Generalized fixation invariant nuclei detection through domain adaptation based deep learning

Mira Valkonen, Gunilla Högnäs, G Steven Bova, and Pekka Ruusuvuori

***Abstract*—Nucleus detection is a fundamental task in histological image analysis and an important tool for many follow up analyses. It is known that sample preparation and scanning procedure of histological slides introduce a great amount of variability to the histological images and poses challenges for automated nucleus detection. Here, we studied the effect of histopathological sample fixation on the accuracy of a deep learning based nuclei detection model trained with hematoxylin and eosin stained images. We experimented with training data that includes three methods of fixation; PAXgene, formalin and frozen, and studied the detection accuracy results of various convolutional neural networks. Our results indicate that the variability introduced during sample preparation affects the generalization of a model and should be considered when building accurate and robust nuclei detection algorithms. Our dataset includes over 67 000 annotated nuclei locations from 16 patients and three different sample fixation types. The dataset provides excellent basis for building an accurate and robust nuclei detection model, and combined with unsupervised domain adaptation, the workflow allows generalization to images from unseen domains, including different tissues and images from different labs.**

***Index Terms*—Deep learning, nuclei detection, digital pathology, domain adaptation, tissue fixation, frozen section, formalin-fixed, PAXgene-fixed.**

I. Introduction

HISTOPATHOLOGICAL examination is an important step in diagnosis of many diseases. Examination usually includes analysis of nuclei morphology, thus, nucleus detection is a fundamental step for many follow up analyses, such as phenotyping on a single-cell level [1], or cancer grading [2]. Machine learning based image analysis provides an efficient, quantitative, and objective way to perform histopathological examination and nuclei detection in a fully automated manner [3]. Nevertheless, building a robust and generalizable nuclei detection model is a challenging task due to the high amount of variability present in histological images. This variability is caused by the underlying biological variation, such as variation in nuclei shape, size, and texture of different tissue types and also by the technical variation introduced during the tissue preparation, such as in fixation process, and scanning procedure [4].

Preparation of tissue depends largely on the type of analysis the tissue is intended for, and for tissue fixation there exists alternatives which have effect on the appearance of tissue components, such as nuclei [5]. While freezing tissue typically provides excellent preservation of biomolecules, freezing also disrupts the structure of the tissue and is therefore not used for routine morphologic analysis. Tissue fixation with formalin is the standard in surgical pathology laboratories due to its low cost and excellent preservation of tissue morphology. Formalin preserves tissue structure by forming crosslinks between molecules [6] and is routinely used for almost all tissue types. PAXgene is an alcohol-based fixative that, in contrast to formalin, simultaneously preserves both tissue morphology and biomolecule integrity. PAXgene has been shown to be suitable for many tissues [7], although artefacts such as increased nuclear staining and tissue shrinkage have been reported. PAXgene-fixed tissue also stains more avidly with eosin, giving H&E (hematoxylin & eosin) stained sections a more intense pink hue compared to formalin fixed specimens. In our earlier study, we studied the feasibility of PAXgene fixation for molecular and diagnostic studies [5], but the fixation effect on modern deep learning based analytics remains unknown.

Deep learning methods, particularly convolutional networks combined with transfer learning [9], [10], have an outstanding ability to learn task specific feature representations and have rapidly become the main approach for microscopy image analysis tasks, including nuclei detection [11], [12]. However, most of the existing methods have been optimised for a specific problem domain using a narrow dataset and fail to generalize to new domains, such as images from different labs or different tissues due to the high variability present in the histological images. A lot of work has been done in order to address the generalization challenge using techniques such as staining normalization [13]–[15], extensive data augmentation [16], [17], or utilization of datasets containing high variability such as multi-tissue datasets [8]. Although these approaches have achieved prominent results, they have still left room for further research, and some variability sources are yet to be studied, such as different tissue preparation techniques. To eventually generalize extensively to diverse patient populations and real world clinical environments, the effects of wider range of variability sources need to be covered.

In digital pathology, labeled training data is not largely available due to the time-consuming manual annotation process performed by an expert. Nuclei detection as an annotation

We want to thank the Academy of Finland (P.R., projects #313921 & #326463 and #314558 & #326364), and ERA PerMed 2019-2022 - ABCAP project (PR) for funding this study. Also KAUTE Foundation, Nokia Foundation, Finnish Foundation for Technology Promotion, and Instrumentarium Science Foundation are gratefully acknowledged (M.V.).

M. Valkonen, G. Högnäs, and G.S. Bova are with the Faculty of Medicine and Health Technology, Tampere University, Finland and are supported by Academy of Finland, Cancer Society of Finland, and Sigrid Juselius Foundation.

P. Ruusuvuori is with the Institute of Biomedicine, University of Turku, Finland and with the Faculty of Medicine and Health Technology, Tampere University, Finland.

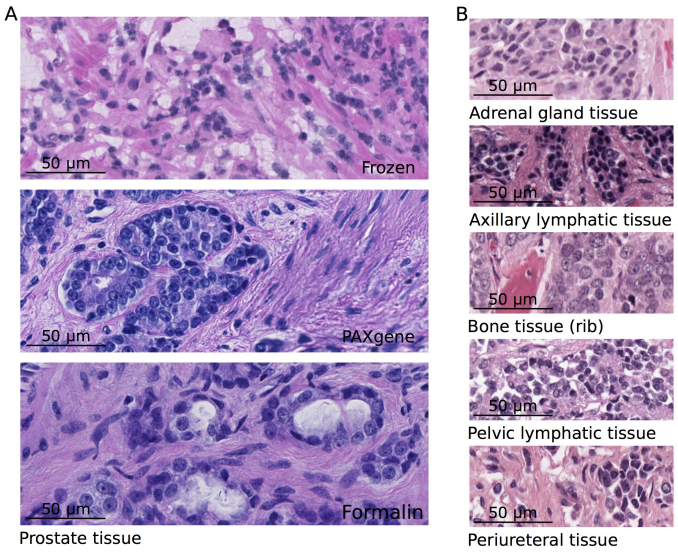

Fig. 1. The materials used in this study include prostate tissue (PT) dataset [5], a holdout test set (5-tissue) and a publicly available MoNuSeg [8] dataset. The PT dataset shown in panel A was collected from radical prostatectomy prostate tissue samples from 16 men, and fixed using three different tissue preparation methods, fresh frozen, formalin-fixed paraffin-embedded, and PAXgene-fixed paraffin-embedded. The 5-tissue validation dataset shown in panel B consists of images collected from five different patients and present 5 different tissue types.

task is particularly laborious due to the large amount of target objects. Consequently, in digital pathology, collecting a vast amount of labeled training data adds to the challenges posed by the heterogeneous nature of histological image data. However, in order to train a deep convolutional neural network in a supervised manner, labeled training data is an absolute requirement. Therefore, in order to build an accurate and robust nuclei detection algorithm that can generalize from one problem domain to another, alternative data labeling methods are needed.

