e-mail: chaozhou@lehigh.edu.

Y. Ma and T. Xu are with the School of Computer Science, Wuhan University, China. X. Wang, C. Li, X. Zeng, B. Wang, Z. Zhang, and X. Zhang

I. INTRODUCTION

CERVICAL cancer is one of the most common cancers among women worldwide, especially in developing nations, and it has relatively high incidence and mortality rates [1]. Fortunately, cervical cancer is mostly preventable with active screening and detection techniques. For example, preventive screening and early detection can decrease the morbidity of cervical cancer by about 70% in the United States [2]. Nowadays, there are a few frequently-used cervical cancer screening techniques, such as high-risk human papillomavirus (HPV) testing, Pap smear cytology testing, colposcopy, and visual inspection of the cervix with acetic acid (VIA), each of which has its advantages and disadvantages.

Although HPV and Pap tests are widely used in women aged 25 and older to identify high-risk types of HPV that are most likely to cause cervical cancer [3] and abnormal cells, they cannot provide test results in real-time and are unable to localize cervical lesions. Instead, a VIA test allows clinicians to observe lesions and other changes in a patient's cervix directly, but it has lower sensitivity and specificity compared with HPV and Pap tests [4]. As the gold standard for diagnosing cervical cancer, colposcopy-directed biopsy with histopathological confirmation [5] is invasive and time-consuming and may cause complications to patients, such as bleeding, infection, and anxiety. Therefore, developing a non-invasive, efficient, and intelligent screening technique with relatively high sensitivity and specificity can significantly improve patient care.

Optical coherence tomography (OCT) [6] is an emerging biomedical imaging technique that utilizes light to obtain micrometer-resolution, cross-sectional images of biological tissue. By using high-resolution, high-speed OCT systems that can image cellular features of tissue samples up to 2 mm in depth in real-time [7], OCT has shown great potential as a non-

are with the Third Affiliated Hospital of Zhengzhou University, China. T. Xu, X. Wang, C. Li, Y. Ning, and X. Zeng were visiting students/scholars in the Department of Electrical and Computer Engineering, Lehigh University, USA. X. Huang is with the College of Information Sciences and Technology, Penn State University, USA. Y. Wang is with the Department of Pathology and Laboratory Medicine, Rhode Island Hospital/Warren Alpert Medical School of Brown University, USA. J. Jerwick and C. Zhou are with the Department of Bioengineering, Department of Electrical and Computer Engineering, Lehigh University, USA.

invasive “optical biopsy” method [8], [9]. Previous studies have demonstrated the feasibility of using OCT for identification of morphological characteristics of the cervix, such as squamous epithelium, basement membrane, cervical stroma, low-grade and high-grade squamous intraepithelial lesions (LSIL and HSIL), and cervical cancers [10]–[15], which makes it possible to use OCT as a diagnostic tool adjunctive to colposcopy for cervical disease screening and detection [16].

Optical coherence microscopy (OCM) [17], known as a combination of the coherent detection method of OCT and confocal microscopy, can provide better axial and lateral resolution than OCT. Moreover, ultrahigh-resolution OCM has recently shown the ability to reveal details of *ex vivo* cervical tissue similar to histology at the cellular level [18], significantly improving the diagnostic accuracy for cervical diseases. For example, in a recent blind test on 297 3-D OCM volumes [18], three human experts achieved an average sensitivity of 80% (95% confidence interval, CI, 72%–86%) and an average specificity of 89% (95% CI, 84%–93%) for identifying high-risk lesions (including HSIL and invasive lesions) using label-free OCM images. The inter-observer agreement value of 0.627 suggests high diagnostic consistency among the three experts.

However, OCM images are foreign to gynecologists and pathologists due to the limited use of the OCM technology in clinics. Clinicians would need to undergo rigorous training, possibly involving viewing thousands of OCM images with different pathologies, to familiarize and recognize diagnostic features in OCM images. This training, however, can be time-consuming and may not be well-received by clinicians due to their busy clinical schedules. Since OCM images are different from those of traditional colposcopy and histological images, the training may create difficulty in clinical acceptance and adoption of this new technology for the screening and diagnosis of cervical diseases. Therefore, a computer-aided diagnosis (CADx) approach to effectively extract diagnostic imaging features and classify OCM images accurately would be highly desirable and can help facilitate broader adoption of the OCM imaging technology for clinical use.

In the past decade, deep learning [19] technologies utilizing deep neural networks have made remarkable progress in computer vision [20] and medical image analysis [21] for their capability to learn implicit or latent features from vast amounts of images and videos. More specifically, Convolutional Neural Networks (CNNs), a popular class of deep neural networks, have been widely used in image classification [22] and object detection [23]. Some recent studies suggest that deep CNNs can obtain results with accuracy comparable to and in some cases better than human experts for tasks such as image-based cancer (or rare disease) detection [24]–[27].

The purpose of this study is to develop a deep learning-based CADx method to evaluate cervical tissue samples using multi-modal feature information extracted from ultrahigh-resolution OCM imagery and routine medical exams such as the HPV test. We strive to accurately classify 3-D OCM images from *ex vivo* cervical specimens into “low-risk” and “high-risk” classes, which would pave the way for *in vivo*, real-time, and intelligent cervical diseases screening and diagnosis. Furthermore, we

TABLE I
STATISTICS OF THE EXPERIMENTAL DATASET

	Normal	Ectropion	LSIL	HSIL	Cancer	Total
Patients	44	25	11	13	9	92
Specimens	71	32	16	18	22	159
3-D volumes	197	79	28	55	138	497
2-D images	55,070	23,150	7,714	16,500	39,033	141,467

attempt to classify cervical lesions and provide their histopathological correlation with OCM imaging features in order to assist clinicians in understanding and interpreting OCM images.

II. MATERIALS AND METHODS

A. Data Collection

The experimental dataset used in this study contains 141,467 grayscale cervical tissue images collected from the Third Affiliated Hospital of Zhengzhou University, China, using a custom-developed ultrahigh-resolution OCM system [18]. Experimental protocol was approved by the Institutional Review Board of the Medical Faculty, Zhengzhou University. All the patients consented to the data collection. There were 159 fresh cervical specimens from 92 female patients obtained from colposcopic biopsy ($n = 79$), conization ($n = 26$), and hysterectomy ($n = 54$). For conization and hysterectomy specimens, cervical tissue was sectioned into multiple (up to 12) ~3-4 mm slices. Before imaging, the cervical specimens were rinsed with saline or fresh water and placed on a thin, wet sponge to keep the surface of the tissue flat and moist. OCM imaging was performed within ~1-2 hours of excision. On average, ~3.1 3-D OCM imaging scans were acquired per specimen. Each OCM scan covers ~800 μm x 800 μm area on the specimen. As described in detail in [18], each of the 497 3-D OCM volumes has a point-specific histology-confirmed diagnosis. Each OCM volume was annotated with a unique class label (see Section II.B). Table I describes the statistics of the experimental dataset. Note that there were 21 squamous cell carcinoma (SCC) specimens with 134 3-D OCM volumes and only one adenocarcinoma specimen with four 3-D OCM volumes. In addition to 3-D OCM volumes, patient information such as age and HPV test results were also collected, which have been clinically proved to be helpful when making a diagnosis for cervical lesions.

