Detection of lobular structures in normal breast tissue

Gregory Apou[☆1], Nadine S. Schaadt[☆2], Benoît Naegel[1], Germain Forestier[3], Ralf Schönmeyer[4], Friedrich Feuerhake[2], Cédric Wemmert[1,*], Anne Grote[2]

[1] *ICube, University of Strasbourg, 300 bvd Sébastien Brant, 67412 Illkirch, France*
[2] *Institute for Pathology, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany*
[3] *MIPS, University of Haute Alsace, 12 rue des Frères Lumière, 68093 Mulhouse, France*
[4] *Definiens AG, Bernhard-Wicki-Strasse 5, 80636 Munich, Germany*

Abstract

Background: Ongoing research into inflammatory conditions raises an increasing need to evaluate immune cells in histological sections in biologically relevant regions of interest (ROIs). Herein, we compare different approaches to automatically detect lobular structures in human normal breast tissue in digitized whole slide images (WSIs). This automation is required to perform objective and consistent quantitative studies on large data sets.

Methods: In normal breast tissue from nine healthy patients immunohistochemically stained for different markers, we evaluated and compared three different image analysis methods to automatically detect lobular structures in WSIs: (1) a bottom–up approach using the cell–based data for subsequent tissue level classification, (2) a top–down method starting with texture classification at tissue level analysis of cell densities in specific ROIs, and (3) a direct texture classification using deep learning technology.

Results: All three methods result in comparable overall quality allowing automated detection of lobular structures with minor advantage in sensitivity (approach 3), specificity (approach 2), or processing time (approach 1). Combining the outputs of the approaches further improved the precision.

Conclusions: Different approaches of automated ROI detection are feasible and should be selected according to the individual needs of biomarker research. Additionally, detected ROIs could be used as a basis for quantification of immune infiltration in lobular structures.

Keywords: Whole Slide Image, Digital Histopathology, Normal Breast Lobule, Image Analysis, Convolutional Neural Network

1. Introduction

Lobular structures are the functional units of the resting mammalian breast that further differentiate into milk–producing glands during lactation. The normal anatomical structures are important because breast cancer and premalignant lesions originate in these epithelial structures and there is evidence that the transition between ductal and lobular structures may be particularly susceptible to oncogenic events [1]. In addition to studies on the origins of cancer, the detection of ducts and lobules is also relevant for an inflammatory condition referred to as *lymphocytic lobulitis* (LLO), which has been observed in the adjacent tissue around breast cancer and in prophylactically removed breast tissue without any evidence for cancer in *BRCA1/2* mutation carriers [2, 3]. This phenomenon is not yet well understood and deciphering its possible link with hereditary breast cancer may lead to better disease understanding, new prognostic indicators, or novel treatment options. In order to perform an objective, repeatable, and statistically reliable quantitative study of LLO on large data sets, the ability to automatically detect relevant structures in histological slides is necessary. We refer to ducts and lobules as lobular structure in the following.

Nowadays, such slides can be routinely digitized; the resulting whole slide images (WSIs) can be processed by automated image analysis techniques with the aim to detect lobular structures and to quantify cell numbers [4, 5]. Many works are based on detecting and automatically counting cells for cancer diagnosis, grading, and prognosis [6, 7, 8]. However, to give more insight to the pathologist, it is necessary to evaluate immune cells beyond estimation of their density, for example by object–based recognition of spatial patterns and interactions at high resolution [9]. Thus, our objective is to pave the way for identifying and classifying cells in those areas in image that are most relevant, like lobules in breast cancer and LLO, finally enabling methods to characterize the spatial distribution of different subtypes of immune cells in relation to these larger image objects.

