DeepPap: Deep Convolutional Networks for Cervical Cell Classification

Ling Zhang , Le Lu, *Senior Member, IEEE*, Isabella Nogues, Ronald M. Summers, Shaoxiong Liu, and Jianhua Yao

Abstract—Automation-assisted cervical screening via Pap smear or liquid-based cytology (LBC) is a highly effective cell imaging based cancer detection tool, where cells are partitioned into "abnormal" and "normal" categories. However, the success of most traditional classification methods relies on the presence of accurate cell segmentations. Despite sixty years of research in this field, accurate segmentation remains a challenge in the presence of cell clusters and pathologies. Moreover, previous classification methods are only built upon the extraction of hand-crafted features, such as morphology and texture. This paper addresses these limitations by proposing a method to directly classify cervical cells—without prior segmentation—based on deep features, using convolutional neural networks (ConvNets). First, the ConvNet is pretrained on a natural image dataset. It is subsequently fine-tuned on a cervical cell dataset consisting of adaptively resampled image patches coarsely centered on the nuclei. In the testing phase, aggregation is used to average the prediction scores of a similar set of image patches. The proposed method is evaluated on both Pap smear and LBC datasets. Results show that our method outperforms previous algorithms in classification accuracy (98.3%), area under the curve (0.99) values, and especially specificity (98.3%), when applied to the Herlev benchmark Pap smear dataset and evaluated using five-fold cross validation. Similar superior performances are also achieved on the HEMLBC (H&E stained manual LBC) dataset. Our method is promising for the development of automation-assisted reading systems in primary cervical screening.

Index Terms—Cell classification, Cervical cytology, deep learning, neural networks, Pap smear.

L. Zhang, L. Lu, R. M. Summers, and J. Yao are with the Imaging Biomarkers and Computer-Aided Diagnosis Laboratory and the Clinical Image Processing Service, Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892 USA (e-mail: zhangling0722@163.com; le.lu@nih.gov; wrms@nih.gov; jyao@nih.gov).

I. Nogues is with the Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892 USA (e-mail: isabella.nogues@nih.gov).

S. Liu is with the Department of Pathology, Peoples Hospital of Nanshan District, Shenzhen 518052, China (e-mail: liusx20088@sina.com).

I. INTRODUCTION

CERVICAL cytology (conventional Pap smear or liquid-based cytology) [1], the most popular screening test for prevention and early detection of cervical cancer, has been widely used in developed countries and has significantly reduced its incidence and number of deaths [2]. However, population-wide screening is still unavailable in underdeveloped countries [2], partly due to the complexity and tedious nature of manually screening abnormal cells from a cervical cytology specimen [3]. While automation-assisted reading techniques can boost efficiency, their current performance is not adequate for inclusion in primary cervical screening [4].

During the past few decades, extensive research has been devoted to the creation of computer-assisted reading systems based on automated image analysis methods [3], [5], [6]. Such systems automatically select potentially abnormal cells in a given cervical cytology specimen, from which the cyto-screener/cytopathologist completes the classification. This task comprises three steps: cell (cytoplasm and nuclei) segmentation, feature extraction/selection, and cell classification.

Accurate cell segmentation is crucial to the success of a reading system. However, despite recent significant progress in this area [7]–[15], the presence of cell clusters (which is even more problematic in Pap smear than in liquid-based cytology), as well as the large shape and appearance variations between abnormal and normal nuclei, remains a major obstacle to the accurate segmentation of individual cytoplasms and nuclei. On the Herlev benchmark dataset [16], [17], the attained nucleus segmentation accuracy ranging between 0.85 [11] and 0.92 [15]. On an overlapping cervical cell dataset [13], the cytoplasm segmentation accuracy ranges from 0.87 to 0.89 [13]. On the other hand, most cell classification studies assume that accurate segmentations of individual cytoplasms and nuclei are already available [17]–[19]. By optimizing features derived from the segmented cytoplasm and nucleus, high classification accuracies (e.g., 96.8%) are achieved on the Herlev dataset, using 5-fold cross validation (CV) [18], [19]. However, these high values would decrease, once the automated segmentation error (deriving mainly from the error-prone abnormal nucleus segmentation [10], [14]), were taken into account.

Several strategies to remove this dependence on segmentation have been investigated. Classification based only on nucleus features (excluding cytoplasm features) is proposed [6], [20], [21]. Comparable results on the Herlev dataset is obtained by using a non-linear dimensionality reduction algorithm and supervised

learning [20]. Another idea is to classify image patches containing full cervical cells [22]–[24]. However, extraction of such patches still requires automated cell detection and segmentation. To avoid the pre-segmentation step, pixel-level classification method is designed to directly screen abnormal nuclei with no prior cytoplasm and nucleus segmentation [25], but reports limited validation results. Alternatively, a technique which classifies the cropped blocks from cell images is proposed [26]. However, arbitrary cropping could potentially separate a full cell into distinct patches.

Current cervical screening systems are hindered by limitations in the feature design and selection components. At present, extracted features fall under the following categories: handcrafted features describing morphology and chromatin characteristics [11], [17]–[20] in accordance with "The Bethesda System (TBS)" rules [27], engineered features representing texture distribution [22], [23], [25] according to previous computeraided-diagnosis experiences, or both combined [6], [9], [26]. The resulting features are then organized, using a feature selection or dimensionality reduction algorithm, for classification. Handcrafted features are compromised by the current limited understanding of cervical cytology. Engineered features are obtained in an unsupervised manner, and thus often encode redundant information. The feature selection process potentially ignores significant clues and removes complementary information. Moreover, considering that the detection of some abnormal cervical cells is challenging even for human experts [17], [27], [28], the hand-crafted features used in previous studies may not be able to represent complex discriminative information. In fact, information describing cell abnormality may potentially lie in latent higher level features of cervical cell images, but this has not yet been investigated.

Representation learning refers to a set of methods designed to automatically learn and discover intricate discriminative information from raw data [29]. Recently, representation learning has been popularized by deep learning methods [30]. In particular, deep convolutional neural networks (ConvNets) [31] have achieved unprecedented results in the 2012 ImageNet Large Scale Visual Recognition Challenge, which consisted in classifying natural images in the ImageNet dataset into 1000 finegrained categories [32]. They have also significantly improved performance in a variety of medical imaging applications [33], [34], such as classification of lung diseases and lymph nodes in CT images [35], [36], segmentation (pixel classification) of brain tissues [37] in MRI, vessel segmentation [38] in fundus images, and detecting cervical intraepithelial neoplasia (CIN, particularly CIN2+) at patient level based on Cervigram images [39] or multimodal data [40]. Additionally, ConvNets have demonstrated superior performance in the classification of cell images, such as pleural cancer [41] and human epithelial-2 cell images [42].

Large datasets are crucial to the high performance of ConvNets. However, there exists a very limited amount of labeled data for cervical cells, as high expertise is required for quality annotation. For instance, the Herlev benchmark dataset [17] only contains 917 cells (675 abnormal and 242 normal). Transfer learning [43]–[45] is an effective method to overcome this

problem. Since the features in the first few ConvNet layers are more generic, they can be appended to various sets of subsequent layers specific to different tasks [43]. For instance, ConvNets trained on large-scale natural image datasets (e.g., ImageNet [46]) can be transferred to various medical imaging datasets, such as CT [35], ultrasound [47] and X-ray [44], [48] datasets, and can subsequently reduce overfitting on small datasets while boosting performance through fine-tuning.

