


## ORIGINAL ARTICLE

# Deep learning based digital cell profiles for risk stratification of urine cytology images

Ruqayya Awan<sup>1</sup>  | Ksenija Benes<sup>2</sup> | Ayesha Azam<sup>1,3</sup> | Tzu-Hsi Song<sup>1,4</sup> |  
 Muhammad Shaban<sup>1</sup> | Clare Verrill<sup>5</sup> | Yee Wah Tsang<sup>3</sup> | David Snead<sup>3</sup> |  
 Fayyaz Minhas<sup>1</sup> | Nasir Rajpoot<sup>1,3,6</sup>

<sup>1</sup>Department of Computer Science, University of Warwick, Coventry, UK

<sup>2</sup>The Royal Wolverhampton NHS Trust, Wolverhampton, UK

<sup>3</sup>Department of Pathology, University Hospitals Coventry and Warwickshire, Coventry, UK

<sup>4</sup>Laboratory of Quantitative Cellular Imaging, Worcester Polytechnic Institute, Worcester, Massachusetts, USA

<sup>5</sup>Nuffield Department of Surgical Sciences and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK

<sup>6</sup>The Alan Turing Institute, London, UK

## Correspondence

Ruqayya Awan, Department of Computer Science, University of Warwick, Coventry, UK. Email: awanruqayya2@gmail.com

## Funding information

Warwick Impact Fund; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/N510129/1; UK Research and Innovation

## Abstract

Urine cytology is a test for the detection of high-grade bladder cancer. In clinical practice, the pathologist would manually scan the sample under the microscope to locate atypical and malignant cells. They would assess the morphology of these cells to make a diagnosis. Accurate identification of atypical and malignant cells in urine cytology is a challenging task and is an essential part of identifying different diagnosis with low-risk and high-risk malignancy. Computer-assisted identification of malignancy in urine cytology can be complementary to the clinicians for treatment management and in providing advice for carrying out further tests. In this study, we presented a method for identifying atypical and malignant cells followed by their profiling to predict the risk of diagnosis automatically. For cell detection and classification, we employed two different deep learning-based approaches. Based on the best performing network predictions at the cell level, we identified low-risk and high-risk cases using the count of atypical cells and the total count of atypical and malignant cells. The area under the receiver operating characteristic (ROC) curve shows that a total count of atypical and malignant cells is comparably better at diagnosis as compared to the count of malignant cells only. We obtained area under the ROC curve with the count of malignant cells and the total count of atypical and malignant cells as 0.81 and 0.83, respectively. Our experiments also demonstrate that the digital risk could be a better predictor of the final histopathology-based diagnosis. We also analyzed the variability in annotations at both cell and whole slide image level and also explored the possible inherent rationales behind this variability.

## KEYWORDS

cell classification, cell detection, cell segmentation, deep learning, digital risk, oversampling, The Paris System, urine cytology

## 1 | INTRODUCTION

Bladder cancer is known to be the ninth most commonly occurring malignancy globally, with around 430,000 new cases reported in 2012

[1]. Urine cytology is considered to be an important detection tool for identifying malignancies in the urinary tracts such as bladder cancer. It is widely used to identify high-grade urothelial cancer (HGUC) and is not encouraged to be used for low-grade carcinoma due to its low

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Cytometry Part A* published by Wiley Periodicals LLC. on behalf of International Society for Advancement of Cytometry.

sensitivity to it. In clinical practice, pathologists observe cytology slides under the microscope and identify atypical and malignant cells. Based on the morphology of these cells, a diagnosis is made leading to decision making for treatment.

Unlike histology, the digital adoption for urine cytology has been impeded due to the lack of scanner's ability for z-stacking along with other limitations related to cytology. The tissue material for histology has a relatively uniform thickness whereas the cytology material is less evenly distributed with variable thickness of different cell clusters in a 3D configuration. For this reason, pathologists would frequently need to focus on different planes to view all the cells. It has been demonstrated that the availability of more than one focal plane on digital cytology slides helps with the diagnostic interpretation [2]. Z-stacking enables the user to look at the sample at different focal planes which is a built-in property of the microscope. With the advancement in whole slide scanners, different vendors have started to provide imaging system with an ability for z-stacking which has motivated the pathologists to scrutinize digital cytology in clinical practice. However, it comes with a cost of much larger image file size and longer scanning time [3].

Similar to histology, diagnosis of cytology cases suffer from high inter and intraobserver variability [4]. In addition to variability in the assessment of urine cytology, different terms for the same entities are being used at both individual pathologists and institutional level. This led to the development of The Paris System (TPS) to provide a consistent and reliable diagnostic tool. An international working group, comprising of expert cytopathologists, urologists, and surgical pathologists, provided criteria for reporting different diagnostic categories including recommendations for HGUC which is the main purpose of urine cytology. TPS was officially released in 2016 [5] and is now accepted worldwide. TPS has shown significant improvement in the assessment of urine cytology specimens with adequate precision for negative cases. However, studies [6–9] conducted on the interobserver variability demonstrated poor interobserver agreement for other categories. In [9], different distribution of categories was reported by five cytopathologists on reviewing 149 cases independently. The interobserver variation makes the automated diagnosis of cytology samples challenging.

In a clinical setting, a cytology specimen is examined manually, under a microscope using a glass slide. Like histology samples, urine cytology slides can be visualized on a computer screen after digitization which is used by occasional labs. The uptake of digital cytology can encourage the assisted assessment of specimen with computer-generated results. This would result in the emergence of quantitative algorithms for analysis, hence enabling the clinicians to obtain non-subjective and reproducible outcomes.

The main goal of this study is to investigate an automated alternative to risk stratification of urine cytology slides. The status quo based on subjective visual analysis is prone to human error and has a large inter- and intraobserver variability. Therefore, there is a need to investigate its limitations wrt the intrinsic difficulty of the problem (in both diagnostic and technical terms). Our contribution in this paper is five-fold. First, we collected cell-level annotations in an iterative way to

improve the generalizability of the model. TPS criteria were used by the expert pathologists for labeling. Second, we explored two different approaches for cell detection and classification and employed the best one for whole slide image (WSI) labeling. Third, we presented a cell count-based approach for identifying high-risk cases. Fourth, we investigated the interobserver agreement at WSI level and intraobserver variability at the cell level. Lastly, we investigated the cytopathology based risk category and our digital risk labeling in correlation with the “gold standard” histopathology based diagnosis.

