DCCL: A Benchmark for Cervical Cytology Analysis

Changzheng Zhang[1], Dong Liu[4], Lanjun Wang[2], Yaoxin Li[1], Xiaoshi Chen[1], Rui Luo[3], Shuanlong Che[4], Hehua Liang[4], Yinghua Li[4], Si Liu[4], Dandan Tu[1], Guojun Qi[3], Pifu Luo[4(✉)], and Jiebo Luo[5(✉)]

[1] Huawei, Shenzhen, China
[2] Huawei Canada, Markham, Canada
[3] Futurewei, Bellevue, USA
[4] KingMed Diagnostics Co., Ltd., Guangzhou, China
gz-luopf@kingmed.com.cn
[5] University of Rochester, Rochester, USA
jluo@rochester.edu

Abstract. Medical imaging analysis has witnessed impressive progress in recent years thanks to the development of large-scale labeled datasets. However, in many fields, including cervical cytology, a large well-annotated benchmark dataset remains missing. In this paper, we introduce by far the largest cervical cytology dataset, called Deep Cervical Cytological Lesions (referred to as DCCL). DCCL contains 14,432 image patches with around $1{,}200 \times 2{,}000$ pixels cropped from 1,167 whole slide images collected from four medical centers and scanned by one of the three kinds of digital slide scanners. Besides patch level labels, cell level labels are provided, with 27,972 lesion cells labeled based on The 2014 Bethesda System and the bounding box by six board-certified pathologists with eight years of experience on the average. We also use deep learning models to generate the baseline performance for lesion cell detection and cell type classification on DCCL. We believe this dataset can serve as a valuable resource and platform for researchers to develop new algorithms and pipelines for advanced cervical cancer diagnosis and prevention.

Keywords: Cervical cancer screening · Liquid-based cytology · Deep learning

1 Introduction

Cervical cancer is one of the most common cancers among women and ranks fourth in terms of incidence and mortality, with approximately 570,000 cases and 311,000 deaths in 2018 worldwide. Moreover, in many developing countries, this disease has been even more commonly diagnosed but has the highest death

C. Zhang and D. Liu—Equal contribution.

H.-I. Suk et al. (Eds.): MLMI 2019, LNCS 11861, pp. 63–72, 2019.
https://doi.org/10.1007/978-3-030-32692-0_8

rate among cancers [3]. Nevertheless, cervical cancer is preventable and can be cured in the early stage as it can be largely detected by cytological screening combined with human papillomavirus virus (HPV) testing. When the patients are infected by HPV, the cervical epithelial cells result in different morphological changes, together with loss of maturity of the squamous epithelial cells and abnormal proliferation. The process is called dysplasia and manifests as loss polarity of the squamous cells, nuclear enlargement, coarse and hyperchromatic nuclei, as well as nuclear condensation. These phenomenons always indicate a higher probability of progressing to cervical cancer.

The application of liquid-based cytology test has greatly improved the diagnosis rate of precancerous and cancerous lesions at the cell level and has become one of the most important methods for cervical cancer diagnosis and prevention. Based on The 2014 Bethesda System for Reporting Cervical Cytology (2014 TBS) [12], the precancerous squamous intraepithelial lesions include four types with an increasing level of severity: atypical squamous cells of undetermined significance (ASC-US), low squamous intraepithelial lesion (LSIL), atypical squamous cell-cannot exclude HSIL (ASC-H) and high squamous intraepithelial lesion (HSIL); while cancerous lesions include mainly two types: squamous cell carcinoma (SCC) and adenocarcinoma (AdC). However, challenges exist when assessing cytology tests. First, it is time-consuming to examine a gigapixel pathological slide that contains thousands of cervical cells. The extremely low ratio of pathologists to patients has become a bottleneck of cervical cancer screening, especially in developing countries. Second, the diagnosis of precancerous and cancerous lesions is highly uncertain, subject to the experiences of pathologists.

To tackle the challenges, it would be highly valuable to develop automatic cervical cytology analysis models. Based on existing cervical cytology datasets [1, 8], several works have been done for lesion cell classification [8,19,20], as well as cytoplasm and nuclei segmentation [16]. Nevertheless, current cervical cytology datasets typically contain a few thousand lesion cells. To facilitate future cervical cytology analysis, we introduce by-far the largest and densely labeled cervical digital pathology dataset, namely DCCL.