In this paper, we apply ConvNets to the classification of cervical cells in cytology images. Our approach directly operates on raw RGB channels sampled from a set of square image patches coarsely centered on each nucleus. A ConvNet pre-trained on ImageNet is fine-tuned to discriminate between patches containing abnormal and normal cells based on deep hierarchical features. For an unseen cell, a set of image patches coarsely centered on the nucleus are classified by the fine-tuned ConvNet. Its classification results are then aggregated to generate the final cell category. Our approach is tested on two cervical cell image datasets: the Herlev dataset consisting of Pap smear images [17]; the HEMLBC (H&E stained manual liquid-based cytology) dataset being used to develop automation-assisted cervical screening system [6]. In our experiments (conducted using five-fold cross-validation(CV)), the fine-tuned ConvNet obtains classification accuracies of 98.3% on Herlev dataset and 98.6% on HEMLBC dataset, surpass the previous best accuracies of 96.8% and 94.3% on the two datasets, respectively.

Our contributions are summarized as follows. 1) To the best of our knowledge, this is the first application of deep learning and transfer learning methods to cervical cell classification. 2) Unlike the previous methods, which rely on cytoplasm/nucleus segmentation and hand-crafted features, our method automatically extracts deep hierarchical features embedded in the cell image patch for classification, as long as a coarse nucleus center is provided. As a result, the classification does not suffer from any accuracy loss caused by inaccurate segmentation, and does not explicitly utilize prior medical knowledge of cervical cytology. 3) Our method generates the highest performances on both the Herlev Pap smear and the HEMLBC datasets, and has the potential to improve the performance of automation-assisted cervical cancer screening systems.

II. METHODS

The proposed method includes a training and a testing stage, as shown in Fig. 1. In the training stage, a ConvNet is first pre-trained on the ImageNet dataset, and data preprocessing is applied on the cervical cell dataset. Next, transfer learning is applied, whereby the pre-trained network parameters are used to initialize a new ConvNet. This ConvNet is then fine-tuned on the preprocessed training samples. In the testing stage, the preprocessed testing images are fed into the fine-tuned ConvNet. The abnormality score is obtained by aggregating the ConvNet's output values. Further details are described below.

A. Data Preprocessing

1) Patch Extraction: Unlike previous patch based cell classification methods [22]–[24], [41], [42], our method does
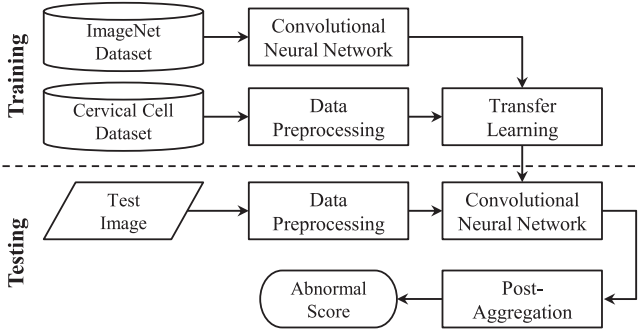
Fig. 1. Overview of the proposed method using convolutional neural networks and transfer learning for classifying cervical cell images.

not directly operate on images containing full cells (like the images in the Herlev dataset), for practical reasons. In particular, obtaining an individual cell requires cell pre-segmentation (at least cytoplasm segmentation), which remains an unsolved, challenging problem [13]. As mentioned in the TBS rules [27], different cervical cytology abnormalities are associated with different nucleus abnormalities. Hence, nucleus features in themselves already include substantial discriminative information. We thus extract image patches of size $m \times m$ centered on the nucleus centroid. This strategy allows for embedding not only the nucleus scale/size information (an important discriminative feature between abnormal and normal cells), but also the contextual clues (e.g., the cytoplasm appearance) in the extracted patches. We acknowledge that automated methods for extracting a nucleus patch, e.g., Laplacian-of-Gaussian (LoG) [49], selective search [50], or ConvNets [51] exist. However, in this paper, we choose to focus on the classification task. We adopt a simple method of directly translating the centroid of the ground-truth nucleus mask to extract a set of image patches as described below.

2) Data Augmentation: Data augmentation improves the accuracy of ConvNets and reduces overfitting [32]. Since cervical cells are rotationally invariant, we perform N_r rotations (with a step size of θ degrees) on each cell image, and thus increase our number of image samples. N_r patches (one per rotated image) of size $m \times m$ centered at the rotated nucleus centers are extracted as the training samples, as shown in the middle (blue) panel in Fig. 2. Note that rotating a cell image may slightly degrade its high frequency contents (could be considered as a lower imaging quality), but should not change its abnormality/normality for most cells. Actually the augmentation step based on image rotation is crucial to the success of the ConvNet [35], and has been demonstrated to be important for improving accuracy of ConvNet-based cell image classification problem [42], given the limited number of images in the Herlev and HEMLBC dataset. Zero padding is used to void regions that lie outside of the image boundary.

Considering that the detected nucleus center may be inaccurate in practice, we randomly translate (by up to d pixels) each nucleus centroid N_t times to obtain N_t points as the coarse nucleus centers. Accordingly, N_t patches of size $m \times m$ centered at these locations are extracted as training samples, as shown in

the right (green) panel in Fig. 2. These patches not only simulate inaccurate nucleus center detection, but also increase the amount of training samples for ConvNets. Other data augmentation approaches such as scale and color transformations are not used, as both the size and intensity of the nucleus are essential features to distinguish abnormal cervical cells from normal ones.

There are ~ 3 times more abnormal cell images than normal cell images in the Herlev dataset. Classifiers tend to exhibit bias towards the majority class (abnormal cells). Although achieving a high sensitivity rate (correct classification of abnormal cells) is ideal from a medical point of view [17], a high false positive rate (mis-classification of normal cells as abnormal) is not desirable from a practical standpoint [6]. A common solution to this dilemma is to balance the proportions of positive and negative training samples [52]. Doing so also improves the accuracy and convergence rate of ConvNets in training [32], [36]. Hence, we create a balanced training set by sampling a higher proportion of normal than abnormal patches.

B. Convolutional Neural Networks

A convolutional neural network (ConvNet) [31], [32] is a deep learning model comprising multiple consecutive stages, namely *convolutional* (*conv*), *non-linearity* and *pooling* (*pool*) layers, followed by more *conv* and *fully connected* (*fc*) layers. The input of the ConvNet is the raw pixel intensity image (in our case, the image obtained by subtracting the mean image over the training set from the original image [32]). The output layer is composed of several neurons each corresponding to one class. The weights (W) in the ConvNet are optimized by minimizing the classification error on the training set using the backpropagation algorithm [53]. Fig. 3 shows two ConvNets. The upper network is trained on the ImageNet dataset, and the lower network is trained on the cervical cell dataset.

1) Convolutional Layer: The *conv* layer takes local rectangular patches across (with offset by *stride* and with/without spatial preservation by *padding*) the input image (for the first layer) or feature maps (for the subsequent layers) as input, on which 2D convolution with a filter is performed. The sum x of the resulting convolutions is fed into a non-linearity function, specifically a rectified linear unit (ReLU) $f(x) = \text{MAX}(0, x)$ [32], in order to increased the speed of training. In a given layer, the same filter is shared in a feature map, while different filters are used for different feature maps. This property of filter sharing in *conv* layer allows for detecting the same pattern in different locations of the feature map.

2) Pooling Layer: The pooling operation down-samples the feature map by summarizing feature responses in each non-overlapping local patch, often by computing the maximum activations (max-pooling). This yields features invariant to minor translations in the data.