In the next section, we review previous work on cytology images. In Section 2, we describe the details of our dataset and our methodology for cell segmentation, detection and classification. In Section 3, we present our results at both cell and WSI level. In Section 4, we discuss our findings while Section 5 concludes this study.

## 1.1 | Related work

In the literature, very few studies can be found on automatic analysis of cytology images in comparison to the work in histology image analysis. Recently, there has been some work on cell detection, classification, and segmentation from cytology images. In [10], GoogleNet and AlexNet models have been used to distinguish between benign and malignant microscopic images of breast cytological samples obtained with fine needle technique. The training and validation dataset was collected by extracting overlapping patches (comprising number of cells) from region of interests (ROIs) selected by the pathologist. Zhang et al. [11] presented a simple convolutional neural network (CNN) to classify cervical cells in a pap-smear cytology image without any prior cell segmentation. Their training set comprises patches of fixed size with nucleus located in the center of the patch. This means that the network was trained with patches containing partial cell content. In another study [12], a simple CNN is used to classify the cells in nasal cytology into one of the seven classes. To train the network with patches containing whole cell content, they perform cell segmentation via the Otsu algorithm followed by morphological operation and watershed algorithm. To overcome the problem of unbalanced classes, they opted random majority undersampling method. Wu et al. [13] employed AlexNet-based network to identify different types of ovarian cancer from cytological images captured from different parts of the tissue sample. These images were then divided and resized into smaller patches for training.

One recent study [14], which integrates deep learning and morphometric approaches, focuses on automating TPS for the analysis of urine cytology images. Deep learning is used to assign atypia score to a given cell while a morphometric approach computes the nucleus to cytoplasmic ratio. They employed thresholding to segment cellular content, followed by connected component analysis for extracting cell patches. Based on their cell classification approach, they have proposed a condensed grid format for an image reconstruction which is less cellular and smaller in size in comparison to the original image. The authors have also illustrated the prediction of high-risk cases based on the cutoff for their employed cell morphological features.

Sanghvi et al. [15] presented a deep learning-based pipeline for classifying urine cytology images into five TPS categories which can further be divided into low and high-risk classes. QuPath was used to detect cells in a WSI and a patch of fixed size was extracted from the center. The authors employed both cell-level and slide-level features for WSI classification and validated it using a large cohort. To the best of our knowledge, [14, 15] are the only studies on risk stratification.

There has been some effort in separating the overlapping cells from both 1-plane and z-stacked cytology images and is not limited to [16–19] and [20]. In our study, we perform segmentation to extract both individual cells and the cluster of cells. This is to ensure that the whole cell or a cluster is captured inside the bounding box. Therefore, separating the overlapping cells is not necessary for our approach.

## 2 | MATERIALS AND METHODS

Atypical and malignant cells are of interest to the pathologist among various types of cells and contaminant found in a urine cytology sample. To discriminate between low and high-risk WSIs, we first identify all the candidate atypical and malignant cells. For our experiments, we employed two deep learning-based approaches for the identification of these cells. We chose the best performing method for our WSI-level classification by setting a threshold on the proportion of the sum of atypical and malignant cells.

### 2.1 | Specimen collection, digitization, and labeled data preparation

The cytology slides used in this study and the associated clinical data were obtained from the University Hospitals Coventry and Warwickshire (UHCW) NHS Trust in Coventry, UK. The dataset was provided after deidentification and informed consent was obtained from the patients. Each slide was labeled as normal, inflammatory, cytological atypia (CA), atypia suspicious for malignancy (ASM), or transitional cell carcinoma (TCC). In this paper, we use the term “reference” for the diagnostic information obtained from the UHCW and does not necessarily mean that it was decided by a single pathologist. All the slides were prepared using a liquid based cytology method, ThinPrep and were scanned at 0.275 mm per pixel. The maximum resolution is 40x. In total, we obtained 398 slides, comprising 243 normal, 13 inflammatory, 76 CA, 38 ASM and 28 TCC. These slides were scanned using an Omnyx VL120 scanner to form a multilayered pyramid enabling the user to visualize the slide at different resolutions.

#### 2.1.1 | Creation of labeled dataset

We obtained cell-level annotations from an experienced pathologist and a recently trained pathologist. Both pathologists followed TPS

criteria for labeling cells as normal, atypical, or malignant urothelial cells. Other cell types present in urine (e.g., squamous, inflammatory, etc.) were also annotated. Degenerated cells and cells that pathologists were uncertain about were annotated as “others.” The variations in annotations affect the performance of a trained classifier. We did the interobserver variability analysis between two pathologists to find out the highly concordant classes. A set of same visual fields were presented to both pathologists for independent annotations. High concordance score was observed in normal, squamous, and inflammatory classes. Considering the variability in the rest of the classes, we expanded our labeled dataset by presenting different visual fields to them. We asked the trained pathologist to annotate samples of normal, squamous, and inflammatory only while the experienced pathologist annotated all the classes.

For network training and validation, annotations were obtained on the WSIs via a web-based interface. The pathologists marked individual cells using a dot in the center of the cell and for the cell clusters, pathologists draw polygon or rectangle around clusters. The annotations were obtained from pathologists at resolution level 40x. The details of the sample split to train our initial network are shown in Supplementary Table 1. More annotations were added to this dataset while verifying the network predictions by the expert pathologist in an iterative manner, as shown in Supplementary Figure 1. During the verification process, 747 normal, 2185 squamous, 2408 others, 2073 debris, 117 inflammatory, 279 atypical, and 88 malignant cells were added to the training set while 163 normal, 544 squamous, 492 others, and 511 debris cells were added to the validation set.