☆Co-first author
*Corresponding author: wemmert@unistra.fr

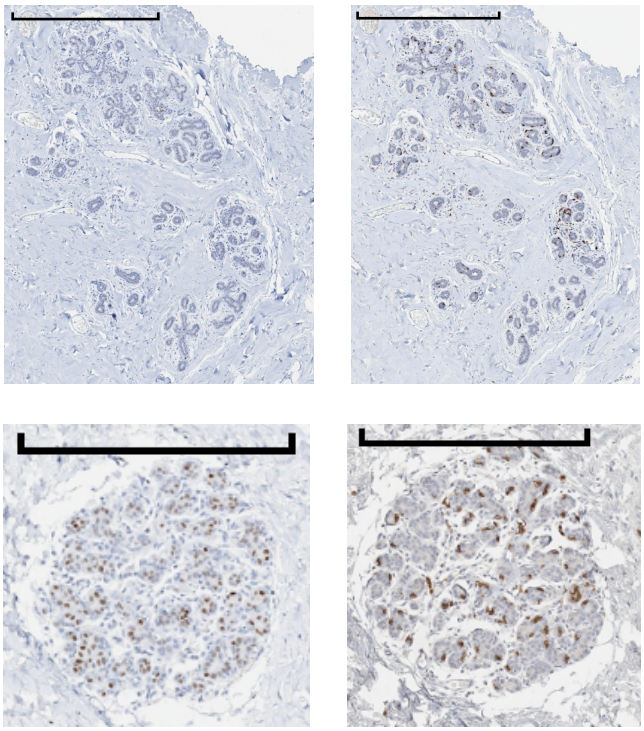

Figure 1: Top row shows a large lobule with sparse branching in two different stainings (left: ER, right: CD8). Bottom row shows a small lobule with dense branching (left: ER, right: CD8). Scale bars are 0.5 mm.

In general, lobular structures are composed of dense areas of epithelial cells in tubular structures, normally with a clear contrast to the lobular stroma. However, lobules can be very different in size, shape, and texture depending on their functional stage (e.g., phase of the menstrual cycle), the degree of immune cell infiltration, and also appear differently according to the used staining method (see Figure 1).

Cancer growth adds to this complexity of tissue structures and makes manual or automated lobule detection even more challenging. As about 75% of invasive breast cancer cases are of ductal type and detection of lobular type is more difficult by mammography [10], it is important to distinguish between ducts and lobules in this context, but we do not provide a way to make this distinction. As a preparatory work before delving into the structural complexity of cancer–affected tissue, we focus on normal breast tissue of healthy patients in order to evaluate and compare three different image analysis methods to automatically detect epithelial structures including ducts and lobules in immunohistochemically (IHC) stained WSIs. Even without solving the cancer–related challenges, the work addresses an important demand because the evaluation of LLO may lead to new biomarker patterns with diagnostic and prognostic value. From a technological point of view, the task of defining regions of interest (ROI) for further analysis tackles a problem that occurs frequently in medical image analysis: The limited availability of experts to perform large-scale manual annotation due to restricted resources of trained pathologists remains an important bottleneck for progress. In the context of LLO, this is evident because the statistical base for experimental evidence has been limited by the available time for pathologists who have to manually select every lobule and therefore could so far only annotate small data sets [11, 12]. Thus, overcoming the limitations of manual annotation of WSIs by automation is highly desirable. This work builds on previous work [13], where detection of normal lobules in the vicinity of breast cancer was optimized for the purpose of analyzing nuclear expression of estrogen receptor (ER) or progesterone receptor (PR). Grote *et al.* detected lobules on several segmentation layers using textural, geometric, and relational features, as well as solid tumor using textural features. Other tissue classification techniques have supported the study of pathologies like odontogenic cysts [14] and various cancers [15, 16]. The identification of general biological structures has received comparatively little attention, although graph–based approaches exist for unsupervised top–down tissue categorization [17] and bottom–up biological object identification [18].

A machine learning algorithm describes how to identify patterns in existing data (learning) and uses this acquired knowledge to make predictions on new data [19]. A deep learning algorithm is a machine learning algorithm that can learn a hierarchical description of the data with multiple sublevels of nonlinear features [20]. In recent years, a type of deep learning architecture optimized for 2D data called convolutional neural networks (CNNs) [21] have provided state–of–the–art results in various applications of machine learning–based image analysis, from general scene labeling [22] to cancer classification [23] and mitosis detection [24].

After presenting our data set and evaluation criterion, we will describe three methods that were developed in the context of lymphocytic lobulitis: (1) a Bottom–Up Method (MBU) starting with cell detection and ending with tissue classification, (2) a Top–Down Method (MTD) starting with texture classification and ending with cell density characterization, and (3) a direct texture classification Method using Deep Learning technology (MDL). We conclude by comparing their strengths and weaknesses using two characteristically different stainings and by assessing the feasibility of combining the methods.