B. Taxonomy of OCM Image Labels

The taxonomy of cervical OCM image labels consist mainly of five fine-grained classes: normal, ectropion, LSIL (CIN1), HSIL (CIN2&3), and cancer (including SCC and adenocarcinoma). This taxonomy takes cervical ectropion (or cervical eversion) into consideration since it is often indistinguishable from early cervical cancer. However, cervical ectropion is a non-cancerous condition that occurs when the endocervix turns outward, exposing the columnar epithelium to the vaginal milieu [28]. As with our previous work [18], we also adopt two distinct general classes, i.e., “low risk” and “high risk.” The general class “low risk” includes normal, ectropion,

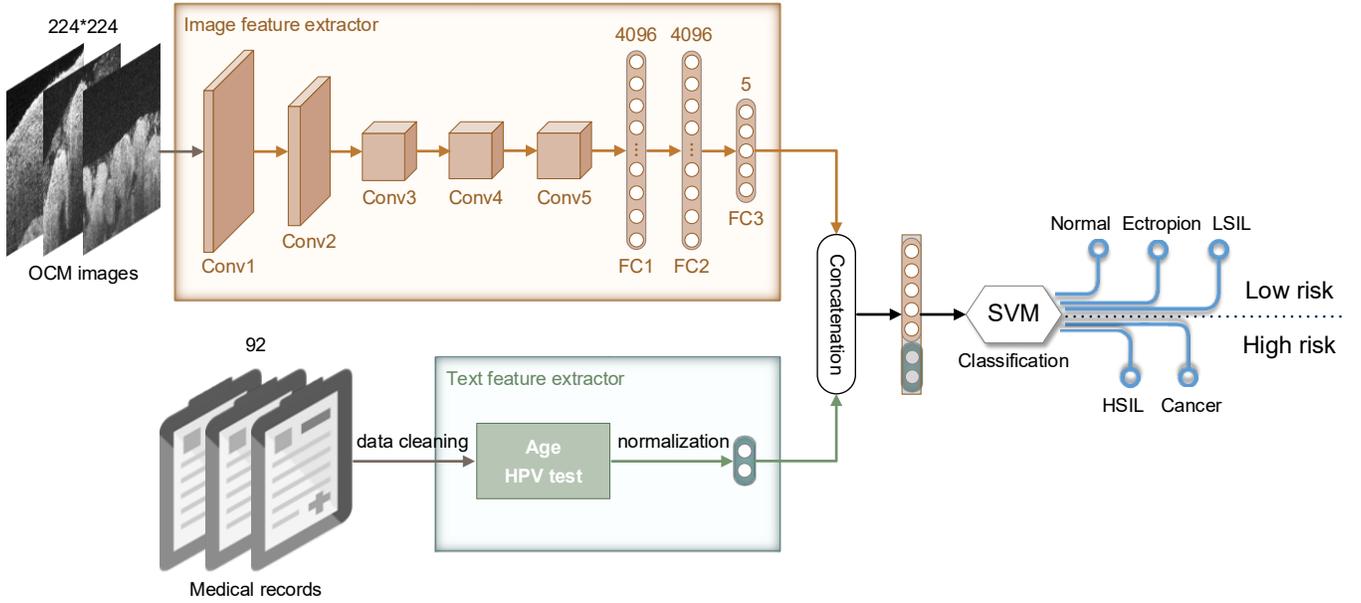


Fig. 1. The overall architecture of our CADx approach. First, we use VGG-16 to train an image feature extractor for ultrahigh-resolution OCM images. In particular, we append a new fully-connected layer (FC3) to the second fully-connected (FC2) layer. The FC3 layer outputs a 5-D feature vector representing the input image. Meanwhile, we build a text feature extractor to process patient information in medical records. Note that only age and HPV test results are available for this study. Second, we take advantage of a 7-D feature vector that concatenates the image and text features obtained to train an SVM-based classifier. Third, for a given OCM image, the SVM-based classifier outputs a predicted label of the five fine-grained classes. Besides, we evaluate the CADx method over two general classes, “low risk” (normal, ectropion, and LSIL) and “high risk” (HSIL and cancer), by inferring the probabilities of the corresponding fine-grained classes. The CADx method makes a decision of classification for each 3-D OCM image volume according to the mechanism of voting based on the majority rule.

and LSIL, while the “high risk” class includes HSIL and cancer.

C. Data Preparation

Each 3-D OCM image volume (~300 megabytes) contained a total of 600 2-D cross-sectional frames, each of which has 901 x 600 (height x width) pixels. The quality of OCM images played a vital role in training deep learning-based classification models (also called classifiers). We removed those 2-D cross-sectional frames that appeared to be “saturated” (i.e. too bright), too dark, or blurry from the original 3-D OCM volumes. According to Table I, the experimental dataset retained 141,467 high-quality cross-sectional 2-D OCM images. A 3 x 3 median filter was used to remove speckle noise from each 2-D OCM image. A single center crop with the size of 600 x 600 pixels was applied to each input image, and then the image was resized to 224 x 224 pixels while maintaining the original aspect ratio. Each 3-D OCM volume was zero-centered by subtracting the average intensity value from all the 2-D cross-sectional images within the volume.

Other patient data that are not images were processed using a text feature extractor. The patient age was converted from text and normalized using a min-max scaling, defined as follows.

$$x^* = \frac{x - \min}{\max - \min}, \quad (1)$$

where x denotes a patient’s age, \min and \max represent the minimum value and the maximum value, respectively, of all patients’ ages. Besides, HPV test results were defined by a Boolean datatype in this study. More specifically, “0” stands for a negative result, “1” for a positive result.