#### 2.1.2 | Balancing of labeled dataset

The dataset obtained after initial annotations suffered unbalanced distribution. Due to unbalanced classes, a classifier does not tend to perform well for the minority classes as it does not get sufficient look at them. To balance the distribution in the training set, we employed oversampling technique known as synthetic minority over-sampling technique (SMOTE) [21]. In our initial dataset, it was atypia and debris class for which most of the annotations were obtained from the expert pathologist. For network training, we kept 500 samples per class in the validation set and the remaining samples in the training set. Except for the atypia class, all other classes were oversampled until the total number of patches per class including original and over-sampled patches were equal to the number of patches in atypia class (4558). Since the dimensionality of patches is not constant due to various sizes of the cell, we applied SMOTE on patches in batches, with each batch having patches of similar sizes.

The SMOTE technique generates new samples by manipulating the feature space by joining the line segments between each minority class sample and its  $k$  nearest neighbors. This is done by computing the difference between the minority sample and its nearest neighbor, followed by multiplying the difference with any random number between 0 and 1. For our experiments, we selected five nearest

neighbors for synthesis. Examples of images of synthetic cells generated by the SMOTE technique are shown in Supplementary Figure 2.

## 2.2 | ROI extraction from whole slide image

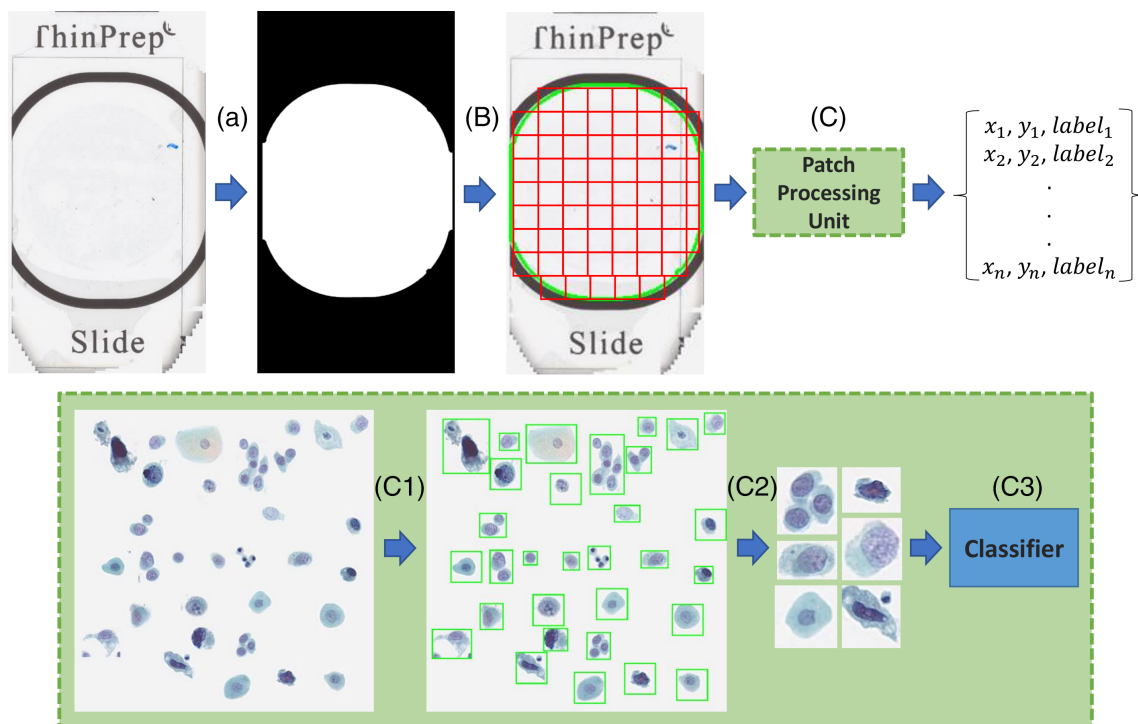
In histology slides, a relatively small region of the slide contains tissue and to reduce the computation time, the tissue region is identified to avoid processing the background white region. To exclude the background region, thresholding can be used for histology images at a low resolution. Like histology images, in the urine WSIs, the cellular content is confined to a limited portion of the slide. However, unlike histology images, thresholding at low resolution would omit cells in WSIs with a fewer number of cells.

In Figure 1(A), an example of urine cytology slide from our dataset is shown at a low-resolution level. The area inside the two fiducial marks contains cells while the remaining area is noncellular. Hence, the region outside these two fiducials should be excluded from processing to reduce the computation time. To achieve this, we adopt the Otsu thresholding [22] which determines a threshold value by maximizing the interclass intensity variance. Specifically, we first convert the RGB images into a grayscale image and then an optimal threshold value is estimated using the Otsu algorithm. A number of other objects such as the text on the slide and other artifacts were

identified using this threshold value. These were excluded based on the area-based threshold. The resulting mask for the ROI is shown in Figure 1(A) and it was carried out at a resolution level of 5 $\times$ .

## 2.3 | Cell segmentation

To identify candidate cells, we separated the cellular content from the background using the thresholding technique. We selected a global cutoff using the Otsu thresholding, resulting in a segmentation map for individual cells and cell clusters. We followed a simple process for obtaining this value which is explained in Algorithm 1. To find an optimal threshold value, the image should contain cells representing the whole population. We employed *k*-medoid clustering [23] to select exemplary cell patches from each class. We set *k* = 20 resulting in 20 clusters per class since we needed 20 exemplar patches from each class. A sample closer to the medoid of the cluster was added to the exemplar bucket. Using these exemplar patches, a big synthetic image was generated by randomly placing the exemplar patches on a plain background image retrieved from one of the WSIs. This image was then converted to HSV from which the saturation channel was used to find the threshold value using Otsu thresholding. A generated segmentation map for an example visual field is shown in Supplementary Figure 3.