2. Materials and Methods

We collected tissue samples from a cohort including nine healthy women who underwent reduction mammoplasty due to hypermasty [25]. The paraffin–embedded samples were cut into 3 µm thick sections and stained for the nuclear marker ER and the immune cell marker cluster of differentiation 8 (CD8) that is expressed in the cell cytoplasm and on the outer cell membranes of cytotoxic

Patient	Age	ER Size	CD8 Size	ER Lobules	CD8 Lobules
NB02	25	20.6×20.6	21.6×21.2	125	120
NB05	29	21.1×20.1	18.1×21.6	32	25
NB11	19	15.6×15.9	14.1×15.8	55	35
NB12	22	17.6×15.7	17.1×14.2	12	11
NB16	27	16.1×15.5	15.1×17.9	54	63
NB20	27	16.6×15.1	17.6×15.0	80	76
NB25	30	20.1×16.6	13.6×19.9	49	34
NB28	21	24.6×20.9	28.7×22.6	19	24
NB34	28	14.1×19.0	16.1×19.7	226	238

Table 1: Overview over data set: Size of image (width×height mm) and number of lobular structures in ground truth. The slight variation in numbers of structures between different stainings could be either due to variable composition of the sections at different levels of the paraffin block or due to slightly different annotation of components of lobules, e.g., merging two elements of the glands to one "lobule".

T lymphocytes, using an automated staining instrument (Ventana Benchmark Ultra). As ER is often expressed in breast cancer, it was chosen as example for a routinely used breast cancer marker important for treatment decisions [26, 27] and CD8 for its relevance in oncoimmunology [28]. Whereas ER^+ cells are a subset of epithelial cells in lobular structures, $CD8^+$ cytotoxic T lymphocytes may spread close to epithelial compartments, over the full lobular stroma, or even in non–lobular stroma. WSIs were acquired by Aperio AT2 scanner at 40X magnification, scanned images have a resolution of 0.253 µm/pixel. For each WSI, a pathologist (FF) performed annotations of the full epithelial compartment (including lobular and ductal regions) and more detailed annotations to distinguish between lobular and non–lobular regions. This was performed by drawing outlines over the digital images using the software tool Aperio ImageScope. Table 1 summarizes the number of annotated lobular structures for each case and each staining.

To evaluate the methods, we calculated a global, pixel–based $F1$ score (equivalent to Dice similarity [29]):

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where TP is the number of true positives, FP the number of false positives, and FN the number of false negatives. We additionally measured the true positive percentages inside the detected and actual lobular objects in order to obtain a more object–based quality estimation.

To the best of our knowledge, there is no equivalent published work that we can use for comparison, but the elementary concepts and tools that we use are well documented [30].

2.1. Method 1: Bottom–Up (MBU)

As a starting point, we developed a workflow (see Figure 3) that applies well–established algorithms using Definiens Developer XD 64 2.4 based on the assumption that cell clusters and therefore highest deviations in RGB occur in lobular structures. In the following, we refer to this workflow as MBU. Due to large size of WSIs (up to 100,000 × 100,000 pixels), we used images down–sampled to a resolution of 8 µm/pixel (∼ 3.2% of the original WSI). Based on a multi–threshold segmentation on a gray layer, tissue regions (remaining: background) are detected, in which afterwards each pixel is classified as lobular or non–lobular area. For this, we generated a color–distribution–range image (CDR) that represents the averaged RGB range in a small part ($s \times s$ pixels) of the down–sampled image cut by a lower and an upper threshold t_l and t_u; i.e., each image is split into tiles $s \times s$ pixel large having three layers red, green, blue each in range $[0, 255]$, in which all new pixel values x are set as defined in equation (2),

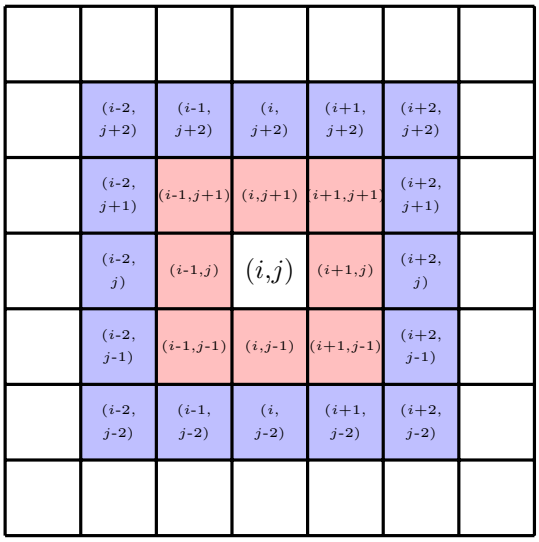

Figure 2: Neighborhood of pixel (i, j) (red: direct neighbors; blue: further neighbors) used in MBU.