D. Image Feature Extraction

We used VGG-16 (Visual Geometry Group 16-layer) [29],

one of the most common CNNs, to train our image feature extractor. VGG-16 has 16 layers and it reduces the number of parameters in such a deep network using small (3 x 3) filters in convolutional layers. We modified the input image dimensions (224 x 224 x 3) for VGG-16 to ensure that the image feature extractor can process grayscale images (224 x 224 x 1). Also, instead of having the last fully-connected, softmax layer of VGG-16 output a vector of 1,000 categories, we replaced that with a fully connected layer that outputs a vector of 5 categories (see FC3 in Fig. 1) to make it suitable for the specific five-class classification task based on the taxonomy defined in Section II.B. Another reason for adding the FC3 layer following the 4,096-D fully-connected layer (see FC2 in Fig. 1) was that our approach could achieve a better balance between the dimensionality of image features and the dimensionality of non-image features (i.e., age and HPV test results).

Because a few previous studies have reported that transfer learning using pre-trained models on the ImageNet dataset [30] is helpful to fine-tune CNN-based classifiers for medical grayscale images [31], [32], we also used pre-trained weights on a subset of the ImageNet dataset to fine-tune the image feature extractor for the sake of efficiency. In addition to the FC3 layer, we re-trained the first convolutional layer using OCM images and the last three convolutional layers (that is, conv5_1, conv5_2, and conv5_3) that capture task-specific features. All the hidden layers mentioned above were fine-tuned using the same global learning rate of 0.002. Moreover, we took advantage of the *Adam* (short for Adaptive Moment Estimation) optimization algorithm [33], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a decay of 0.0002.

We trained, validated, and tested the image feature extractor using Keras (<https://keras.io>) and Google’s TensorFlow

(<https://www.tensorflow.org>) deep learning framework on a computer equipped with Intel Core i7 7200HQ processor, 64 GB RAM (Random-access Memory), and an NVIDIA GeForce GTX 1080 GPU (Graphics Processing Unit). The operating system was 64-bit Microsoft Windows 10.

E. Classification Algorithm

Since support vector machines (SVMs) [34] can efficiently perform linear classification and non-linear classification tasks, they are helpful in text categorization and image classification. In this study, we developed an SVM-based classifier that utilized the multi-modal feature information (see Fig. 1). As stated above, we trained a CNN-based model to extract a 5-D feature vector from each OCM image. At the same time, we processed the age and HPV test results to obtain a 2-D non-image feature vector. The concatenation of the two types of features, which are 7-D feature vectors, are used as the input for the SVM-based classifier. The classifier outputs the probability of a given test sample (including an OCM image and its corresponding patient information) belonging to each of the five fine-grained classes. Note that the SVM-based classifier we utilized is based on an open-source tool called scikit-learn (<http://scikit-learn.org>), with default settings.

Recall that each 3-D OCM volume contains many 2-D cross sectional images, thus in order to obtain a probability distribution of a given 3-D OCM volume \mathbf{y} over the five fine-grained classes, we calculate the probability using the following equation.

$$P(\mathbf{y} = j) = \frac{1}{|\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} I(y_i = j), \quad (2)$$

where \mathbf{y} denotes the whole OCM volume consisting of a set of 2D cross-sectional images, $|\mathbf{y}|$ is the total number of 2-D images, j represents the class label ranging between 0 and 4, and y_i is the class label predicted by the SVM-based classifier for the i^{th} 2-D image included in \mathbf{y} . Here, $I(x)$ is an indicator function defined as below.

$$I(x) = \begin{cases} 1 & \text{if } x \text{ is TRUE,} \\ 0 & \text{if } x \text{ is FALSE.} \end{cases} \quad (3)$$

After calculating the probability that a given 3-D OCM volume belongs to a specific fine-grained class, we then infer the likelihood that it belongs to a general class by summing up the probabilities over all the subclasses of the general class.

$$P(m) = \sum_{n \in S(m)} P(n), \quad (4)$$

where m is a general class in the taxonomy mentioned above (i.e., “low risk” or “high risk”), n is a fine-grained class, and the function $S(m)$ returns all the subclasses of m . Note that, in both fine-grained five-class and general two-class classification tasks, the SVM-based classifier made decisions according to the mechanism of voting based on the majority rule.

We evaluated the performance of the CADx method using ten-fold cross-validation. More specifically, we partitioned the experimental dataset into ten subsets of equal size, one of which was retained as the validation data for testing the CADx method and the other nine subsets were used as training data. This cross-validation process was repeated ten times, with each of the ten

subsets used only once as the validation data, and the final evaluation was the average of the ten results. When partitioning experimental data into a training set and a test set for each iteration of the cross-validation process, all the OCM images from the same specimen were placed wholly into either the training set or the test set, to ensure that the two sets were mutually exclusive. This strategy helped prevent overfitting in the CADx method.

F. Blind Test for Human Experts

To compare the difference between human and machine in classification result, two experienced OCM researchers (Investigators 1 and 2) and two experienced pathologists (Investigators 3 and 4) from different institutes participated in this study. By using a three-step method including training, pre-testing, and blinded-testing [18], the four experts evaluated 297 3-D OCM volumes (i.e., a subset of the experimental dataset) separately and independently. Patient age and HPV results were provided to the human experts and taken into consideration when the diagnosis were made. For the three-step training and testing, each investigator carefully reviewed the same training dataset that comprised 100 3-D OCM volumes and corresponding hematoxylin and eosin (H&E) histological slides in digital form. Next, each investigator made a diagnosis for each of the 100 3-D OCM volumes in the pre-testing dataset. After the diagnosis was recorded, feedback was provided to each investigator to allow them to review correctly diagnosed and misdiagnosed samples in order to provide additional training. At last, the remaining 297 3-D OCM volumes were presented to each investigator to perform the final test. Details of the three-step training and testing method were reported in [18].

G. Evaluation Metrics

Accuracy (or *Trueness*) is a descriptor of systematic errors, also known as a measure of statistical bias. *Sensitivity* (also called the true positive rate) measures the proportion of actual positives that are identified as such correctly, while *specificity* (also called the true negative rate) measures the percentage of actual negatives identified as such correctly. The following equations formulate their respective definitions in a binary classification task.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}, \quad (5)$$

$$Sensitivity = \frac{TP}{TP+FN}, \quad (6)$$

$$Specificity = \frac{TN}{TN+FP}, \quad (7)$$

where TP , TN , FP , and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

We used the estimated probability and actual class label to calculate the sensitivity and specificity of “low-risk” and “high-risk” diagnosis. In this study, a “low-risk” diagnosis includes

TABLE II
PERFORMANCE COMPARISON BETWEEN OUR CADX METHOD AND FOUR HUMAN EXPERTS

	Accuracy		Sensitivity			Specificity	
	Five classes	L/H	L/H	CI	L/H	CI	
Investigator 1	0.808	0.889	0.922	0.857~0.964	0.868	0.810~0.914	
Investigator 2	0.832	0.906	0.878	0.804~0.932	0.923	0.874~0.957	
Investigator 3	0.721	0.828	0.652	0.558~0.739	0.940	0.894~0.969	
Investigator 4	0.700	0.845	0.817	0.735~0.883	0.863	0.804~0.909	
Average (95% CI)	0.765	0.867	0.817	0.739~0.880	0.899	0.846~0.937	
CADx method	0.883(± 0.049)	0.913(± 0.051)	0.867(± 0.114)	/	0.935(± 0.038)	/	

Five classes: normal, ectropion, LSIL, HSIL, and Cancer; L/H: low risk and high risk; CI: confidence interval. Confidence intervals for sensitivity and specificity are “exact” Clopper-Pearson confidence intervals [36] at the 95% confidence level. The classification results of the first, third, and fourth investigators are cited from [18].