**FIGURE 1** The illustration of our proposed method for cell detection and classification from a whole slide image (WSI). (A) region of interest (ROI) detection (B) patches of size 5000  $\times$  5000 are extracted from ROI (C) unit which process every patch and output the coordinates and predicted label of each candidate cell (C1) cell segmentation followed by connected component analysis (C2) patch extraction (C3) label prediction using a trained classifier [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

### Algorithm 1

#### Threshold selection for candidate cell segmentation

Result: threshold

```

1 k ← 20
2 exemplars_bucket ← []
3 synthetic_image ← get_random_bg_image()
4 for iClass ← 1 to num_classes do
5   X ← rgb2gray(resize(all_patches, 128))
6   Xpca ← pca(vectorize(X), 100)
7   clusters, medoids ← kmedoid(Xpca, k)
8   for iCluster ← 1 to k do
9     temp ← min_distance(medoids(iCluster), clusters
      (iCluster))
10    exemplars_bucket ← add_to_bucket
      (reshape(temp))
11  end
12 end
13 synthetic_image ← generate_synthetic_image
   (exemplars_bucket, synthetic_image)
14 synthetic_imagehsv ← rgb2hsv(synthetic_image)
15 threshold ← otsu_thresholding(synthetic_imagehsv)
  
```

## 2.4 | Cell detection and classification

In our study, we applied two different approaches for identifying different types of cells. These approaches comprise of first detecting the candidate cells either by thresholding or using CNN. The candidate cells are then classified using different CNNs.

### 2.4.1 | Our approach

#### Training data preparation

The annotations were obtained at WSI level and the patches of different sizes were extracted from the images depending on the size of the candidate cells at 40× resolution. For cell clusters, the whole region surrounded by the polygon or rectangle was extracted while individual cells for which a dot was placed around the center of the cell were captured in a different way. For a given dot, cell segmentation mask was generated for a patch of size 500 × 500 with a dot in its center, followed by a connected component analysis. A component having a dot inside or on its boundary was considered as a candidate cell. A patch capturing the whole candidate cell was extracted and was saved to the hard drive as an input to the classification network along with its label information.

#### Methodology details

In our approach, we applied global thresholding to segment the candidate cells, as explained in Section 2.3. The generated mask was further

processed with hole-filling and area-based object removal to avoid artifacts. The connected component analysis was performed to compute the bounding box for each identified object in the mask. The bounding box was then used to collect input data for the classification network. For classification, we employed Xception which is the extension of inception network [24], with depthwise separable convolution operations replacing inception modules. The network architecture is shown in Supplementary Figure 4. The input image to the network was resized to a size of 256 × 256 pixels and was normalized by subtracting mean from the images. We trained the network for 392 epochs with a batch size of 20 images. The accuracy and loss curves for the training and validation set are shown in Supplementary Figure 5. The network was configured by setting focal loss as a loss function and Adam function as an optimizer. The broad illustration of our proposed pipeline for risk stratification is shown in Supplementary Figure 6 while the overflow diagram for cell detection and classification is shown in Figure 1. Our code for processing a WSI of urine cytology is publicly available (<https://warwick.ac.uk/fac/sci/dcs/research/tia/software/urinecyto>).

### 2.4.2 | RetinaNet detection and classification

#### Training data preparation

The training of an object detector requires a training set with either dense annotations or an approach to nullify the effect of unannotated objects from the loss function. In liquid-based cytology samples, cells do not often confine to a compact region. Therefore, it results in regions with sparse annotations, not suitable for the training of object detectors. To mitigate this problem, we generated synthetic regions of dense annotations with the cells extracted from the different WSIs. First, we randomly extracted a background image of size 5000 × 5000 from one of the WSIs; then, the cell patches used in our previous approach were randomly placed on it. The background white patches were excluded while training this network.

#### Methodology details

In our second approach, we employ an object identification method for simultaneous detection and classification of cells. There are a number of one-stage and two-stage object detectors, not limited to [25–32]. We use a one-stage detector which has been shown to perform well in terms of both speed and accuracy, known as RetinaNet [32]. One-stage detectors are faster than two-stage detectors but do not perform well comparatively due to the class-imbalance problem. In [32], the class imbalance problem is tackled using a novel focal loss. We used ResNet as a backbone network for our experiments. We have used a publicly available code for RetinaNet for our experiments (<https://github.com/fizyr/keras-retinanet>).

## 2.5 | WSI-level classification

The clinical data used in this study comprises TPS categories assigned by our cytopathologists to each WSI of a cytology slide. The ground

truth (GT) risk-based labels are derived from the relative risk associated with categories outlined in [4]. It is defined in relation to the extent of follow-up needed which segregates the cases with a high risk of malignancy for more aggressive follow-up. We considered the stated percentage of risk to generate the GT information for classification of samples into low and risk cases. We put all the cases with risk less than 50% to be in low-risk class and the cases with a risk higher than 50% to be in high-risk class. The low-risk class comprises Normal, Inflammatory, CA cases while high-risk class contains ASM and TCC cases. There were some images in our dataset that were not scanned properly and were not in focus. We excluded these images by setting a threshold on the number of all identified cells except debris in relation to the count of cells predicted as debris. Using our system, we stratified these cases with the count of atypical and malignant cells. We also conducted some additional experiments with different cell profiling which are listed in the Supplementary Material Document.

### 3 | RESULTS

#### 3.1 | Cell-level classification

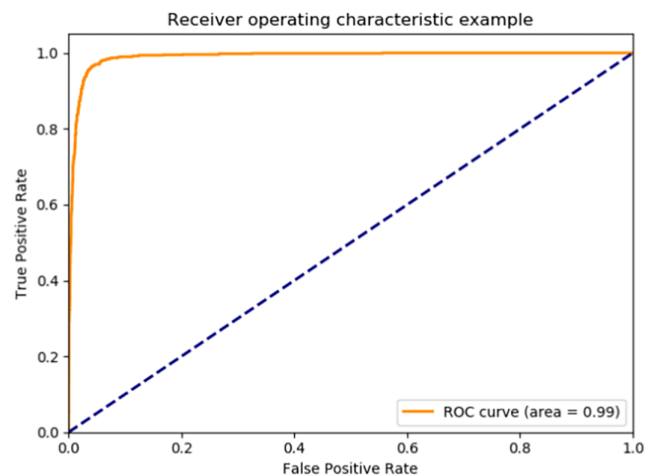
We evaluated our results obtained with Xception and RetinaNet using commonly used measures, along with results of some other CNNs, that is, VGG, MobileNet, Inception, and ResNet. All these networks were initiated with the pre-trained weights for ImageNet. The RetinaNet detected more than one cell with slightly different bounding boxes against a single cell. For evaluation, we computed these measures for those predictions which lie inside the GT bounding boxes. The predicted label of a cell with the highest probability was considered as a final label if there were more than one prediction against a single GT bounding box.

