Original Article

DeepCIN: Attention-Based Cervical histology Image Classification with Sequential Feature Modeling for Pathologist-Level Accuracy

Sudhir Sornapudi[1], R. Joe Stanley[1], William V. Stoecker[2], Rodney Long[3], Zhiyun Xue[3], Rosemary Zuna[4], Shellaine R. Frazier[5], Sameer Antani[3]

[1]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO, USA, [2]Stoecker and Associates, Rolla, MO, USA, [3]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA, [4]Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA, [5]Department of Surgical Pathology, University of Missouri Hospitals and Clinics, Columbia, MO, USA

Submitted: 16-Jun-2020 Revised: 02-Sep-2020 Accepted: 21-Oct-2020 Published: 24-Dec-2020

Abstract

Background: Cervical cancer is one of the deadliest cancers affecting women globally. Cervical intraepithelial neoplasia (CIN) assessment using histopathological examination of cervical biopsy slides is subject to interobserver variability. Automated processing of digitized histopathology slides has the potential for more accurate classification for CIN grades from normal to increasing grades of pre-malignancy: CIN1, CIN2, and CIN3. **Methodology:** Cervix disease is generally understood to progress from the bottom (basement membrane) to the top of the epithelium. To model this relationship of disease severity to spatial distribution of abnormalities, we propose a network pipeline, DeepCIN, to analyze high-resolution epithelium images (manually extracted from whole-slide images) hierarchically by focusing on localized vertical regions and fusing this local information for determining Normal/CIN classification. The pipeline contains two classifier networks: (1) a cross-sectional, vertical segment-level sequence generator is trained using weak supervision to generate feature sequences from the vertical segments to preserve the bottom-to-top feature relationships in the epithelium image data and (2) an attention-based fusion network image-level classifier predicting the final CIN grade by merging vertical segment sequences. **Results:** The model produces the CIN classification results and also determines the vertical segment contributions to CIN grade prediction. **Conclusion:** Experiments show that DeepCIN achieves pathologist-level CIN classification accuracy.

Keywords: Attention networks, cervical cancer, cervical intraepithelial neoplasia, classification, convolutional neural networks, digital pathology, fusion-based classification, histology, recurrent neural networks

Introduction

Cervical cancer prevention remains a big global challenge. It is estimated that in 2020 in the US, 13,800 women will be diagnosed with invasive cervical cancer, and among them, 4290 will die.[1] This cancer ranks second in fatalities among 20–39-year-old women.[1] Screening has helped to decrease the incidence rate of cervical cancer by more than half since the mid-1970s through early detection of precancerous cells,[2] yet 300,000 women die every year worldwide.[3] As a public health priority in 2018, the WHO director general made a global call for the elimination of cervical cancer.[4]

If clinically indicated, the cervix is further examined by taking a sample of cervical tissue (biopsy). The tissue sample is transferred to a glass slide and observed under magnification (histopathology). Cervical dysplasia or cervical intraepithelial neoplasia (CIN) is the growth of abnormal cervical cells in the epithelium that can potentially lead to cervical cancer. CIN is usually graded on a 1–3 scale. CIN

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_50_20

Address for correspondence: Dr. R. Joe Stanley, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, 127 Emerson Electric Co. Hall, 301 W. 16th Street, Rolla, MO 65409-0040, USA.
E-mail: stanleyj@mst.edu

How to cite this article: Sornapudi S, Stanley RJ, Stoecker WV, Long R, Xue Z, Zuna R, *et al*. DeepCIN: Attention-based cervical histology image classification with sequential feature modeling for pathologist-level accuracy. J Pathol Inform 2020;11:40.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2020/11/1/40/304779

1 (Grade I) is mild epithelial dysplasia, confined to the inner one third of the epithelium. CIN 2 (Grade II) is moderate dysplasia, usually spread within the inner two-third of the epithelium. CIN 3 (Grade 3) is carcinoma *in situ* (severe dysplasia) involving the full thickness of the epithelium.[5] A diagnosis of Normal indicates the absence of CIN. Figure 1 depicts the localized regions with all four classes.

Our previous work on computational approaches for digital pathology image analysis has relied mostly on extraction of handcrafted features based on the domain expert's knowledge. Guo *et al*.[6] manually extracted traditional nuclei features for CIN grade classification. The images were split into ten equal vertical segments for extraction of local features and classified using voting fusion with support vector machine (SVM) and linear discriminant analysis (LDA). Huang *et al*.[7] used the LASSO algorithm for feature extraction with SVM ensemble learning for classification of cervical biopsy images. Automated CIN grade diagnosis was also performed through analyzing Gabor texture features with K-means clustering[8] and slide-level classification with texture features.[9] Kayser *et al*.[10] proposed a tool that can integrate the digital image content information with a system that understands the context for digitized tissue-based diagnosis. The classification accuracy with above mentioned approaches fell short of that needed for clinical or laboratory use. In the past decade, success of deep learning approaches for image segmentation and classification in the health domain has attracted more research.[11] Toward that, AlMubarak *et al*.[12] developed a fusion-based hybrid deep learning approach that combined manually extracted features and convolutional neural network (CNN) features to detect the CIN grade from histology images. Li *et al*.[13] proposed a transfer learning framework with the Inception-v3 network for classifying cervical cancer images. An excellent review of computer vision approaches for cervical histopathology image analysis was presented in Li *et al*.[14]

A critical problem with manual CIN grading by pathologists is the variability among general pathologists in CIN determination. Stoler *et al*.[15] found an agreement for the general community pathologist with the expert pathologist panel assignment to range from 38% to 68%: 38.2%, 38%, and 68% for CIN Grades 1, 2, and 3, respectively. The overall Cohen's kappa value (κ) was 0.46 for four grades, these three CIN grades and cervical carcinoma. Cai *et al*.[16] found close agreement among expert pathologists. For four expert pathologists, with 8–30 years of grading CIN slides, a weighted κ range of 0.799–0.887 was found. If automated CIN grading results can be made as close to expert readings as the variability among expert pathologist readings, automated CIN grading may become feasible.