3) Fully Connected Layer: *conv* and *pool* generate feature maps of smaller dimensions than the input image, which are then passed through several *fc* layers. The first few *fc* layers fuse these feature maps into a feature vector. The last *fc* layer contains two neurons which compute the classification
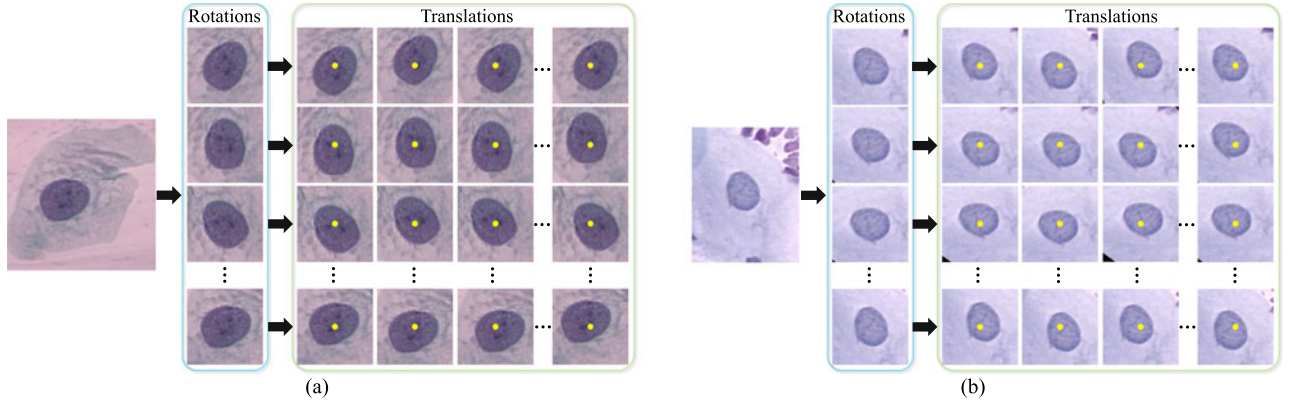
Fig. 2. Two set of image patches are generated from (a) an abnormal and (b) a normal cervical cell image by rotations and translations. The centroids of translated image patches are shown as yellow, indicating the potentially inaccurate detection of nucleus centers.
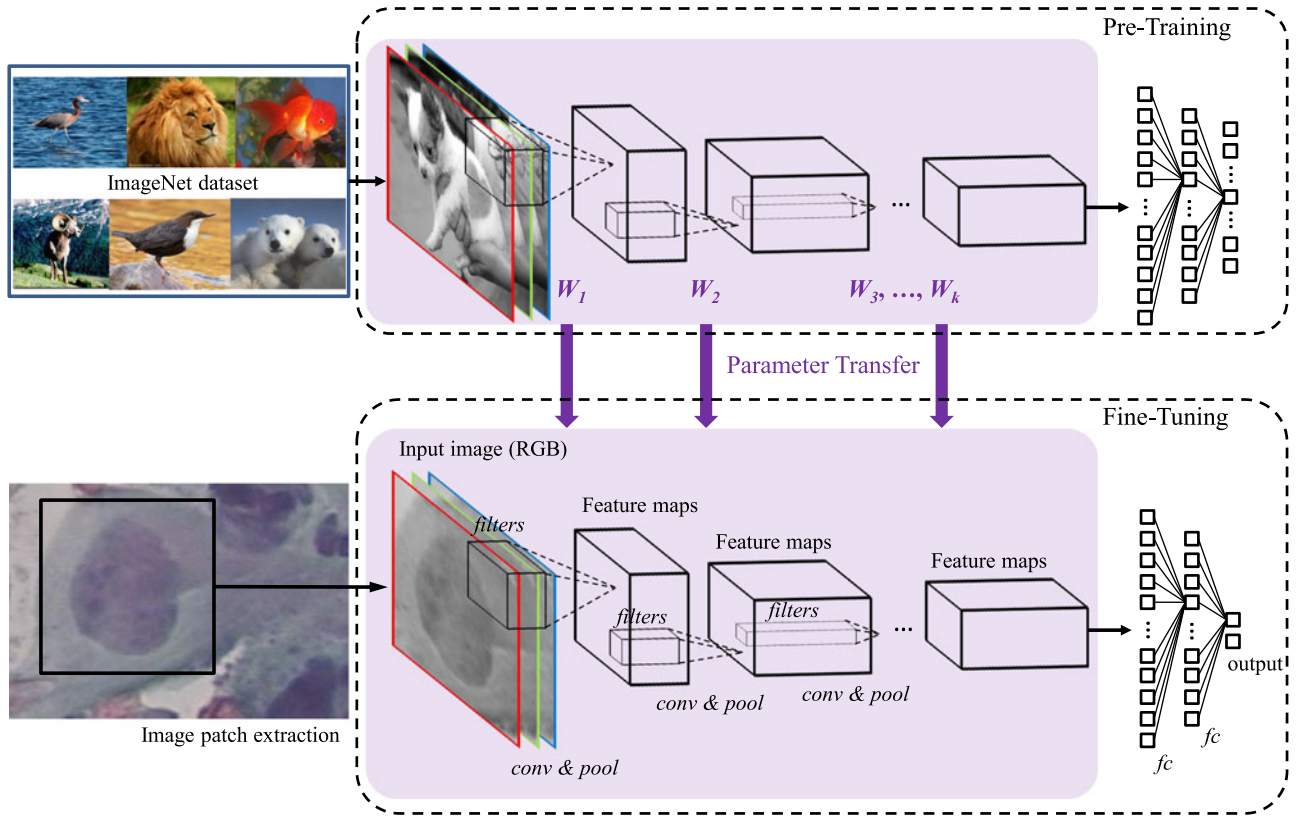

Fig. 3. A ConvNet is fine-tuned on the cervical cell dataset with parameters transferred from another ConvNet pre-trained on the ImageNet dataset. In this example, the parameters (W_1, W_2, ... W_k) in the purple region in the pre-trained model (upper panel) are transferred to the same locations in another model (lower panel) for fine-tuning on cervical cell dataset.

probability for each class using softmax regression. To reduce overfitting, "dropout" [54] is used to constrain the fully-connected layers.

4) Network Training: The weights W in ConvNets are initialized with values from the Gaussian distribution. During training, these weights are iteratively updated with the gradients of the loss function, computed via stochastic gradient descent (SGD) over a mini-batch (size of 256) of training samples. The initial learning rate is decreased after certain epochs. As in Ref. [32], momentum and weight decay are used to speed up the

learning and reduce overfitting. The training process is terminated after a pre-determined number of epochs. The model with the lowest validation loss value is selected as the final network.

C. Transfer Learning

Transfer learning refers to the fine-tuning of deep learning models that are pre-trained on other large-scale image datasets. In this study, the first few *conv* and *pool* layers of a ConvNet pre-trained on the ImageNet classification dataset (ILSVRC2012)

Category	Class	Cell type	Num.
Normal	1	Superficial squamous epithelial	74
Normal	2	Intermediate squamous epithelial	70
Normal	3	Columnar epithelial	98
Abnormal	4	Mild squamous non-keratinizing dysplasia	182
Abnormal	5	Moderate squamous non-keratinizing dysplasia	146
Abnormal	6	Severe squamous non-keratinizing dysplasia	197
Abnormal	7	Squamous cell carcinoma in situ intermediate	150

(purple region in the upper part in Fig. 3) are used as the base of our network, on top of which several task-specific fc layers with random initialized weights are attached. In order to facilitate the transfer of features, the same network layers ($conv$ and $pool$) with the BVLC CaffeNet [55] are transferred to the same locations in our model (purple regions in Fig. 3). Like our network, the CaffeNet also takes RGB channels as input. All of these layers are jointly trained (fine-tuned) on our cervical cell dataset, for which a learning rate 10 times smaller than the original CaffeNet value is used to fine-tune the transferred layers, and the original learning rate is used to train the fc layers from scratch.