|                 |   | Predicted |     |     |     |     |     |     |  |
|-----------------|---|-----------|-----|-----|-----|-----|-----|-----|--|
|                 |   | N         | S   | I   | O   | A   | M   | D   |  |
| Original Labels | N | 529       | 35  | 0   | 43  | 51  | 3   | 2   |  |
|                 | S | 53        | 903 | 0   | 58  | 2   | 3   | 25  |  |
|                 | I | 6         | 0   | 443 | 43  | 5   | 0   | 3   |  |
|                 | O | 75        | 21  | 2   | 806 | 57  | 13  | 18  |  |
|                 | A | 7         | 2   | 0   | 25  | 408 | 58  | 0   |  |
|                 | M | 3         | 3   | 0   | 11  | 163 | 320 | 0   |  |
|                 | D | 3         | 35  | 0   | 35  | 2   | 1   | 935 |  |

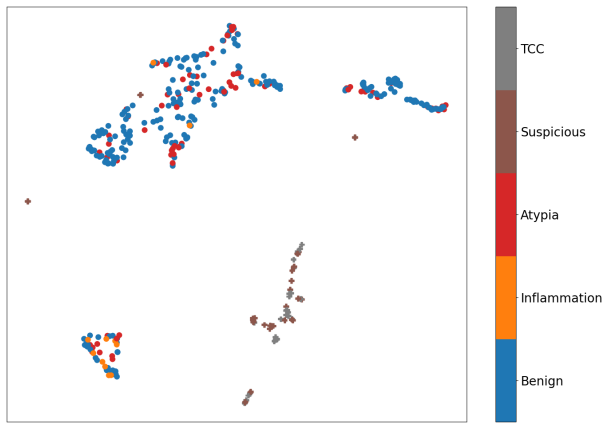
Among all the models, Xception outperforms all the methods on our validation set. The evaluation results of all these models are shown in Supplementary Table 2. Figure 2 shows confusion matrix and receiver operating characteristic (ROC) using Xception on our validation set. For ROC, atypia and malignant cells are considered as positive classes while all other classes are considered as negative classes. The area under the curve is found to be 0.99.

#### 3.2 | WSI-level classification/risk assessment

The UMAP projection of the count of all seven categories of cells is shown in Figure 3. The UMAP plot shows a clear separation between high-risk and low-risk classes. However, subclasses have significant overlap with one another and do not separate clearly. It also demonstrates that there is variability in low-risk data points in terms of features and it is higher in comparison to samples from the high-risk class which are more closely clustered, except for some outliers. A low-risk cluster on the bottom left side of the plot shows a significant overlap between atypia and inflammation which is in line with the fact that the inflammatory samples tend to show atypical features induced as a reaction to the treatment. We performed classification of WSIs on the basis of diagnostically important cells rather than considering all the cells. Hence we restrict to (1) the count of malignant cells and (2) the total count of atypical and malignant cells. We found the total count of atypical and malignant cells to be more discriminating as compared to the count of malignant cells only. This is demonstrated in Figure 4(A) and (B). The area under the ROC curve with the count of atypical and malignant cells is 2% better as compared to that obtained with the count of malignant cells. The results obtained with additional cell profiling are demonstrated in the Supplementary Material Document.



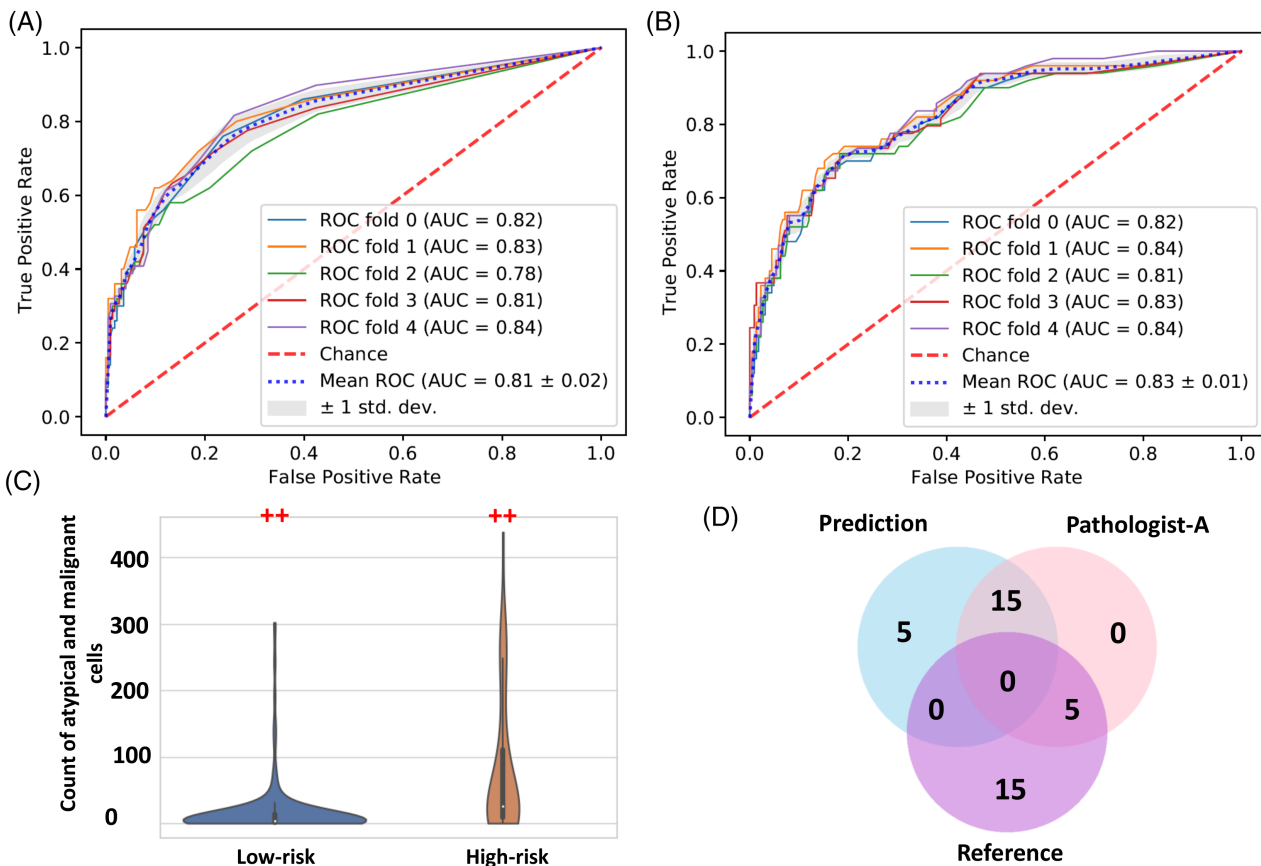
**FIGURE 2** Results of cell classification using the Xception network. The left column shows the confusion matrix and the right column shows the region of interest (ROC) for binary classification. For binary cell classification, malignant, and atypical cells are considered in a positive class while all the other classes are kept in a negative class [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 3** Supervised UMAP projection of the count of all seven types of cells predicted by the Xception network. Plus and dot markers are used for high-risk and low-risk cases, respectively [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