Domain adaptation provides tools for overcoming the requirement of labeled data. In domain adaptation, the representations learned from labeled source data are utilized in a classification problem in an unlabeled target domain [18]. As the goal in domain adaptation is to enable domain shift from the problem domain of the original training data into a new domain from which no labeled data for re-training exists, it is a potential solution for generalizing histopathological image analysis methods from one tissue domain to another. Domain adaptation can be utilised in histopathological classification problems to address the requirement of labeled data in unsupervised [19], [20] or weakly supervised [21] manner. We have also shown in our previous studies how domain adaptation can be successfully utilised in model generalization to unseen cell lines from brightfield images in an unsupervised manner [22].

Deep learning methods have shown great success in many machine vision tasks, however their lack of transparency has attracted an increasing amount of attention and research [23]–[25]. Especially in medical domain applications, interpretability and transparency are seen as a necessity, since these can provide insights into the functioning of a deep learning model and can be used to verify and comprehend the predictions by a human expert. One aspect particularly of interest in classifier interpretability is the contribution of patterns in specific spatial locations in input data to classifier decision or outcome. For deep neural networks, methods such as Layer-wise Relevance Propagation (LRP) [26] have recently enabled such analysis, and their availability as tools for explainable AI [27] help giving insight in the classifier decision process. Interpretability tools can also enable verifying that a model learns relevant and similar information from training data even with subtle differences present in the data, caused for example by differences in staining or fixation.

In this study, we trained a convolutional neural network baseline model for nuclei detection using supervised transfer learning. As the second step of the workflow, we applied unsupervised domain adaptation to allow generalization to images from unseen domains, including different tissues and images from different laboratories, without the need for labeled data. Our main contributions are, 1) to study the effect of sample fixation on the accuracy of nuclei detection by using hematoxylin and eosin (H&E) stained training images prepared with three fixation methods; PAXgene, formalin and frozen, and 2) to provide the presented extensive multi-fixation dataset with manually obtained annotations for cell locations that allows further method development. The code and data are available at https://github.com/BioimageInformaticsTampere/NucleiDetection. The implemented workflow allows generalisation and adaptation to external unlabeled datasets in the field of digital pathology utilising unsupervised domain adaptation based on pseudo-labels and hard positive mining.

II. Materials and Methods

A. Data

The materials used in this study included prostate tissue (PT) dataset [5], a holdout test set (5-tissue), and a publicly available Multi-Organ Nuclei Segmentation (MoNuSeg) [8] dataset. Examples from the PT dataset and 5-tissue test set are shown in the Figure 1. In addition, number of annotated nuclei, number of different tissue types and number of patients in each dataset are shown in the Table I. The following sections will describe the used materials in more detail.

Dataset	N patients	N tissue types	N annotated nuclei
PT	16	1	67070
5-tissue	5	5	9011
MoNuSeg	30	7	16966

TABLE I
NUMBER OF ANNOTATED NUCLEI, NUMBER OF DIFFERENT TISSUE TYPES AND NUMBER OF PATIENTS IN EACH DATASET.

Prostate tissue (PT) dataset

The image data was collected from radical prostatectomy prostate tissue samples from 16 men. The samples were collected and studied under Tampere University Hospital Ethical Committee Approval R03203. From each patient, three cores

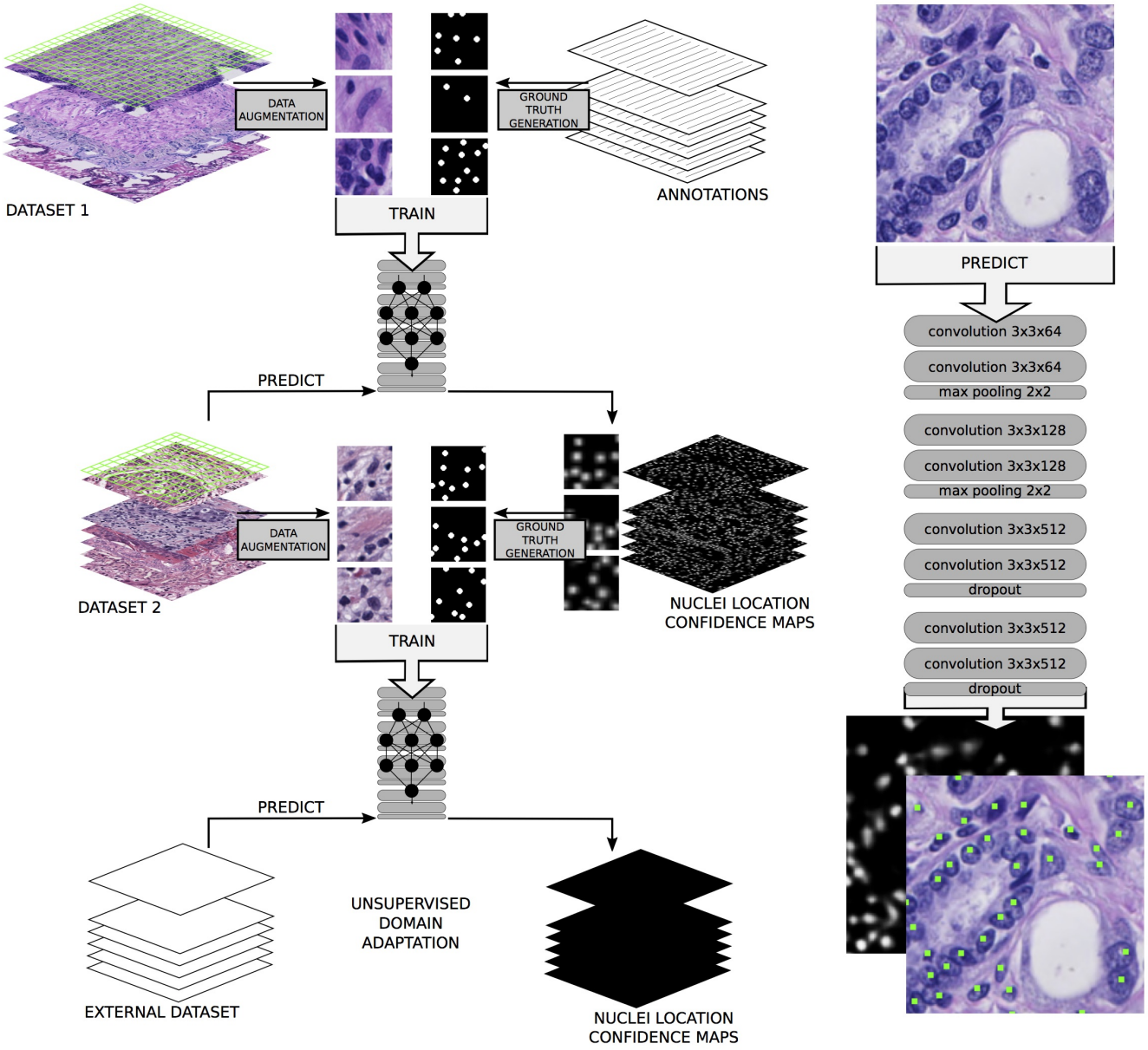

Fig. 2. The nuclei detection workflow. Upper half presents the baseline model training step which is followed by an option to utilise unsupervised domain adaptation to detect nuclei from an external dataset (new domain) without annotations. The convolutional neural network consists of four convolutional base layers from a pre-trained VGG-16 network appended with four additional convolutional layers.

were collected from posterior side of the prostate. Each of the three cores was fixed using one of three different tissue preparation methods: fresh frozen, formalin-fixed paraffin-embedded, and PAXgene-fixed paraffin-embedded. The tissue sections were stained with H&E and scanned with a Hamamatsu Photonics Nano Zoomer XR C12000 automated scanner, using pixel resolution $0.23m$. More detailed data acquisition is presented in a study by Högnäs et al. [5]. From every H&E stained core whole-slide image (WSI) one $550x550$ μm image was randomly selected and nuclei were manually annotated, resulting in a total of 67 070 nuclei in 48 images. Compared to the original study where 50 images were used, two images were excluded in order to balance the number of nuclei in each dataset with respect to number of patients and fixations.