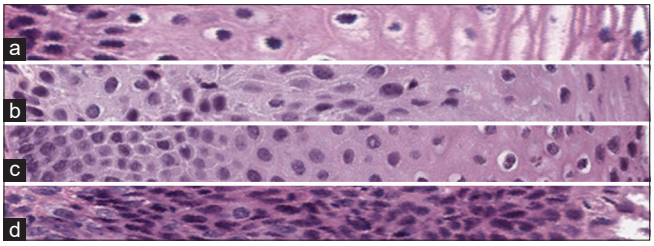

Figure 1: Sections of epithelium region with increasing cervical intraepithelial neoplasia severity (from [b-d]) showing delayed maturation with an increase in immature atypical cells from bottom to top. The sections can be categorized as (a) Normal, (b) CIN1, (c) CIN2, and (d) CIN3. In these images, left to right corresponds to bottom to top of the epithelium

Our proposed DeepCIN pipeline draws inspiration from the way pathologists examine epithelial regions under the microscope. They do not scan the entire slide at once; instead, they analyze local regions across the epithelium to understand the bottom-to-top growth of atypical cells and to compare the relative sizes of the cell nuclei in local neighborhoods. They use this local information to decide the CIN grade globally for the whole epithelium region. We developed a pathologist-inspired automated pipeline analogous to human study of histopathology slides, where we first localize the epithelial regions, then we analyze the features across these regions in both directions; finally, we fuse the feature information to predict the CIN class label and estimated the contribution of these local regions toward the global class result.

In this article, we present DeepCIN to automatically categorize high-resolution cervical histology images into Normal or one of the three CIN grades. Images used in this work are manually segmented epithelium regions extracted from digitized whole slide images (WSIs) at × 10 magnification. The classification is carried out through hierarchical analysis of local epithelial regions by focusing on individual vertical segments and then combining the localized feature information in spatial context by introducing recurrent neural networks (RNNs).

The use of RNNs[17,18] has been found to be successful in solving time-series and sequential prediction problems. Their use has led to a better understanding of contextual features from images when combined with CNN-based models. Typically, CNNs act as a feature extractor, and RNNs learn the contextual information. Shi *et al*.[19] proposed a convolutional RNN for scene text (sequence-to-sequence) recognition. Attention mechanisms[20] were incorporated later to improve performance.[21,22] Attention-based networks have been used in speech, natural language processing, statistical learning, and computer vision.[23]

A key aspect of our model is that it focuses on differentially informative vertical segment regions. This is crucial for deciding the level of CIN because the variation of CIN grade in the local region could impact the overall CIN assessment of the epithelium.[24] The major contributions of this article are:

1. Hierarchical image analysis from localized regions to the whole epithelium image
2. Capturing the varying nuclei density across the epithelium region by vertically splitting the region into standard width segments with reference to the medial axis
3. Weakly supervised training scheme for vertical segments
4. Image-to-sequence two-stage encoder model for extracting localized segment level information
5. Attention-based fusion (many-to-one model) for the whole epithelium image CIN classification

6. Identifying local segment contributions toward the whole image CIN classification.

Methodology

DeepCIN incorporates a two-fold learning process [Figure 2]. First, generated vertical segments from the epithelial image are fed to a two-stage encoder model for weak supervision training to constrain the segment class to the image class. Second, an attention-based fusion network is trained to learn the contextual feature information from the sequence of segments and classify the epithelial image into one of the four classes. The remainder of this section of the paper is organized as follows: Section II. A discusses cross-sectional vertical segment generation within an epithelium image; Section II. B and Section II. C present the two parts of the model: a segment-level sequence generator and an image-level classifier; Section II. D describes the model training approach.

Localization

Initially, we process the manually segmented epithelium regions to find the medial axis and reorient the epithelium to be aligned horizontally, as performed by Guo *et al*.[6] Guo's methods are modified to generate standard-width vertical segments with reference to the medial axis. This helps in better understanding the pattern of atypical cells under uniform epithelium sections and generating more image data for training our deep learning model. We approximate the medial axis curve as a piece-wise linear curve by iteratively drawing a series of circles (left to right) of radii equal to the desired segment width. The center of each successive circle is the right-most intersection point of the previously drawn circle and the medial axis curve. All the consecutive intersection points along the medial axis curve are joined to form a polygonal chain. At the midpoint of each line segment, we compute the slope corresponding to an intersecting perpendicular line. At the endpoints of the line segment, we draw vertical lines parallel to this midpoint perpendicular. This creates rectangular vertical regions of interest, as shown in Figure 3. Using these individual vertical regions, we compute a bounding box, which we apply to the original image to crop a refined vertical segment. The heights and counts of vertical segments created in this manner vary with the shapes and sizes of the epithelial images. The height and width of the segments are empirically chosen to be 704 pixels and 64 pixels, respectively (Section III. A). The RGB image segments are further processed by channel-wise normalizing the pixel intensities with 0 mean and standard deviation of value 1 and rotating counterclockwise by 90°. This facilitates the classification of localized epithelial regions.

Formally, we assume that an epithelial image I_{epth} has N vertical segments I_{vs_i} stacked up in a sequence by their spatial positioning from left to right such that

$$I_{\text{epth}} = \left\{ I_{vs_1}, I_{vs_2}, \ldots, I_{vs_N} \right\} \quad (1)$$

Segment-level sequence generation

The segment-level sequence generator network is built as a two-stage classifier model. The main objective of this network is to generate logit vectors to serve as localized sequence information for further image-level analysis. Since ground-truth labels for our vertical segments are not available, the network is trained against the image-level CIN grade. Since we expect variability in the true CIN grades across the vertical segments, use of the single image-level grade for all segments within an image introduces noisy labeling for the segments, and this may be expected to affect our training. Hence, we consider this a weakly supervised learning process.

We tackle this classification problem as a sequence recognition problem. As shown in Figure 4, the stage I is constructed with a CNN that can extract the convolutional feature maps. These spatial features are then reduced to have a height of 1 with maximum pooling operation. It is further transformed into a feature sequence by splitting along its width and concatenation of vectors formed by joining across the channels, similar to Shi *et al*.[19] The RNN acts as a stage II encoder model that further encodes the sequential information to predict the class value (many-to-one model). It is important to understand that the vertical segments carry valuable localized feature information, including varying nuclei density, which is crucial in the decision process. Therefore, it is well represented as a feature sequence and a bidirectional RNN focuses on the intrinsic details within these vertical segment regions from left to right and right to left.