$$x = \begin{cases} 0 & \textbf{if } \frac{1}{3}\left(r_R + r_G + r_B\right) < t_l \\ 255 & \textbf{if } \frac{1}{3}\left(r_R + r_G + r_B\right) > t_u \\ \frac{1}{3}\left(r_R + r_G + r_B\right) & \textbf{otherwise} \end{cases} \tag{2}$$

where $r_X = \max_X - min_X$ represents the range of layer X in the corresponding small tile. Thus, a homogeneous area with similar colors in all pixels is set to black, whereas a texture–rich area with large color deviations in these tiny tiles is set to white. Lobular structures in IHC images have a high contrast between epithelial cells in dark blue and very bright lobular stroma. The remaining image is composed of bright stroma with less contrasts and almost white fatty tissue, which is mostly classified as background. Since we do not use a sliding window approach for simplicity, we afterwards split the image into tiles $l \times l$ pixel large to redefine the pixel values depending on their neighborhood (defined in Figure 2).

In tiles with less than t_{w_l} white pixels or in tiles where more than 50% (parameter t_b) of the pixels are black and less than 25% (parameter t_{w_u}) are white, each pixel value

Param.	Description	Used value			
		$B \cup C$	$A \cup C$	$A \cup B$	$A \cup B \cup C$
s	size of small tiles	8	6	8	8
l	size of large tiles	25	25	25	25
t_l	lower threshold to cut CDR	0	25	20	20
t_u	upper threshold to cut CDR	150	150	150	150
t_{w_l}	lower threshold for white pixels	0.002	0.021	0.002	0.002
t_{w_m}	middle threshold of white pixels	0.010	0.125	0.013	0.012
t_{w_u}	upper threshold of white pixels	0.020	0.250	0.025	0.025
t_b	threshold of black pixels	0.040	0.500	0.050	0.050

Table 2: Parameters of MBU with highest $F1$ score for different folds of cross–validation, where A={NB02, NB05, NB11}, B={NB12, NB16, NB20}, and C={NB25, NB28, NB34}.

x is set to $\frac{x}{2}$ (darker). For the remaining tiles, the decision is separately done for each individual pixel based on its eight direct neighbors and 16 further neighbors, for details see equation (3),

$$x_{\text{new}} = \begin{cases} \frac{x}{2} & \textbf{if } (|B| > t_b \cdot l^2 \wedge |W| \geq t_{w_u} \cdot l^2 \wedge \neg N_8) \\ x + \frac{x}{2} & \textbf{if } \big(|B| \leq t_b \cdot l^2 \wedge |W| > t_{w_m} \cdot l^2 \wedge (N_8 \vee N_{16})\big) \\ & \quad \vee \big(|B| > t_b \cdot l^2 \wedge |W| \geq t_{w_u} \cdot l^2 \wedge N_8\big) \\ x & \textbf{otherwise} \end{cases} \tag{3}$$

where $|B|$ and $|W|$ are the number of black and white pixels and N_i is true when at least one of i neighbors is 255 (obviously, we set values larger than 255 to 255). The influence of each parameter on the $F1$ score as well as the selection of the parameters is given in Supplementary Materials (images NB02–NB34 are used in cross–validation for parameter fit). All parameter values used for the different cross–validation folds (A={NB02, NB05, NB11}, B={NB12, NB16, NB20}, and C={NB25, NB28, NB34}) are listed in Table 2; the values are then applied to corresponding test group (e.g., results of $A \cup B$ applied to C).

The cut CDR was included as additional layer into our Definiens rule set and averaged with a layer which represents clusters of well–shaped nuclei. For this combination, the two layers are normalized to the same range to allow an equal weighting of both. If a pixel in this combined layer is non–dark, it is classified as lobular. The nuclei layer is generated by a k–means clustering to find areas with high epithelial cell densities using a robust nucleus detection by Definiens [31] and excluding stromal and immune cells based on size and shape. Due to small size of cells, the clustering is done on a higher resolution of 1.14 µm/pixel ($\sim$ 22.1% of the original WSI). Here, we assume that the nuclei of epithelial cells in lobular structures are usually round and 20–50 pixels large (i.e., a diameter of about 5–10 µm).