Manual annotation was carried out by an experienced histology expert using the Cell Counter plugin in ImageJ software [28] by visually inspecting the images and by manually clicking coordinates of every nucleus in the image. The coordinates for each nucleus in were saved in xml files, one file per image. Further, we validated the accuracy of the manual annotation for a randomly chosen image by reproducing coordinate markings by two independent persons, yielding very high agreement both in terms of number of nuclei and numerical accuracy (F-score 0.9 for both annotators with the ground truth). The detailed validation results as well as visual representation of the validation image are provided as Supplementary Figure 1 and Supplementary Table I. This

dataset, along with the annotations, is available at: https://github.com/BioimageInformaticsTampere/NucleiDetection/

5-tissue dataset

The 5-tissue dataset included formalin-fixed paraffin-embedded H&E stained metastatic tissue images from 5 different tissue types. The tissue types included periurethral tissue, bone tissue from rib, axillary lymphatic tissue, adrenal gland tissue, and pelvic lymphatic tissue. The samples were collected from 5 patients and from each WSI a $500x500$ μm image was randomly selected. All nuclei were annotated from the images using Multi-point Tool in ImageJ software [28] in a similar, fully manual fashion as described for the prostate dataset. In total, the test set included 5 images and 9011 annotated nuclei coordinates.

Multi-organ nuclei dataset

Publicly available multi-organ dataset (MoNuSeg) [8] included 30 images captured from 30 different H&E stained WSIs. These images were collected from The Cancer Genome Atlas and originally prepared in 18 different hospitals. The manually annotated nuclei represent a diversity of nuclear appearances from several patients, disease states, and organs. The dataset consisted of seven different tissue types, including breast, liver, kidney, prostate, bladder, colon, and stomach tissue. In total, the dataset included 16 966 nuclei mask annotations. As the images are publicly available from https://monuseg.grand-challenge.org, this dataset enables benchmarking and provides further insight about generalization of the methods.

B. Nuclei detection model

A convolutional neural network model was built to detect cell nuclei locations from histopathological images. To achieve better detection accuracy and to reduce the computational costs of optimisation, transfer learning approach was used. Transfer learning allows the utilisation of a pre-trained network that is already optimised to classify images from some other domain [9]. The lower level features of a pre-trained network tend to be more generic and the later layer features become more specific to the details of the original classification task in the training data domain. Therefore, we utilised four base layers from VGG-16 architecture pre-trained on the ImageNet dataset [29]. These base layer weights were fixed during training, and four additional convolutional layers were added on top of them. Each convolutional layer was followed by Rectified Linear Unit (ReLu) activation and every two convolutional layers were followed by dropout in order to avoid overfitting of the model [30]. Sigmoid activation function was used at the model output layer to provide a nuclei location confidence map. The nuclei detection model was implemented using Python programming language and Keras [31] module with Tensorflow [32] backend. The model workflow and architecture are visualized in the Figure 2.

Our choice of the base deep neural network architecture was motivated by the relatively simple adaptability of the VGG-16 architecture, which can be done simply by adding problem domain spedific layers into a generic VGG-16 network. We have successfully used similar transfer learning strategy of extending the generic VGG-16 network into a specific problem domain in histopathology earlier in a study where image-to-image transform from immunohistochemichal staining to cytokeratin staning mask was done using VGG-16 based architecture [33]. In the current study, the use of a generic, well tested architecture underlines the applicability of the proposed domain adaptation approach. The use of other, more developed implementations or different architectures, is possible in a similar manner, but optimizing their accuracy for the nuclei detection task is out of the scope of this study.

Model training specifications

A convolutional neural network (CNN) consists of a sequence of layers that maps an input vector $\mathbf{x}$ to an output vector $\mathbf{y}$.

$$\mathbf{y} = f(\mathbf{x}, \mathbf{w}), \tag{1}$$

where $\mathbf{w}$ is the weight and bias vector that define the network layers. During the training phase, the network variables are estimated by solving an optimization problem. In supervised learning, where labeled training data is a prerequisite, a set of input vectors $\mathbf{x}_n$ have corresponding target vectors $\mathbf{t}_n$ ($n = 1, .., N$). In which case, the optimization problem can be defined as

$$\underset{w}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} L(f(x_i, \mathbf{w}), t_i) \tag{2}$$

where, L is a task-fitting loss function. Here, we used binary crossentropy as a loss function.

$$L(\mathbf{y}, \mathbf{t}) = -\frac{1}{N} \sum_{i=1}^{N} (t_i log(y_i) + (1 - t_i) log(1 - y_i)) \tag{3}$$

The optimization was performed by using an Adam optimizer [34] which is a stochastic gradient descent method. For Adam algorithm, learning rate was initialized to 0.0001, the exponential decay rates for the moment estimates ($\beta 1$, $\beta 2$) were set to 0.9 and 0.999, respectively, and the fuzz factor was set to 1×10^{-8} to prevent null division.

A set of model hyperparameters was optimised by using grid-search to maximize the nuclei detection accuracy in the training phase using pixel size of 0.5 μm. Test data was never used in hyperparameter optimization. These hyperparameters included number of epochs ($n_{epochs} = 2$), learning rate ($lr = 0.0001$), batch size ($bs = 16$) and dropout for regularization ($drop\_rate = 0.5$). Also optimal input image size (64x64 pixels) and nuclei location mask structuring element shape and size (round, $R = 4$) were searched to maximize the nuclei detection accuracy on the PT dataset.

In order to reduce variation caused by the different staining procedures and to focus on the morphological differences in the tissue caused by the different fixations, a data augmentation step was included. The step included color augmentation on HSV space and adding Gaussian noise ($\mu = 0$, $\sigma = 0.01$).