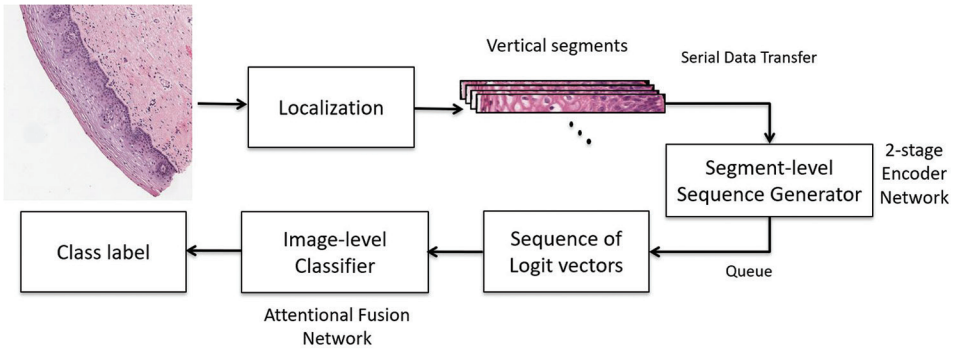

Figure 2: Overview of DeepCIN model

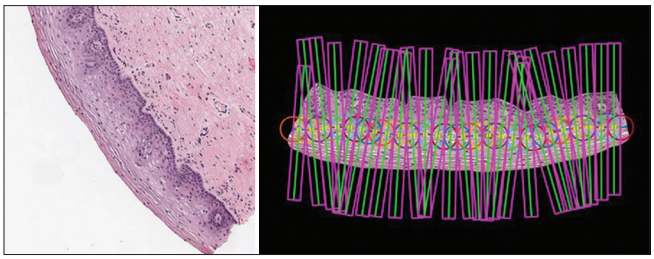

Figure 3: Localized vertical segment generation from an epithelial image

The architecture of the proposed segment-level sequence generator is given in Table 1. The stage I encoder is built with first 87 layers of the DenseNet-121 model.[25] A max-pooling layer is added to this last layer such that the feature map has the height of 1. This can be considered as a feature sequence generated from left to right. Note that the convolutions always operate on local regions and hence are translationally invariant. Hence, the pixels in the feature maps from left to right correspond to a local region in the original image (receptive field) from left to right, that is, the elements in the feature sequence are image descriptors in the same order. Importantly, they preserve the bottom-to-top spatial relationships in the original epithelium image. To further analyze this feature context, the generated feature sequence is fed to a stage II model built of RNNs. Specifically, we employed Bidirectional Long-Short-Term Memory (BLSTM)[26] networks to analyze and capture the long-term dependencies of the sequence from both the directions. For the stage II encoder, two sets of BLSTM and single-layer neural networks (NN) were appended to the last max-pooling layer of the stage I encoder. The final classification result is extracted from the logit vector of the last element in the output sequence generated at the stage II encoder. These logit vectors summarize the information of all the vertical segments and when combined, form an information sequence that is fused to determine the image-level CIN classification.

Assuming an epithelial image with N vertical segments I_{vsi}, we have created logit sequence vectors vs_i obtained with a segment-level sequence generator $f_s(;\theta)$:

$$vs_i = f_s\left(I_{vs_i};\theta\right) \tag{2}$$

where θ represents the model parameters.

Image-level classification

The image-level classifier network is designed as an attention-mechanism based fusion network, as shown in Figure 5. We aim to capture the dependencies among vertical segments with a gated recurrent unit (GRU).[18] The input sequences are picked up by GRU, which tracks the state of the sequences with a gating mechanism. The output is a sequence vector that represents the image under test. We use a small classifier with an attentional weight for each GRU cell output to encode the sequence of the vertical segments as:

$$h_i = GRU\left(vs_i; h_{i-1}\right) \tag{3}$$

where $i \in [1, N]$ and h_i is the hidden state that summarizes the information of the vertical segment I_{vsi}.

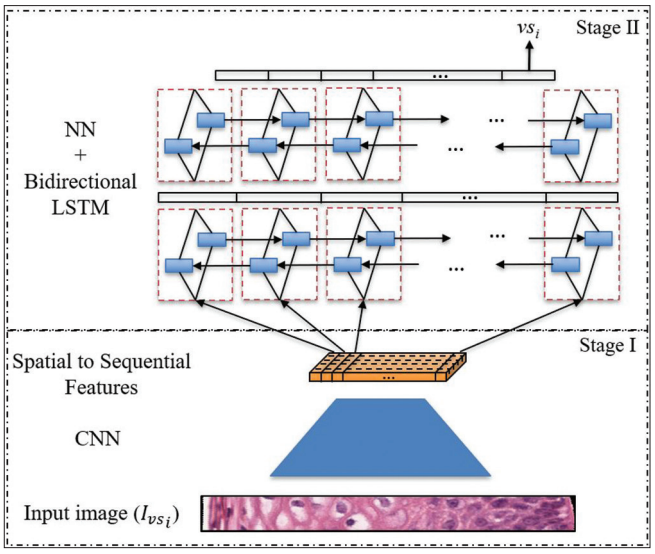

Figure 4: Segment-level sequence generator network with two-stage encoder structures

The vertical segments may not contribute equally to epithelial image classification. We use an attention mechanism with a randomly initialized segment-level context vectorapplied on the outputs of the GRU units that were subjected to tanh activated NN. This vector is used to generate the attentional weights which analyze the contextual information and give a measure of importance of the vertical segments. The following equations explain the employed attention mechanism:

$$e_i = w^T \tanh\left(W_{vs} h_i + b_{vs}\right) \tag{4}$$

$$\alpha_i = \frac{\exp\left(e_i\right)}{\sum_{i=1}^{N} \exp\left(e_i\right)} \tag{5}$$

$$v_I = \sum_{i=1}^{N} \alpha_i h_i \tag{6}$$

where W_{vs} and b_{vs} are trainable weights and bias. v_I is the image feature vector that summarizes all the information of vertical segments in an epithelial image. The image-level classification is determined by:

$$p_I = \text{softmax}\left(W_0 v_I + b_0\right) \tag{7}$$

Training

We trained the proposed networks independently with stratified K-fold cross-validation split at the image level. First, the segment-level sequence generator is trained to generate the logit vectors of all the segments and then concatenated to form a sequence to further train the image-level classifier.