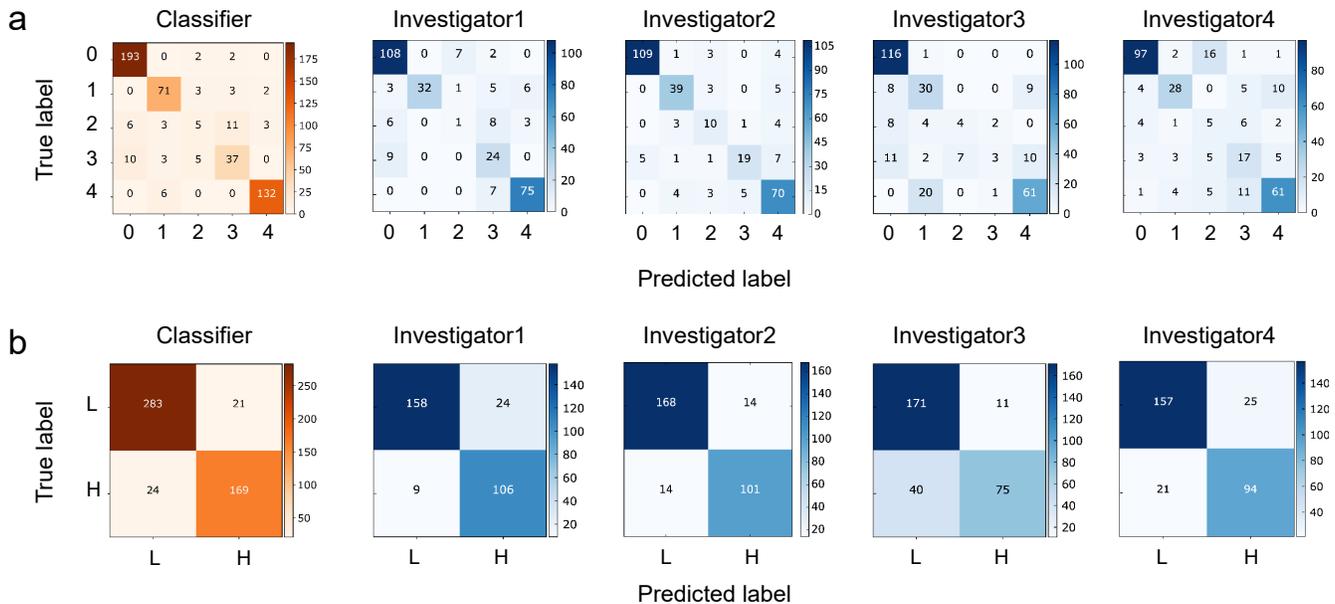


Fig. 2. Performance comparisons between the CADx method and four human experts. (a) Confusion matrices for the five-class classification task. “0,” “1,” “2,” “3,” and “4” represent normal, ectropion, LSIL, HSIL, and cancer, respectively. (b) Confusion matrices for the binary classification task. “L” and “H” represent low-risk and high-risk diagnosis, respectively.

normal, ectropion and LSIL, while a “high-risk” diagnosis includes HSIL and cancer. A 3-D OCM image is classified as “positive” if its estimated probability of being “high-risk” is larger than a given probability threshold t , and “negative” if the estimated probability is below the threshold. We then plotted the receiver operating characteristic (ROC) curve for the two general classes when varying the threshold value between 0 and 1. The area under the ROC curve (AUC) was also used to measure how well the CADx method performed in the binary classification task.

H. Classification Performance Visualization

A confusion matrix, also known as an error matrix [35], is used to visualize the performance of a classification algorithm. Each row of a confusion matrix denotes the instances (or class labels) of an actual class while each column indicates the cases of a predicted class, or vice versa. In this study, confusion matrices were used to show the misclassifications made by the CADx method and the four investigators on different (fine-grained or general) classes in two classification tasks. Each cell (i, j) in these confusion matrices represents the empirical probability of predicting class j given that the actual class label is i . The darker the color are off-diagonal elements, the higher the error rate. Therefore, a darker off-diagonal element

indicated that the CADx method or the human expert under discussion had trouble distinguishing between the two given classes.

I. Image Feature Visualization

We utilized two commonly-used feature maps to interpret how pixel-level image features “perceived” by the CNN-based image feature extractor differ among the five fine-grained classes. Springenberg *et al.* [37] proposed a method of guided backpropagation (GB). In this method, all the neurons (or called nodes in an artificial neural network) act like detectors of particular image features, thus explicitly visualizing diagnostic OCM image features for different classes. We also used a saliency map [38] to highlight visually dominant pixels based on the unique quality of each pixel, such as gray level intensity and image texture, etc.

III. RESULTS

A. Comparison between Human Experts and Machine on Five-Class classification

As shown in Table II, in the five-class classification task, the CADx method achieved $88.3 \pm 4.9\%$ (mean \pm s.d.) overall accuracy on the whole experimental dataset, while the four

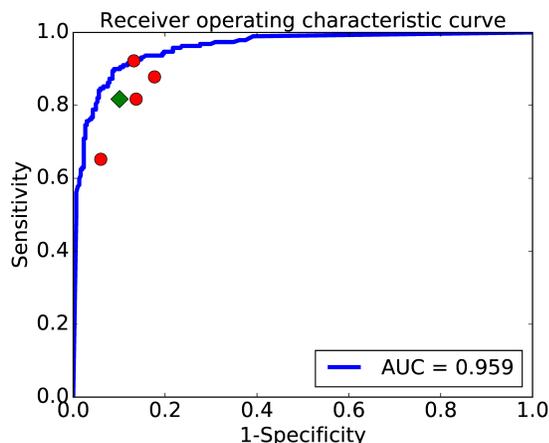


Fig. 3. ROC curve for the binary classification task. The X-axis stands for the false positive rate ($1 - \text{specificity}$), while the Y-axis represents the true positive rate (sensitivity). The (blue) ROC curve indicates the performance of the CADx method, each solid (red) circle denotes the performance of a human expert, and the (green) rhombus represents the average level of the four human experts.

human experts obtained 80.8%, 83.2%, 72.1%, and 70.0% accuracy, respectively, on a subset of the experimental dataset, including 297 3-D OCM volumes.