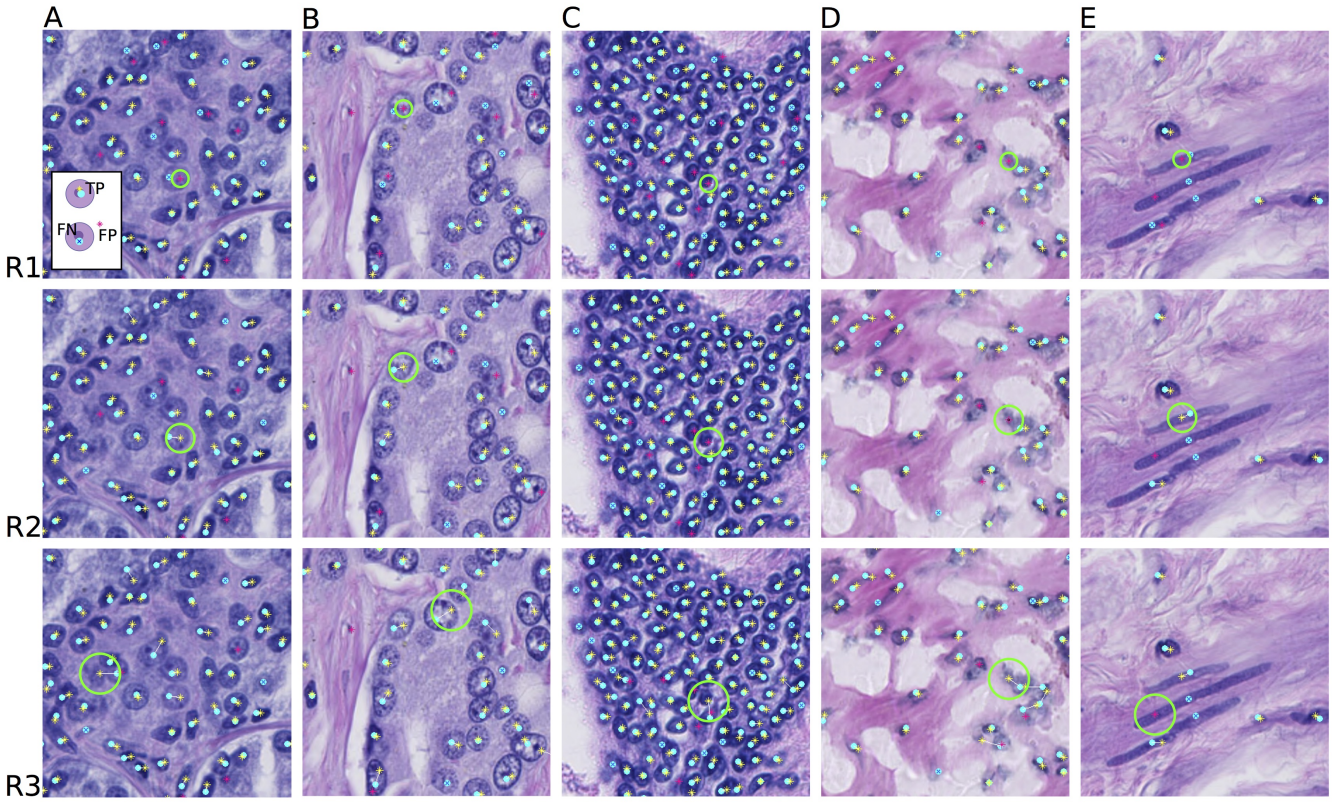

Fig. 3. Examples (A-D) of predicted nuclei locations using three different values for radius R (R1-R3). The ground truth annotation is marked with a light blue circular marker and the predictions are marked with a star shaped marker. The color coding for true positive (TP), false negative (FN) and false positive (FP) cases are shown in the upper left subfigure. Light green circle around a predicted nucleus (magenta star) visualises the different R values; R1=6 , R2=10, R3=14. In addition, a light line is drawn between each ground truth coordinate annotation and the corresponding true positive prediction. The problems that originate from having coordinate annotations as ground truth detections and defining true predictions are also visible in the figure, such as having a seemingly conflicting annotations with FN and FP in a single nucleus (column E). These problems are addressed in the discussion chapter.

Color augmentation was implemented by first converting the RGB image into HSV space and then adding a constant value the hue channel. The constant value was randomly drawn from normal distribution with mean of $\mu = 0.1$ and standard deviation of $\sigma = 0.01$. After the hue shift, the sample image was converted back to RGB space. Every third input image block was left in the original form and the other samples were augmented using either HSV shift or by adding noise.

Prediction

The trained nuclei detection model can be used to predict the nuclei locations of an input image with pixel size of 0.5 μm. The model takes an arbitrary sized RGB image as an input and predicts a confidence map as an output, where values close to 1 denote a high probability for a nucleus location and values close to 0 indicate a background pixel in the corresponding location of the input image.

The confidence map is post-processed in order to find single pixel locations for each detected nucleus. The confidence map is first converted into a binary image using threshold value of 0.5. Different objects in the binary image are labeled and the connectivity of the objects is defined by a centrosymmetric 3x3 structuring element. Finally, the center of mass from each object is selected as the coordinate for a detected nucleus.

Ground truth generation from annotations

The ground truth nuclei location masks were generated from nucleus coordinates annotated manually by an expert. The coordinates were read from a csv file and first scaled to match the operating resolution of the model. A binary mask image was generated with a single pixel representing the coordinates of one annotated nucleus. This mask image was then dilated using a circular structuring element with radius R=4 in order to expand each nuclei location area. This process corresponds to adding computationally a small level of uncertainty in the annotation coordinates, as the manual marking is practically impossible to be done on a pixel level accuracy for thousands of objects. The value of R was defined experimentally to be small enough such that the whole ground truth marker would remain inside the nuclei and that no overlap between ground truth objects occur. Examples of ground truth nuclei location masks can be seen from the workflow Figure 2.

Unsupervised domain adaptation using pseudo-labels

The ground truth nuclei location masks for an external dataset without any annotations can be generated by using the trained baseline model. Any dataset can be run through the trained baseline model and from the predicted nuclei location

confidence maps a set of positive examples can be extracted based on a thresholding rule. These hard positive examples can be then used as new training examples to adapt the nuclei detection model to a new data domain.

The thresholding rule includes two different thresholds; higher for determining hard positives, which are confident detections, and lower for detecting other nuclei around the hard positives. To generate new training samples, an image block of the size 64×64 pixels is extracted around each hard positive example. In order to generate the ground truth nucleus location mask for the whole training sample block and not to miss any nuclei within the image, the second threshold (the lower one) is applied. The detection controlled by the two thresholds is an elemental part of the pseudo-label domain adaptation step. Thus, we examined the effect of the threshold values on detection accuracy in a grid search (Supplementary Table II), and chose the threshold values (0.8 higher, 0.5 lower) based on this experiment. From the generated binary image, the nuclei locations are generated similarly as in the prediction step and the final mask image is generated similarly as in the ground truth generation from the annotation step as explained in previous paragraphs.

R	mean F1-score	mean precision	mean recall
6	0,800	0,823	0,780
8	0,861	0,885	0,839
10	0,879	0,904	0,857
12	0,886	0,911	0,864
14	0,891	0,917	0,869
16	0,895	0,921	0,873
18	0,898	0,924	0,876
20	0,902	0,928	0,880
22	0,905	0,931	0,883

TABLE II
THE ACCURACY METRICS FOR PT DATASET USING DIFFERENT VALUE FOR RADIUS R. THE R SPECIFIES A DISTANCE BETWEEN A GROUND TRUTH COORDINATE AND A PREDICTED NUCLEUS LOCATION THAT IS CONSIDERED AS A TRUE POSITIVE.