## 4 | DISCUSSION

The automatic cell classification in urine cytology images is a challenging task due to various reasons. This may include inconsistent annotations, classes with a subset of similar features, how the cell samples are categorized into different classes and changes in the cell appearances due to the treatment. The confusion matrix as shown in Figure 2 demonstrates that the atypical and malignant cells are mostly confused with each other and this is due to their overlapping visual features with respect to TPS criteria. Most of the misclassified cells belong to the “others” class and this is mainly due to the nature of the samples, we have placed in this class. It comprises of degenerating cells however these cells could belong to a normal, atypical, or malignant class. Normal cells change their appearance when the patient is on medication and are termed as reactive cells and may resemble atypical cells. Therefore, there are some normal cells misclassified as



**FIGURE 4** The top row shows the performance of binary classification (low-risk vs. high-risk) at whole slide image (WSI) level using (A) a count of malignant cells (M) and (B) total count of atypical and malignant cells (A + M). Region of interest (ROC) is shown with 5-cross validation and the average area under curve using M and A + M is 0.81 and 0.83, respectively. (C) A violin plot displaying the count of predicted atypical and malignant cells in WSIs belonging to low and high-risk classes. (D) A Venn diagram presenting interobserver variability in labeling the 20 WSIs. These cases were misclassified by the method wrt the reference but 15 out of those predictions agreed with another pathologist-A. To generate these results, cells were classified using the Xception network [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

atypical cells. We also verified the network predictions for some benign cases for which the number of malignant cells predicted by the network was greater than 10. We found that some of these cases had reactive normal cells and cells with fluffy cytoplasm. This could be improved by adding these challenging cases to the training set.

In this study, we demonstrated the potential promise of an automated risk stratification method. There are some limitations of the proposed method related to how the data were obtained. We obtained WSIs and their corresponding data from a single center which may have introduced a bias in our proposed approach. Therefore, our findings need to be validated with a large-scale study. Additionally, sourcing annotations from two pathologists may also have introduced a bias into the machine learning model. However, involving more pathologists is not necessarily the solution to this problem due to the potentially larger degree of disagreement between multiple pathologists, as shown in previous studies such as Reid et al. [33].

#### 4.1 | Annotation variability

We sourced cell-level annotations from two pathologists. The inconsistency in their annotations can undermine the performance of the model, given the model has a tendency to learn the complexity. To inquire about the inconsistency in the labeled dataset, we randomly selected some cells from our validation set and asked the expert pathologist to reannotate them. We selected these cells from our more concerned classes, normal, atypia, and malignant. We selected atypia and malignant classes since these are important in terms of making a diagnosis. The normal class was selected since it was mostly misclassified as atypia by the network. The variability in the annotations of the same pathologist is demonstrated in Table 1. In addition to the slide quality and the lack of multiple focal planes, the intraobserver variability could be due to pathologists' lack of experience with the digital slides for urine cytology. The intraobserver variability is a recognized issue in cytology. However, sourcing the labeling with consensus among different pathologists in an effort to improve the variability will improve the performance of the model.

#### 4.2 | Performance of RetinaNet

There is a huge difference between the performance of RetinaNet with ResNet and a ResNet followed by the cell segmentation. This is partially due to the limitation of the detector in the RetinaNet, missing

several cells. Additionally, the detector resulted in many bounding boxes for a single candidate object. On choosing a bounding box with predicted label with the highest probability, further increases the number of missing cells. In our validation set, we had 5175 cell samples, out of which 68 cells were missed when no detected object was ignored. However, selecting the predictions with a probability greater than 50% resulted in 692 cells to be missed. Contrary to it, the threshold-based segmentation does not miss any cell, except that it may fail to segment the whole cell, particularly squamous cell.

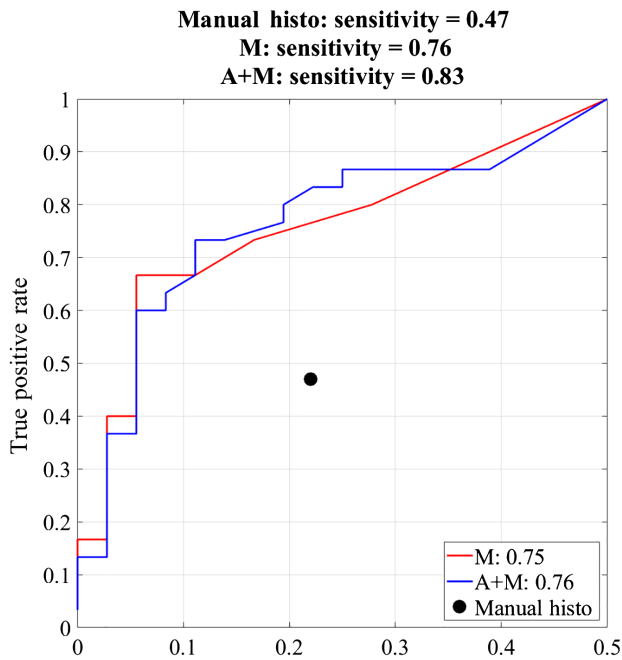
#### 4.3 | Correlation between cytology and histology