D. Testing

To classify an unseen image, we combine random-view aggregation [36] and multiple crop testing [32] to produce the final prediction score. In particular, our data augmentation method generates N_v image patches (rotations and translations about the nucleus centroid). From each of these patches, N_c sub-patches are cropped (including its corner, center and mirrored patches). Hence, for each test cervical cell image, $N_v \times N_c$ sub-patches are fed into the ConvNet. The final prediction score is obtained by averaging the scores of these $N_v \times N_c$ predictions.

III. EXPERIMENTAL METHODS

A. Data Set

The cell data used to train and test the ConvNets comes from two datasets with two types of cervical cytology images, which were acquired by different slide preparation, staining methods, and imaging conditions.

1) Herlev Dataset: The first one is from a publicly available dataset (http://mde-lab.aegean.gr/downloads) collected at the Herlev University Hospital by a digital camera and microscope [17]. The image resolution is 0.201 μm per pixel [16]. The specimens are prepared via conventional Pap smear and Pap staining. The Herlev dataset consists of 917 images – each containing one cervical cell – with ground truth segmentation and classification. There are a total of seven different classes – diagnosed by two cyto-technicians and a doctor, in order to maximize certainty of the diagnosis. These seven classes belong to two categories: class 1-3 are normal, and class 4-7 are abnormal, as shown in Table I. Examples of some cells are provided in Fig. 4(a). As can be seen, most abnormal cells have larger

nucleus size than normal cells. However, the normal columnar nucleus may have similar size (also maybe similar chromatin distribution) as severe and/or carcinoma nuclei, which makes the classification challenging.

For each abnormal cell image in the Herlev dataset, $N_r = 10$ rotations ($\theta = 36°$) and $N_t = 10$ translations (up to 10 pixels) are performed. For each normal cell, we use $N_r = 20$ ($\theta = 18°$) and $N_t = 14$, resulting in 100 and 280 image patches for each abnormal and normal cell image, respectively. This yields a relatively balanced data distribution. Note that such different steps of rotation/translation for abnormal and normal cells are only for training not testing set. The image patch size is set to $m = 128$ pixels to cover some cytoplasm region for most cells, and to contain most of the full nucleus region for the largest one. These image patches are then up-sampled to a size of $256 \times 256 \times 3$ pixels via bi-linear interpolation, in order to facilitate the transfer of pre-trained ConvNet model [35].

2) HEMLBC Dataset: The second one is from our own dataset collected at the People's Hospital of Nanshan District by using our previously developed autofocusing system (Olympus BX41 microscope with $20\times$ objective, Jenoptik ProgRes CF Color 1.4 Megapixel Camera, and MS300 motorized stage) [6]. Each pixel has a size of 0.208 μm^2. The specimens are prepared by manual liquid-based cytology with H&E staining. The dataset used in this paper is a subset used to train the abnormal/normal nucleus classifier for our automation-assisted cervical screening system [6]. There are totally 989 abnormal cells from 8 biopsy-confirmed CIN slides and 1381 normal cells from another 8 NILM (negative for intraepithelial lesion and malignancy) slide available. To create a balanced data distribution, 989 normal cells are randomly selected. The abnormal cells are diagnosed by two experienced pathologists. Most of them are segmented by an automated algorithm [6] and the ill-segmented ones are manually segmented by a pathologist. The normal cells are formed by two subsets: the first subset is collected by a pathologist with automated segmentation; the second subset is some false positive cells (e.g., cells with large nuclei, atrophic cells, etc.) collected during bootstrap process from validation images with manual segmentation for the ill-segmented ones by an engineer. More details are described in [6]. Examples of some cells are shown in Fig. 4(b).

For both abnormal and normal cell image in the HEMLBC dataset, $N_r = 10$ rotations ($\theta = 36°$) and $N_t = 10$ translations (up to 10 pixels) are performed, resulting in 100 image patches for each cell image. The image patch size is also set to $m = 128$ pixels and then up-sampled to a size of $256 \times 256 \times 3$ pixels as in Herlev dataset.

B. Network Architectures and Implementation

Fig. 3 illustrates our network architecture. The base ConvNet (denote as ConvNet-B) is pre-trained on the ImageNet classification dataset. ConvNet-B contains five $conv$ layers ($conv1 - conv5$), three $pool$ layers ($pool1$, $pool2$, $pool5$), and three fc layers ($fc6 - fc8$). Layers from $conv1$ to $pool5$ are transferred to the same locations in our model (denote as ConvNet-T). In other word, the first 5 weight layers ($conv1$
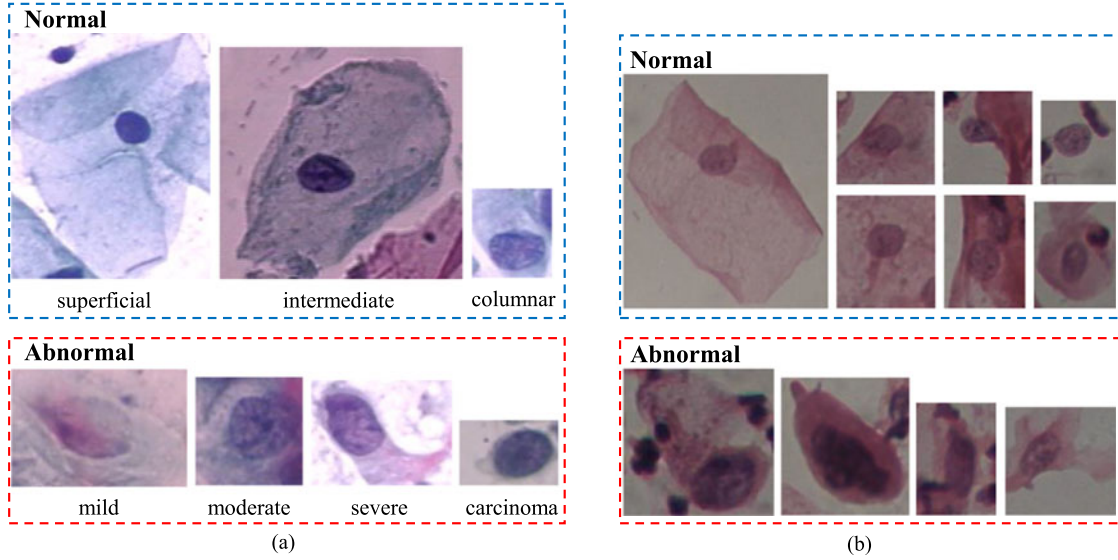
Fig. 4. Example images of normal and abnormal cervical cells from the (a) Herlev and (b) HEMLBC dataset. All these examples keep their originally relative scales for better illustrating the different characteristics (mainly nucleus size) between normal and abnormal cells.

TABLE II
CONVNET-T ARCHITECTURES FOR CERVICAL CELL CLASSIFICATION

	Input	$conv1$	$pool1$	$conv2$	$pool2$	$conv3$	$conv4$	$conv5$	$pool5$	$fc6$	$fc7$	$fc8$
Filter size	-	11×11	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	
Channels	3	96	96	256	256	384	384	256	256	1024	256	2
Stride	-	4	2	1	2	1	1	1	2	-	-	-
Padding	-	-	-	2	-	1	1	1	-	-	-	-

to $pool5$) of ConvNet-T are copied from the pre-trained ConvNet-B, and $fc6 - fc8$ layers of ConvNet-T are initialized with random Gaussian distributions. The detailed configurations of our ConvNet-T are listed in Table II. Local response normalization is used for $conv1$ and $conv2$ layers using the same setting as [32], and all hidden layers are equipped with the ReLU activation function. Note that the ConvNet-B and ConvNet-T share the same network structure from $conv1$ to $pool5$, but the number of neurons of the three fc layers in ConvNet-B and ConvNet-T are 4096-4096-1000 and 1024-256-2, respectively. The 1024 and 256 are set based on our empirical evaluation, and 2 is to accommodate the new object categories in our 2-class (abnormal/normal) classification problem. Actually, setting the number of neurons of $fc6$ and $fc7$ layers in the range of 1024∼256 will result in similar accuracy, while more number of neurons (e.g., 4096-4096) tend to have slightly lower accuracy (1%-2% lower) on our data, which is more compact and specific compared to ImageNet. ConvNet-T is run on the Caffe platform [55], using a Nvidia GeForce GTX TITAN Z GPU with 12 GB of memory.