C. Evaluation

Accuracy metrics

For numerical evaluation of the nuclei detection model accuracy, F1-score was used that rely on precision and recall. Precision measures the fraction of correctly classified positive instances among all retrieved positive instances. Precision is also referred to as the positive predictive value and it can be defined using true positive (TP) and false positive (FP) counts.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall measures the fraction of correctly classified positive instances among the actual positive instances, and it is defined using true positive (TP) and false negative (FN) counts.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The F1-score is defined as the harmonic mean of precision and recall.

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

In order to analyse the accuracies of the nuclei detection models, a rule was needed to compare the ground truth coordinate annotations with the predicted nuclei locations. A predicted nuclei location was considered as a true positive detection when a ground truth annotation was within a certain radius from the prediction. An optimal radius (R) was selected based on analysing the accuracy results of PT dataset using multiple different R values. The R values and corresponding F1-scores are presented in the Table II. However, the F1-score alone was not sufficient measure for selecting optimal R value, since higher values of R generated persistently higher F1-score. The reason for this is fundamentally in the inability to select completely correct metric for true detections while having wide scale of different sizes and shapes of nuclei and a coordinate annotation for a ground truth detection. Therefore, additional visual examination of different R values was performed to discover the optimal value for R. Three different R values are visualised in Figure 3. Overall analysis resulted in selecting R=10. This decision and the problems concerning coordinate annotations are further addressed in the discussion chapter. All of the accuracy results presented in this paper are calculated using the selected optimal R value, excluding the results in Table II.

For MoNuSeg dataset, the groundtruth nuclei segmentations were provided instead of coordinates, therefore, a true positive was considered to be a detection that hits a segmented nuclear area.

Model evaluation and interpretation

In order to ensure that the model decisions are based on meaningful patterns in the input data, we utilised Layer-wise Relevance Propagation [26], [35]. Here we used the LRP implementation provided by the iNNvestigate toolbox [27]. The LRP is a technique for propagating the prediction backward in a neural network based on certain propagation rules. The method provides a heatmap that visualises positively and negatively relevant areas in the input image with respect to the classification task. We randomly selected a set of 64x64 sized fields of views around a detected nuclei from each of the processed image datasets. The set of sample images were analysed using LRP method and the generated relevance heatmaps were visually assessed in order to shed light on the meaningful patterns related to model decisions according to LRP analysis.

III. EXPERIMENTAL RESULTS

A. Nuclei detection from prostate tissue

First, we consider the deep learning based nuclei detection from prostate tissue, for which we have an extensive annotated dataset of 67070 nuclei.

Each of the different fixation models were trained on the data that is defined by the model name. PAXgene model was

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3039414, IEEE Journal of Biomedical and Health Informatics

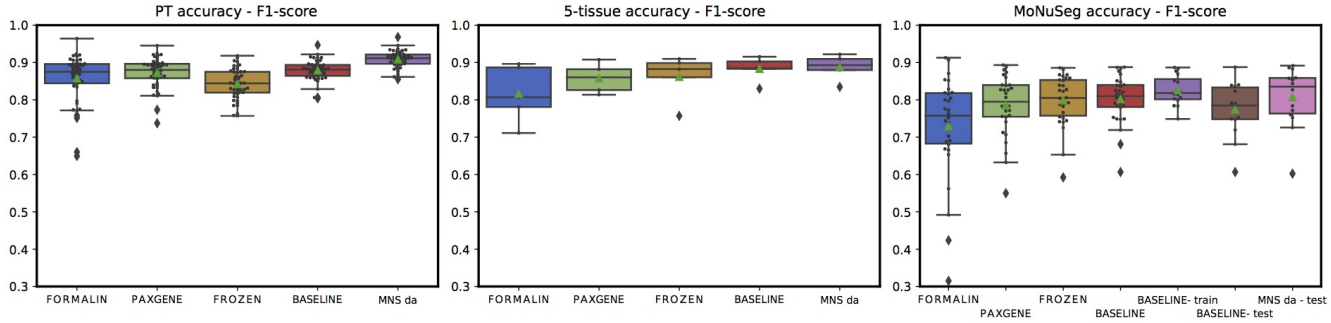

Fig. 4. The nuclei detection F1-scores visualized using boxplots. Title of a subfigure denotes the dataset that was used as a test set and the model names in the x-axis denote the modelname that was used for prediction. The line within a boxplot presents the median F1-score, the green triangle presents the mean F1-score, boxplot lines visualise the 25th and 75th percentiles and the values that settle between these limits, and the outlier F1-scores are presented with diamond shape marker

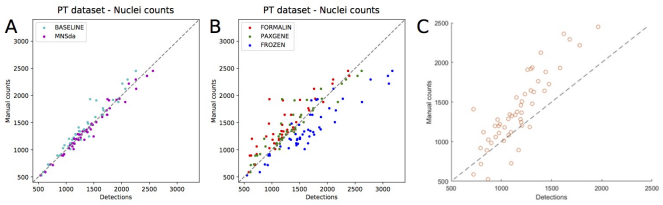

Fig. 5. Nuclei counts of the PT dataset. On the A plot, the annotated nuclei counts are plotted against the detected nuclei using the baseline model and MNSda model. On the B plot, the annotated nuclei counts are plotted against the detected nuclei using the different tissue fixative models. The plot C shows annotated nuclei counts plotted against a 2-step segmentation algorithm based nuclei detection presented in [5].

trained on PAXgene fixed image data, and similarly formalin and frozen models were trained on their respective image data. In total 17 models were trained, each model by leaving out all the data from one of the 16 patients and a final fixation model trained on the whole dataset of one fixation data. The 16 leave-one-patient-out models were used for calculating the results for a left-out image in a cross validation manner, and the reported result is the average from the 16 test images. The final fixation model trained with all data from the fixation type was then used to analyse the accuracy when detecting nuclei from other fixations.

We trained a baseline model with the whole PT dataset. The baseline model was used as an initial model in the domain adaptation step. In addition to the baseline model, 16 models were trained by leaving out all the data from one of the 16 patients, the nuclei from this left out patient data was then predicted using this leave-one-patient-out-model (LOPO-model). In total, 17 models were trained.

The numerical accuracy results are collected in Table III. Each row presents the detection accuracy of one model and the corresponding training and test data are specified in the following columns.

In rows 1-4, we present the results for the whole prostate tissue dataset with fixation specific models and with the baseline model. The F1-scores for fixation specific models range from 0.843-0.873, whereas the model trained with multiple fixations reaches 0.879 on the PT dataset. When comparing to the F1-score 0.78 reported in [5] for a two-step segmentation algorithm, all of the the deep learning based results presented here show clear improvement.

It can be observed that better F1-score is achieved when nuclei detection model training is carried out using image data with multiple fixations compared to the fixation specific models.