During segment-level sequence generation, the problem of class imbalance is solved by upsampling the vertical segment images with image augmentations: randomly flipping vertically and horizontally, rotating with a range of 180°–180° angles, changing hue, saturation, value and contrast, and applying blur and noise. The objective is to minimize the cross-entropy loss (equation 8) calculated directly from the

Table 1: Segment-level sequence generator model architecture

	Layers	Configurations	Size
Stage I	Input	-	3×64×704
	Transition layer 0	*k*:7 × 7, *s*:2, p:3 *mp*:3 × 3, *s*:2, p:1	64×32×352 64×16×176
	Dense block 1	$\begin{bmatrix} k:1\times1, s:1, p:1 \\ k:3\times3, s:1, p:1 \end{bmatrix}\times6$	256×16×176
	Transition layer 1	$\begin{bmatrix} k:1\times1, s:1 \\ ap:2\times2, s:2 \end{bmatrix}$	128×8×88
	Dense block 2	$\begin{bmatrix} k:1\times1, s:1, p:1 \\ k:3\times3, s:1, p:1 \end{bmatrix}\times12$	512×8×88
	Transition layer 2	$\begin{bmatrix} k:1\times1, s:1 \\ ap:2\times2, s:2 \end{bmatrix}$	256×4×44
	Dense block 3	$\begin{bmatrix} k:1\times1, s:1, p:1 \\ k:3\times3, s:1, p:1 \end{bmatrix}\times24$	1024×4×44
	Pooling	*mp*:4 × 1, *s*:1	1024×1×44
Stage II	BLSTM+ NN	*nh*:256 *nh*:256	512×44 256×44
	BLSTM+ NN	*nh*:256 *nh*:4	512×44 4×44
	Output	-	4×1

k, *s*, *p*, *mp*, *ap*, and *nh*, are kernel, stride size, padding size, max pooling, average pooling, and number of hidden layers, respectively. "BLSTM" and "NN" stands for bidirectional LSTM and single-layer neural network, respectively. BLSTM: Bidirectional Long-Short-Term Memory, NN: Neural network, LSTM: Long-Short-Term Memory

vertical segment image and its restricted ground-truth label given by

$$L_k = -\Sigma_{vs} \log\left(\frac{\exp(y_k)}{\sum_j \exp(y_j)}\right) \quad (8)$$

where k is the class label of vertical segment image vs an y_k is the kth label element value in the logit vector. We use ADADELTA[27] for optimization since it automatically adapts the learning rates based on the gradient updates. The initial learning rate was set to 0.01.

For image-level classification, we use the weighted negative log-likelihood of correct labels to compute the cost function and back propagate the error to update the weights with a stochastic gradient descent optimizer (learning rate was fixed at 0.0001). Training loss is given by:

$$L'_k = -q_k \sum_I \log\left(p_{I_k}\right) \quad (9)$$

where k is the class label of epithelial image I and q_k is the weight of the label k.

Experiments

We conducted experiments on our cervical histopathology image database to evaluate the effectiveness of the proposed classification model and compared its performance with other state-of-the-art methods.

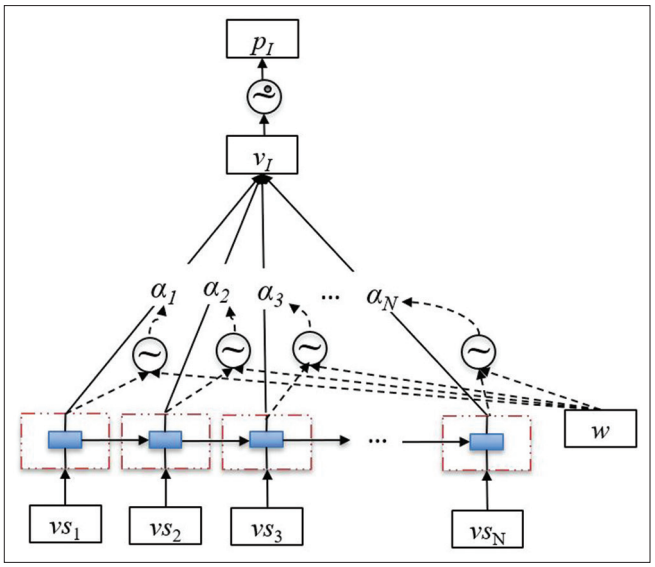

Figure 5: Attention-based fusion network for epithelial image-level classification. The input sequences are fed to GRU cells. ϴDenote a two-layer neural network with hyperbolic tangent and softmax activation functions, respectively to generate attentional weights. ϴDenotes a single layer NN with softmax activation function that produces the classification output

Dataset and evaluation metrics

For all the cross-validation experiments, we use a dataset that contains 453 high-resolution cervical epithelial images extracted from 146 hematoxylin and eosin-stained cervical histology WSIs. In addition, we use independent 224 high-resolution epithelium images as a hold-out test data. These WSIs were provided by the Department of Pathology at the University of Oklahoma Medical Center in collaboration with the National Library of Medicine. The WSIs were scanned at ×20 using Aperio ScanScope slide scanner in a pyramidal tiled format and saved with the file extension svs. Each pixel in the WSI has a size of 0.25 μm^2. The pyramidal tile level varies from 0 to 2/3/4. In this study, ×20 magnification images (pyramid level 0) downsampled to ×10 magnification are referred to as high-resolution images. All images have corresponding ground-truth labels. These annotations were carried out by an expert pathologist. The epithelial images have varying sizes which range from about 550 × 680 pixels (smallest) to 7500 × 1500 pixels (largest). This varying size affects the number of vertical segments generated from an image, typically ranging from 6 to 118. Although the vertical segments are generated such that the widths are 64 pixels wide and the height of these segments ranges from 160 to 1400 pixels, We address this problem by resizing the images to their median height: 704 pixels. This height was chosen empirically as a multiple of 32 to apply convolutions for feature extraction.

The segments were preprocessed such that they are RGB images of standard size: 64 × 704 × 3. We have created a total of 11,854 vertical segment images from 453 epithelial images.

