Medical Image Analysis 58 (2019) 101547

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# Learning to detect lymphocytes in immunohistochemistry with deep learning

Zaneta Swiderska-Chadaj[a,*], Hans Pinckaers[a], Mart van Rijthoven[a], Maschenka Balkenhol[a], Margarita Melnikova[a,c,d], Oscar Geessink[a], Quirine Manson[e], Mark Sherman[f], Antonio Polonia[g], Jeremy Parry[h], Mustapha Abubakar[i], Geert Litjens[a], Jeroen van der Laak[a,b], Francesco Ciompi[a]

[a] Department of Pathology, Radboud University Medical Center, The Netherlands
[b] Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden
[c] Department of Clinical Medicine, Aarhus University, Denmark
[d] Institute of Pathology, Randers Regional Hospital, Denmark
[e] Department of Pathology, University Medical Center, Utrecht, The Netherlands
[f] Mayo Clinic, Jacksonville, Florida, USA
[g] Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal
[h] Fiona Stanley Hospital, Murdoch, Perth, Western Australia
[i] Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA

ARTICLE INFO

*Article history:*
Received 16 January 2019
Revised 12 August 2019
Accepted 20 August 2019
Available online 21 August 2019

*Keywords:*
Deep learning
Immune cell detection
Computational pathology
Immunohistochemistry

ABSTRACT

The immune system is of critical importance in the development of cancer. The evasion of destruction by the immune system is one of the emerging hallmarks of cancer. We have built a dataset of 171,166 manually annotated $CD3^+$ and $CD8^+$ cells, which we used to train deep learning algorithms for automatic detection of lymphocytes in histopathology images to better quantify immune response. Moreover, we investigate the effectiveness of four deep learning based methods when different subcompartments of the whole-slide image are considered: normal tissue areas, areas with immune cell clusters, and areas containing artifacts. We have compared the proposed methods in breast, colon and prostate cancer tissue slides collected from nine different medical centers. Finally, we report the results of an observer study on lymphocyte quantification, which involved four pathologists from different medical centers, and compare their performance with the automatic detection. The results give insights on the applicability of the proposed methods for clinical use. U-Net obtained the highest performance with an F1-score of 0.78 and the highest agreement with manual evaluation ($\kappa = 0.72$), whereas the average pathologists agreement with reference standard was $\kappa = 0.64$. The test set and the automatic evaluation procedure are publicly available at lyon19.grand-challenge.org.



## 1. Introduction

The immune system is of critical importance in preventing cancer development. The immune system can kill cells that are dividing uncontrollably, resulting in prevention of cancerous growth. However, cancers can have specific mutations that help it evade this immune destruction, which is one of the hallmarks of cancer development (Hanahan and Weinberg, 2011). As such, understanding the ability of immune cells to prevent cancer development or kill cancer cells is an active topic in cancer research. In particular, recent advances in immunotherapy strategies (Khalil et al., 2016; Rosenberg and Restifo, 2015; Xie et al., 2017) have further increased the interest in understanding the mechanism of immune response to cancer. One of the current hypotheses states that the balance between lymphocytes, a subset of immune cells, with pro- and anti-inflammatory function is important for disease progression (Galon et al., 2006). Specifically, lymphocytes that occur within the tumor area and with the tumor-associated stroma are of interest. These lymphocytes are called tumor infiltrating lymphocytes (TILs). Studies have shown that the presence of TILs is related to patient prognosis after undergoing surgery or immunotherapy (Coussens et al., 2013; Syn et al., 2017; Vánky et al., 1986). Therefore, detection and quantification of lymphocytes has the potential

* Corresponding author.
*E-mail address:* zaneta.swiderska@radboudumc.nl (Z. Swiderska-Chadaj).

https://doi.org/10.1016/j.media.2019.101547
1361-8415/

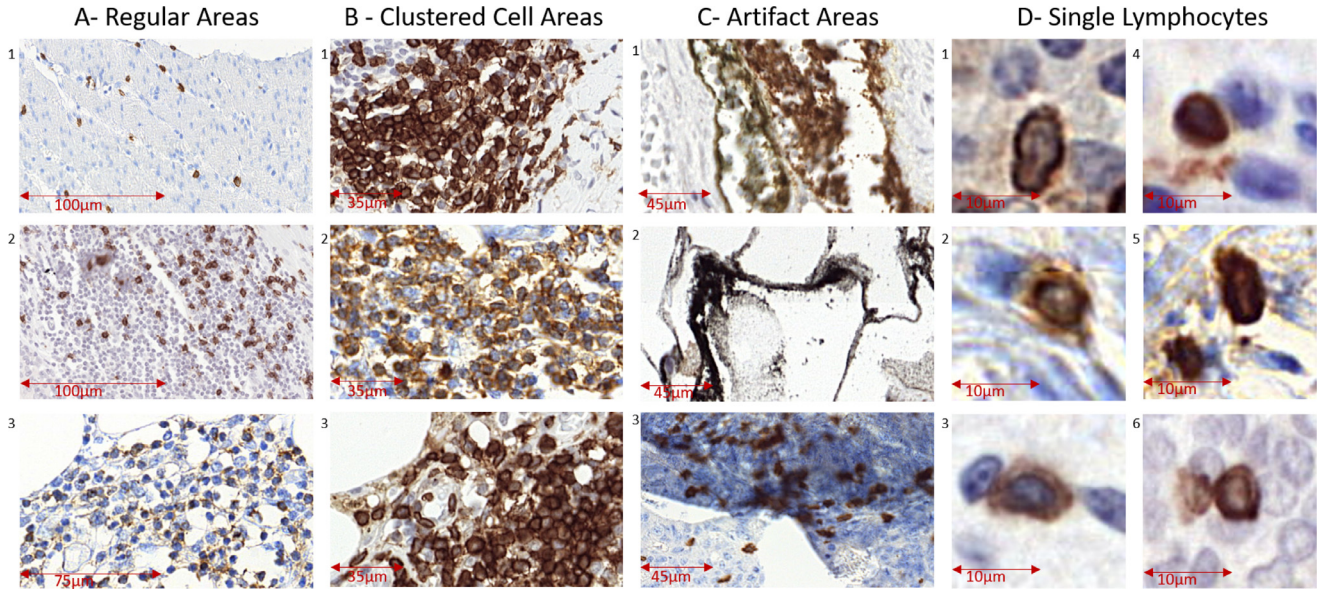


**Fig. 1.** Examples of image regions used in this study, containing areas with a regular lymphocyte distribution (A), areas with lymphocyte clusters (B) and areas with artifacts or damaged tissue (C), where first row: stain artifacts; second row: ink; third row: tissue folds. Patches containing a single lymphocyte are depicted as well (D), in order to show the difference in appearance.

to provide biomarkers with strong prognostic and predictive power for cancer progression and therapeutic efficacy (Varn et al., 2017).

An important tool to detect and quantify specific cell populations in histopathology is immunohistochemistry (IHC). IHC is a technique that allows to stain specific cell types, including lymphocytes, by attaching a colored label to a specific antigen that is expressed by a cell, making it distinguishable from other types of cells. In the context of TILs, widely used immune cell markers are CD3 (general T-cell markers) and CD8 (cytotoxic T-cell marker). Both CD3 and CD8 are membrane markers, meaning that they target an antigen in the cells' membrane, resulting in a colored ring in positive cells (Fig. 1).

To quantify the immune cells in immmunohistochemistry, visual assessment via light microscopy is the standard approach in research. This procedure requires training by pathologists and suffers from inter- and intra-observer variability (Klauschen et al., 2018). The rise of digital pathology has fostered the development of computer algorithms based on machine learning for the analysis of histopathology whole-slide images (WSI). These methods have the potential to make the transition from subjective visual estimation to reproducible accurate quantification of cells via automatic detection. Furthermore, moving from overall quantification to detection of each lymphocyte in the slide allows analysis of complex spatial patterns such as cell density distributions and cell-to-cell interactions, which are currently not assessed due to a lack of standardized methodology, time and difficulty in making such assessment (Klauschen et al., 2018).

It is easy to show that a large variety of challenges are present in tissue samples stained with lymphocyte markers, making detection a non-trivial task. In this study, we define three different types of areas containing T-cells that can be distinguished in CD3 and CD8 stained slides (see Fig. 1(a)–(c)), namely (a) *regular tissue* areas, which are areas with a regular lymphocyte distribution without artifacts, damaged or large areas of cell clusters; (b) *lymphocyte cluster* areas, which contain significant number of clustered T-cells with vague cell boundaries; (c) *artifact* areas, which include various types of staining artifacts, i.e., areas with a range of non-specific stain, damaged regions or ink. Quantification of T-cells is relatively straightforward in regions of category (a), whereas in categories (b) and (c) detection and accurate quantification of lymphocytes can be very challenging. Such regions are often not considered or discussed in scientific literature (Garcia et al., 2017) but are highly relevant for procedures that aim to fully automatically analyze immunohistochemistry.