The contributions of this work are summarized as follows:

- We introduce a large-scale cervical cytology dataset called DCCL. To the best of our knowledge, this is the largest cervical cytology dataset. We crop a total of 14,432 image patches from 1,167 whole slides. The number of slides is ten times larger than the previous benchmark datasets [13,18].
- We release 27,972 lesion cell bounding boxes ranging from low-grade precancerous lesions to cancerous lesions, and 6,420 semi-supervised labeled negative cell samples (model results). The overall number of cells is three times larger than the previous benchmark datasets [8,13,18].
- We provide benchmark performance on lesion cell classification and detection by leveraging widely used deep neural network models.
- We visualize the lesion cell similarity map to facilitate the understanding of inter-class and intra-class relationships via t-SNE.

2 Related Work

We roughly divide current datasets for cervical cytology analysis into two groups based on their target usages: (i) for lesion cell classification, and (ii) for cytoplasmic and nuclei segmentation.

Lesion Cell Classification. The most well-known dataset is Papanicolaou (Pap) smear based Herlev Dataset [2,8] collected by microscopes and digital cameras. In Herlev, a cell image is categorized into four types: NILM (negative for intraepithelial lesion or malignancy), LSIL, HSIL and SCC based on Bethesda standard [2]. Besides, there are CerviSCAN dataset results from CerviSCAN project [18] for low-cost cervical cancer screening, and HEMLBC Dataset [19] relies on liquid-based cytology. Details of these datasets are shown in Table 3.

We observe that the performance of recent lesion classification algorithms become saturated with these datasets [4,20]. However, those classifiers cannot achieve comparable performance in practice, because the existing datasets have limited variations on lesion types, cell morphological structures, and background clutters. Thus, a challenging dataset is required to facilitate future cervical cytology analysis for clinical applications.

Cytoplasm and Nuclei Segmentation. For cytoplasm segmentation, Shenzhen University (SZU) Dataset [16] consists of 21 cervical cytology images in seven clumps with 6.1 cells per clump on the average. For nuclei segmentation, in ISBI 2015 challenge [1], there are eight cervical cytology images, each of which has 20–60 Pap stained cervical cells.

A recent study leverages the above two to formulate the segmentation problem with a graphical model, which can take more geometric information into account and generate an accurate prediction [16]. However, due to the small size, these datasets have not been used to design any deep learning model.

3 DCCL

3.1 Collection Methodology

There are 1,167 specimens of cervical cytology from participants whose ages are in the range of 32 to 67. The specimens are prepared by Thinprep methods stained with Papanicolaou stain, which were collected by four provincial medical centers from 2016 to 2018. The gathered slides to generate DCCL include 933 positive patients and 234 normal cases. The slide labels are from pathology reports. All of the slides are scanned evenly by one of three kinds of digital slide scanners (Nanozoomer2.0HT, KFBIO KF-RPO-400, and AperioAT2) all with 200× zoom and in 24-bit color.

We cut each slide image in grids, where rectangular areas are around 1200 × 2000 pixels (with a physical size of 1011.6 μm × 606.96 μm). Generally, a slide is converted to 700–800 patches. In detail, DCCL is composed of 14,432 image patches without cell-free, out-of-focus, blur, fade or any bubble inside. It includes 9,930 positive image patches which have precancerous and/or cancerous lesions

and 4,502 negative image patches which are from normal cases. The details about the slide distribution and patch distribution are shown in Table 1. It is noted that: (i) all of the data used in our study are strictly anonymized; (ii) the types of slides, as well as patches, are from the diagnosis of pathologists; (iii) "unlabeled" and "labeled" in Table 1 indicates whether a patch is labeled manually on the cell level, the process of which is to be explained in Sect. 3.2.