The class distribution of these data is shown in Table 2. There are two main challenges with this epithelial image dataset. First, the cervical tissues have irregular epithelium regions, with color variations, intensity variations, red stain blobs, variations in nuclei shapes and sizes, and noise and blurring effects created during image acquisition. These effects tend to have large inter- and intraclass variability across the four classes we seek to label. Second, even though our database is labeled by experts and may be considered of high quality, it is relatively small. This is a common and recognized problem in the biomedical image processing domain.

The scoring metrics used for the performance evaluation are precision (P), recall (R), F1-score (F1), classification accuracy (ACC), area under the receiver operating characteristic curve, average precision, and Matthews correlation coefficient. Cohen's kappa score (κ) is used for the evaluation of the scoring schemes described in Section III. D. The percentage weighted average scores were reported due to the inevitable imbalance in the data distribution.

Implementation details

Although the entire DeepCIN model can be implemented end to end, we have split the process into two independent training steps. This model was chosen to overcome the GPU memory limitation to process these large input images and network architectures.

Details about the segment-level sequence generator network and image-level classifier network are given in Table 1 and Figure 4, respectively. Both the networks output four classes. The first network is trained with weak supervision to determine the logit sequence vectors of each vertical segment. The class outputs of the final network comprise our major concern.

A transfer learning technique was incorporated in the stage I encoder of the segment-level sequence generator. The convolution filters were initialized with ImageNet[28] pretrained weights and were left frozen since the stage I encoder is built with initial layers of the DenseNet-121 model, which presumably has weights already set to extract low-level image features such as edges, colors, and curves. All the CNN layers are activated with the rectified linear unit (ReLU) function, and the single layer NN, followed by BLSTM layers in the stage II encoder, which does not impose any nonlinearity to get logit vector sequence. The latter network consists of GRU cells (with 128 hidden units), a two-layer NN with hyperbolic tangent and softmax activation functions to generate attentional weights and a single-layer NN with softmax activation function to produce the classification output from the image feature vector.

Table 2: Class label distribution from 453 epithelial images

Class	Count (%)	
	Epithelial images	Segments
Normal	244 (53.8)	6836 (57.7)
CIN1	57 (12.6)	1433 (12.1)
CIN2	79 (17.5)	2039 (17.2)
CIN3	73 (16.1)	1546 (13.0)
Total	453 (100.0)	11,854 (100.0)

CIN: Cervical intraepithelial neoplasia

We trained and validated the models using stratified fivefold cross-validation. We split training and validation data at the image level and maintained the same distribution across both the models. To address the class imbalance problem, we have upscaled the less populated class images with image augmentations for the segment-level sequence generation and in the image level classification, we employed a weighted loss function.

Each individual fold for both the models was trained for 200 epochs with a batch size of 56 with early stopping to avoid overfitting.

We implemented our localized vertical segment generation in MATLAB[29] running on an Intel Xeon CPU @ 2.10GHz which took 3.42 s on average to process one epithelial image. The deep learning models are trained under CUDA 10.2 and CuDNN v7.6 backend on an NVIDIA Quadro P4000 8GB GPU and 64GB RAM with a PyTorch v1.4[30] framework. The time taken for validation is about 0.68 s per epithelial image. Thus, the entire DeepCIN pipeline takes 4.10 s on average to process and validate one epithelial image.

Ablation studies

In this section, we perform classifier ablation studies on the DeepCIN pipeline to understand its key aspects. The experiments include a comparison with different segment widths, stage I and stage II encoder variants, different fusion techniques, and benchmark models.

The proposed model takes standard size image inputs. Resizing images will cause image distortions. We observe that this has a minor effect on the performance, expected since both the training and testing images are similarly resized, which would result in the model's capability of handling such distortions. However, the segment width is to some extent a free variable whose setting may modulate the amount of local spatial information contained in a vertical segment. Recognizing this, we experimented with segment widths of 32, 64, and 128. According to Table 3, we observe that a segment width of 64 pixels is an optimal choice (in our experimental search space) compared to the segments with 32 pixels wide and 128 pixels wide.

The stage I encoder in the segment-level sequence generator acts like a spatial feature extractor. Since our biomedical digital image environment is not data rich for training deep learning models, we have experimented with various published models which have been pretrained with the benchmark ImageNet database. Only a set of initial layers that extract low-level features from the input image are considered in building the stage I encoder. The top-performing Stage I encoder model results were recorded, as shown in Table 4.

We observed that DenseNet-121 was better at extracting the crucial epithelial information, compared to ResNet-101[31] and Inception-v3[32] models. The DenseNet-121 model is better at feature reuse and feature propagation throughout the network with reduced parameters. Both DenseNet-121 and ResNet-101 are good at alleviating vanishing gradient problems; however, DenseNet-121 with its feed-forward interconnections among layers helps in better feature understanding. Inception-v3 uses models that are wider rather than deeper to prevent overfitting with factorizing convolutions to reduce the number of parameters without compromising network efficiency.

The stage II encoder further encodes the feature sequence that is mapped from the translationally invariant feature information available from the encoder. Our efforts to use bidirectional LSTM as a stage II encoder delivered better performance on the segment-level sequence generation that reflects on generating essential and better logit feature vectors. Table 5 shows that bidirectional analysis enables understanding of the context of the feature information; this aided in upsampling the segment data by flipping the input images horizontally. The use of attention was not helpful for understanding the feature sequence in the vertical segments with almost 1% decrease in performance across all the metrics [Table 5]. This indicates that the entire feature sequence is equally important to interpret the localized information, as shown by the equal distribution of attentional weights. The use of vanilla NNs (fully connected layers) was comparatively less efficient because LSTMs contain internal state cells that act as long-term and short-term memory units and manage to learn by remembering the important information and forgetting the unwanted. NNs lack this ability and focus only on the very last input.

We observed that attentional weights help to analyze the valuable information from the contribution of each segment towards the image-level classification. Table 6 confirms this observation, showing nearly a 2% improvement in performance with the inclusion of attention. Techniques like maximum voting and average voting of segment-level sequence generation results are simple and straight forward but fail to provide the additional information about the localized segment data.