The precision and recall values for each model and test set are presented in the last two columns in the Table III. It can be noted that the models trained with visually better quality images (formalin, PAXgene) reach higher precision than a model trained with noisy frozen data. The precision values for PT data for formalin and PAXgene models are 0.925 and 0.903 respectively, whereas the frozen model precision reaches only 0.780. However, model recall is higher for a frozen model (0.924) compared to the formalin model (0.810) and the PAXgene model (0.850) recall values.

B. Generalization to multi-organ datasets

In order to test the generalization ability of our approach, we conducted experiments on a publicly available MoNuSeg dataset [8], and on another dataset with tissue from five organs (5-tissue dataset).

The results for the multi-organ data are shown in lines 14-17 of Table III. When testing the detection performance with MNS data, the F1-scores are 0.730-0.799 and 0.802, revealing a clear drop in accuracy when compared to the prostate tissue dataset. For precision and recall, a similar pattern can be seen when testing with the MNS dataset as for the PT dataset; frozen model yields lower precision and higher recall when comparing to those by formalin and PAXgene models. Further, we applied a division to train and test sets according to the split applied in [8], and the results obtained with the baseline prostate model are listed in Table III, rows 18-19. The divergence in these results show that the dataset has significant variation in image characteristics.

Next, we further tested the generalization to other tissue types with our 5-tissue dataset. The results are shown in Table III lines 20-23. When testing with 5-tissue set, the F1-scores

Model name	TRAIN data	TEST data	F1-score	Precision	Recall
FORMALIN	PT-formalin	PT	0,858	0,925	0,810
PAXGENE	PT-PAXgene	PT	0,873	0,903	0,850
FROZEN	PT-frozen	PT	0,843	0,780	0,924
baseline model	PT	PT	0,879	0,904	0,857
FORMALIN-CV	PT-formalin	PT-formalin	0,885	0,892	0,879
PAXGENE	PT-PAXgene	PT-formalin	0,886	0,878	0,895
FROZEN	PT-frozen	PT-formalin	0,814	0,718	0,941
PAXGENE-CV	PT-PAXgene	PT-PAXgene	0,894	0,899	0,890
FORMALIN	PT-formalin	PT-PAXgene	0,894	0,922	0,868
FROZEN	PT-frozen	PT-PAXgene	0,835	0,755	0,935
FROZEN-CV	PT-frozen	PT-frozen	0,879	0,867	0,896
FORMALIN	PT-formalin	PT-frozen	0,794	0,961	0,683
PAXGENE	PT-PAXgene	PT-frozen	0,838	0,931	0,766
FORMALIN	PT-formalin	MoNuSeg	0,730	0,887	0,647
PAXGENE	PT-PAXgene	MoNuSeg	0,785	0,840	0,758
FROZEN	PT-frozen	MoNuSeg	0,799	0,763	0,852
baseline model	PT	MoNuSeg	0,802	0,797	0,821
baseline model	PT	MoNuSeg-train	0,826	0,803	0,859
baseline model	PT	MoNuSeg-test	0,775	0,790	0,778
FORMALIN	PT-formalin	5-tissue	0,816	0,943	0,727
PAXGENE	PT-PAXgene	5-tissue	0,858	0,898	0,827
FROZEN	PT-frozen	5-tissue	0,862	0,813	0,927
baseline model	PT	5-tissue	0,883	0,864	0,907
MNS DA	PT+MoNuSeg	PT	0,908	0,903	0,915
MNS-train DA	PT+MoNuSeg-train	MoNuSeg-test	0,807	0,781	0,851
MNS DA	PT+MoNuSeg	5-tissue	0,888	0,846	0,939

TABLE III

THE EXPERIMENTAL RESULTS FOR NUCLEI DETECTION ACCURACY OF EACH TRAINED MODEL USING F1-SCORE, PRECISION AND RECALL. THE COLUMNS SPECIFY THE MODEL NAME, USED TRAINING DATA AND THE TEST DATA. THE RESULTS ARE GROUPED BY THE TEST DATA, AND THE LAST THREE ROWS ARE THE DOMAIN ADAPTATION RESULTS. THE CV MODELS STAND FOR LEAVE-ONE-PATIENT-OUT CROSS VALIDATION, WHERE THE REPORTED RESULT IS AN AVERAGE OF RESULTS WITHIN CROSS VALIDATION LOOP - TEST DATA IS ALWAYS LEFT OUT IN TRAINING PHASE.

are 0.816-0.862 for the fixation specific models, and 0.883 for the baseline model. Again, similar patterns are observable in precision and recall, but this time the baseline model does not outperform PAXgene and frozen models in F1-score.

Improved generalization through unsupervised pseudo-label domain adaptation

To enhance generalization of the deep learning models from PT dataset to other domains, we applied pseudo-label domain adaptation step. The baseline model trained on all PT data was used as a starting point for domain adaptation to the MoNuSeg data domain. The MoNuSeg data was divided into train and test datasets based on the division on the original paper. From the MoNuSeg train set, hard positive examples were collected using the detections using baseline model. The training data was generated as described previously in the Unsupervised generation of training samples from confidence map -section. Thus, the annotations provided with the dataset were merely utilized in the evaluation of the model - the DA step was fully unsupervised.

The resulting model after domain adaptation from PT to the MoNuSeg dataset is called MNS DA, and the results for all three datasets are listed on rows 24-26 in Table III. The results show that the unsupervised pseudo-label domain adaptation step enhances detection accuracy in all three cases. Specifically, the sensitivity is improved through domain adaptation; as the model gets samples from the new domain, it starts to detect more nuclei (i.e., the sensitivity increases). While the adaptation to the MoNuSeg data domain could be expected to improve accuracy for the MoNuSeg dataset, it was also the case for the 5-tissue dataset, and perhaps surprisingly, also for PT dataset.

For the sake of clarity and comparability, the F1-scores are also presented as boxplots in the Figure 4. The nuclei counts predicted by different models were plotted against the manual counts and are shown in the Figure 5, where (A) baseline and MNSda models show clear correlation with manual ground truth, (B) fixative-specific models yield more variation and divergence from manual ground truth, and (C) reference result using the two-step segmentation from [5] shows increased variance and bias which does not exist in the deep learning based results.

Effect of cell density on detection accuracy

Further, in order to show that the accuracy is not severely limited by the challenge caused by areas densely populated by nuclei, we conducted the following experiment: the prostate tissue dataset was divided pixelwise into five groups based on spatial density of nuclei, and the detection accuracy for baseline model as well as for baseline + MNS DA model

was determined for each density group. The results, shown in Figure 6 reveal there is only a minor drop in overall accuracy (F1-score) when moving from low-density (< 20 nuclei per $50\mu m$) to high-density (> 40 nuclei per $50\mu m$) areas. While the recall drops due to part of nuclei not being detected, the precision increases as there are less false detections. Visual inspection of the detections (see Supplementary Figure 2) supports the results presented in 6.

Interpretability analysis using layerwise relevance propagation

As a final experiment, we investigate the classification model performance from interpretability viewpoint in order to gain insight into the connection between the spatial patterns in input data and classifier outcome. The results from the model interpretation analysis with LRP method are shown in the Figure 7, which presents typical examples of the areas in an input image considered to be important by the classifier in nucleus detection.