### 1.1. Related work

There has been a large corpus of methods for cell detection in digital histopathology slides based on classical image analysis and machine learning approaches, such as morphological operations, region growing, analysis of hand-crafted features and image classifications.

In recent years, deep learning (LeCun, 2015) has brought a revolution to the field of pattern analysis and machine learning, by providing algorithms with the capacity to learn complex representations from the raw data itself, achieving human and even super-human level performance in some fields, including medical image analysis (Ehteshami Bejnordi et al., 2017; Esteva et al., 2017; Gulshan et al., 2016). A review on recent work on automated detection of tumor-infiltrating lymphocytes was recently published (Klauschen et al., 2018), which covers both classical and more recent machine learning approaches, including deep learning.

Automatic cell detection in digitized histopathology tissue sections can be tackled by many different approaches. However, in the context of deep learning, we can define two main categories: (1) *"learning to segment cells"* (segmentation approach) and (2) *"learning to regress cell location"* (regression approach).

*Learning to segment cells.* This category contains methods that aim at detecting cells by object segmentation, or in the form of patch classification. The location of cells is predicted by post-processing of the segmentation map. In deep learning, patch classification is mostly done by convolutional neural networks (CNN), and in particular by fully-convolutional networks (FCN) (Long et al., 2015), which can be efficiently used in segmentation tasks. Additionally, much of research on pixel classification methods in the context of semantic segmentation in medical imaging relies on the U-Net architecture or derivatives (Ronneberger et al., 2015).

There has been a substantial amount of research on automatic cell detection in histopathology images (Chen, 2014; Garcia et al., 2017; Janowczyk and Madabhushi, 2016; Saltz et al., 2018; Xing and Yang, 2016; Xu et al., 2016). Most methods are based on patch classification with convolutional networks (Chen (2014); Garcia et al. (2017); Janowczyk and Madabhushi (2016); Saltz et al. (2018)) or with an auto-encoder architecture (Xu et al. (2016)). It should be noted, that most of those methods are developed for the analysis of H&E stained specimens (Janowczyk and Madabhushi (2016); Saltz et al. (2018); Xing and Yang (2016)).

*Learning to regress cell location.* This category contains methods that aim at predicting the location of cells directly, for example by predicting the position of nuclei, or by predicting bounding boxes that contain the entire cell. One way to formulate such an approach with deep learning is by predicting the coordinates of a cell directly as the output of a deep learning network. In this context, the Locality Sensitive Method (LSM) proposed in Sirinukunwattana et al. (2016) represents a seminal work in computational pathology, where several types of nuclei in colon cancer specimens stained with H&E were both detected and classified. For the detection part, the method used convolutional networks to directly predict the location of target cells, instead of predicting labels of input patches. Xie et al. (2015b) addressed the detection of $Ki67^+$ cells in neuroendocrine tumors with a system that computes a map of detected cells by using a structural regression approach based on FCN. A similar approach was proposed by Xie et al. (2015a) where network is learning an offset vector referring to cell locations.

In the computer vision community, methods able to predict bounding boxes of target objects have become very popular in recent years. The published methods strive to be both effective and efficient. In particular, methods like Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2016) and You Only Look Once (YOLO) (Redmon et al., 2015) have quickly become reference approaches for object detection in natural images. Recently, improvements both in terms of accuracy and efficiency have been proposed for YOLO (Redmon and Farhadi, 2016; Redmon and Farhadi, 2018). However, to the best of our knowledge, the only work on cell detection in computational pathology using deep learning methods that predict bounding boxes is our recent work on lymphocyte detection (van Rijthoven et al., 2018; Swiderska-Chadaj et al., 2018), which also represents preliminary work of this paper.

### 1.2. Our contribution

In this paper, we address the problem of automatic detection of $CD3^+$ and $CD8^+$ immune cells in a multi-center set of whole-slide images of breast, prostate and colon cancer using deep learning. Preliminary results of these research were presented in van Rijthoven et al. (2018) and Swiderska-Chadaj et al. (2018). This paper represents a significant extension of previous work, which includes: pathologist evaluation, comparison of deep learning results with pathologist evaluation, large multicenter dataset, modification of applied network architectures, analysis of the results at multiple levels, as well as release of a web platform containing the test set and automatic evaluation procedure. In presented work we can distinguish four main contributions.:

First, we developed and compared four different deep learning approaches that address both the $CD3^+$ and the $CD8^+$ cell detection problem at whole-slide image level from different angles. In particular, we developed methods based on *"learning to segment cells"* based on the fully-convolutional networks (Long et al., 2015) and U-Net (Ronneberger et al., 2015), as well as a methods to *"learning to regress cell locations"* based on You Only Look Once approach (Redmon et al., 2015), and a locality sensitive method (Sirinukunwattana et al., 2016). All methods were trained and validated using exactly the same data sets and comparable amounts of trainable parameters, in order to provide a fair overview of potential and limitations of the considered approaches. Together with overall performance, we also investigate and report the robustness of each developed method in the presence of challenging areas, therefore reporting detailed performance for the sub-categories depicted in Fig. 1. Moreover, we stratify results based on medical centers, staining type (CD3, CD8) and organ (breast, colon, prostate).

Second, in order to train and validate the considered deep learning methods, we have built a unique dataset of 83 whole-slide images collected from 9 different pathology laboratories in the Netherlands, in which we selected 932 Region of Interests (ROIs) and manually annotated 171,166 lymphocytes in these ROIs. To the best of our knowledge, this is by far the largest set of data that has been built in the context of developing and validating deep learning methods for the detection of lymphocytes, which in terms of ROIs represents an increase of 70x compared to previous work (Garcia et al., 2017). Furthermore, differently from previous work, we specifically focus our attention and report results on regions that are known to be challenging in the context of TIL detection in IHC, namely regions containing dense clusters of lymphocytes, regions with abundant background staining, and regions with stain artifacts and ink, from which false positive detections are known to originate.

Third, in order to assess the performance of developed methods with respect to experienced pathologists, we designed an observer study and involved four pathologists from four different medical centers. We report the results of the comparison of the best deep learning method and the pathologists.

Last, we made the 441 ROIs of the test set and an evaluation metric for LYmphocyte detectiON (LYON) publicly available at lyon19.grand-challenge.org. This allows the scientific community to compare the performance of other approaches with the results presented in this paper using exactly the same test set and the same evaluation procedure. This is the first publicly available test set to assess the performance of lymphocyte detection in immunohistochemistry.

## 2. Materials

*Whole-slide images.* For this study, we collected 83 glass slides of breast (33 slides), prostate (22 slides), and colon (28 slides) cancer specimens from nine different medical centers in the Netherlands. All slides were stained with an antibody against CD3 or CD8. In order to introduce stain variability, tissue samples were stained in the local lab of the nine participating medical centers. In this way, we covered a wide range of staining protocols and provide our deep learning models with an heterogeneous appearance of tissue samples. *Slides were subsequently digitized using a Pannoramic 250 Flash II scanner (3DHistech, Hungary), resulting in WSIs with a pixel size of 0.24 μm/px (magnification 20x).*

*Manual annotations.* In order to train and validate deep learning methods, approximately 11 regions of interest (ROI) per WSI were selected by a trained human analyst, with the aim of making exhaustive annotations of lymphocytes in each ROI (Fig. 2). For this task, the in-house developed open-source ASAP software (Litjens et al., 2018) was used. Selected regions were distributed across the areas of interest: (1) areas with regular lymphocyte distribution, (2) clustered cells, and (3) staining or tissue artifacts (see Fig. 1). Within the selected ROIs, lymphocytes were manually annotated exhaustively. Annotations were made by three human analysts ($E_1$, $E_2$ and $E_3$) trained to execute this task. A set of locations $\{(x_l, y_l)\}$ corresponding to the center of the nucleus of