In general, the four human experts and the machine learning classifier can differentiate normal cervix (see class 0 in Fig. 2a) from the other four cervical diseases. The CADx method only had a classification error rate of 2.0% for normal cervical tissue, comparable to that of the best human investigator (that was, 0.9%). Besides, it has less misclassification error in cervical cancer (see class 4 in Fig. 2a) than the four human experts. The classification error rates of the CADx method and the best human expert were 4.3% and 8.5%, respectively on cancer detection. Although cervical ectropion (see class 1 in Fig. 2a) is not an abnormality, the four human experts each misclassified a few 3-D OCM volumes of this class into LSIL (see class 2 in Fig. 2a), HSIL (see class 3 in Fig. 2a), or cancer, with classification error rates of 31.9%, 17.0%, 36.2%, and 40.4%, respectively. Instead, the CADx method’s classification error rate for ectropion was only 10.1%, much less than that of human experts. Because $\sim 21\%$ of misclassifications made by the four human experts occurred between ectropion and cancer, the above result indicated that the CADx method had a greater ability to distinguish between these two classes’ irregular features in OCM images.

B. Comparison between Human Experts and Machine on Binary Classification

Fig. 2b presents the confusion matrices of the CADx method and the four human experts over the two general classes, showing that they can identify low-risk and high-risk 3-D OCM volumes with low misclassification rate. For example, the classification error rate of the CADx method for “low risk” was 6.9%, which was very close to that of the best human expert (that was, 6.0%). As shown in Table II, the CADx method achieved $91.3 \pm 5.1\%$ overall accuracy, while the four human experts obtained 88.9%, 90.6%, 82.8%, and 84.5% accuracy, respectively, in the binary classification task. The inter-observer agreement for the four investigators, characterized by

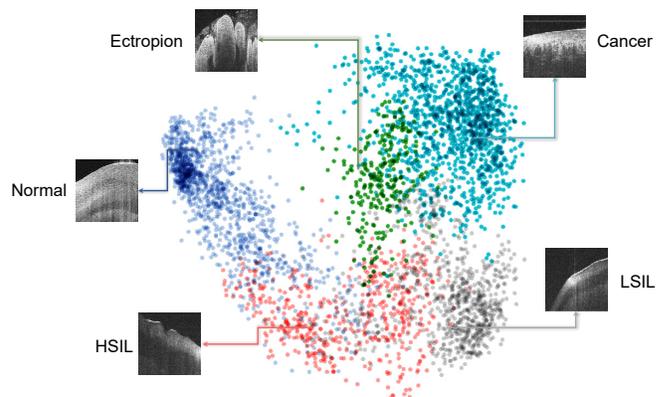


Fig. 4. Visualization of high-level representations of OCM images for the five fine-grained classes. Here, we illustrate 4,096-D feature vectors learned by the CNN-based image feature extractor for 2,500 randomly-selected OCM images in a Cartesian plane using PCA, a commonly-used dimension reduction method. Each colored point cloud represents a fine-grained class. Insets show images corresponding to various points in the point clouds. PCA embeds high-dimensional features in a low-dimensional space while preserving the pairwise distances of all the points that belong to different classes.

a Fleiss’ kappa [39] value of 0.633, suggested a substantial agreement. However, the sensitivity values varied from person to person. This may be due to the differences among individual experts in skills, experience, and working state. For example, Fig. 2b shows that the first expert (an OCM researcher) appeared to be good at discerning high-risk 3-D OCM images and tended to misclassify those uncertain test samples into “high risk,” and the third expert (a pathologist) appeared to be quite the opposite. Consequently, the former had a high false positive rate of 13.2%, and the latter had a high false negative rate of 34.8%.

Fig. 3 depicts a ROC curve of the CADx method in the binary classification task. It was evident from Fig. 3 that the CADx method outperformed the four human experts who were denoted by solid (red) circles lying below the (blue) ROC curve in this binary classification task. As a result, the (green) rhombus, which represents the average of the four human experts, was also below the ROC curve. The AUC value of the SVM-based classifier reached 0.959, with a sensitivity of $86.7 \pm 11.4\%$ and a specificity of $93.5 \pm 3.8\%$, indicating that it performed well overall and could provide diagnoses like experts.

Furthermore, unsurprisingly, the CADx method was far more efficient than human experts in both the binary and multi-class classification tasks. It took the CADx method ~ 4.8 milliseconds on average to complete diagnosis of a cross-sectional OCM image, and ~ 3 seconds to make a diagnosis of a 3-D OCM volume containing 600 2-D cross-sectional images. In contrast, it took a human expert several minutes to finish viewing a 3-D OCM volume and record a diagnosis.

C. Visualization of High-Level Representations of OCM Images

As shown in Fig. 1, the FC2 layer of the CNN-based image feature extractor outputs a 4,096-D feature vector for each input