2.2. Method 2: Top–Down (MTD)

The second approach (Figure 4), in the following called MTD, has two basic steps where first candidate regions are identified in low resolution and then refined in medium resolution. In the first step, lobular candidate regions are detected from a texture image in low resolution (5.06 µm/pixel or 5% of the original image). The texture image is produced by a texture–based classification using local binary patterns (LBP) and local variance as features, as described in [16]. In [16], LBP/local variance features were compared with other texture features (Gabor filters and Haralick features) on epithelium/stroma classification of colorectal cancer images and achieved the best accuracy. Since the tissue properties in our samples are similar, we also used LBP and local variance. A support vector machine (SVM) is used for the classification. For training of the SVM model, small image subsets (from images not contained in our test set but belonging to the same series) containing only lobular tissue or only other tissue are used. The output of the classification is a gray value image (values 0–255) showing the probability of lobular tissue. This image is thresholded to obtain the candidate regions with a threshold (60 — corresponds to a lobular tissue probability of at least 0.24) chosen so that most of the lobules in the images from which the training subsets were taken are inside the candidate regions.

The second step, the refinement of lobular candidates, is done using algorithms from Definiens Developer XD 64 2.4; it is based on nuclear density, which is higher in lobular areas than in the surrounding tissue. In medium resolution (0.84 µm/pixel or 30%), the image in the candidate regions is filtered using a Laplacian of Gaussian (LoG) filter whose parameters (9 × 9 kernel, $\sigma = 1.768$) are set to give maximum response for blobs of the average size of epithelial nuclei. The LoG image is thresholded using a

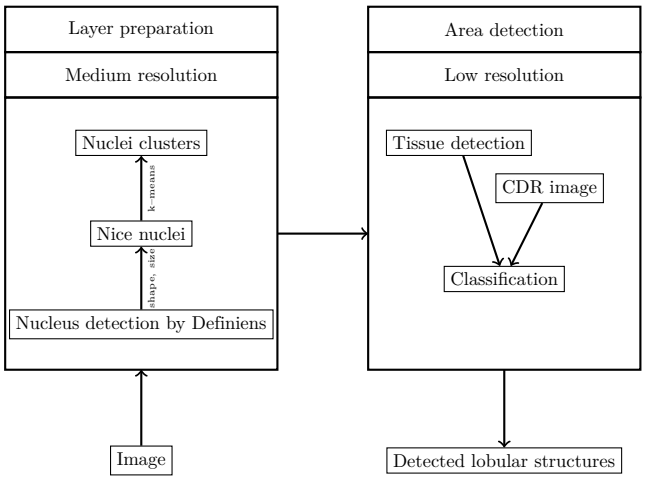

Figure 3: Overview of MBU processing a single WSI. The classification into lobular and non–lobular structures is based on a combination of nuclei clusters in a medium resolution (1.14 µm/pixel $\sim$ 22.1% of original WSI) and a cut color–distribution–range (CDR) image, which represents the range of RGB in tiles of $s \times s$ pixels, in a low resolution (8 µm/pixel $\sim$ 3.2% of original WSI).

fixed threshold to select an initial set of nuclei. The detected nuclei are filtered according to elliptic fit (≥ 0.8) so that only round nuclei remain. This nuclei detection does not attempt to detect all nuclei; the goal is rather to reliably detect a sufficient subset of epithelial nuclei for nuclear density estimation. The detected nuclei are used to produce a nuclear density channel using a sliding window approach (window size 41×41 pixel). Using the density channel, the candidate regions are shrunk iteratively until the remaining objects meet a minimum density criterion (average density ≥ 0.05). For further refinement, stromal and epithelial areas inside the lobular structures are separated by thresholding a smoothed gray value image. As threshold, the average of gray values in all refined lobular regions is used. Structures that contain very little epithelial tissue ($\leq 10\%$) after this step are eliminated. In a post–processing step, the borders of regions are smoothed by growing and shrinking, small regions are eliminated and small holes are filled.