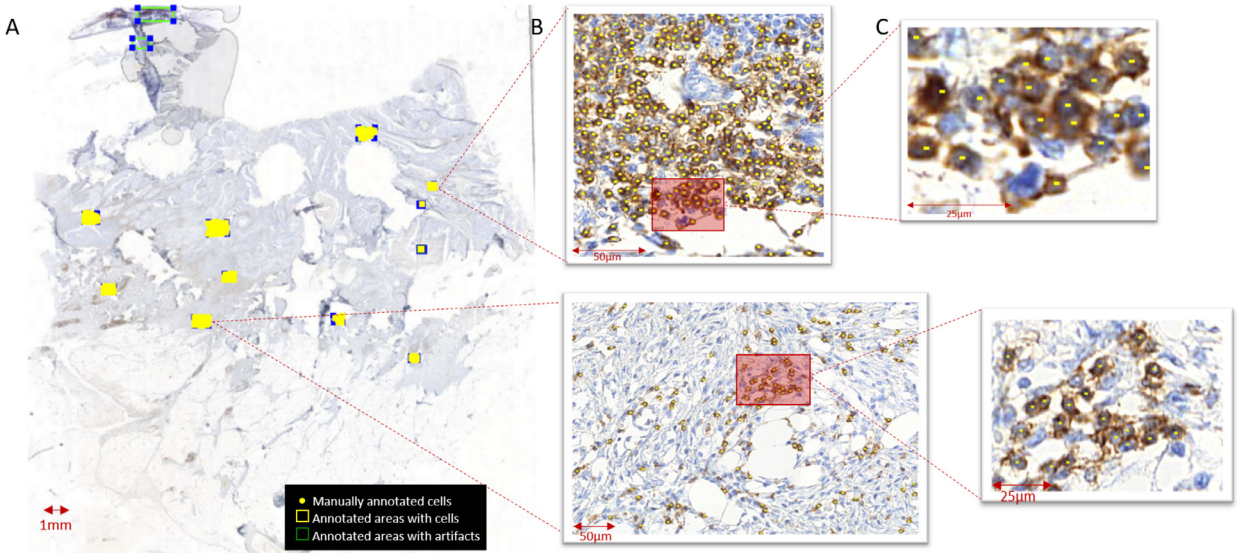


**Fig. 2.** Example of manually annotated lymphocytes in ROIs of a WSI, where: A- whole slide image with marked on a yellow annotated ROIs with cells and on green ROIs with artifacts; B and C - zoomed areas where manually annotated cells are marked in yellow.

each lymphocyte was established, where $l = 1, \ldots, N$. As a result, $N = 171{,}166$ lymphocytes in 932 ROIs were annotated, which were used as reference standard.

*Datasets.* The 83 whole-slide images were divided into a training (n=37), validation (n=6) and test set (n=40). Training and validation slides were selected from two medical centers, whereas the independent set of test slides was created using data from eight centres. Data from one lab was shared across two training and test, but care was taken that the same patient did not appear in both sets. The validation set contained two images of each considered organ (breast, colon, prostate), one stained with CD3 and one stained with CD8. The test set contained fifteen images of colon cancer and breast cancer, and ten images of prostate cancer, with the same proportion of slides stained with CD3 and CD8.

## 3. Learning to detect lymphocytes

Two categories of cell detection methods based on deep learning techniques were investigated, namely *learning to segment cells* and *learning to regress cell location*. For each category, two different approaches for lymphocyte detection were developed: (1) patch classification using a fully-convolutional network, and (2) semantic segmentation using U-Net, for *"learning to segment cells"*; (3) bounding box detection based on the YOLO network, and (4) prediction of cell center locations by the LSM network, for *"learning to regress cell location"*.

### 3.1. Learning to segment cells

We formulate the pixel classification problem as a multi-class problem. For each provided lymphocyte location $(x_l, y_l)$, a target map with three classes was generated (Fig. 3). For each manually annotated location $(x_l, y_l)$, regions approximating the *cell body* and the *membrane* of each lymphocytes were defined, with radii $r_B$ and $r_M$, respectively (see Fig. 3). Additionally, we also considered the *background* class. The value of $r_M$ was determined based on the average diameter of a lymphocyte, which is in the range of 6–8 $\mu m$. The values of $r_M$ and $r_B$ were established as $2.88\mu m$ and $2.4\mu m$ respectively. This target map is used in the learning procedure of both FCN and U-Net.

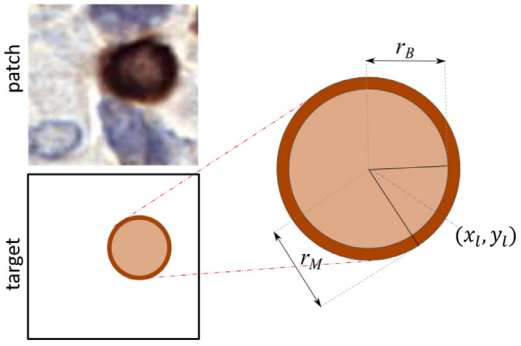


**Fig. 3.** Example of target map generated from the annotation of a location $(x_l, y_l)$ in the input patch that contains one lymphocyte. This target map is used to train U-Net and the FCN, where $r_B = 2.4\ \mu m$ is the radius of the cell body; $r_M = 2.88\ \mu m$ is the radius of the whole cell.

*FCN: Fully Convolutional Networks.* The FCN architecture we used in this paper contains twelve convolutional layers. The first two layers are interleaved with pooling layers to improve spatial invariance. Every convolutional layer is followed by a rectified linear unit, except the last layer where a softmax nonlinearity is applied. Training was done using the mask depicted in Fig. 3 as a 3-class problem.

At inference time, a dense prediction is obtained by means of the shift-and-stitch technique (Long et al., 2015; Sermanet et al., 2014), which allows to prevent loss of resolution due to the pooling layers. In practice, the implementation described in Long et al. (2015) is used to efficiently produce results identical to shift-and-stitch using *filter dilation*: all layers $D_i$ with stride $s_i > 1$ are applied using now stride $\hat{s}_i = 1$, and filters of subsequent layers $L_j$ (both convolutional and pooling, $j > l$) are dilated by a factor $\hat{d}_j = \prod_{i<j} s_i$. As a result, a likelihood map at the resolution of the input image is obtained.

In order to obtain the final detection result, likelihood maps are post-processed by smoothing with a Gaussian filter ($\sigma = 1\mu m$) and subsequently detecting local maxima within a neighbourhood of $2.9\mu m$. These values were established based on typical lymphocyte size.

*U-Net.* In our study, we extended the original U-Net (Ronneberger et al., 2015) architecture by adding a spatial dropout layers between convolutional layers, with the aim of reducing overfitting. In order to train the U-Net, we used target segmentation masks as the one depicted in Fig. 3. The output of the U-Net network is post-processed using the same algorithm as for the FCN.

### 3.2. Learning to regress cell location

We developed two methods for predicting the location of lymphocytes, namely YOLLO (You Only Look at Lymphocyte Once) and LSM. In particular, a recently presented optimization of the YOLLO method, tailored to detect lymphocytes in immunohistochemically stained whole-slide images is used.

*YOLLO: You Only Look atLymphocytes Once.* We adapted the original YOLO architecture for the problem of lymphocyte detection in IHC (van Rijthoven et al., 2018). YOLLO processes an image by first dividing it into a grid, and then predicting bounding boxes that should contain lymphocytes for each grid cell. For each bounding box, the network outputs a confidence level $C$, which is a measure of how sure the model is that a lymphocyte is captured by that bounding box, instead of background.

The loss function presented in Redmon and Farhadi (2016) was initially used for training, which was simplified as proposed in van Rijthoven et al. (2018). A prediction was considered correct if the predicted bounding box had an intersection over union (IoU) $\geq$ 0.5 with a true bounding box. During inference, predicted bounding boxes with overlap are considered as detecting the same lymphocyte. Therefore, non-maximum suppression is applied in order to keep only one prediction per lymphocyte at test time. The output confidence $C$ of the selected bounding box is considered as the lymphocyte score.

*LSM: Locality Sensitive Method.* The goal of the *locality sensitive method* presented in Sirinukunwattana et al. (2016) is to learn a mapping function $\mathcal{F}$ from a 2D input domain (i.e., an image patch) $I(x, y)$ to $M$ output locations $(x_m, y_m)$, which correspond to locations of centers of lymphocytes, as well as a likelihood parameter $h_m$ for each location: $\{x_m, y_m, h_m\} = \mathcal{F}(I(x, y))$, for $m = 1, \ldots, M$. In this way, $M$ represents the *maximum* number of lymphocytes that are expected to be found in an input patch $I(x, y)$, and the output variable $h_m$ allows to control the likelihood of each predicted location to contain an actual lymphocyte. If a patch contains $T$ lymphocytes and $T < M$, then some values of $h_m$ will be $\approx 0$. Differently from Sirinukunwattana et al. (2016), we implemented a *spatially constrained layer* that predicts an output map which has the same size as the input patch.

The output of the network is a 2D map of Gaussian-like profiles that indicate the predicted locations of lymphocytes. In order to improve the robustness of predictions, as in Sirinukunwattana et al. (2016), we process each input patch multiple times and accumulate the predicted profiles. For this purpose, we shifted the patch from its central position $n_s = 4$ times per image dimension, therefore accumulating $(n_s + 1)^2 = 25$ predictions per patch. The final set of locations $(x, y)$ of predicted lymphocytes is extracted by detecting local maxima on prediction maps. The value of the maximum is used as lymphocyte score.