Results

We finally compare the performance of the proposed model with the state-of-the-art CIN classification models. The models used for the comparison are proposed by Guo *et al*.[6] and AlMubarak *et al*.[12] The best model of Guo *et al*.,[6] LDA, was trained with 27 handcrafted features extracted from vertical image segments. The epithelium was split into ten equal parts to create these segments and fusion was performed through a voting scheme. AlMubarak *et al*.[12] used the same vertical segments and divided them into three sections: top, middle, and bottom. 64 × 64 size Lab color space image patches were extracted to train three CNN models.

Table 3: Ablation study on segment widths

Segment width	*P*	*R*	F1	ACC	AUC	AP	MCC
32	82.9	82.3	81.2	82.3	93.5	85.3	72.3
64*	88.6	88.5	88.0	88.5	96.5	91.5	82.0
128	85.3	85.6	84.9	85.6	95.9	89.8	77.1

P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under Receiver Operating Characteristic curve, ACC: Classification accuracy, *Indicates the best performing model

Table 4: Ablation study on stage I encoder models

Stage I encoder	*P*	*R*	F1	ACC	AUC	AP	MCC
DesnseNet-121*	88.6	88.5	88.0	88.5	96.5	91.5	82.0
ResNet-101	87.1	86.9	86.4	86.9	95.0	88.9	79.6
Inception-v3	85.5	85.4	85.1	85.4	94.8	87.8	77.1

P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under Receiver Operating Characteristic curve, ACC: Classification accuracy

Table 5: Ablation study on stage II encoder models

Stage II encoder	*P*	*R*	F1	ACC	AUC	AP	MCC
BLSTM*	88.6	88.5	88.0	88.5	96.5	91.5	82.0
BLSTM+attention	87.9	87.6	87.7	87.6	95.2	88.9	80.1
FC	85.3	85.0	84.2	85.0	94.7	87.4	76.3

BLSTM: Bidirectional Long-Short-Term Memory, P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under Receiver Operating Characteristic curve, ACC: Classification accuracy, FC: Fully-connected layer, * indicates the best performing model

Table 6: Ablation study on fusion techniques

Fusion	*P*	*R*	F1	ACC	AUC	AP	MCC
GRU	86.3	86.1	85.6	86.1	96.3	90.4	78.0
GRU+attention*	88.6	88.5	88.0	88.5	96.5	91.5	82.0
Max vote	87.6	87.2	87.0	87.2	-	-	79.9
Avg vote	88.0	87.6	87.4	87.6	-	-	80.6

GRU: Gated recurrent unit

The resulting confidence values from these sections were treated as features, and the 27 features were concatenated to form a hybrid approach for training an SVM classifier. The final classifiers of both these models were trained with a leave-one-out approach.

For a direct comparison, we have retrained Guo *et al*.[6] and AlMubarak *et al*.[12] models on the 453 high-resolution epithelial histopathology image data. Table 7 shows that the proposed model performs best for the CIN classification task. In addition, our model provides the significance of individual local regions toward the whole image classification. The results for sample images from the proposed DeepCIN model are shown in Figure 6. We observed that the performance was uniform among different sizes of epithelium images. The distribution of the entire data and the predictions for all

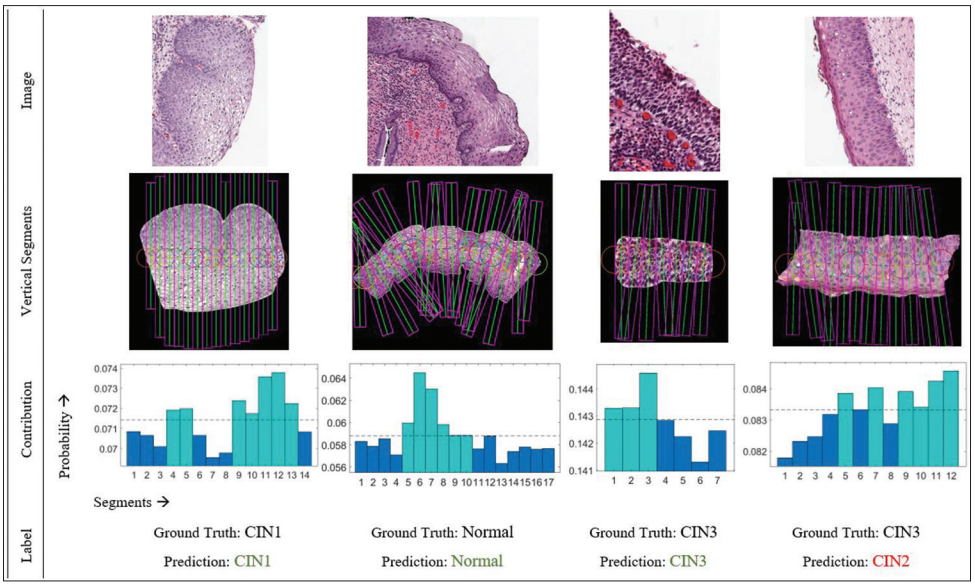

Figure 6: Results of DeepCIN. From top to bottom, each column presents original image, localized vertical regions, contribution of segments within an image toward the image-level CIN classification (represented as probability distribution over the segments [attentional weights], the dotted lines indicate mean value and segments above the mean value, highlighted in green, are contributing the most), and corresponding ground truth and prediction labels, respectively

Table 7: Comparison with state-of-the-art models

Model	*P*	*R*	F1	ACC	AUC	AP	MCC
Guo *et al*.[6]	67.5	73.3	69.4	73.4	-	-	56.5
AlMubarak *et al*.[11]	66.1	75.6	70.4	75.5	90.9	78.1	60.3
Ours*	88.6	88.5	88.0	88.5	96.5	91.5	82.0

P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under receiver operating characteristic curve, ACC: Classification accuracy

5-folds is depicted in the Sankey diagram in Figure 7, which shows the proportion of images that are correctly classified and misclassified. Image samples belonging to the CIN1 class were mostly misclassified as a normal class. Two reasons may explain this: (1) CIN1 images closely resemble normal images and (2) the number of CIN1 class images is small, relative to the number of Normal class images.