Table 1. Statistics on slides and patches by types

	Train			Val			Test			Total			Ratio		
	Slide	Patch		Slide	Patch		Slide	Patch		Slide	Patch		Slide	Patch	
		Unlabeled	Labeled		Unlabeled	Labeled		Unlabeled	Labeled		Unlabeled	Labeled		Unlabeled	Labeled
NILM	117	2301	0	46	812	0	71	1389	0	234	4502	0	20%	31%	0%
ASC-US	67	0	203	27	0	61	41	0	105	135	0	369	12%	0%	3%
ASC-H	84	1227	295	34	264	278	51	493	389	169	1984	962	14%	14%	7%
LSIL	197	3	1683	79	9	579	119	12	884	395	24	3146	34%	0%	22%
HSIL	90	988	466	36	48	210	55	40	253	181	1076	929	16%	7%	6%
SCC	14	406	154	5	158	42	8	152	111	27	716	307	2%	5%	2%
AdC	13	130	115	5	45	23	8	81	23	26	256	161	2%	2%	1%
Total	582	5055	2916	232	524	1193	353	778	1765	1167	8558	5874	100%	59%	41%

3.2 Image Annotation

We extract 5,874 patches from DCCL to label manually, and the rest are unlabeled. The details of the labeled data distribution and unlabeled data distribution, as well as the patch distribution on training, validation and testing sets, are shown in Table 1. It is noted that DCCL contains unlabeled patches, because they may help to promote semi-supervised and unsupervised learning, even transfer learning for intelligent cervical cytology analysis.

In each patch, a lesion cell is annotated with its type based on 2014 TBS [12] along with a bounding box. There are 27,972 lesion cells labeled manually, including ASC-US, LSIL, ASC-H, HSIL, SCC and AdC, a total of six types of lesion cells. The details of the labeled lesion cells are in Table 2 and some examples are shown in Fig. 1. There are six board-certified pathologists participated in the annotation task. In the initial blind reading phase, each patch is annotated by two pathologists independently to mark all suspicious lesion cells. Next, every pair of two bounding boxes with consistent lesion type from different readers are merged by average if the Intersection over Union (IoU) is larger than 0.3. That is to say, we skip the bounding boxes marked by only one pathologist to keep high quality, and the merged bounding box together with its lesion type is the final annotation on a cell.

Nevertheless, if DCCL only contains lesion cell bounding boxes, the analysis result may have a bias on positive samples. Hence, DCCL also considers negative cells by leveraging those 4,502 negative patches where none lesion cell exists. However, it would introduce large data bias to add all negative cells in the dataset because they are in a huge amount and easy to be distinguished from lesion cells. In order to be beneficial for cervical cytology analysis algorithms, what we need are hard cases which are negative cells but easy to be recognized as positives, such as NILMs in Fig. 1. With these hard negative cells, the community

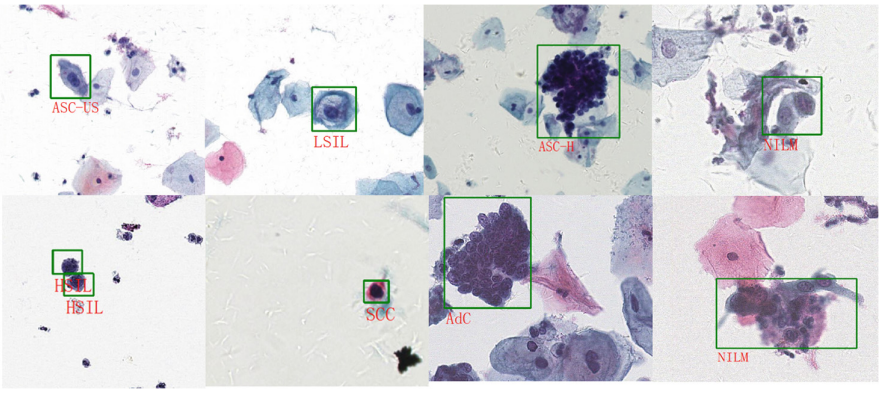

Fig. 1. Examples of cells

can focus on some inherent challenges of cervical cytology recognition, such as intra-class variance (e.g., parts of LSIL cells have a sharply defined perinuclear cavity, but the rests do not have) and inter-class similarity (e.g., both HSIL and SCC have high nuclear to cytoplasmic ratios).