### 3.3. Model parameters

Methods were developed using both Tensorflow (Abadi et al., 2015) and Keras (Chollet et al., 2015), or Theano (Al-Rfou et al. (2016)) and Lasagne (Dieleman et al., 2015). All methods were trained in a supervised fashion using exactly the same ROIs from the same training and validation slides. All network architectures were designed in order to have a comparable amount of trainable parameters. We considered each approach as consisting of several parts, where one was considered as the main component for learning data representation, and the rest was a model-specific and a task-specific component. Applied architectures were designed in order to actually have a comparable number of trainable parameters in the part of the model that is mostly responsible for learning data representation. The contraction path of U-Net (4M) was used as a reference and the architecture of all other models was adapted to have a representation learning component comparable to the one of U-Net. The difference in the final number of trainable parameters is related to the final layers of each model (number and size of convolutional layers, pooling layers or spatial constrained layers), which represent the model-specific and task-specific components. During training, the performance of each network on the validation set were monitored. Hyperparameters (input image size, loss function, optimizer, batch size, learning rate, dropout factor) were optimized for each network independently. This is needed because no default set of hyperparameters exist that is optimal for all different architectures.

**Table 1**
Overview of parameters for each developed approach. CP: contraction path of U-Net model; SGD: stochastic gradient descent; CCR: categorical crossentropy; BCR: binary crossentropy; NM: Nesterov Momentum. Note that the number of trainable parameters for U-Net is reported separately for the feature extraction and for the segmentation part.

| | FCN | U-Net | YOLLO | LSM |
|---|---|---|---|---|
| Trainable parameters | 7M | 8M(CP 4M) | 6M | 4M |
| Input size (px) | 284X284 | 256x256 | 256x256 | 27x27 |
| Output size (px) | 1x1 | 256x256 | 256x256 | 27x27 |
| Dropout factor | 0 | 0.25 | - | 0.5 |
| Optimizer | ADAM | SGD | ADAM | NM |
| Loss function | CCR | CCR | yolo loss | BCR |
| Learning rate | 0.0001 | 0.0005 | 0.00005 | 0.05 |
| Batch size | 32 | 1 | 4 | 32 |
| Input pixel size ($\mu m$) | 0.24 | 0.49 | 0.49 | 0.49 |

### 3.4. Training deep learning models

Applied deep learning models were trained in an end-to-end fashion using an individual hyperparameter optimization to obtain the best possible F1-score on the validation set. The overview of parameters is presented in Table 1. The input tiles were in the RGB color scale with pixel size of 0.49$\mu m$ for U-Net, YOLLO and LSM methods, and with pixel size of 0.24$\mu m$ for FCN method. The same image augmentation techniques, including modification of brightness, contrast, saturation, rotation and noise adding, were applied for each network.

The FCN training was performed using the Adam optimization algorithm and the Categorical Cross Entropy loss function. The batch size was 32 with input patches of size 284x284 pixels. The learning rate was set to 0.0001 with decay of 0.75 when no improvement on validation loss was observed for 25 consecutive epochs, which resulted in 45,000 mini-batch iterations.

The U-Net was optimized using stochastic gradient descent with a categorical cross entropy loss function. The batch size was set to 1, with input patch size of 256x256 pixels. The training was performed by 100 epochs of 200 mini-batches with a learning rate equal to 0.0005. The YOLLO method was trained with Adam optimization technique and the *yolo* loss function. The sampling strategy as proposed in van Rijthoven et al. (2018) was not applied in this paper, in order to keep the YOLLO method comparable with the other presented methods. The batch size was set to 4 with the input patch size of 256x256. The training was performed for 200 epochs with a learning rate equal to 0.00005. Additionally, the

bounding box size of YOLLO, was determined based on the average size of a lymphocyte and was established as 12 pixels (5,88$\mu m$) for a pixel size of 0.49$\mu m$, and parameters of the loss function were set to $\lambda_{coord} = 5$ and $\lambda_{noobj} = 1$. For effective elimination of unsure and redundant boxes, thresholds for object confidence and non maximum suppression were set to 0.2 and 0.1, respectively. For the LSM approach, the parameters of the spatial constraints layer were set to $d = 4$ and $M = 2$.

The LSM network was trained with Nesterov Momentum optimization algorithm and Binary Crossentropy as the loss function. The batch size was set to 32 for the 27x27 input images. The training was performed for 1000 epochs with a learning rate equal to 0.05.

## 4. Observer study

In order to assess the potential clinical applicability of the proposed deep learning methods for cell detection, we designed an observer study to compare the performance of our best-performing automatic algorithm with pathologists on the task of lymphocyte quantification. However, counting all cells in whole-slide images is generally not done by pathologists. Therefore, we limited the task to *visual estimation* of the amount of stained lymphocytes within given regions of interest, which more closely mimics how IHC is used in clinical practice.

For this purpose, we designed a web-based user interface in which all the ROIs belonging to the 40 WSIs in the test set were visualized. Subsequently we asked four pathologists (MS, AP, JP, MA), each from a different medical center, to visually assess the amount of stained lymphocytes in each ROI. Each region was categorized into one of the seven following classes: (a) no stained lymphocyte, (b) 1–5 stained lymphocyte(s), (c) 6–10 stained lymphocytes, (d) 11–20 stained lymphocytes, (e) 21–50 stained lymphocytes, (f) 51–200 stained lymphocytes, (g) $>200$ stained lymphocytes. All pathologists independently scored all ROIs in the test set.

## 5. Experimental results

In this section, we first report the quantitative results of the comparison of four developed automatic cell detection algorithms in terms of F1-score and FROC analysis. FROC is similar to ROC analysis, except that the false positive rate on the x-axis is replaced by the number of false positives per image. The FROC is more appropriate for detection tasks as it more naturally deals with false negatives (e.g. limited sensitivity), whereas the ROC does not account for the 'not found' category (Moskowitz, 2017). Second, we analyze the inter-observer variability in estimating the amount of lymphocytes by the pathologists. Finally, we compare the best performing automatic detection algorithm with human performance. In all cases, manual annotations of lymphocytes as described in Section 2 were used as a reference standard, both for cell detection and for cell counting.

### 5.1. Hit criterion

In order to address the evaluation of cell detection performance, we defined a *hit criterion* by considering a circular area of radius $r$=4 $\mu m$ centered on each manually annotated cell location as a valid region for the detection of that specific lymphocyte. The value of $r$ was established based on the typical radius of lymphocytes. If a detected cell is at a valid distance from the reference standard, it is counted as a true positive, otherwise it is counted as a false positive. If a manually annotated cell was not detected, it was regarded as a false negative. Based on these criteria, the following metrics were computed: Precision, Recall, F1-score and the Free-response Receiver Operating Characteristic (FROC) curve.

**Table 2**
Detection performance of developed deep learning algorithms on the test set. Bold indicates the best method in terms of F1-score.

| Area type | Method | F1-score | Precision | Recall |
|---|---|---|---|---|
| Regular tissue | FCN | 0.71 | 0.69 | 0.74 |
| | LSM | 0.74 | 0.63 | 0.90 |
| | YOLLO | 0.78 | 0.70 | 0.88 |
| | U-Net | **0.82** | 0.83 | 0.81 |
| Cell clusters | FCN | 0.71 | 0.80 | 0.64 |
| | LSM | 0.70 | 0.76 | 0.65 |
| | YOLLO | 0.79 | 0.77 | 0.81 |
| | U-Net | **0.81** | 0.87 | 0.77 |
| Artifacts | FCN | **0.47** | 0.35 | 0.73 |
| | LSM | 0.25 | 0.14 | 0.84 |
| | YOLLO | 0.19 | 0.11 | 0.86 |
| | U-Net | **0.47** | 0.34 | 0.77 |
| All areas | FCN | 0.69 | 0.67 | 0.70 |
| | LSM | 0.63 | 0.52 | 0.80 |
| | YOLLO | 0.64 | 0.51 | 0.85 |
| | U-Net | **0.78** | 0.76 | 0.79 |

### 5.2. Lymphocyte detection

*Performance of automatic cell detection.* The four developed deep learning algorithms were applied to all ROIs of the 40 WSIs in the test set. In total, 441 ROIs covering areas between 0.2 $mm^2$ and 23.8 $mm^2$ (on average 12 $mm^2$) were exhaustively annotated in the test set.