C. Training and Testing Protocols

From each 256×256 training image patch or its mirrored version, a 227×227 sub-patch is randomly cropped, from which the mean image over the training set is then subtracted. Stochastic Gradient Descent (SGD), with a mini-batch size of 256, is used to train the ConvNet-T for 30 epochs. The learning rates of layers $conv1 - pool5$ and layers $fc6 - fc8$ start from 0.001 and 0.01, respectively, and are decreased by a factor of 10 at every tenth epoch. Weight decay and momentum are set to 0.0005 and 0.9. A dropout ratio of 0.5 is used to constrain the $fc6$ and $fc7$ layers.

In testing, we obtain the final score by averaging the scores of the $fc8$ output on 1000 patches ($N_v = 100$ augmentations each with $N_c = 10$ sub-crops).

D. Evaluation Methods

We evaluate the cervical cell classification using five-fold CV on both Herlev and HEMLBC datasets, to facilitate comparison with most previously reported results. In each of the 5 iterations of the ConvNet, 4 of 5 folds are used as training data, and the remaining one as validation. It's worth mentioning that data augmentation is after the training/validation splitting of cell population. We obtain the model's final performance values by averaging results from the 5 validation sets. The performance evaluation metrics include sensitivity ($Sens$), specificity ($Spec$), accuracy (Acc), harmonic mean ($H\text{-}mean$), $F\text{-}measure$, and area under the ROC curve (AUC), where $Sens$ measures the proportion of correctly identified abnormal cells, and $Spec$ the proportion of correctly identified normal cells; Acc is the global percentage of correctly classified cells; $H\text{-}mean = 2 \times \frac{Sens \times Spec}{Sens + Spec}$, used in [20], takes into account the
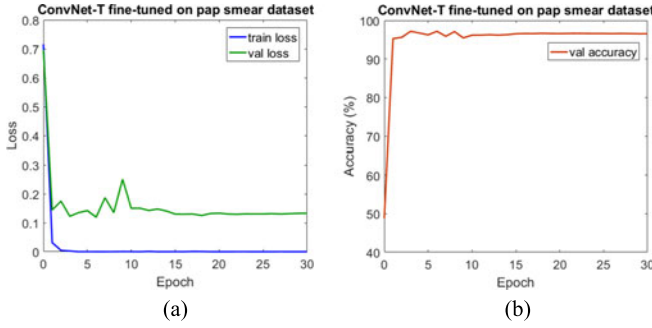
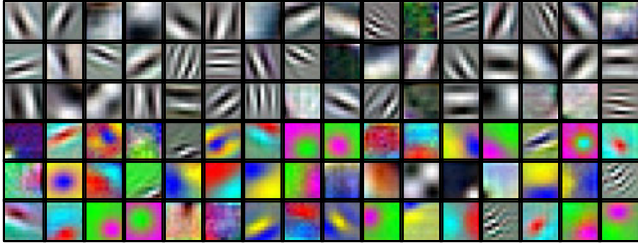Fig. 5. (a) Training and validation loss, and (b) validation accuracy versus number of training epochs.

Fig. 6. Visualization of the 96 filters with size of $11 \times 11 \times 3$ in the first convolutional layer of ConvNet-T fine-tuned on Pap smear dataset.

imbalanced data distribution; $F\text{-}measure$, the harmonic mean of precision and recall, is used in [21]. The ROC curve is computed by varying thresholds on the final classification scores (each final score is the average score of 1000 predictions). To test the robustness of our method against localization error of nucleus center, we randomly translate the ground truth centers of the test cells up to 5 or 10 pixels in both x and y directions, and the resulting performances on Herlev dataset are reported. In addition, the numbers of correct classification (normal vs. abnormal) and the distribution (shown by box plots) of the predicted abnormal scores of all cells for each of the seven cell classes (listed in Table I) are reported. Finally, we further consider the 7-class classification problem by simply modifying the number of neurons in the last fc layer from 2 to 7, and report the overall error (OE)% as in [17], [19].

IV. RESULTS

A. ConvNet Learning Results

Fig. 5 illustrates a fine-tuning process of ConvNet-T during 30 training epochs on the Herlev dataset. As shown in the figure, after 6 epochs, the validation loss reaches its minimum value (0.119), with a corresponding validation accuracy of 0.972. Fig. 6 shows the learned filters of the first convolutional layer of ConvNet-T trained on the Herlev Pap smear dataset. These automatically learned filters mainly consist of gradients of different frequencies and orientations and blobs of color, which are necessary for the cervical cell classification task. Along with these learned filters, the activations (feature maps) of an example cell at different pooling layers ($pool1$, $pool2$, and $pool5$) are provided in Fig. 7. One can observe that the pooling operation

summarizes the input cell image or previous feature maps by highlighting the activated spatial locations, and that the features become increasingly abstract in deeper layers of the ConvNet.

B. Qualitative Results

Figs. 8 and 9 contain examples of correctly classified abnormal and normal cell patches from the validation Herlev dataset, respectively. Fig. 10 provides examples of misclassified cervical cells from both Herlev and HEMLBC datasets, including both false negatives and false positives. The first two false negatives are instances of carcinoma, and the third one is an example of severe dysplasia. All false positives are columnar epithelial cells.

C. Quantitative Results on Herlev Dataset

Table III shows the classification performance ($Sens$, $Spec$, Acc, $H\text{-}mean$, $F\text{-}measure$, and AUC) of our method in comparison with previous methods [11], [17]–[20], [22], [23] on the Herlev dataset. The mean values of $Sens$, $Spec$, Acc, $H\text{-}mean$, $F\text{-}measure$, and AUC from our method (ConvNet-T) are 98.2%, 98.3%, 98.3%, 98.3%, 98.8%, and 0.998, respectively. We thus outperform previous methods in all metrics but $Sens$, which is slightly below others. Among these metrics, our $Spec$ (98.3%) substantially surpasses the previous highest result (92.2%). Also note that certain degree of localization error (up to 10 pixels) of nucleus center only results in a small reduction of performances of our method (e.g., Acc from 98.3% to 97.8%).

Table IV provides the numbers (and corresponding percentages) of correct classification for each of the seven cell classes. Our method shows perfect performance on two types of normal cell (superficial and intermediate squamous epithelial), as well as one type of abnormal cell (mild dysplasia). While the performances are relatively lower for columnar epithelial and severe squamous non-keratinizing dysplasia (both are 95.9%). The distribution of the abnormal scores of all cells for the seven cell classes are shown as box plots in Fig. 11. The proposed method returns abnormality-scores close to 0 or 1 for most normal and abnormal cells, respectively. The few misclassifications mainly occur to normal columnar and severe squamous cells, given the probability threshold at 0.5.

Finally, for the 7-class problem, an overall error (OE) of 1.6% is achieved on average. Such an error is lower than errors of previous methods, such as 7.9% [17] and 3.9% [19].