In the study, we leverage the widely used detection methods trained with labeled positive patches and regard those bounding boxes on negative patches with high false positive probabilities to be hard negatives. In detail, Faster R-CNN and RetinaNet are trained on 5,874 labeled positive patches, respectively. We test them on 4,502 negative patches and get 6,420 bounding boxes whose average positive probability obtained by two methods is over 0.2. The distribution of the generated hard negatives (a.k.a. NILM) is also shown in Table 2.

Table 2. Statistics on training/validation/test sets by types

	Train	Val	Test	Total		Ratio	
NILM	2588	1540	2292	6420	6420	19%	19%
ASC-US	2471	838	1378	4687	27972	14%	81%
ASC-H	1147	543	591	2281		7%	
LSIL	1739	356	595	2690		8%	
HSIL	5890	1807	3482	11179		33%	
SCC	3006	1225	2731	6962		20%	
AGC	122	20	31	173		1%	
Total	16963	6329	11100	34392	34392	100%	100%

3.3 Dataset Statistics

We analyze the properties of DCCL by comparing with other widely used datasets, including CerviSCAN [18], Herlev Dataset [8], ISBI 2015 Challenge

Dataset [1], HEMLBC [19], Shenzhen Second People's Hospital (SSPH) Dataset [11] and Cervix93 [13]. Table 3 shows that they are different in the targeted task type, data size and diversity, as well as types of lesions and accessibility. Take the task type as an example, CerviSCAN, Herlev Dataset and Cervix93 are only designed for cell type classification, where samples are cut from original slides without context information, while ISBI 2015 focuses on cytoplasm and nuclei segmentation; SSPH and HEMLBC are targeted lesion cell detection. However, DCCL can be used for both cell type classification and lesion cell detection.

The volume of DCCL is also illustrated in Table 3. In total, there are 1,167 patients, 14,432 image patches and 34,392 bounding boxes. Table 3 shows that the number of patients of DCCL is more than 10× larger compared with CerviSCAN and Cervix93. Besides, the number of lesion types of DCCL is also larger than the others. To sum up, DCCL is challenging due to diversities on geographical data sources, types of digital slides scanners, patient ages, types of lesion cells and background clutters. More importantly, all of these diverse factors are essential to building a robust and reliable clinical application system. Furthermore, it is noted that DCCL will be open upon the acceptance of this paper.

Table 3. Properties of cervical cytology datasets.

Dataset	Num. of Patients	Num. of labelled patch	Num. of labelled cells	Num. of lesion cell types	Cell Classification	Cell Detection	Cell Segmentation	Open
CerviSCAN [18]	82	>900	12043	3	✓	×	×	✓
ISBI 2015 [1]	–	961	–	–	×	×	✓	✓
SSPH [11]	500	5721	10307	5	✓	✓	×	×
Herlev [8]	–	–	917	3	✓	×	×	✓
HEMLBC [19]	200	–	2370	4	✓	✓	✓	×
Cervix93 [13]	93	–	2705	2	✓	×	×	✓
DCCL	**1167**	**14432**	**34392**	**6**	✓	✓	×	✓

4 Experiments Results

4.1 Lesion Cell Detection

Baseline Detection Model. Our baseline detectors are Faster R-CNN [14] and RetinaNet [10] that represent two-stage algorithms and one-stage algorithms, respectively. Both of them are based on a ResNet-50 [6] backbone network. More detailed implementations are referred to Appendix 6.1.

Evaluation Metrics. We follow the evaluation metric used for Pascal VOC [5] and MS COCO [9], which is mean average precision (mAP). Since DCCL cell detection is a much more challenging task, the performance of the trained detectors in our experiments is evaluated at lower IoU values: mAP@0.1:0.2:0.5.

Results. Table 4 illustrates the results for fine-grained cell detection based on Faster R-CNN and RetinaNet. In the experiments, we extract positive cell

anchors as positive anchors, while two types of negative anchors: negative cell anchors and background anchors (without any cells). We observer that negative cell anchors and positive cell anchors tend to have some smaller variations. However, the percentage of negative cell anchors are extremely low as compared with the percentage of background anchors, which makes the model hard to distinguish them from positive cell anchors, even using the focal loss function in RetinaNet, and causes high false positive detection. As a result, Table 4 shows mAP@0.5, mAP@0.3, and mAP@0.1 do not present a huge boosting with IoU values decreasing, which indicates DCCL is a challenging benchmark.