Detection performance was first evaluated in terms of F1-score. Performance was computed both for all ROIs and for subsets of ROIs belonging to different types of areas, namely (a) regular areas, (b) cells clusters, (c) artifacts and damaged tissue. Quantitative results are presented in Table 2, and qualitative results are depicted in Fig. 4.

Overall, F1-scores were in the range of 0.71-0.82 for regular tissue areas. The FCN and LSM approaches achieved F1-score values in range 0.71-0.74. The highest performances were achieved by U-Net and YOLLO with F1-scores of 0.82 and 0.78 respectively. In the presence of clusters of cells, F1-scores were in the range of 0.70-0.81, while in the presence of artifacts, lower F1-score values were observed, in the range of 0.19-0.47. The best result for all analyzed regions was achieved with the U-Net approach (F1-score of 0.78). It should be noted that results for YOLLO and U-Net approaches were similar for regular ROIs and ROIs with clustered cells. The U-Net approach seems to work considerably better than YOLLO and LSM approaches in the presence of artifacts (F1-score of 0.47). Table 2 and Figs. 4–6 present results for each of the considered approaches. The Bland-Altman plots (Fig. 6) visualize an agreement between deep learning models and manual annotations. We can observed that U-Net achieved the highest result (F1-score equal 0.82) for regular tissue areas, and relatively high F1-scores were also obtained for difficult areas such as cell clusters (highest F1-score of 0.81) and artifacts (highest F1-score of 0.47).

Moreover, results were stratified based on medical centers (Table 3), staining types (Table 4) and organs (Table 5). In all cases, we computed the average, the maximum difference and the standard deviation of F1-scores across multiple settings (i.e., labs, stainings, organs). Low standard deviation and max difference are indicators of robustness of a method across different settings. We can observe that the standard deviation (SD) for medical centers is in a range of 0.08-0.17, where the best result was achieved by the U-Net method (SD=0.08). Other methods (FCN, YOLLO and LSM) had a higher SD (in a range 0.14-0.17). Results for different staining types (CD3, CD8) present a high robustness across all approaches (SD value was in a range 0.01-0.05). Results stratified

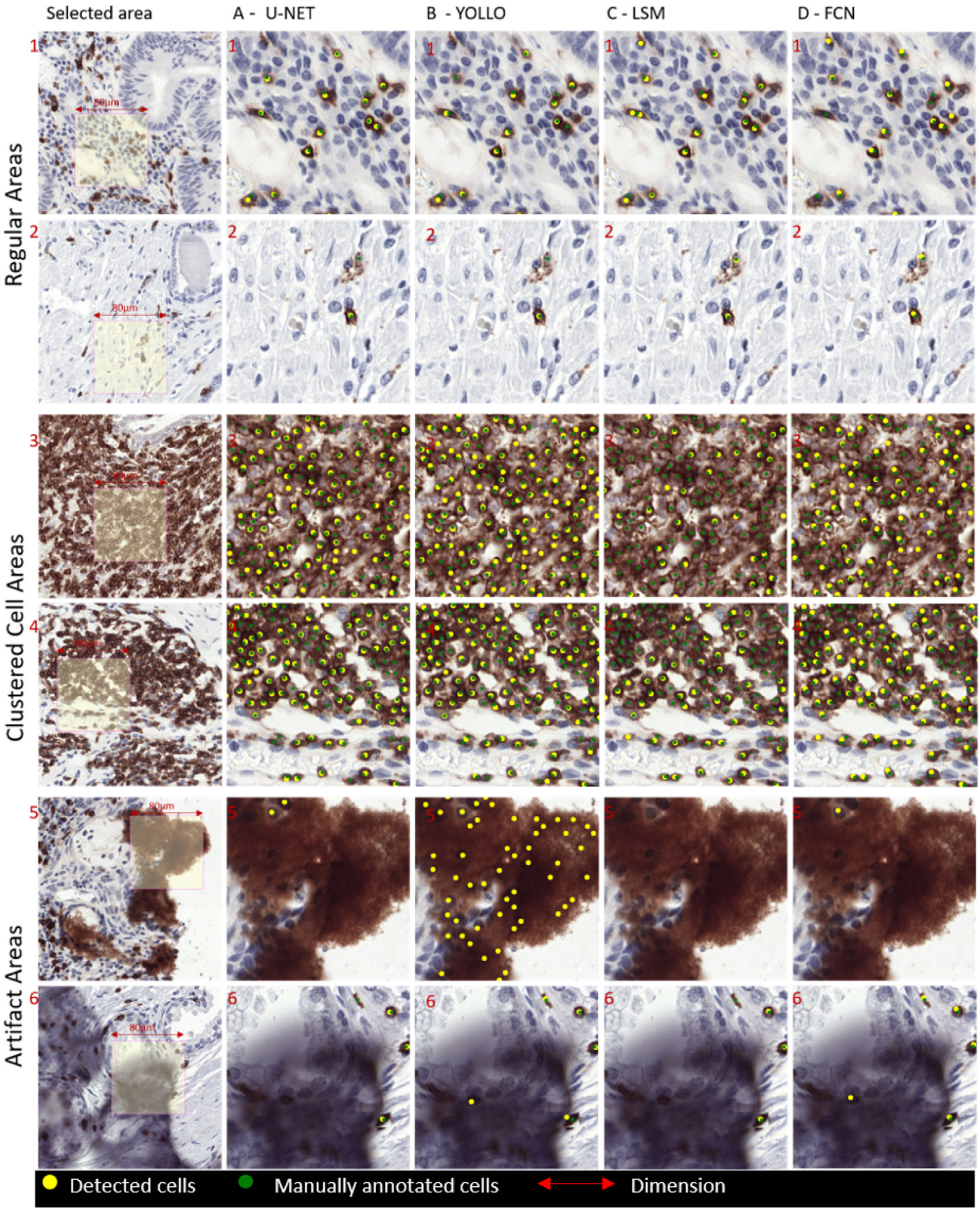


**Fig. 4.** Presentation of the deep learning methods' performance on the test slides, where: green- manually annotated cells, yellow- detected cells. Notice that U-Net method achieved the best performance for all areas, and we can observe the most accurate cell detection for all categories (A). YOLLO method achieved the lower F1-score for artifact areas, where we can observe many false positive detections (B5). LSM and FCN got the lower performance for clustered cell areas, where we can observe many missing detections (C3, D3).

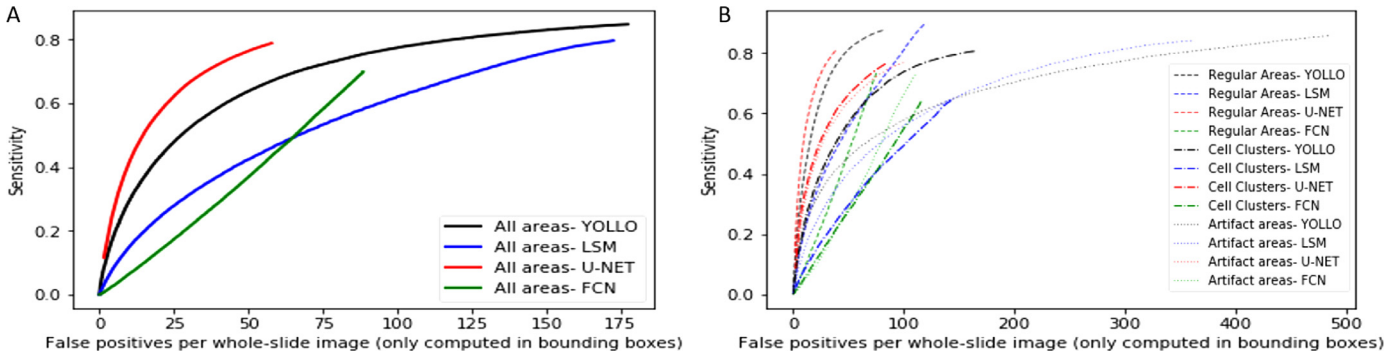


**Fig. 5.** FROC curves for DL models for the test set, where: A. FROC-curve for all tissue areas; B. FROC-curves for individual tissue areas.