D. Quantitative Results on HEMLBC Dataset

Table V compares the classification performance between the proposed deep ConvNet-based method and our previous MLP (multilayer perceptron)-based method [6] on the HEMLBC dataset. Although the dataset used in this paper is a subset slightly smaller than that used in [6], an obvious trend of performance improvement can be observed.

E. Computational Speed

The average training time of a ConvNet-T running over up to 30 epochs is about 4 hours. Using the $N_v \times N_c = 1000$
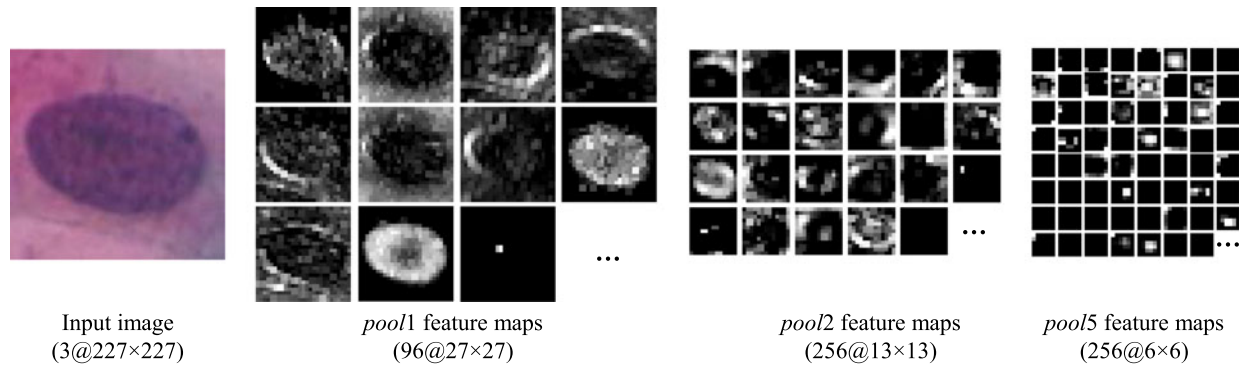
Input image   *pool*1 feature maps   *pool*2 feature maps   *pool*5 feature maps
(3@227×227)   (96@27×27)   (256@13×13)   (256@6×6)

Fig. 7.   Visualization of the activations (feature maps) of three pooling layers, *pool*1, *pool*2, and *pool*5 for an input cervical cell image.
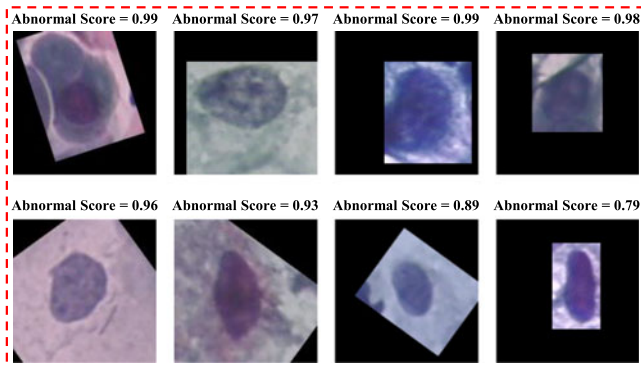

Fig. 8.   Examples of correctly classified abnormal cervical cells from Herlev dataset. Column 1 to column 4 are mild dysplasia, moderate dysplasia, severe dysplasia and carcinoma, respectively. Score = 1 corresponds to a 100% probability of representing an abnormal cell.
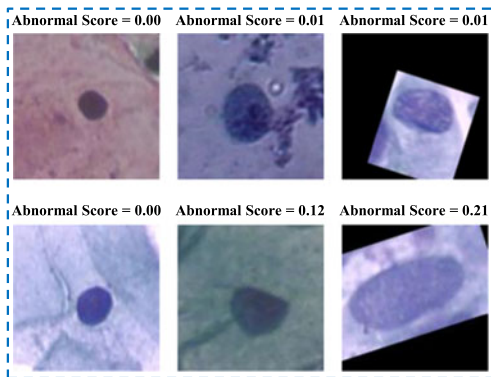

Fig. 9.   Examples of correctly classified normal cervical cells from Herlev dataset. Column 1 to column 3 are superficial, intermediate and columnar epithelial, respectively. Score = 1 corresponds to a 100% probability of representing an abnormal cell.

classification strategy, the testing time for one cervical cell is 3.5 seconds on average.

V. DISCUSSION

A. Comparison With Previous Methods

The methods in [17]–[21] follow the traditional cell classification pipeline – with features derived from manually segmented cytoplasms/nuclei. The techniques presented in [22],
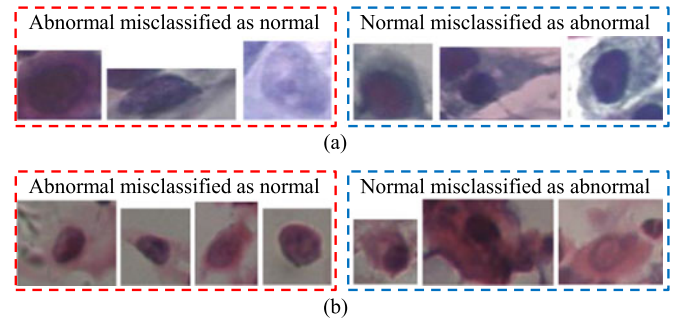

Fig. 10.   Examples of misclassified cervical cells from (a) Herlev dataset; (b) HEMLBC dataset. All images are shown at their original scales.

[23] perform direct texture classification of the input image. In contrast, our method automatically learns from the input image patch, and thus is not limited by the shortcomings of cell segmentation or feature design. The $Sens$ values of previous methods are slightly higher than those from our method (99.0% vs. 98.2% under 5-fold CV). Such high $Sens$ results mainly from the imbalanced data distribution – number of abnormal cells ~ 3 X higher than the number of normal cells – which induces the classifier to predict more cells as abnormal. High $Sens$ even at the expense of fairly low $Spec$ is required for specimen level diagnosis, as all positives will be reexamined by human experts. However, considering the abundance of normal cells (up to 300,000) in a Pap smear slide, the resulting lower $Spec$ will generate many false positives in clinical practice. For example, a 92% specificity [18], [19] will result in about 24,000 false positive cells. As a result, extensive and tedious targeted reading from a human observer will be necessary to refine the accuracy of the screening. Our approach substantially decreases the number of false positives. Although there are still about 1.7% false positives, they only come from columnar epithelial cells (Table IV). Actually, the differentiation between some columnar epithelial cells and (severe) abnormal cells are also challenging for experienced pathologists. Our method perfectly eliminates the majority types of normal cells (superficial and intermediate epithelial) in a specimen, and thus alleviates the labor burden of targeted reading and potentially reduces screening errors. Compared to our previous MLP method [6] on HEMLBC dataset, the deep ConvNet method achieves both higher $Sens$