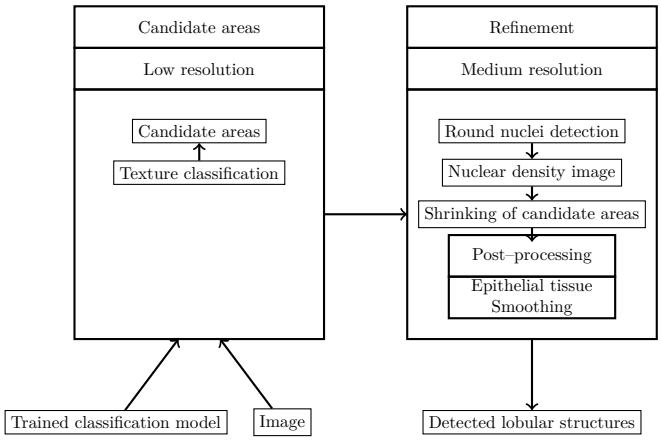

Figure 4: Overview of MTD processing a single WSI. Image and previously trained model are used as inputs. The first step uses texture–based classification in low resolution, the second step takes place in medium resolution and is based on nuclear density.

2.3. Method 3: Deep Learning (MDL)

A CNN has the ability to learn a hierarchical description of visual patterns from a set of annotated examples, and then make accurate predictions for new visual inputs. By combining patch–based image analysis with CNNs, we are able to automatically detect lobular structures in normal breast histological images. The process is summarized in Figure 5.

For patch–based image analysis, we assume that it is possible to predict the class of a pixel by observing its neighboring pixels. In order to keep computation time low, a pathologist visually estimated the lowest level of detail at which he could reliably distinguish epithelial tissue from other elements to be 4 µm/pixel. Likewise, we settled on a square neighborhood of size 128 µm which is

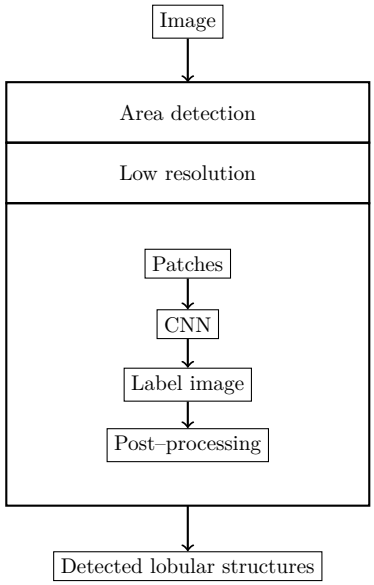

Figure 5: Overview of the CNN–based lobular detection at a fixed resolution of 4 µm/pixel: a CNN for binary classification is trained with an equal number of positive and negative RGB patches (32×32 pixels) randomly sampled from images annotated by a senior pathologist; the trained network can then be used to predict the class of patches extracted from a new image, resulting in a binary mask; after a simple post–processing step (removal of small connected components), the binary mask can be used to locate lobular structures in the image.

enough to cover a full cross–section of a duct with its surroundings. It is then possible to generate, for each pixel, a "raw" or "featureless" description of the patch centered on this pixel. In other words, to each pixel we associate a square RGB patch described by $32 \times 32 \times 3 = 3072$ values. In images annotated by a pathologist, each patch can be considered *positive* or *negative* according to its central pixel: if a pixel is in an annotated lobular region, then its patch is positive, otherwise it is negative.

Our CNN is a function that can predict the class of a patch (positive or negative) based on its featureless description. This function has a fixed form defined in the "CIFAR10 Quick" example provided with the software Caffe [32]. The associated deep network architecture (Table 3) is designed to perform multiclass classification of small RGB images[33], which is conveniently similar to the task of patch classification.

We can use it as a black box although its parameters (weights) need to be learned during supervised training. In order to estimate the prediction error of the method on unseen data for a given staining, we perform a 3–fold cross–validation using nine annotated images arbitrarily divided in three groups: A={NB02, NB05, NB11}, B={NB12, NB16, NB20}, and C={NB25, NB28, NB34} (see Table 1 for general information on ground truth contents). In fold 0, the method is trained on $B \cup C$ and evaluated on A;

Input	$32 \times 32 \times 3$
Convolution layer	kernels: 32; kernel size: 5×5; padding: 2
Max–pooling layer	kernel size: 3×3; stride: 2
Activation layer	Rectified Linear Unit (ReLU)
Convolution layer	kernels: 32; kernel size: 5×5; padding: 2
Activation layer	ReLU
Average-pooling layer	kernel size: 3×3; stride: 2
Convolution layer	kernels: 64; kernel size: 5×5; padding: 2
Activation layer	ReLU
Average–pooling layer	kernel size: 3×3; stride: 2
Fully connected layer	neurons: 64
Fully connected layer	neurons: 2