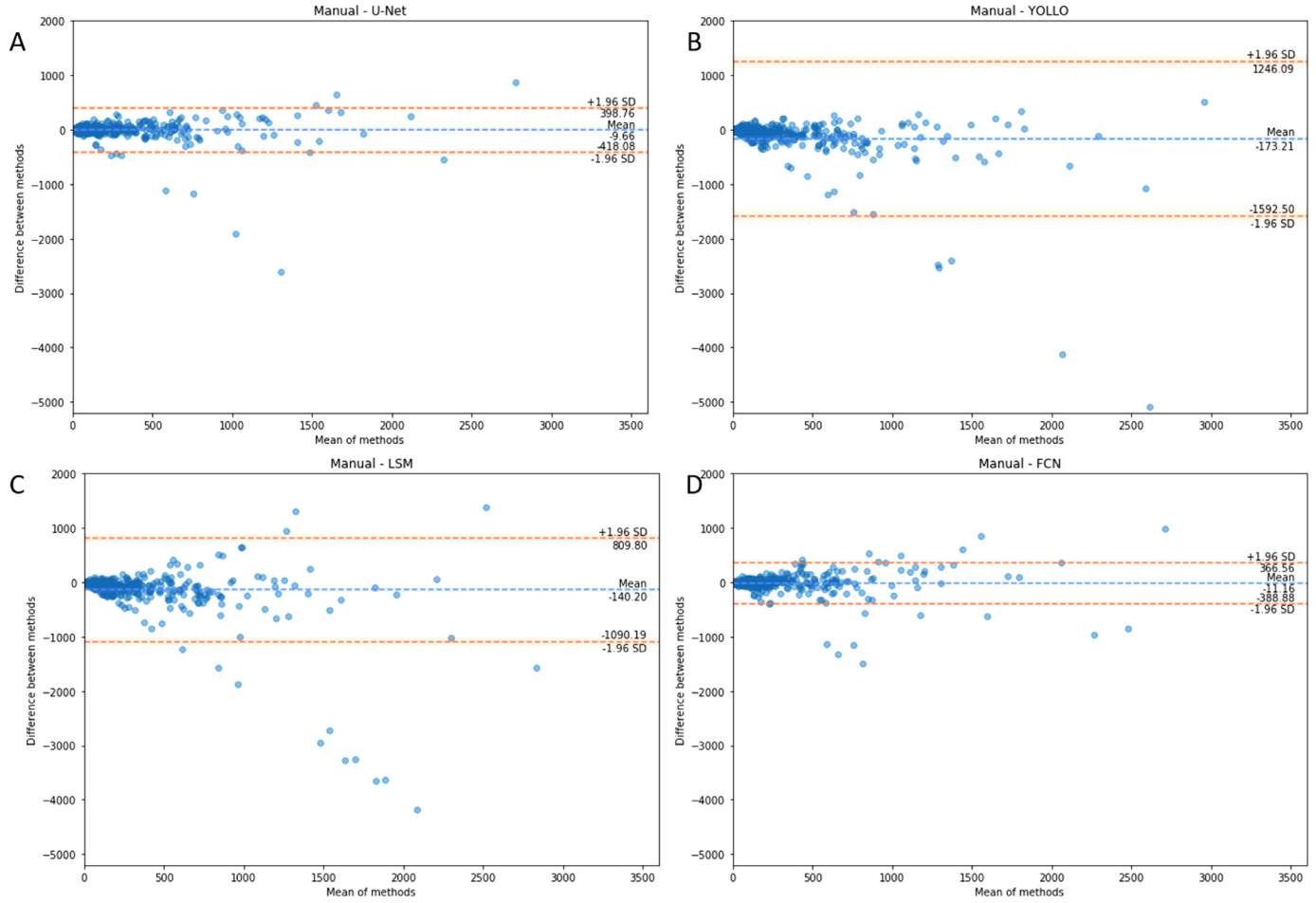


**Fig. 6.** Bland-Altman plot for deep learning models and manual annotations (reference standard) for the test set, where: A. plot for Manual and U-Net; B. plot for Manual and Yollo; C. plot for Manual and LSM; D. plot for Manual and FCN. Notice the superior performance of the U-Net method, with the smaller standard deviation range and smaller number of outliers.

based on organs (colon, breast, prostate) also showed fairly good performance for all methods, with SDs in the range of 0.02-0.09. In all cases the U-Net approach was the most robust, with an SD equal to 0.08, 0.01 and 0.02 for medical centers, staining types and organs, respectively. As a result, the U-Net approach has been used in the comparison with human performance, where we refer to it as the "automatic method".

### 5.3. Lymphocyte assessment

In order to get insight into the effectiveness of the developed algorithms to perform automated analysis in clinical practice, we compared the performance of the automatic method to four pathologists in categorizing ROIs based on number of lymphocytes.

We used the weighted linear Cohen's Kappa score to measure (1) the agreement among pathologists, computed as the average agreement (Kappa $\kappa$) between each pair of pathologists; (2) the agreement between each pathologist and the reference standard, (3) the agreement between pathologists and automatic results, (4) the agreement between automatic results and the reference standard.

The results of this analysis are reported in Table 6. We found high agreement among pathologists, with an average $\kappa$ of 0.76. However, slightly lower agreement is found when comparing

**Table 3**
Results per medical center: Detection performance of developed deep learning algorithms on the test set. Where: *- data from centers participated in the training process, MD- max difference between labs, AVG- average value, STD- standard deviation. Bold indicates the best method.

| Lab id | U-Net | | | FCN | | | YOLLO | | | LSM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision |
| All Labs | 0.78 | 0.79 | 0.76 | 0.69 | 0.70 | 0.67 | 0.64 | 0.85 | 0.51 | 0.63 | 0.80 | 0.52 |
| Lab 1* | 0.76 | 0.78 | 0.75 | 0.72 | 0.76 | 0.68 | 0.63 | 0.84 | 0.50 | 0.63 | 0.78 | 0.53 |
| Lab 2 | 0.66 | 0.54 | 0.87 | 0.64 | 0.51 | 0.85 | 0.64 | 0.61 | 0.67 | 0.66 | 0.68 | 0.64 |
| Lab 3 | 0.82 | 0.86 | 0.77 | 0.64 | 0.68 | 0.61 | 0.69 | 0.93 | 0.55 | 0.66 | 0.91 | 0.52 |
| Lab 4 | 0.58 | 0.88 | 0.43 | 0.27 | 0.31 | 0.24 | 0.26 | 0.94 | 0.15 | 0.29 | 0.88 | 0.17 |
| Lab 5 | 0.74 | 0.90 | 0.62 | 0.54 | 0.65 | 0.47 | 0.60 | 0.95 | 0.44 | 0.53 | 0.92 | 0.37 |
| Lab 6 | 0.70 | 0.83 | 0.60 | 0.62 | 0.81 | 0.50 | 0.32 | 0.90 | 0.20 | 0.38 | 0.87 | 0.24 |
| Lab 7 | 0.83 | 0.89 | 0.77 | 0.61 | 0.62 | 0.60 | 0.70 | 0.93 | 0.57 | 0.68 | 0.85 | 0.57 |
| Lab 8 | 0.81 | 0.77 | 0.86 | 0.74 | 0.70 | 0.79 | 0.77 | 0.83 | 0.71 | 0.71 | 0.77 | 0.65 |
| AVG | **0.74** | | | 0.60 | | | 0.58 | | | 0.57 | | |
| STD | **0.08** | | | 0.14 | | | 0.17 | | | 0.14 | | |
| MD | **0.25** | | | 0.47 | | | 0.51 | | | 0.42 | | |

**Table 4**
Results per staining type: Detection performance of developed deep learning algorithms on the test set. Where: MD- max difference between labs, AVG- average value, STD- standard deviation. Bold indicates the best method.

| Staining | U-Net | | | FCN | | | YOLLO | | | LSM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision |
| CD3 | 0.78 | 0.79 | 0.77 | 0.71 | 0.72 | 0.70 | 0.66 | 0.84 | 0.54 | 0.64 | 0.77 | 0.56 |
| CD8 | 0.76 | 0.78 | 0.73 | 0.62 | 0.65 | 0.60 | 0.60 | 0.87 | 0.46 | 0.60 | 0.87 | 0.46 |
| AVG | **0.77** | | | 0.67 | | | 0.63 | | | 0.62 | | |
| STD | **0.01** | | | 0.05 | | | 0.03 | | | 0.02 | | |
| MD | **0.02** | | | 0.09 | | | 0.06 | | | 0.04 | | |

**Table 5**
Results per organ type: Detection performance of developed deep learning algorithms on the test set. Where: MD- max difference between labs, AVG- average value, STD- standard deviation. Bold indicates the best method.

| Organ | U-Net | | | FCN | | | YOLLO | | | LSM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision |
| breast | 0.78 | 0.88 | 0.70 | 0.62 | 0.69 | 0.56 | 0.55 | 0.92 | 0.39 | 0.56 | 0.87 | 0.41 |
| colon | 0.80 | 0.74 | 0.86 | 0.73 | 0.67 | 0.79 | 0.75 | 0.80 | 0.71 | 0.70 | 0.76 | 0.65 |
| prostate | 0.76 | 0.74 | 0.77 | 0.73 | 0.74 | 0.72 | 0.68 | 0.82 | 0.58 | 0.67 | 0.76 | 0.60 |
| AVG | **0.78** | | | 0.69 | | | 0.66 | | | 0.64 | | |
| STD | **0.02** | | | 0.05 | | | 0.09 | | | 0.06 | | |
| MD | **0.04** | | | 0.11 | | | 0.20 | | | 0.14 | | |