TABLE III
PERFORMANCE COMPARISON OF OUR METHOD WITH PREVIOUS METHODS ON THE HERLEV DATASET

Methods	k-fold CV	$Sens$ (%)	$Spec$ (%)	Acc (%)	$H\text{-}mean$ (%)	$F\text{-}measure$	AUC
Benchmark [17]	10	98.8 ± 1.3	79.3 ± 6.3	93.6 ± 1.9	$(88.0 \pm NA)$	-	-
PSO-1nn [18]	5	$98.4 \pm NA$	$92.2 \pm NA$	$96.7 \pm NA$	$(95.2 \pm NA)$	-	-
GEN-1nn [19]	5	$98.5 \pm NA$	$92.1 \pm NA$	$96.8 \pm NA$	$(95.2 \pm NA)$	-	-
ANN [11]*	LOO	$99.9 \pm NA$	$96.5 \pm NA$	$99.3 \pm NA$	$(98.2 \pm NA)$	-	-
K-PCA + SVM [20]	10	-	-	-	$96.9 \pm NA$	-	-
Ensemble [21]	5	$\mathbf{99.0} \pm NA$	$89.7 \pm NA$	$96.5 \pm NA$	-	$93.1 \pm NA$	-
ENS [22]	5	-	-	-	-	-	0.884
$dis(S + M)$ [23]	5	-	-	-	-	-	0.964
ConvNet-T	5	98.2 ± 1.2	$\mathbf{98.3} \pm 0.9$	$\mathbf{98.3} \pm 0.7$	$\mathbf{98.3} \pm 0.3$	$\mathbf{98.8} \pm 0.5$	$\mathbf{0.998}$
ConvNet-T$^{\Delta 5}$	5	98.4 ± 1.3	97.1 ± 2.3	98.0 ± 0.8	97.7 ± 1.0	98.7 ± 0.6	0.994
ConvNet-T$^{\Delta 10}$	5	98.2 ± 1.2	96.7 ± 3.1	97.8 ± 0.5	97.4 ± 1.2	98.5 ± 0.4	0.994

PSO-1nn: particle swarm optimization for feature selection and 1-nearest neighbor as the classifier. GEN-1nn: genetic algorithm for feature selection and 1-nearest neighbor as the classifier. ANN: artificial neural networks. K-PCA: kernel principal component analysis for dimensional reduction. Ensemble: majority voting of three classifiers. ENS: ensemble classifiers based on Local Binary Pattern (LBP) with different configurations. $dis(S + M)$: discriminative LBP with concatenated sign and magnitude components. In the $H\text{-}mean$ (%) column, numbers in () indicate that no such result is present in the literature, so approximate results are calculated based on the corresponding $Sens$ and $Spec$ to enable comparison. *: The method in [11] uses leave-one-out cross validation (LOOCV) and excludes the columnar cells, which is not directly comparable to 10-fold ([17], [20]) or 5-fold CV ([18], [19], [21]–[23] and our proposed method) that involve all types of cells. $^{\Delta d}$: randomly translate the ground truth nucleus center up to d pixels in both x and y directions. Bold indicates the highest value in each column.

TABLE IV
NUMBERS (AND PERCENTAGES) OF CORRECT CLASSIFICATION (NORMAL VS. ABNORMAL) FOR EACH OF THE SEVEN CELL CLASSES IN HERLEV DATASET

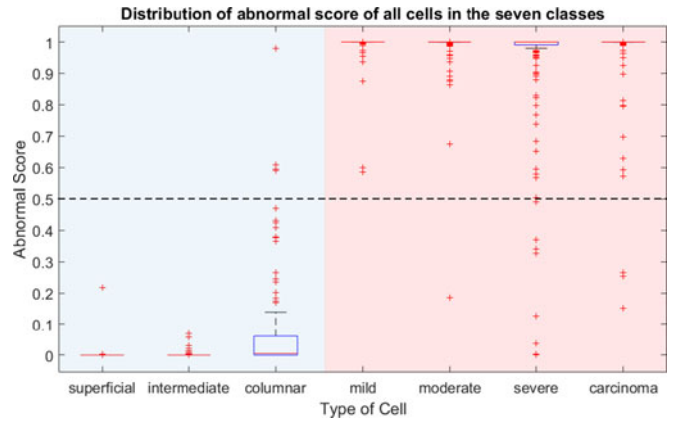| Cell type	Correct classification
Superficial squamous epithelial	74 (100%)
Intermediate squamous epithelial	70 (100%)
Columnar epithelial	94 (95.9%)
Mild squamous non-keratinizing dysplasia	182 (100%)
Moderate squamous non-keratinizing dysplasia	145 (99.3%)
Severe squamous non-keratinizing dysplasia	189 (95.9%)
Squamous cell carcinoma in situ intermediate	147 (98.0%)

Fig. 11. Box plots shown the distribution of abnormal scores of all cells in the seven classes from Herlev dataset. Blue and red backgrounds indicate normal and abnormal cell categories, respectively. Black dash line is the score threshold 0.5.

and $Spec$ at cell level. Actually, the automation-assisted screening system [6] built upon the MLP method has a satisfyingly high $Sens = 88\%$ and a perfect $Spec = 100\%$ at slide level by pathologist's targeted reading. Therefore, our new method has a high potential to further improve the $Sens$ of screening system while reducing the labor burden of targeted reading.

B. Advantages of the Proposed Method

1) The proposed method is designed for robust automated screening applications, since it only requires a coarse nucleus centroid as input (no cytoplasm/nucleus segmentation is needed). Our experiments indicate that the proposed image patch based cell classification is robust to inaccurate detection of nucleus centroid (refer to ConvNet-T$^{\Delta 5}$ and $^{\Delta 10}$ in Table III). Some examples can be seen in Figs. 8 and 9, where most of the image patches are not centered on the exact nucleus centroids, but are still assigned reasonable abnormal scores by our method. 2) Moreover, our method is able to distinguish the abnormal and normal cells even for some "difficult" cases. For example, the two columnar epithelial cells in the third column in Fig. 9 appear to exhibit a far greater level of abnormality than a severe dysplasia cell (lower one in the third column in Fig. 8),

as these columnar cells have much larger nuclei or nonuniform chromatin distributions. Unlike traditional morphology/texture based classification methods, which simply classify both cells as abnormal, our method provides a much higher abnormal score (0.89) for the severe dysplasia cell than the two columnar cells (0.01 and 0.21). This indicates that the ConvNet-T captures some latent but essential features embedded in the cell images. 3) Finally, the deep learning based method has high $Sens$ and especially high $Spec$, and produces the highest performances on a Pap-stained Pap smear (Herlev) and a H&E stained liquid-based cytology (HEMLBC) datasets. Such a strong performance has the potential to boost the development of automation-assisted reading systems in primary cervical cancer screening.

C. Limitations

Despite its high performance, our method demonstrates a few limitations hindering its inclusion in existing cervical screening

Methods	k-fold CV	$Sens$ (%)	$Spec$ (%)	Acc (%)
MLP [6]	5	92.5 ± NA	96.0 ± NA	94.3 ± NA
ConvNet-T*	5	**98.3 ± 0.7**	**99.0 ± 1.0**	**98.6 ± 0.3**

ConvNet-T*— It's worth mention that the 989/989 abnormal/normal cells
used in this paper is a subset of 1,126/1,126 abnormal/normal cells used in
[6] due to data missing of some abnormal cells.

systems. 1) Classification of a single patch requires 3.5 seconds, which is far too slow in a clinical setting. One could address this issue by eliminating the image patch augmentation step (100 variants per patch) from the testing phase, thus reducing speed to 0.035 seconds while compromising accuracy by only ~1%. 2) Despite high classification accuracy on the Herlev dataset, our method misclassify a few severe dysplasia (4.1%) and carcinoma (2%) cells as normal (Table IV and Fig. 11). As shown in Fig. 10(a), two dark stained carcinoma nuclei and a very large severe dysplasia nucleus are incorrectly classified as normal. An ideal screening system should not miss such severe abnormalities. To better handle such mis-classifications, cytoplasm/nucleus segmentation based features could be integrated into the system, either via deep learning or via "TBS" rules. Furthermore, both Herlev and HEMLBC are mainly consisted of expert-selected "typical" cells. The real life situation is more complex so more investigations are needed before transferring the results of this study to practice. For example, refer to the last two false positive cells in Fig. 10(b), they are from NILM slides, but it's hard to tell whether: the first one is an abnormal or normal cell due to poor imaging quality, and the second one is a normal atrophic cell or an abnormal cell. 3) A nucleus center is pre-required for applying our method, and is obtained from the ground truth segmentation in this paper. However, screening of abnormal cells within a given field-of-view requires automated detection of nucleus centers. Our ongoing study shows that this may easily be achieved using the fully convolutional networks (FCN) [15], [56] for semantic segmentation of cervical cells. And we already show that our method is robust to certain amount of inaccurate nucleus center detection. 4) The current experiments are conducted on a majority of images with individual cells. The effect of overlapping nuclei, cell clumps and artifacts on classification accuracy needs to be analyzed more extensively in the future investigation, since a screening system is expected to able to avoid misclassifying such objects as abnormal cells. Task-specific classifiers (mostly likely relying on deep learning) may be needed to handle these problems [3], [5], [6].