IV. Conclusions and Discussion

In this paper, we implemented a workflow for nuclei detection that utilizes pseudo-label based unsupervised domain adaptation in order to generalize to images from new domains, including different tissues and images from different labs. In addition, we studied the effect of three histopathological sample fixation types on the accuracy of a nuclei detection trained with H&E stained images.

In order to allow further model development, we have shared the workflow implementation and the dataset, these are available at https://github.com/BioimageInformaticsTampere/NucleiDetection. The number of annotated nuclei in the provided dataset is considerably high compared to other public datasets. In addition, it includes three types of tissue fixation and processing and therefore enables development of fixation agnostic nuclei detectors or further study of the topic.

In order to study the effect of sample fixation on the accuracy of a nuclei detection, we trained multiple convolutional neural networks with varying training data. The numerical results are collected in the Table III. The detection results suggest that better accuracy can be achieved when more variation in the sample fixation is present in the training data. This can be concluded when comparing F1-scores of the baseline model, trained with all three fixation types, and the models trained only with a single fixation data. The effect is similar when testing with each of the dataset. Correspondingly, after applying the unsupervised domain adaptation step using MoNuSeg dataset, the F1-score continues to increase.

Furthermore, the similarity between training and test data fixation can be perceived based on the results in the Table III. Detection in images from both PAXgene fixed and formalin fixed tissue sections is distinctly of better in quality compared to the images from frozen tissue sections. The similarity of image quality improves generalization of a machine learning algorithm. Consequently, PAXgene and formalin models detect nuclei nearly equally well nuclei from both PAXgene and formalin fixed tissue section images, yet, these models score low accuracies on the frozen tissue section image data. Similar effect can be concluded based on the precision and recall values. The model trained with noisy frozen data detects quite accurately nuclei from better quality images (PAXgene, formalin). Yet, conversely the PAXgene and formalin models fail (low recall) to detect the majority of the nuclei in images from frozen tissue sections. However, high precision indicates that when the nuclei is found it most often is a true positive. In the PT dataset, frozen model scores lower F1-scores compared to the formalin and PAXgene models. However, a contrary effect can be seen when testing with the 5-tissue data and MoNuSeg data. This seems to also indicate that the formalin and PAXgene fixed tissue sample images are quite similar concerning the image quality and therefore model generalization between these two datasets is decent. It also indicates that in order to generalise across tissue types, a frozen fixed tissue section images provide more variability in the image data domain.

When directly comparing the numerical results of the study, the problem with the metric itself, as well as the challenge caused by using a single coordinate as the ground truth, should be kept in mind. This was considered when analysing the effect of selected radius in the final accuracy values. The Table II presents how the F1-score reaches higher values when the R is increased and thus based on the Table alone one might argue that an even larger value for R should be utilized. The reality however is very different when looking into the examples of different R values in the Figure 3. In the figure, white line is connecting the ground truth coordinate and the corresponding true positive detection. Here, a chain effect can be seen when one false negative is falsely detected as true positive by a prediction that is actually a signal from closeby nuclei (see example D - R3). Thus, when selecting an optimal R, instead of looking at F1-score, the physical size and shape of nuclei present in the datasets should be considered.

In addition to selecting an optimal value of R, the location of a ground truth coordinate needs to be considered. An obvious challenge can be seen in Figure 3 (E), where a benign elongated nucleus is shown. The location of a ground truth coordinate annotation can vary throughout the dataset, complicating the evaluation of true positive samples. Similar problem is faced when ground truth annotation is marked on the edge of a nucleus (see Figure 3 (B) for an example).

Overall results confirm that sample fixation is a significant factor in the variability present in histological image data, and this should be considered when developing a robust and generalizable nuclei detection methods. Based on our results, increased cross domain generalization is achieved when multiple sample fixation methods are present in the training data. Our proposed pseudo-label based unsupervised domain adaptation step was shown to be beneficial for detection accuracy. However, the fully unsupervised domain adaptation step contains the risk of failed adaptation in cases where false positives are detected by the baseline model. In the experiments presented here, such problem did not occur. However, with ambiguous domain changes, e.g. when moving from HE to immunohistochemical staining, the unsupervised pseudo-labeling step may not provide adequate support for the

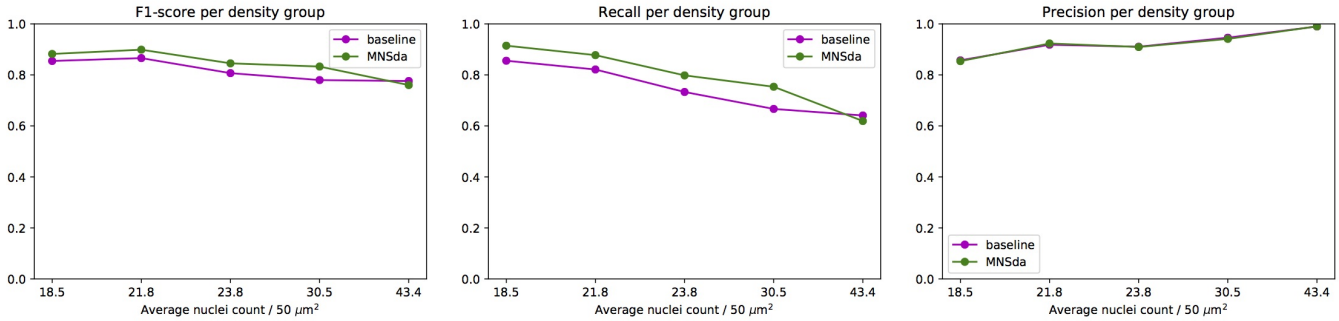

Fig. 6. The nuclei detection F1-scores for PT dataset as a function of cell density. Images were segmented pixelwise into five density groups, and detection accuracy by the baseline and MNS-DA models is presented as F1-score, precision and recall. Note that precision values by the two detection models are almost identical, and not clearly visible due to overlap of curves.