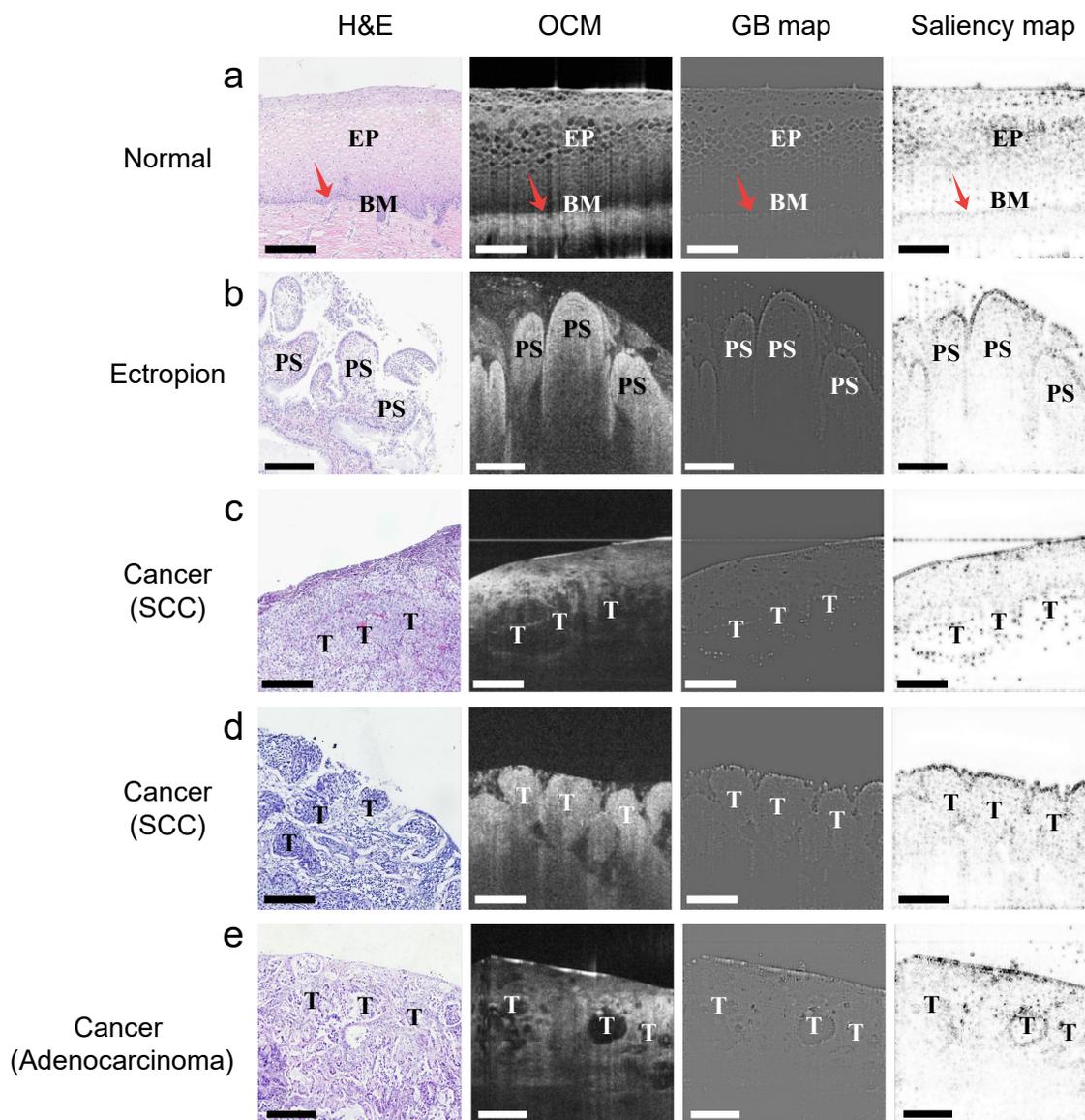


Fig. 5. Visualization of pixel-level morphological characteristics in OCM images extracted by our CADx method for three types of cervical tissue. (a) Normal tissue. (b) Cervical ectropion. (c&d) Cervical cancer: squamous cell carcinoma (SCC). (e) Cervical cancer: adenocarcinoma. The four panels in each row correspond to the H&E histologic section, OCM image, GB map, and saliency map for each specimen, respectively. GB and saliency maps highlight pixel-level morphological representations learned by the CNN-based image feature extractor. EP: squamous epithelium; BM: the basal membrane; PS: the papillary structure with interpapillary ridges; T: Tumor. Scale bars: 200 μ m.

OCM image. We visualized this high-dimension feature of OCM images using principal component analysis (PCA) [40] in Fig. 4. Colored point clouds denote the five fine-grained classes, showing how the CADx method groups 2,500 randomly-selected OCM images into different classes (or clusters). OCM data in the same class are represented with the same color and clustered close to each other than to data from other classes. A 3-D view of the data clouds is shown in Video S1.

Generally speaking, the high-level representations obtained were semantic and had high intra-class similarity and apparent inter-class difference, which indeed facilitated the accurate classification of OCM images. Take the classification of cervical ectropion and cervical cancer samples as an example. Although it was not an easy task for the four investigators to differentiate cervical ectropion from cervical cancer in OCM

image features (see the result introduced in Section III.A), the CADx method was able to detect distinct differences in the high-level feature space, with some minor overlaps between the two clusters (see Fig. 4). Thus, the CADx method only misclassified 2.5% of cervical ectropion samples as cervical cancer and grouped 4.3% cervical cancer samples into cervical ectropion incorrectly. This result was much better compared to those of the four human experts.

D. Visualization of Pixel-Level Morphological Features in OCM Images

Since deep learning is often considered as a “black box” [41], another major challenge in the development of the CADx method was to extract feature representations from OCM images and associate them with established morphological

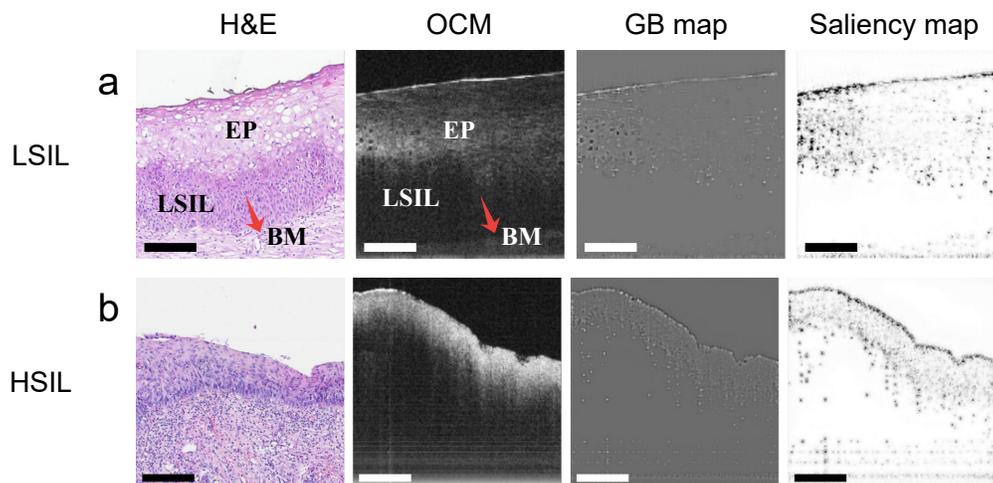


Fig. 6. Two confusing examples that were misclassified by the CADx method. (a) LSIL. (b) HSIL. The four panels in each row correspond to the H&E histologic section, OCM image, GB map, and saliency map for each specimen, respectively. EP: squamous epithelium; BM: the basal membrane. Scale bars: 200 μm .

characteristics in histology. In Fig. 5, we utilized GB and saliency maps to visualize and highlight characteristic OCM image features for normal cervical tissue, cervical ectropion, and cervical cancers. In addition to the original OCM image, we also presented the corresponding H&E histologic section to correlate features observed from OCM images.