VI. CONCLUSION

This paper proposes a convolutional neural network-based method to classify cervical cells. Unlike previous methods, which rely on cytoplasm/nucleus segmentation and hand-crafted features, our method automatically extracts deep features embedded in the cell image patch for classification. It consists in extracting image patches coarsely centered on the nucleus as network input, transferring features from another pre-trained model into a new ConvNet for fine-tuning on the cell image patches, and aggregating multiple predictions to form the final network output. The proposed method yields the highest performance on both the Herlev Pap smear and the HEMLBC liquid-based cytology datasets, compared to previous methods. We anticipate that a segmentation-free, highly accurate cervical cell classification system of this type is promising for the development of automation-assisted reading systems for primary cervical screening.

REFERENCES

[1] E. Davey *et al.*, "Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid-based versus conventional cervical cytology: A systematic review," *Lancet*, vol. 367, no. 9505, pp. 122–132, 2006.

[2] 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. Clinicians*, vol. 62, no. 3, pp. 147–172, 2012.

[3] G. G. Birdsong, "Automated screening of cervical cytology specimens," *Human Pathol.*, vol. 27, no. 5, pp. 468–481, 1996.

[4] H. C. Kitchener *et al.*, "Automation-assisted versus manual reading of cervical cytology (MAVARIC): A randomised controlled trial," *Lancet Oncol.*, vol. 12, no. 1, pp. 56–64, 2011.

[5] E. Bengtsson and P. Malm, "Screening for cervical cancer using automated analysis of pap-smears," *Comput. Math. Method Med.*, vol. 2014, pp. 1–12, 2014.

[6] L. Zhang *et al.*, "Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining," *Cytom. Part A*, vol. 85, no. 3, pp. 214–230, 2014.

[7] A. Gençtav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images," *Pattern Recognit.*, vol. 45, no. 12, pp. 4151–4168, 2012.

[8] M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4568–4580, Nov. 2012.

[9] Y.-F. Chen *et al.*, "Semi-automatic segmentation and classification of pap smear cells," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 1, pp. 94–108, Jan. 2014.

[10] L. Zhang *et al.*, "Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts," *Comput. Med. Imag. Graph.*, vol. 38, no. 5, pp. 369–380, 2014.

[11] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Automatic cervical cell segmentation and classification in pap smears," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 539–556, 2014.

[12] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multi-scale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, Oct. 2015.

[13] Z. Lu *et al.*, "Evaluation of three algorithms for the segmentation of overlapping cervical cells," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 441–450, Mar. 2017.

[14] L. Zhang, H. Kong, S. Liu, T. Wang, S. Chen, and M. Sonka, "Graph-based segmentation of abnormal nuclei in cervical cytology," *Comput. Med. Imag. Graph.*, vol. 56, pp. 38–48, 2017.

[15] L. Zhang, M. Sonka, L. Lu, R. M. Summers, and J. Yao, "Combining fully convolutional networks and graph-based approach for automated segmentation of cervical cell nuclei," in *Proc. 2017 IEEE 14th Int. Symp. Biomed. Imag.*, 2017.

[16] E. Martin, "Pap-smear classification," M. S. Thesis, Tech. Univ. Denmark, Kongens Lyngby, Denmark, 2003.

[17] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," *Nature Inspired Smart Inf. Syst.*, pp. 1–9, 2005. [Online]. Available: http://orbit.dtu.dk/en/publications/papsmear-benchmark-data-for-pattern-classification(2308e608-62e6-47ee-8ec8-828ba479df0a)/export.html

[18] Y. Marinakis, M. Marinaki, and G. Dounias, "Particle swarm optimization for pap-smear diagnosis," *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1645–1656, 2008.

[19] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Comput. Biol. Med.*, vol. 39, no. 1, pp. 69–78, 2009.

[20] M. E. Plissiti and C. Nikou, "On the importance of nucleus features in the classification of cervical cells in pap smear images," in *Proc. Int. Workshop Pattern Recognit. Health Anal.*, 2012, p. 11.

[21] K. Bora, M. Chowdhury, L. B. Mahanta, M. K. Kundu, and A. K. Das, "Automated classification of pap smear images to detect cervical dysplasia," *Comput. Methods Programs Biomed.*, vol. 138, pp. 31–47, 2017.

[22] L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis," *Artif. Intell. Med.*, vol. 49, no. 2, pp. 117–125, 2010.

[23] Y. Guo, G. Zhao, and M. Pietikälnen, "Discriminative features for texture description," *Pattern Recognit.*, vol. 45, no. 10, pp. 3834–3843, 2012.

[24] B. Sokouti, S. Haghipour, and A. D. Tabrizi, "A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and thinPrep histopathological cell image features," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 221–232, 2014.

[25] J. Zhang and Y. Liu, "Cervical cancer detection using svm based feature screening," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Springer, New York, NY, USA, 2004, pp. 873–880.

[26] M. Zhao, A. Wu, J. Song, X. Sun, and N. Dong, "Automatic screening of cervical cells using block image processing," *Biomed. Eng. Online*, vol. 15, no. 1, 2016. [Online]. Available: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-016-0131-z

[27] D. Solomon and R. Nayar, *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. New York, NY, USA: Springer Science & Business Media, 2004.

[28] M. Desai, "Role of automation in cervical cytology," *Diagnostic Histopathol.*, vol. 15, no. 7, pp. 323–329, 2009.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[33] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[34] R. M. Summers, "Progress in fully automated abdominal CT interpretation," *Amer. J. Roentgenol.*, vol. 207, no. 1, pp. 67–79, 2016.

[35] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[36] H. R. Roth *et al.*, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.

[37] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[38] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.

[39] T. Xu *et al.* "Multi-feature based benchmark for cervical dysplasia classification evaluation," *Pattern Recognit.*, vol. 63, pp. 468–475, 2017.

[40] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Springer, New York, NY, USA, 2016, pp. 115–123.

[41] P. Buyssens, A. Elmoataz, and O. Lézoray, "Multiscale convolutional neural networks for vision–based classification of cells," in *Proc. Asian Conf. Comput. Vis,*. Springer, New York, NY, USA, 2012, pp. 342–352.

[42] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "Hep-2 cell image classification with deep convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 416–428, Mar. 2017.

[43] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[44] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Proc. 2015 IEEE 12th Int. Symp. Biomed. Imag.*, 2015, pp. 294–297.

[45] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[46] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[47] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[48] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Springer, New York, NY, USA, 2015, pp. 652–660.

[49] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, 1998.

[50] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[51] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 550–566, Feb. 2016.

[52] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[53] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. New York, NY, USA: Springer, 2012, pp. 9–48.

[54] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580, 2012.

[55] Y. Jia *et al.* "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[56] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

Authors' photographs and biographies not available at the time of publication.