Table 4 also illustrates the results for coarse-grained cell detection, in which Faster R-CNN and RetinaNet output two classes of lesion cells and negative cells, as well as the corresponding bounding boxes. The same reason for low mAPs is that methods both generate a large number of false positives.

To improve low mAPs in Table 4, a potential method is to attach a false positive reduction module after the lesion cell detection module, which has been proved to be feasible in LUNA2016 challenge [15]. How to combine these two models would be our future work.

Table 4. Detection evaluation on DCCL

Method	Metrics	Fine-grained cell							Coarse-grained cell
		mAP	ASCUS	LSIL	ASCH	HSIL	SCC	AGC	mAP
Faster R-CNN	mAP@0.1	21.16	25.91	27.08	17.46	16.68	14.03	25.81	26.16
	mAP@0.3	19.8	24.3	24.95	16.58	14.09	13.1	25.81	22.18
	mAP@0.5	17.1	21.01	20.46	14.1	10.73	10.41	25.71	19.35
Retina-Net	mAP@0.1	19.61	23.41	25.79	14.11	16.64	15.09	22.64	24.01
	mAP@0.3	18.37	21.98	23.48	13.22	14.01	14.89	22.64	21.85
	mAP@0.5	15.93	18.71	19.89	11.86	10.08	12.67	22.39	18.07

4.2 Cell Type Classification

Baseline Classification Model. In order to improve cell type classification, we include six lesion cell types plus the negative cell types to conduct deep learning based classification experiments.

The experiments are conducted on three different CNN architectures: Inception-v3 [17], ResNet-101 [6] and DenseNet-121 [7]. According to 2014 TBS, determining the type of a cell depends on multiple factors. Besides the properties of an individual cell (e.g., cell size, nuclear size, nuclear to cytoplasmic ratio, irregularity of cytoplasm, chromatin and nuclear, etc.), the contextual factors by comparing the cell with the other cells in the same slide also play an important role. Thus, we set up a series of experiments on DenseNet-121 to explore local, context as well as geometric information, where "local" only includes the bounding box area of cytologic, "context" indicates a dilated area around the bounding box, and "deform" indicates the network considers an extra deformable layer to model geometric transformations of cells. The other detailed implementations are referred to Appendix 6.2.

Table 5. Classification results on test set with different training mode

Model	Local/context information	Accuracy	F-score	Precision	Recall
ResNet-101	Local	86.92%	48.32%	48.70%	47.94%
Inception-v3	Local	87.38%	50.89%	50.30%	51.50%
DenseNet-121	Local	87.87%	52.22%	51.40%	53.07%
	Local+Deform	87.73%	54.51%	54.35%	54.66%
	Context	87.90%	53.66%	51.02%	56.59%
	Context+Deform	88.52%	57.33%	53.78%	61.39%
	Local+Deform+Context+Deform	88.84%	59.96%	58.88%	61.08%

Evaluation Metrics. We follow the evaluation metrics in [2], including both accuracy and F-score. Precision and recall are also included for reference.

Results. As shown in Table 5, for models trained based on local information, we observe that DenseNet-121 [7] obtains the best result compared with Inception-v3 [17] and ResNet-101 [6]. Meanwhile, by leveraging the context information, DenseNet-121 only enhances the accuracy and F-score by 0.13% and 1.44%, respectively. By adding the deformable layer on "local" and "context", F-scores increase 2.31% and 3.67% comparing with no deformable layer, respectively. Moreover, the "Local+Deform" and "Context+Deform" ensemble can achieve the F-score as high as 59.96%. These results suggest the context and geometric information provide useful clues to improve cell type classification.