As shown in Fig. 5a, the OCM image of a normal cervical tissue sample showed a layered architecture of stratified squamous cells (EP) and the stroma, separated by the basement membrane (BM). These features matched well with the corresponding H&E image. Both the GB map and the saliency map highlighted the mesh-like epithelial cells and the layered structure with a clear and smooth interface between the basal layer of the epithelium and the stromal layer, which was one of the most striking morphological characteristics recognized in previous studies [11], [13], [18].

The OCM image in Fig. 5b demonstrated an example of cervical ectropion. The layered architecture of the epithelium was lost, and papillary structures with hyper-scattering boundaries were formed. The hyper-scattering boundaries of the papillary structures and interpapillary ridges were visualized in the OCM image as well as in the GB and saliency maps.

Fig. 5c and 5d present two examples of invasive cervical cancer (more specifically, SCC). In Fig. 5c, the GB and saliency maps captured a common morphological characteristic of SCC. The epithelium became unstructured and disorganized, and the basement membrane was no longer observed, thus leading to a complete lack of architectural polarity. Besides, the two maps identified some sheets/nests of heterogeneous regions composed of epithelial cells and tumor cells. In Fig. 5d, the GB and saliency maps highlighted another common diagnostic feature of SCC, that is, the microstructure of the cervical tumor sample disappeared, and oval-shaped clusters of homogeneous regions composed of nests of tumor cells were observed.

Although cervical adenocarcinoma is much less common than SCC, its incidence has been increasing in the past two to three decades. Fig. 5e presents an example of adenocarcinoma featuring destructive stromal invasion. Glandular structures

with different sized lumens and intraluminal infoldings were observed in the H&E image in Fig. 5e. These glandular lumens, which scattered in a disorderly manner, displayed a clear hypo-scattering feature in the corresponding OCM image and thus looked like dense, round to oval cavities. The GB and saliency maps highlighted a characteristic vacuole appearance with a smooth contour, which is a marked morphological feature of adenocarcinoma.

IV. DISCUSSION

In this paper, we applied a deep-learning based CADx method to diagnose cervical diseases based on label-free and non-destructive OCM images. The CADx method was trained to combine multi-modal feature information extracted from OCM imagery and patient information such as age and HPV test results to make a diagnosis. Moreover, it worked similarly to pathologists who often make a diagnosis based on the histology slides of biopsy specimens, taking into consideration of patient information. Since it usually takes two or three days to get a pathology report, we argue that the proposed OCM imaging and CADx method may simplify the workflow for cervical disease screening and generate diagnostic reports in a timely fashion.

Previously, Kang *et al.* developed a CADx algorithm to classify OCT images of the human cervix using a linear discriminant analysis [42]. Nevertheless, it depended heavily on handcrafted image features, for example, the thickness of the epithelium and the contrast between the epithelium and the stroma, leading to a low average sensitivity of 51% (95% CI, 36%–67%) on a small-scale dataset including 152 cross-sectional OCT images. In this study, we validated the effectiveness of the CADx method in two classification tasks using ten-fold cross-validation. Compared with four human experts, the CADx method achieved higher overall accuracy (> 88%) in both the two-class and five-class classification tasks. Besides, in the binary classification task, the CADx method achieved an AUC value of 0.959 with $86.7 \pm 11.4\%$ sensitivity and $93.5 \pm 3.8\%$ specificity, showing its impressive

performance in classifying 3-D OCM volumes of human cervical tissue. Compared with the traditional CADx algorithm [42], our method had a greater ability to effectively extract diagnostic imaging features from ultrahigh resolution OCM images.

The proposed CADx method still has room for improvement, especially in distinguishing between LSIL and HSIL. Even for expert pathologists, it was not an easy task to distinguish LSIL and HSIL from OCT images [12], [15]. Fig. 6 illustrates two examples misclassified by the CADx method. The low-grade lesion (see Fig. 6a) was incorrectly classified as HSIL, while the high-grade lesion (see Fig. 6b) was misclassified as LSIL. In Fig. 6a, the OCM image presented a layered architecture, including the superficial layer of the epithelium, a lesion of LSIL (at about the lower third of the epithelium), and the stroma. Moreover, the low-grade lesion displayed a hypo-scattering (poorer scattering) feature in the OCM image. According to [15], [18], one of the most common optical feature of HSIL lesions on OCM was that more than one-third of the epithelium became irregular and thicker (for example, the OCM image in Fig. 6b). Sometimes, such a cervical lesion may span about two-thirds of the epithelium, and the basement membrane may not be visible. Unfortunately, the saliency and GB maps in Fig. 6 showed that the CNN-based image feature extractor, which was trained with limited samples of LSIL and HSIL, failed to capture the morphological features mentioned above that are meaningful to human experts. As a result, the CADx method produced similar probabilities for LSIL and HSIL. Hence, our CADx method had a limited ability to differentiate HSIL from LSIL (also see the overlap between the clusters of LSIL and HSIL in Fig. 4) at the moment due to limited samples (for example, only 28 LSIL and 55 HSIL 3-D OCM image volumes were included in the current study). Similarly, the diagnostic performance of the CADx method for adenocarcinoma remained to be determined due to limited OCM data sets available from adenocarcinoma specimens.

Our future work is to tackle the above problem from two aspects. On the one hand, we will collect more LSIL, HSIL and cancer (both SCC and adenocarcinoma) specimens to train a better image feature extractor. Furthermore, we will take advantage of *en face* OCM images obtained from the 3-D OCM volume when training the image feature extractor. Our previous study has showed that cervical ectropion had very different image features compared to invasive lesions in an *en face* OCM image [18]. On the other hand, we plan to enhance the learning ability of the image feature extractor using deep reinforcement learning [43] in combination with human knowledge and skills as well as instant feedback.

The clinic niche for the CADx method is to objectively provide *in vivo* diagnosis of cervical diseases in real-time. To achieve this goal, we are developing an endoscopic OCM probe suitable for imaging the human cervix *in vivo*. Potential complications, such as motion artifacts, bleeding, and abnormal vaginal secretion, associated with *in vivo* experiments will be taken into consideration when design the hardware and software system. The CADx method will be further optimized in term of speed and accuracy and integrated with the OCM

image acquisition software for *in vivo* applications. In addition to tissue classification, an improved version of the CADx method will be able to rapidly flag suspicious disease regions on the cervix. This function will be helpful in guiding biopsies under colposcopy examination. Finally, since OCM is a new imaging technology for gynecologists and pathologists, clinical utilities of the OCM imaging and CADx methods for cervical disease diagnosis will need to be carefully evaluated through a multi-center clinical trial in order to facilitate the adoption of the new methods in clinical practice.