As an extension, we have tabulated the performance model with exact class labels, CIN versus Normal, CIN3-CIN2 versus CIN1-Normal, CIN3 versus CIN2-CIN1-Normal, and off-by-one class [Table 8]. For the exact class label scheme, the predicted class label should exactly match the expert ground-truth class label. The CIN versus Normal scheme is an abnormal–normal grouping of the predicted labels. The CIN3-CIN2 versus CIN1-Normal and CIN3 versus CIN2-CIN1-Normal interclass grouping schemes resemble the clinical decisions for treatment. The off-by-one scheme emphasizes the possible disagreement between the expect pathologists while labeling the CIN class which is usually observed to be one grade off.[33]

We have ensembled our five models from the fivefold cross-validation with maximum voting system to test the model performance on unseen data. The results from the hold-out 224 image data are shown in Table 9. The results when compared with Table 8 indicate that the proposed model is good at generalizing on unseen data. We noticed that the kappa score with CIN3 versus CIN2-CIN1-Normal scoring scheme is affected due to small portion of CIN 3 images were miss predicted as CIN 2 class.

Discussion

The main objective of the DeepCIN model is to classify the high-resolution epithelium images into normal or precancerous transformation of cells of the uterine cervix. We generate classification results by fusing localized information, forming a sequence of logit feature vectors in the same order of the vertical segments from the epithelium image. The number of vertical segments created varies since the epithelium images have arbitrary shapes. Traditional NNs are limited to fixed-length input, but RNNs have the capability to read varying input sequences along with memorization. We employ a GRU to read the arbitrarily shaped input sequences. GRU with attention helps in better understanding the differentially informative localized data. Unlike the stage II encoder from the segment-level sequence generator, incorporation of attention helped the model to better fuse the segment data and identify localized regions that are significantly important in the classifying the epithelial image.

It is now four decades since Marsden Scott Blois presented a paradigm for medical information science to distinguish domains in medicine in which humans are essential from those in which computation is essential and computers are likely to play a primary role.[34] He emphasized the importance of human judgment in the former domain, which includes most of clinical medicine but does not include the evaluation and interpretation of physiological parameters, for example, blood

Table 8: Fivefold cross-validation results with different scoring schemes

Scoring scheme	*P*	*R*	F1	ACC	AUC	AP	MCC	κ
Exact class label	88.6	88.5	88.0	88.5	96.5	91.5	82.0	81.5
CIN versus Normal	94.6	94.1	94.0	94.1	93.8	97.7	88.5	87.9
CIN3-CIN2 versus CIN1-normal	96.8	96.7	96.7	96.7	96.0	98.9	92.7	92.5
CIN3 versus CIN2-CIN1-normal	96.2	96.0	96.0	96.0	88.4	98.3	85.3	84.8
Off-by-one	-	-	-	98.9	-	-	-	-

P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under receiver operating characteristic curve, ACC: Classification accuracy

Table 9: Cervical intraepithelial neoplasia classification results on 224 image-set

Scoring scheme	*P*	*R*	F1	ACC	AUC	AP	MCC	κ
Exact class label	90.2	88.4	88.2	88.4	98.0	93.1	80.5	80.0
CIN versus normal	97.3	97.3	97.3	97.3	97.2	99.7	94.4	94.4
CIN3-CIN2 versus CIN1-Normal	95.7	95.6	95.5	95.5	94.0	99.1	90.3	90.0
CIN3 versus CIN2-CIN1-Normal	93.0	92.4	91.5	92.4	78.2	97.0	71.9	68.1
Off-by-one	-	-	-	98.2	-	-	-	-

P: Precision, R: Recall, F1: F1-score, AP: Average precision, MCC: Matthews correlation coefficient, AUC: Area under receiver operating characteristic curve, ACC: Classification accuracy

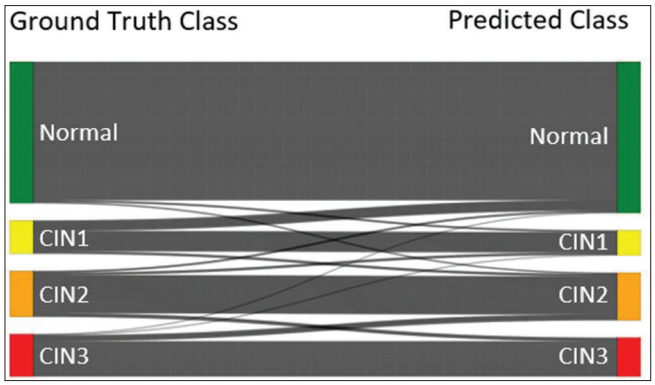

Figure 7: Sankey diagram – based on the combined test results from the fivefold cross-validation. The height of each bar is proportional to the number of samples corresponding to each class

gases, which is the proper domain of computers. With regard to the Blois paradigm, we propose that computer processing of histopathology images falls within the computational domain, and computers are likely to play a primary role.

Conclusion

In this study, we address the CIN classification problem by focusing on localized epithelium regions. The varying atypical nuclei density which is crucial in CIN determination is better analyzed by sequence mapping of the deep learning features. This sequence is interpreted in both directions under weak supervision with the long-term and short-term memory of the feature information. We employed an attention-based fusion approach to carry out an image-level classification. This hierarchical approach not only produces the image-level CIN classification labels but also provides the contribution of each individual vertical segment of the epithelium toward the whole image classification. We conjecture that this information highlights the highest-risk areas; this serves as an automated check for the pathologist's assessment.

We observed that our proposed model, DeepCIN, has outperformed state-of-the-art models in classification accuracy. The final image-level classification accuracies and Cohen's kappa score are {88.5% (± 2.2%), 81.5%}, {94.1% (± 2.0%), 87.9%}, {96.7% (±1.6%), 92.5%}, {96.0% (1.7%), 84.8%}, and {98.9% (± 0.0%)-}, for exact class label, CIN versus Normal, CIN3-CIN2 versus CIN1-Normal, CIN3 versus CIN2-CIN1-Normal, and leave-one-out schemes, respectively. These results significantly exceed the variability of community pathologists when measured against the gold standard and are in the range of inter-pathologist variability for expert pathologists as measured by the κ statistics.

Limitations of this work include use of a database that is not publicly available, which precludes validation by other researchers. Ground truth for the entire set was based on only one expert pathologist. Part of the set was scored by two pathologists; accuracies obtained for the two sets are similar.