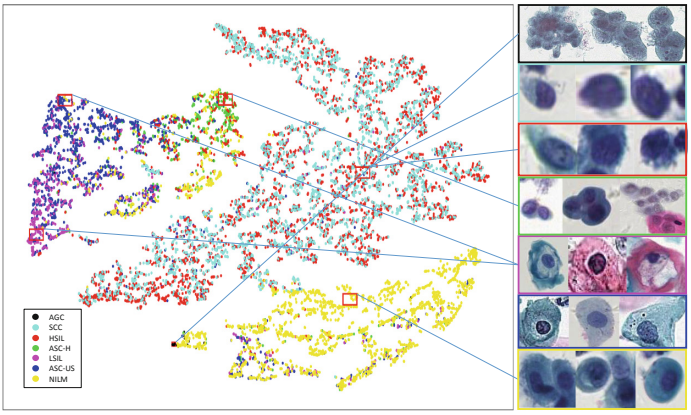

Fig. 2. t-SNE visualization of the lesion cell embeddings

Moreover, lesion cell clusters are visualized via t-SNE in Fig. 2 by deriving test set feature embeddings based on DenseNet-121 with "Context+Deform" training mode. This figure shows that the cells at adjacent lesion levels tend to have similar visual appearances, thus are hard to be distinguished: e.g. NILM

vs. ASC-US, ASC-US vs. LISL, and HSIL vs. SCC. This is predictable given the progression of a cell lesion is gradual without clear boundaries. A better approach to analyzing fine-grained cervical cytology remains an open problem for future research.

5 Conclusions

In this paper, we present a new benchmark dataset to enable cervical cytology analysis for future research and clinical studies on cervical cancer screening. The presented DCCL is highlighted with three distinguishing features. First, compared with the existing cervical cytology datasets, DCCL has a larger scale with diverse fine-grained types of lesion cells. Second, it provides full annotations on six types of lesion cells, ranging from low-grade precancerous lesions to cancerous lesions. A large volume of hard negative cell bounding boxes is also provided. Finally, the baseline lesion cell classification and detection results on DCCL are reported. This dataset presents inherent challenges for cervical cytology recognition, such as intra-class variance (e.g., parts of LSIL cells have a sharply defined perinuclear cavity, but the rests do not have) and inter-class similarity (e.g., both HSIL and SCC have high nuclear-to-cytoplasmic ratios), which are ubiquitous in real clinical studies. These labels on fine-grained types of cell lesions play a crucial role in the screening and diagnosis of cervical cancer. The dataset will be released upon the acceptance of this paper.

References

1. https://cs.adelaide.edu.au/simcarneiro/isbi15_challenge/
2. Bora, K., et al.: Pap smear image classification using convolutional neural network. In: 10th ICVGIP, p. 55. ACM (2016)
3. Bray, F., et al.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **68**(6), 394–424 (2018)
4. Chen, Y.F., et al.: Semi-automatic segmentation and classification of pap smear cells. IEEE J. Biomed. Health Inform. **18**, 94–108 (2014)
5. Everingham, M., et al.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
6. He, K., et al.: Deep residual learning for image recognition. In: CVPR (2016)
7. Huang, G., et al.: Densely connected convolutional networks. In: CVPR (2017)
8. Jantzen, J., et al.: Pap-smear benchmark data for pattern classification. NiSIS (2005)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Lin, T.-Y., et al.: Focal loss for dense object detection. In: CVPR (2017)
11. Meiquan, X., et al.: Cervical cytology intelligent diagnosis based on object detection technology (2018)
12. Nayar, R., et al.: The Pap Test and Bethesda 2014. Acta Cytologica (2015)

13. Phoulady, H.A., et al.: A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection. arXiv preprint arXiv:1811.09651 (2018)
14. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks (2015)
15. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Med. Image Anal. **42**, 1–13 (2017)
16. Song, Y., et al.: Automated segmentation of overlapping cytoplasm in cervical smear images via contour fragments. In: AAAI 2018 (2018)
17. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
18. Tucker, J.: CERVISCAN: an image analysis system for experiments in automatic cervical smear prescreening. Comput. Biomed. Res. **9**(2), 93–107 (1976)
19. Zhang, L., et al.: Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining. Cytometry Part A **85**, 214–230 (2014)
20. Zhang, L., et al.: DeepPap: deep convolutional networks for cervical cell classification. IEEE J. Biomed. Health Inform. **21**, 1633–1643 (2017)