We also studied a correlation between the cytopathology-based diagnosis and histopathology-based diagnosis. We obtained histopathology diagnosis for 48 cases along with their cytology and histology reports. These cases comprises 26 CA and 37 ASM, diagnosed using the cytology slides. We hypothesized that the cases for which network predicted more number of atypical and malignant cells would be diagnosed as malignant on performing histology. We observed a trend of association between cell count and histopathology-based diagnosis as shown in Supplementary Figure 7. We compared the results of cytopathology based risk category and our digital risk labeling against the "gold standard" histopathology-based diagnosis. The confusion matrix for manual cytopathology-based risk versus manual histopathology based diagnosis is shown in Supplementary Table 3. As can be seen in Figure 5, the digital risk could be considered a better predictor of the histopathology-based diagnosis. However, this needs to be validated with a large-scale multicenter study. To study it further, we looked into the cytology and histology reports of some of these cases to understand the grounds for the possible discrepancies between cytology and histology diagnosis. We came up with the following rationales for the discrepancies: (1) urine of patients with bladder cancer can be negative (i.e., no shedding of malignant cells in urine). (2) In cases of the instrumented urine sample, at least some abnormal looking groups of cells can be expected due to this sampling technique. Also, an intervention or surgical procedure can lead to the appearance of granulation tissue, inflammation, and reactive atypia. However, if information about sampling technique and relevant history is not available to a pathologist, these cells can then get wrongly labeled as atypical or suspicious. (3) Information about female genital tract, kidney, or prostate pathology is relevant and should be available to pathologists. Otherwise, malignant cells from these organs (which can sometimes be found in urine) could be misdiagnosed as malignant

**TABLE 1** Table presenting intraobserver variability of cell-level annotations. The Cohen's kappa is 0.15 showing a slight agreement for these classes; mainly the disagreement is in between atypical and malignant classes and in between atypical and others classes

|                 | Normal | Atypia | Malignant | Inflammatory | Others |
|-----------------|--------|--------|-----------|--------------|--------|
| Normal (157)    | 108    | 18     | 0         | 14           | 17     |
| Atypical (407)  | 85     | 138    | 7         | 9            | 168    |
| Malignant (397) | 29     | 212    | 104       | 0            | 52     |





**FIGURE 5** Region of interest (ROC) curves for digital cytology-based risk versus manual histopathology-based diagnosis. The sensitivity of all the predictors is computed for a fixed corresponding specificity of 0.56. To determine the digital cytology based risk, urothelial cells were classified using the Xception network. The confusion matrix for manual cytopathology based risk vs manual histopathology based diagnosis is provided in the Supplementary Material Document (Supplementary Table 3) [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

urothelial carcinoma cells (i.e., bladder carcinoma). (4) The presence of calculi/stones or the BCG treatment also results in the appearance of abnormal-looking cells, thus affecting the cytology diagnosis. (5) Malignant cells can be missed in specimens contaminated by fungal or bacterial overgrowth. (6) In histology reports, both high and low-grade tumors are reported while in cytology detecting low-grade TCC is not appreciated. (7) A long interval between cytology and histology can be one of the reasons for the difference in diagnosis.

#### 4.4 | Interobserver variability wrt WSI labeling

All pathologists involved in labeling are fellows of the Royal College of Pathologist (RCPath) with experience in assessing urine cytology slides under a light microscope. However, it does not suggest a lack of disagreement in their diagnosis. Keeping this in mind, we investigated the concordance of our digital risk with the diagnosis of an independent pathologist. We selected WSIs which were misclassified (low-risk and high-risk cases with high and less number of atypical and malignant cells, respectively) by our method wrt the reference labeling. We selected 20 such cases from low-risk and high-risk categories and asked the independent pathologist to assign labels to these WSIs. Figure 4(D) shows the interobserver variability between the reference and independent pathologist-A labeling. Out of 20 cases, the

agreement was found for only 5 cases. In other words, our predictions make a concordance of 0% and 75% with reference and pathologist-A, respectively. The readers are referred to Supplementary Table 4 for more details on interobserver variability. This disagreement could be because the reference labeling was carried out in a clinical setting where pathologists had access to other clinical data whereas the pathologist P-2 made their decision solely on the basis of image content. Also, voided and aspirated slides are interpreted slightly differently by the pathologists. P-1 pathologists knew how the sample was obtained while the pathologist P-2 did not have this information and interpreted all the samples as if they were voided samples. We observed interrater variability (as shown in Supplementary Table 4) more between cytological atypia and suspicious cases, both of which are already considered as contentious and borderline, rather than between malignant and normal cells. This is similar to findings reported in [6–8]. Considering this variation, the ROC obtained in this study could vary on testing the proposed method with WSI labels obtained from a different pathologist. The interobserver variation in labeling cells and WSIs makes the automated diagnosis of cytology samples challenging.

## 5 | CONCLUSION

In this study, we found that the count of atypical and malignant cells is more robust in discriminating between low and high-risk cases as compared to the count of malignant cells only. The difference between the clinical study and our finding is due to several interdependent factors including intraobserver variability in annotations for atypical and malignant cells, leading to poor performance of the classifier in discriminating between the atypical and malignant cells. Since the cytology material is less evenly distributed (even in LBC samples), pathologists frequently need to focus on different planes to view all the cells in a cell cluster. Therefore, due to the intrinsic nature of cytology samples, z-stacking feature can potentially help. We believe that the availability of different planes in the images could improve the annotations and network performance at the same time. The proposed method for automated risk stratification of urine cytology slides has demonstrated clear promise. However, before we can deploy such a system in clinical practice, we will need to conduct large-scale multicentric trials for establishing the efficacy of the proposed method.