Future work could improve results by including more annotated image data with balanced class distribution for training. There is also a possibility for improvements if the entire model could be trained end to end, which requires greater GPU resources. Our future research will focus on WSI-level classification with end-to-end automation which combines the proposed model with our previous work on automated epithelium segmentation[35] and automated nuclei detection[36] for extracting enhanced feature information.

Financial support and sponsorship

This research was supported (in part) by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

Conflicts of interest

There are no conflicts of interest.

References

1. American Cancer Society. Cancer Facts & Figures 2020. Atlanta, GA: American Cancer Society; 2020.
2. Islami F, Fedewa SA, Jemal A. Trends in cervical cancer incidence rates by age, race/ethnicity, histological subtype, and stage at diagnosis in the United States. Prev Med 2019;123:316-23.
3. World Health Organization. Sexual and Reproductive Health, Prevention and Control of Cervical Cancer; 2019. Available from: https://www.who.int/reproductivehealth/topics/cancers/en/. [Last accessed on 2020 Jun 09].
4. World Health Organization. Sexual and Reproductive Health, WHO Director-General Calls for all Countries to Take Action to Help end the Suffering Caused by Cervical Cancer; 2018. Available from: https://www.who.int/reproductivehealth/call-to-action-elimination-cervical-cancer/en/. [Last accessed on 2020 Jun 09].
5. Melnikow J, Nuovo J, Willan AR, Chan BK, Howell LP. Natural history of cervical squamous intraepithelial lesions: A meta-analysis. Obstet Gynecol 1998;92:727-35.
6. Guo P, Banerjee K, Joe Stanley R, Long R, Antani S, Thoma G, *et al*. Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. IEEE J Biomed Health Inform 2016;20:1595-607.
7. Huang P, Zhang S, Li M, Wang J, Wang B, Lv X. Classification of cervical biopsy images based on LASSO and EL- SVM. IEEE Access 2020;8:24219-28.
8. Rahmadwati R, Naghdy G, Ros M, Todd C, Norahmawati E. Cervical cancer classification using Gabor filters. Proc.-2011 1st IEEE Int. Conf. Healthcare Informatics, Imaging Syst. Biol. HISB; 2011. p. 48-52.
9. Wang Y, Crookes D, Eldin OS, Wang S, Hamilton P, Diamond J. Assisted diagnosis of cervical intraepithelial neoplasia (CIN). IEEE J Sel Top Signal Process 2009;3:112-21.
10. Kayser K, Brokenfeld S, Kayser G. Digital image content and context information in tissue-based diagnosis. Diagn Pathol 2018;4: 269.
11. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
12. AlMubarak H, Stanley J, Guo P, Long R, Antani S, Thoma G, *et al*. A hybrid deep learning and handcrafted feature approach for cervical cancer digital histology image classification. Int J Healthc Inf Syst Inform 2019;14:66-87.
13. Li C, Xue D, Zhou X, Zhang J, Zhang H, Yao Y, *et al*. Transfer learning based classification of cervical cancer immunohistochemistry images. Proc of the third Int Sym on Image computing and Digital medicine 2019 Aug; Xi'an, China. New York: ACM; 2019.
14. Li C, Chen H, Li X, Xu N, Hu Z, Xue D, *et al*. A review for cervical histopathology image analysis using machine vision approaches. Artif Intell Rev 2020; 53:4821-62.
15. Stoler MH, Ronnett BM, Joste NE, Hunt WC, Cuzick J, Wheeler CM, *et al*. The Interpretive Variability of Cervical Biopsies and Its Relationship to HPV Status. Am J Surg Pathol 2015;39:729-36.
16. Cai B, Ronnett BM, Stoler M, Ferenczy A, Kurman RJ, Sadow D, *et al*. Longitudinal evaluation of interobserver and intraobserver agreement of cervical intraepithelial neoplasia diagnosis among an experienced panel of gynecologic pathologists. Am J Surg Pathol 2007;31:1854-60.
17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-80.
18. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, *et al*. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1724-34.
19. Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell 2017;39:2298-304.
20. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. ICLR; 2015 May 7-9; San Diego, USA. San Diego:Conference Track Proceedings; 2015.
21. Shi B, Yang M, Wang X, Lyu P, Yao C, Bai X. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. IEEE Trans Pattern Anal Mach Intell 2019;41:2035-48.
22. Luo C, Jin L, Sun Z. MORAN: A Multi-Object Rectified Attention Network for scene text recognition. Pattern Recognit 2019;90:109-18.
23. Chaudhari S, Polatkan G, Ramanath R, Mithal V. An Attentive Survey of Attention Models. New York:IEEE; 2019.
24. Wentzensen N, Walker JL, Gold MA, Smith KM, Zuna RE, Mathews C, *et al*. Multiple biopsies and detection of cervical cancer precursors at colposcopy. J Clin Oncol 2015;33:83-9.
25. Huang G, Liu Z, Maaten L, Weinberger K. Densely Connected Convolutional Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York:IEEE;2017. p. 2261-9.
26. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 2009;31:855-68.
27. Zeiler MD. ADADELTA: An adaptive learning rate method. arXiv 2012;1212. https://arxiv.org/abs/1212.5701.
28. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09; 2009. New York: IEEE; 2009.
29. Higham DJ, Higham NJ. MATLAB guide. 3rd ed. Manchester:Siam; 2016.
30. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, *et al*. Automatic Differentiation in PyTorch. 31st Conference on Neural Information Processing Systems; 2017 Dec 4-9; Long Beach, USA. New York:Curran Associates Inc; 2017.
31. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. New York: IEEE; 2016.
32. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. New York: IEEE; 2016.
33. Guo P. Data Fusion Techniques for Biomedical Informatics and Clinical Decision Support. Theses. Ann Arbor: Pro Quest Diss; 2018. p. 135.
34. Blois MS. Clinical judgment and computers. N Engl J Med 1980;303:192-7.
35. Sornapudi S, Hagerty J, Stanley RJ, Stoecker WV, Long R, Antani S, *et al*. EpithNet: Deep regression for epithelium segmentation in cervical histology images. J Pathol Inform 2020;11:10.
36. Sornapudi S, Stanley RJ, Stoecker WV, Almubarak H, Long R, Antani S, *et al*. Deep learning nuclei detection in digitized histology images by superpixel. J Pathol Inform 2018;9:10.