V. CONCLUSION

In summary, we developed a deep-learning based CADx method and applied it to diagnose cervical diseases based on 3-D OCM image volumes and patient information. The CADx method was shown to be effective in binary and multi-class classification tasks, demonstrating classification accuracies better than four human experts. Using guided backpropagation and saliency maps, we further identified morphological characteristics in OCM imagery to provide histopathological correlation of OCM image features. With the assistance of the CADx method, ultrahigh resolution OCM technology holds the potential to become a promising complement to existing technologies for non-invasive, label-free and real-time screening and diagnosis of human cervical diseases.

ACKNOWLEDGMENT

We thank Wei Zhang, Yiming Du, Di Meng, Xian Cui, Linyu Li, and Zhaoning Jiang for helpful discussions and assistance in collecting and imaging the specimens.

REFERENCES

- [1] C. Fitzmaurice *et al.*, "Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study," *JAMA Oncol.*, vol. 3, no. 4, pp. 524–548, 2017.
- [2] R. L. Siegel *et al.*, "Cancer Statistics, 2017," *CA: Cancer J. Clin.*, vol. 67, no. 1, pp. 7–30, 2017.
- [3] G. Ronco *et al.*, "Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials," *Lancet*, vol. 383, no. 9916, pp. 524–532, 2014.
- [4] V. Singh *et al.*, "Performance of a low cost magnifying device, magnivisualizer, versus colposcope for detection of pre-cancer and cancerous lesions of uterine cervix," *J. Gynecol. Oncol.*, vol. 25, no. 4, pp. 282–286, 2014.
- [5] D. Saslow *et al.*, "American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer," *CA: Cancer J. Clin.*, vol. 62, no. 3, pp. 147–172, 2012.
- [6] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.
- [7] S. A. Boppart *et al.*, "In vivo cellular optical coherence tomography imaging," *Nat. Med.*, vol. 4, no. 7, pp. 861–865, 1998.
- [8] J. G. Fujimoto *et al.*, "Optical biopsy and imaging using optical coherence tomography," *Nat. Med.*, vol. 1, no. 9, pp. 970–972, 1995.
- [9] G. J. Tearney *et al.*, "In vivo endoscopic optical biopsy with optical coherence tomography," *Science*, vol. 276, no. 5321, pp. 2037–2039, 1997.
- [10] A. M. Sergeev *et al.*, "In vivo endoscopic OCT imaging of precancer and cancer states of human mucosa," *Opt. Express*, vol. 1, no. 13, pp. 432–440, 1997.

- [11] P. F. Escobar *et al.*, “Diagnostic efficacy of optical coherence tomography in the management of preinvasive and invasive cancer of uterine cervix and vulva,” *Int. J. Gynecol. Cancer*, vol. 14, no. 3, pp. 470–474, 2004.
- [12] S.-W. Lee *et al.*, “Optical diagnosis of cervical intraepithelial neoplasia (CIN) using polarization-sensitive optical coherence tomography,” *Opt. Express*, vol. 16, no. 4, pp. 2709–2719, 2008.
- [13] J. Gallwas *et al.*, “Optical coherence tomography as a non-invasive imaging technique for preinvasive and invasive neoplasia of the uterine cervix,” *Ultrasound Obst. Gyn.*, vol. 36, no. 5, pp. 624–629, 2010.
- [14] R. Wessels *et al.*, “Optical biopsy of epithelial cancers by optical coherence tomography (OCT),” *Lasers Med. Sci.*, vol. 29, no. 3, pp. 1297–1305, 2014.
- [15] J. Gallwas *et al.*, “Detection of cervical intraepithelial neoplasia by using optical coherence tomography in combination with microscopy,” *J. Biomed. Opt.*, vol. 22, no. 1, 016013, 2017.
- [16] P. F. Escobar *et al.*, “Optical coherence tomography as a diagnostic aid to visual inspection and colposcopy for preinvasive and invasive cancer of the uterine cervix,” *Int. J. Gynecol. Cancer*, vol. 16, no. 5, pp. 1815–1822, 2006.
- [17] J. A. Izatt *et al.*, “Optical coherence microscopy in scattering media,” *Opt. Lett.*, vol. 19, no. 8, pp. 590–592, 1994.
- [18] X. Zeng *et al.*, “Ultrahigh-resolution optical coherence microscopy accurately classifies precancerous and cancerous human cervix free of labeling,” *Theranostics*, vol. 8, no. 11, pp. 3099–3110, 2018.
- [19] Y. LeCun *et al.*, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [20] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” *Neural Netw.*, vol. 61, pp. 85–117, 2015.
- [21] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [22] A. Krizhevsky *et al.*, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [23] R. B. Girshick *et al.*, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE CVPR*, 2014, pp. 580–587.
- [24] V. Gulshan *et al.*, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [25] E. Long *et al.*, “An artificial intelligence platform for the multihospital collaborative management of congenital cataracts,” *Nat. Biomed. Eng.*, vol. 1, 0024, 2017.
- [26] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [27] D. S. Kermany *et al.*, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.
- [28] P. M. Casey *et al.*, “Abnormal Cervical Appearance: What to Do, When to Worry?” *Mayo Clin. Proc.*, vol. 86, no. 2, pp. 147–151, 2011.
- [29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*, 1409.1556, 2014.
- [30] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [31] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: full training or fine tuning?” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, 2016.
- [32] M. Havaei *et al.*, “Brain tumor segmentation with Deep Neural Networks,” *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.
- [33] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv*, 1412.6980, 2014.
- [34] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [35] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” *Remote Sens Environ.*, vol. 62, no. 1, pp. 77–89, 1997.
- [36] A. Agresti and B. A. Coull, “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *Am. Stat.*, vol. 52, no. 2, pp. 119–126, 1998.
- [37] J. T. Springenberg *et al.*, “Striving for simplicity: The all convolutional net,” *arXiv*, 1412.6806, 2014.
- [38] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proc. ACM MM*, 2006, pp. 815–824.
- [39] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.
- [40] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.
- [41] D. Golovin *et al.*, “Google Vizier: A Service for Black-Box Optimization,” in *Proc. ACM KDD*, 2017, pp. 1487–1495.
- [42] W. Kang *et al.*, “Diagnostic efficacy of computer extracted image features in optical coherence tomography of the precancerous cervix,” *Med. Phys.*, vol. 38, no. 1, pp. 107–113, 2011.
- [43] V. Mnih *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.