#### ACKNOWLEDGMENTS

Clare Verrill, David Snead, Fayyaz Minhas, and Nasir Rajpoot are part of the PathLAKE digital pathology consortium. These new Centres are supported by a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI). Nasir Rajpoot was also partly funded by the Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was partly supported by a Warwick Impact Fund (WIF) award. Clare Verrill was also supported by the National Institute for Health Research (NIHR)

Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

## AUTHOR CONTRIBUTIONS

**Ruqayya Awan:** Data curation; formal analysis; investigation; methodology; software; validation; visualization; writing-original draft; writing-review and editing. **Ksenija Benes:** Data curation; investigation; resources; validation; writing-review and editing. **Ayesha Azam:** Data curation; resources; writing-review and editing. **TZU-HSI Song:** Data curation; resources; software. **Muhammad Shaban:** Formal analysis; validation. **Clare Verrill:** Data curation; resources; validation; writing-review and editing. **Yee Wah Tsang:** Data curation; resources. **David Snead:** Conceptualization; funding acquisition; resources. **Fayyaz Minhas:** Validation; writing-review and editing. **Nasir Rajpoot:** Conceptualization; data curation; funding acquisition; project administration; resources; supervision; validation; writing-review and editing.

## CONFLICT OF INTEREST

The authors declared no potential conflicts of interest.

## ORCID

Ruqayya Awan  <https://orcid.org/0000-0002-2223-329X>

## REFERENCES

- Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder cancer incidence and mortality: a global overview and recent trends. *Eur Urol*. 2017;71(1):96–108.
- Donnelly AD, Mukherjee MS, Lyden ER, Bridge JA, Lele SM, Wright N, et al. Optimal z-axis scanning parameters for gynecologic cytology specimens. *J Pathol Inform*. 2013;4:38.
- Thomas S. Pathologist versus artificial pathologist: what do we really want (need) from machine learning. 2020.
- Barkan GA, Wojcik EM, Nayar R, Savic-Prince S, Quek ML, Kurtycz DF, et al. The Paris system for reporting urinary cytology: the quest to develop a standardized terminology. *Acta Cytol*. 2016;60(3):185–97.
- Rosenthal DL, Wojcik EM, Kurtycz DF. The Paris system for reporting urinary cytology. Switzerland: Springer; 2016.
- Long T, Layfield LJ, Esebua M, Frazier SR, Giorgadze DT, Schmidt RL. Interobserver reproducibility of the Paris system for reporting urinary cytology. *Cytojournal*. 2017;14:17.
- Kurtycz DF, Barkan GA, Pavelec DM, Rosenthal DL, Wojcik EM, VandenBussche CJ, et al. Paris interobserver reproducibility study (PIRST). *J Am Soc Cytopathol*. 2018;7(4):174–84.
- Bakkar R, Mirocha J, Fan X, Frishberg DP, de Peralta-Venturina M, Zhai J, et al. Impact of the Paris system for reporting urine cytopathology on predictive values of the equivocal diagnostic categories and interobserver agreement. *CytoJournal*. 2019;16:21.
- Sahai R, Rajkumar B, Joshi P, Singh A, Kumar A, Durgopal P, et al. Interobserver reproducibility of the Paris system of reporting urine cytology on cytocentrifuged samples. *Diagn Cytopathol*. 2020;48(11):979–85.
- Żejmo M, Kowal M, Korbicz J, Monczak R. Classification of breast cancer cytological specimen using convolutional neural network. *Journal of physics: conference series*. Volume 783. Lille, France: IOP Publishing; 2017. p. 12060.
- Zhang L, Lu L, Nogue I, Summers RM, Liu S, Yao J. DeepPap: deep convolutional networks for cervical cell classification. *IEEE J Biomed Health Inform*. 2017;21(6):1633–43.
- Dimauro G, Ciprandi G, Deperte F, Girardi F, Ladisa E, Latrofa S, et al. Nasal cytology with deep learning techniques. *Int J Med Inform*. 2019;122:13–9.
- Wu M, Yan C, Liu H, Liu Q. Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks. *Biosci Rep*. 2018;38(3):BSR20180289.
- Vaickus LJ, Suriawinata AA, Wei JW, Liu X. Automating the Paris system for urine cytopathology—a hybrid deep-learning and morphometric approach. *Cancer Cytopathol*. 2019;127(2):98–115.
- Sanghvi AB, Allen EZ, Callenberg KM, Pantanowitz L. Performance of an artificial intelligence algorithm for reporting urine cytopathology. *Cancer Cytopathol*. 2019;127(10):658–66.
- Phoulady HA, Goldgof DB, Hall LO, Mouton PR. A new approach to detect and segment overlapping cells in multi-layer cervical cell volume images. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE; 2016. p. 201–204.
- Win K, Choomchuay S, Hamamoto K, Raveesunthornkiat M. Detection and classification of overlapping cell nuclei in cytology effusion images using a double-strategy random forest. *App Sci*. 2018;8(9):1608.
- Wan T, Xu S, Sang C, Jin Y, Qin Z. Accurate segmentation of overlapping cells in cervical cytology with deep convolutional neural networks. *Neurocomputing*. 2019;365:157–70.
- Lu Z, Carneiro G, Bradley AP, Ushizima D, Nosrati MS, Bianchi AG, et al. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE J Biomed Health Inform*. 2016;21(2):441–50.
- Lee H, Kim J. Segmentation of overlapping cervical cells in microscopic images with superpixel partitioning and cell-wise contour refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2016. p. 63–69.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
- Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62–6.
- Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Vol 344. New Jersey: John Wiley & Sons; 2009.
- Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1251–1258.
- Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:13126229. 2013.
- Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 580–587.
- Girshick R. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 1440–1448.
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: single shot multibox detector. European conference on computer vision. Cham: Springer; 2016. p. 21–37.
- Dai J, Li Y, He K, Sun J. R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems; 2016. p. 379–387.
- Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 779–788.
- Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems; 2015. p. 91–99.
- Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2980–2988.

33. Reid MD, Osunkoya AO, Siddiqui MT, Looney SW. Accuracy of grading of urothelial carcinoma on urine cytology: an analysis of inter-observer and intraobserver agreement. *Int J Clin Exp Pathol.* 2012; 5(9):882–91.

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Awan R, Benes K, Azam A, et al. Deep learning based digital cell profiles for risk stratification of urine cytology images. *Cytometry.* 2021;99:732–742. <https://doi.org/10.1002/cyto.a.24313>