**Table 6**
Assessment of methods agreement with using the weighted linear Cohen's Kappa metric ($\kappa$).

| | $\kappa$ |
|---|---|
| Agreement among pathologists (Average) | 0.76 |
| $Pathologist_1$ vs. Manual reference standard | 0.64 |
| $Pathologist_2$ vs. Manual reference standard | 0.66 |
| $Pathologist_3$ vs. Manual reference standard | 0.58 |
| $Pathologist_4$ vs. Manual reference standard | 0.67 |
| Pathologists vs. Manual reference standard (Average) | 0.64 |
| Automatic vs. Manual reference standard | 0.72 |
| Automatic vs. Pathologists (Average) | 0.57 |

pathologists with the reference standard ($\kappa \in [0.58, 0.67]$). At the same time, a high agreement is observed between automatic detection and the reference standard ($\kappa$ 0.72), which is on the same level as the average agreement between pathologists. Moreover, comparison with the reference standard shows higher agreement for the automatic method ($\kappa = 0.72$) than for the pathologist evaluation (average $\kappa = 0.64$). The lowest agreement result (average $\kappa = 0.57$) is observed between automatic method and pathologists, but is generally still considered reasonable agreement.

Finally, we broke down the analysis of the agreement by investigating the *"per-class"* and overall accuracy at assessing lymphocytes for both pathologists and automatic method. The results are reported in Table 7. Pathologists can accurately distinguish areas without lymphocytes (class 0) from areas with lymphocytes, with a performance that is superior to the automatic method (0.87 vs. 0.30 class accuracy). This is mainly due to the presence of false positives detected by the automatic method. On the other hand, the computer outperforms pathologists at detecting the presence of small numbers of lymphocytes (classes 1–5 and 6–10), although still suffering from the presence of false positives, and show a trend of increasing performance when increasing numbers of lymphocytes are present (0.92 accuracy for the "$>200$ lymphocytes" class). Is should be noted, that agreement values increase probably due to the increase of the range of possible values between the different categories (the agreement increases when the interval of each category is larger) for both, both pathologists and automatic evaluation.

## 6. Discussion

Most of the research on automatic cell detection published so far is based on data from a single center. A key strength of the presented study is the use of a multi-center cohort. In our research, the test set includes slides from eight different medical centers, where data from seven centers were not used to train automatic algorithms. This allowed us to assess the robustness of

**Table 7**
Sensitivity at estimating the number of lymphocytes as compared to counting based on manual annotations. Results of four pathologists ($P_1$-$P_4$) and of an automatic algorithm are reported. The average sensitivity over all pathologists is also reported to ease the comparison with the automatic method. Bold indicates best method.

| | Predicted labels | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1–5 | 6–10 | 11–20 | 21–50 | 51–200 | >200 | All |
| $P_1$ | 0.78 | 0.11 | 0.25 | 0.15 | 0.32 | 0.71 | 0.54 | 0.41 |
| $P_2$ | 0.96 | 0.17 | 0.20 | 0.15 | 0.27 | 0.58 | 0.73 | 0.44 |
| $P_3$ | 0.78 | 0.28 | 0.15 | 0.20 | 0.32 | 0.48 | 0.35 | 0.37 |
| $P_4$ | 0.96 | 0.33 | 0.25 | 0.15 | 0.55 | 0.65 | 0.43 | 0.47 |
| Average | **0.87** | 0.22 | 0.21 | 0.16 | 0.37 | 0.60 | 0.51 | 0.42 |
| Auto | 0.30 | **0.44** | **0.30** | **0.35** | **0.54** | **0.76** | **0.92** | **0.52** |

the proposed deep learning methods. In Table 3 detailed results are reported for each of the centers involved in the study. The highest generalization capacity, as well as the highest robustness across different laboratories was achieved by the U-Net model, where standard deviation and the maximum difference in terms of F1-score across laboratories were respectively 0.08 and 0.25. The method that showed the least robustness to different staining was the YOLLO approach, where the standard deviation was 0.17 and the maximum difference in F1-score across centers was 0.51. The same holds across organ (breast, colon, prostate) and staining type (CD3, CD8) (Tables 4 and 5).

Another unique characteristic of the data set established in this work is the large amount of manually annotated cells (171,166 lymphocytes). To the best of our knowledge, this is the largest set of manually annotated lymphocytes to date, which represents an increase of 70x compared to the amount of manually annotated cells used in recent work (Garcia et al., 2017).

We also specifically look at areas with cell clusters and artifacts. Previous work has mostly focused on cell detection in regular tissue areas. This leads to an optimistically biased result, and makes it hard to assess usefulness of methods for the analysis of whole-slide images. As an example, the worst and best method on regular tissue areas have a difference of 0.11 in F1-score, whereas in areas with artifacts the difference is 0.28.

In this paper, four different deep learning techniques were applied to lymphocyte detection. Results show that all methods achieved performance in a range 0.63–0.78 of F1-score for all tissue areas. The best lymphocyte detection was observed in the regular tissue areas, where F1-score was in a range 0.71–0.82. Nonetheless, large differences between methods are observed for clustered cells and artifact areas. Cells in clusters were distinguished well by YOLLO and U-Net methods (F1-score in the range 0.79–0.81). However, artifact areas were a challenge for all considered approaches, where the lowest results were observed (F1-score in the range 0.19–0.47). An alternative to deal with areas with artifacts is to perform a pre-processing, step, for example by applying convolutional networks specifically trained to detect artifacts. This might be able to remove most of the artifacts from consideration.

Despite the fact that the LSM method was originally designed to perform cell detection (Sirinukunwattana et al. (2016)), in our experiments we observed lower performance than what was reported in the original paper. A key difference is that we focus on lymphocytes, compared to several different classes in the original paper. Most of the classes in the original paper (e.g. epithelial cells) do not cluster together in the same manner as lymphocytes. Furthermore, the membranous nature of the CD3 and the CD8 staining makes it challenging to see the boundaries of closely clustered cells, in contrast to the H&E stain used in the original paper.

The application of the three-class cell mask method, that was used for FCN and U-Net, makes it easier to separate cells and accurately localize the cell centers, which was much more difficult in the case of binary masks. The multi-class cell mask was inspired by biology, and developed based on analysis of stained lymphocytes (see Fig. 3). This is a novel approach for cell detection, which leads to correct cell center detection using neural networks. This approach could further be improved by using segmentations of cell nuclei to derive the borders and the center instead of a circle with predetermined radius.

Another aspect that can be taken into account is the time-efficiency of proposed approaches. The number of parameters and complexity of the model, as well as post-processing operations have a direct impact on the computation time of a single patch and consequently of a whole-slide image. The average time of a single tile classification (size 256x256 pixels) by U-Net and YOLLO methods is 25ms and 17ms, respectively. Both methods need post-processing operations, which take additional time. However, in this paper we did not optimize the post-processing pipelines for computation time.

The algorithms developed in this paper open doors to new studies on evaluating the number of tumor infiltrating lymphocytes. Moreover, proposed solutions could be applied to other tissue and staining types. Preliminary studies show that lymphocytes are correctly detected for slides stained with CD45RO and FoxP3. This would make the proposed solution independent from tissue type (organ) and IHC staining method, resulting in a spectrum of possible applications. The reference standard used in this study are manual annotations of lymphocytes (see Fig. 2). It should be noted, that the task of manual annotation is time-consuming and tedious. In practice, it means that the references standard is biased by subjectivity in manual annotations made by observers, and it is not perfect. A better reference standard could be obtained when annotations from multiple experts are available and are combined, as for example done in public challenges on detection of mitotic figures (AMIDA13, TUPAC16).

Four pathologists with experience in visual estimation of lymphocytes inside of each ROI were involved in the presented observer study. The agreement among pathologists (average value) is Kappa=0.76, which is classified as an excellent agreement by Fleiss (1981) and as a substantial agreement by Landis and Koch (1977). It should be noted that the main task that pathologists have to perform is visual estimation, rather than cell counting. On the one hand this makes the task less time consuming, but on the other hand it makes it less objective, therefore subject to inter-observer variability. In particular, areas such as clustered cells or weakly stained regions can be challenging for a task purely based on visual assessment. The observer study shows that, compared to the reference standard, the automated method achieves a higher Kappa score than the average over

the pathologists. Detailed analysis of evaluation results for both methods show differences for the "no- lymphocytes" class, where the automatic evaluation performs less well than the pathologists in the study. This results from single false positive detection in the automatic evaluation, leaving room for improvement on the automatic cell detection method. Nonetheless, for other classes, especially for a high immune-cell density the automatic method is more accurate than pathologists.