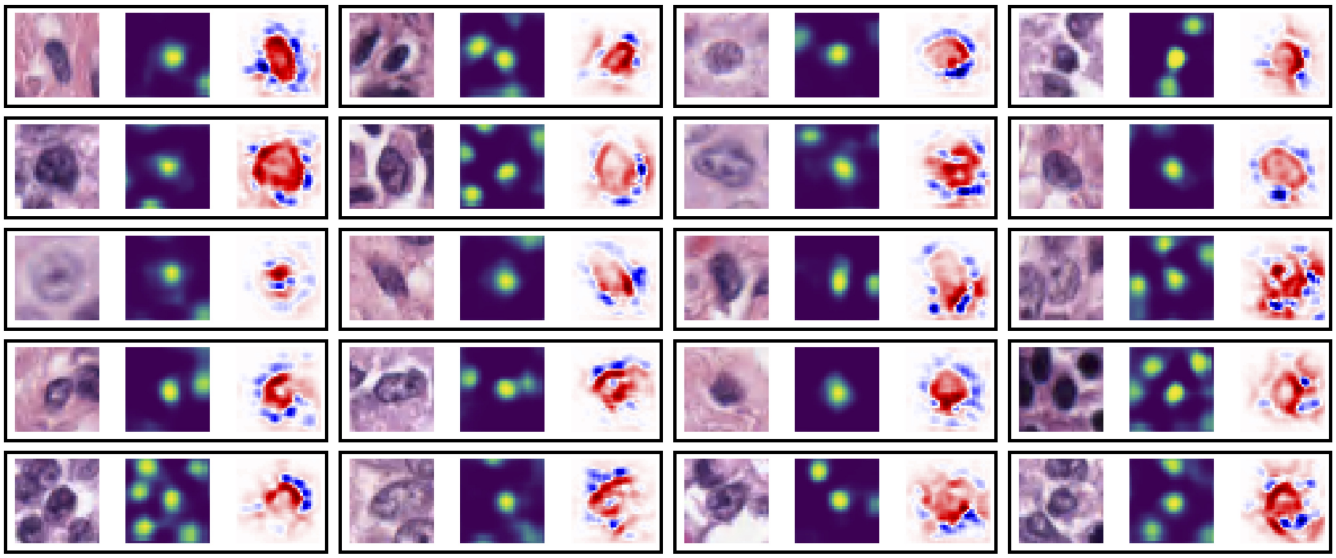

Fig. 7. Examples of the important areas in an input image resulting in nucleus detection using LRP method. Each example visualises a 32x32 image block around a detected nucleus and the corresponding confidence map presenting the network output, and a heatmap for relevant areas provided by LRP method. The red areas present positively relevant areas related to nucleus detection and the blue areas present the negatively relevant areas.

new domain (see Supplementary Figure 3 for such examples using the algorithm presented here and IHC data from [33].

In order to interpret the deep learning models and to discover the reasons behind model decisions, we visually assessed the model decisions using LRP method implemented in the iNNvestigate toolbox [27]. Few examples are shown in the Figure 7. Based on these examples and visual assessment of multiple similar samples, the nuclei detection model seems to find reasonable areas important related to nucleus detection, such as nuclei edges. In addition, if nucleoli are visible in case of a vesicular nucleus, those are often detected as important areas related to nuclei detection. Overall, the relevant areas provided by the LRP algorithm seem to correspond to the areas that a human observer would find relevant as well.

To conclude, this study addresses the question on the importance of the variability present in histological image data that is caused by the tissue fixation process, and the effects this variability has on the accuracy of a nuclei detection algorithm. We have shown with our experiments that the tissue fixation variability in the training data can cause significant differences between nuclei detection accuracies obtained by deep neural network models. The results of study are encouraging, and therefore, call for further research. A suitable next step would be to conduct experiments with bigger and more diverse datasets. In addition, more quantitative analysis of model explanation and interpretation methods are needed to build trust on the deep learning based approaches on these important biological questions. These steps will eventually enable development of a model that can generalize to real world clinical environments.

V. Acknowledgements

We are grateful to Noora Salokorpi for her skillful assistance in validation.

References

[1] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, “Locality sensitive deep learning for detection

and classification of nuclei in routine colon cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[2] H. W. Jackson, J. R. Fischer, V. R. Zanotelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber *et al.*, "The single-cell pathology landscape of breast cancer," *Nature*, vol. 578, no. 7796, pp. 615–620, 2020.

[3] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 10, pp. 4550–4568, 2017.

[4] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *Journal of pathology informatics*, vol. 9, 2018.

[5] G. Högnäs, K. Kivinummi, H. M. Kallio, R. Hieta, P. Ruusuvuori, A. Koskenalho, J. Kesseli, T. L. Tammela, J. Riikonen, J. Ilvesaro *et al.*, "Feasibility of prostate paxgene fixation for molecular research and diagnostic surgical pathology," *The American journal of surgical pathology*, vol. 42, no. 1, pp. 103–115, 2018.

[6] R. Thavarajah, V. K. Mudimbaimannar, J. Elizabeth, U. K. Rao, and K. Ranganathan, "Chemical and physical basics of routine formaldehyde fixation," *Journal of oral and maxillofacial pathology: JOMFP*, vol. 16, no. 3, p. 400, 2012.

[7] M. Kap, F. Smedts, W. Oosterhuis, R. Winther, N. Christensen, B. Reischauer, C. Viertler, D. Groelz, K.-F. Becker, K. Zatloukal *et al.*, "Histological assessment of paxgene tissue fixation and stabilization reagents," *PLoS One*, vol. 6, no. 11, 2011.

[8] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[9] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 17–36.

[10] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2017, pp. 1–6.

[11] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential," *IEEE reviews in biomedical engineering*, vol. 7, pp. 97–114, 2013.

[12] S. Tripathi and S. K. Singh, "Cell nuclei classification in histopathological images using hybrid olconvnet," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, pp. 1–22, 2020.

[13] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. De Souza, A. Baidoshvili, G. Litjens, B. Van Ginneken, I. Nagtegaal, and J. Van Der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 160–163.

[14] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 953–956.

[15] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, J. Shi, and C. Xue, "Adaptive color deconvolution for histological wsi normalization," *Computer methods and programs in biomedicine*, vol. 170, pp. 107–120, 2019.

[16] I. Arvidsson, N. C. Overgaard, K. Åström, and A. Heyden, "Comparison of different augmentation techniques for improved generalization performance for gleason grading," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 923–927.

[17] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical image analysis*, vol. 58, p. 101544, 2019.

[18] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[19] Y. Huang, H. Zheng, C. Liu, X. Ding, and G. K. Rohde, "Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images," *IEEE journal of biomedical and health informatics*, vol. 21, no. 6, pp. 1625–1632, 2017.

[20] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Unsupervised domain adaptation for classification of histopathology whole-slide images," *Frontiers in bioengineering and biotechnology*, vol. 7, 2019.

[21] N. Brieu, A. Meier, A. Kapil, R. Schoenmeyer, C. G. Gavriel, P. D. Caie, and G. Schmidt, "Domain adaptation-based augmentation for weakly supervised nuclei detection," *arXiv preprint arXiv:1907.04681*, 2019.

[22] K. Liimatainen, L. Kananen, L. Latonen, and P. Ruusuvuori, "Iterative unsupervised domain adaptation for generalized cell detection from brightfield z-stacks," *BMC bioinformatics*, vol. 20, no. 1, p. 80, 2019.

[23] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 5–22.

[24] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 55–76.

[25] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 169–191.

[26] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[27] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[28] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH image to imagej: 25 years of image analysis," *Nature methods*, vol. 9, no. 7, p. 671, 2012.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] F. Chollet *et al.*, "Keras," 2015, available from https://github.com/fchollet/keras.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[33] M. Valkonen, J. Isola, O. Ylinen, V. Muhonen, A. Saxlin, T. Tolonen, M. Nykter, and P. Ruusuvuori, "Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for er, pr, and ki-67," *IEEE transactions on medical imaging*, 2019.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.