## 7. Conclusion

In this study the effectiveness of a deep learning approaches was investigated for automatic detection of lymphocytes in immunohistochemically stained tissue sections of breast, colon and prostate cancer. We tested four algorithms for the problem of automatic immune cells detection, and found that especially the U-Net based approach performs well and even exceeds the performance of human observers for this task. Moreover, we evaluated the clinical impact of the developed methods by performing an observer study with four pathologists and comparing their performance with automatic detection. Achieved results show that deep learning techniques can be applied to detect positively stained cells in immunohistochemistry, with great promise for immuno-oncology. The fact that we can now reliably quantify these cells opens an avenue of research in which we relate immune cell quantities to tumor progression and treatment response. Furthermore, the technique is not limited to CD3 and CD8 stained images and could readily be applied to other immunohistochemistry markers, such as CD45RO or FOXP3, or other cell membrane markers.

Finally, we made the test set and an evaluation metric publicly available. This allows the scientific community to compare the performance of other approaches with the results presented in this paper using exactly the same test set and the same evaluation procedure.

## Declaration of Competing Interest

None.

## CRediT authorship contribution statement

**Zaneta Swiderska-Chadaj:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Hans Pinckaers:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Mart van Rijthoven:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Maschenka Balkenhol:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Margarita Melnikova:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Oscar Geessink:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Quirine Manson:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Mark Sherman:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Antonio Polonia:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Jeremy Parry:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Mustapha Abubakar:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Geert Litjens:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Jeroen van der Laak:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Francesco Ciompi:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing.

## Acknowledgments

The authors would like to thank Sophie van den Broek for her support in the process of lymphocyte annotation. This project was supported by the Alpe d'HuZes/Dutch Cancer Society Fund (Grant Number: KUN 2014–7032, KUN 2015–7970), the Netherlands Organization for Scientific Research (NWO) (project number 016.186.152), the Stichting IT Projecten (project PATHOLOGIE 2), and partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode, http://www.examode.eu/).

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2019.101547.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv: 1603.04467(2016).

Al-Rfou, R., Alain, G., Almahairi, A., et al., 2016. Theano: a python framework for fast computation of mathematical expressions. arXiv e-prints.

Chen, T., Chefd'hotel, C., 2014. Deep learning based automatic immune cell detection for immunohistochemistry images. In: Wu, G., Zhang, D., Zhou, L. (Eds.), Machine Learning in Medical Imaging. MLMI 2014. Lecture Notes in Computer Science, 8679. Springer, Cham.

Chollet, F., et al., 2015. Keras. GitHub repository. https://github.com/fchollet/keras.

Coussens, L.M., Zitvogel, L., Palucka, A.K., 2013. Neutralizing tumor-promoting chronic inflammation: a magic bullet? Science (New York, N.Y.) 339, 286–291. doi:10.1126/science.1232227.

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., de Almeida, D.M., McFee, B., Weideman, H., Takács, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K., Liu, C., French, G., Degrave, J., 2015. Lasagne: first release doi:10.5281/zenodo.27878.

Ehteshami Bejnordi, B., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., the CAMELYON16 Consortium, Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M.U., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer 318, 2199–2210. doi:10.1001/jama.2017.14585.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.

Fleiss, J.L., 1981. Statistical methods for rates and proportions. No. 04; QA279, F5 1981.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P.-H., Trajanoski, Z., Fridman, W.-H., Pagès, F., 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science (New York, N.Y.) 313 (5795), 1960–1964. doi:10.1126/science.1129139.

Garcia, E., Hermoza, R., Castanon, C.B., Cano, L., Castillo, M., Castanñeda, C., 2017. Automatic lymphocyte detection on gastric cancer IHCimages using deep learning. In: Proc. IEEE 30th Int. Symp. Computer-Based Medical Systems (CBMS), pp. 200–204. doi:10.1109/CBMS.2017.94.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. arXiv: 1504.08083v2.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316, 2402–2410. doi:10.1001/jama.2016.17216.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. Cell 144, 646–674. doi:10.1016/j.cell.2011.02.013.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inf. 7, 29. doi:10.4103/2153-3539.186902.

Khalil, D.N., Smith, E.L., Brentjens, R.J., Wolchok, J.D., 2016. The future of cancer treatment: immunomodulation, cars and combination immunotherapy. Nature Rev. Clin. Oncol. 13, 394. doi:10.1038/nrclinonc.2016.65.

Klauschen, F., Müller, K.-R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S., Michiels, S., Nielsen, T.O., Adams, S., Savas, P., Symmans, F., Willis, S., Gruosso, T., Park, M., Haibe-Kains, B., Gallas, B., Thompson, A.M., Cree, I., Sotiriou, C., Solinas, C., Preusser, M., Hewitt, S.M., Rimm, D., Viale, G., Loi, S., Loibl, S., Salgado, R., Denkert, C., Group, I.I.-O.B.W., 2018. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. Seminar. Cancer Biol. 52, 151–157. doi:10.1016/j.semcancer.2018.07.001.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J., 2018. 1399 H&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 7. doi:10.1093/gigascience/giy065.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition.

Moskowitz, C.S., 2017. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. JAMA 318, 2250–2251. doi:10.1001/jama.2017.18686.

Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A., 2015. You only look once: unified, real-time object detection. arXiv: 1506.02640.

Redmon, J., Farhadi, A., 2015. YOLO9000: Better, faster, stronger. arXiv: 1612.08242.

Redmon, J., Farhadi, A., 2018. Yolov3: an incremental improvement. arXiv: 1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv: 1506.01497v3.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 234–241.

Rosenberg, S.A., Restifo, N.P., 2015. Adoptive cell transfer as personalized immunotherapy for human cancer. Science (New York, N.Y.) 348, 62–68. doi:10.1126/science.aaa4967.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., Van Arnam, J., Network, C.G.A.R., Shmulevich, I., Rao, A.U.K., Lazar, A.J., Sharma, A., Thorsson, V., 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 23, 181–193.e7. doi:10.1016/j.celrep.2018.03.086.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv: 1312.6229v4.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imag. 35 (5), 1196–1206. doi:10.1109/TMI.2016.2525803.

Swiderska-Chadaj, Z., Pinckaers, H., van Rijthoven, M., Balkenhol, M., Melnikova, M., Geessink, O., Manson, Q., Litjens, G., van der Laak, J., F., C., 2018. Convolutional neural networks for lymphocyte detection in immunohistochemically stained whole-slide images. MIDL.

Syn, N.L., Teng, M.W.L., Mok, T.S.K., Soo, R.A., 2017. De-novo and acquired resistance to immune checkpoint targeting. Lancet Oncol. 18, e731–e741. doi:10.1016/S1470-2045(17)30607-1.

van Rijthoven, M., Swiderska-Chadaj, Z., Seeliger, K., van der Laak, J., Ciompi, F., 2018. You only look on lymphocytes once. MIDL.

Varn, F.S., Wang, Y., Mullins, D.W., Fiering, S., Cheng, C., 2017. Systematic pan-cancer analysis reveals immune cell interactions in the tumor microenvironment. Cancer Res. 77, 1271–1282. doi:10.1158/0008-5472.CAN-16-2490.

Vánky, F., Klein, E., Willems, J., Böök, K., Ivert, T., Péterffy, A., Nilsonne, U., Kreicbergs, A., Aparisi, T., 1986. Lysis of autologous tumor cells by blood lymphocytes tested at the time of surgery. correlation with the postsurgical clinical course. Cancer Immunol. Immunother 21, 69–76.

Xie, S., Chen, J., Zhang, M., Wu, Z., 2017. Allogenic natural killer cell immunotherapy of sizeable ovarian cancer: a case report. Mol. Clin. Oncol. 6, 903–906. doi:10.3892/mco.2017.1230.

Xie, Y., Kong, X., Xing, F., Liu, F., Su, H., Yang, L., 2015a. Deep voting: a robust approach toward nucleus localization in microscopy images. In: Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 9351, pp. 374–382. doi:10.1007/978-3-319-24574-4_45.

Xie, Y., Xing, F., Kong, X., Su, H., Yang, L., 2015b. Beyond classification: structured regression for robust cell detection using convolutional neural network. In: Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 9351, pp. 358–365. doi:10.1007/978-3-319-24574-4_43.

Xing, F., Yang, L., 2016. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. IEEE Rev. Biomed. Eng. 9, 234–263. doi:10.1109/RBME.2016.2515127.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A., 2016. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. IEEE Trans. Med. Imag. 35, 119–130. doi:10.1109/TMI.2015.2458702.