Table 3: Network architecture adapted from "CIFAR10 Quick" example to two classes, i.e. the number of neurons in the last layer is reduced to two instead of ten in the original.

in fold 1, the method is trained on $A \cup C$ and evaluated on B; and in fold 2, the method is trained on $A \cup B$ and evaluated on C. For any fold, training consists of 20 iterations using error backpropagation with adaptive learning rate (ADAGRAD [34]) on a dataset composed of 600,000 patches (50,000 for each class and each image). In order to reduce computation times, the number of iterations was fixed by experimenting on a preliminary dataset; as a consequence, we do not have a validation phase to optimize for this particular parameter (the apparent lack of overfitting in our test results suggests that an optimal number of iterations would be above 20). Given as input an image at 4 µm/pixel, this method will output a binary image where positive and negative pixels have different values. An additional post–processing step is used to denoise the result by removing small elements below a size threshold. In each fold, this threshold is learned by maximizing the macro–$F1$ score (average of $F1$ scores for each class) on the six training images.

2.4. Combination

We notice that the misclassified objects are not the same for each method. This suggests that they potentially offer complementary information about the patterns to be classified. As stated in [35, 36], if we have many different classifiers it is sensible to consider using them in a combination in the hope of increasing the overall accuracy.

In order to evaluate whether a combination of our methods will improve the results, we use the detection results of every single method in the form of binary masks to produce a combined result. In a pixel–based combination of masks (in low resolution, 5.06 µm/pixel or 5%), all pixels which belong to detected areas for at least two of the three methods are labeled as detected. This majority voting procedure is useful to eliminate false detections that are made by only one method.

3. Results and discussion

The methods were tested and reviewed thoroughly on a set of nine cases (test set: NB02–NB34); for these cases, $F1$ scores and object–based evaluations were produced.

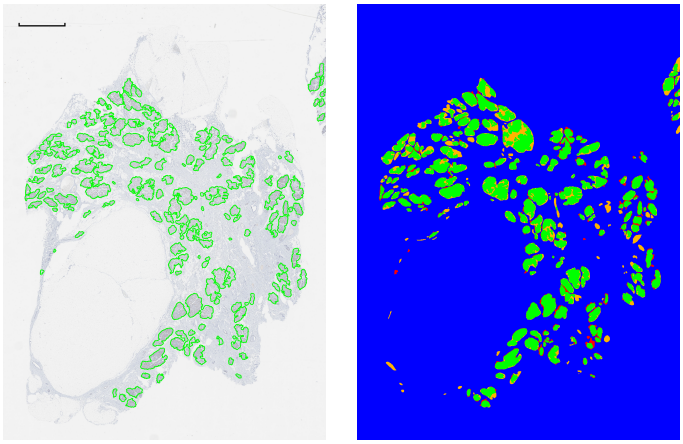

Figure 6: Image with best result for MBU. For the MBU, the best result was in NB34 ER, with a $F1$ score of 0.84. Left outlines of detected lobular areas in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Scale bar is 2 mm.

Obviously, all shown $F1$ scores are based on test results (not from training).

The overall performance of all methods for the stainings ER and CD8 is given in Table 4, as $F1$ score (measured pixel–based) for each of the 18 test images as well as the average for each staining. The highest $F1$ scores averaged over both stainings were obtained by MTD (0.59) and MDL (0.60) for the test set.

Staining	Method	NB02	NB05	NB11	NB12	NB16
ER	MBU	.77	.17	.49	.42	.62
	MTD	.73	.51	.54	**.73**	.76
	MDL	.69	**.67**	.56	.24	**.80**
	Combination	**.80**	**.67**	**.62**	.69	**.80**
CD8	MBU	.36	.25	.60	.61	.71
	MTD	.65	.42	.58	.68	.57
	MDL	.59	.51	.60	.50	.77
	Combination	**.68**	**.67**	**.67**	**.69**	**.78**

Staining	Method	NB20	NB25	NB28	NB34	Average
ER	MBU	.41	.41	.68	.84	.53±.20
	MTD	.71	.57	.61	.84	.67±.11
	MDL	.62	**.76**	.59	.80	.64±.16
	Combination	**.74**	.62	**.70**	**.87**	**.72±.08**
CD8	MBU	.20	.37	.17	.74	.45±.21
	MTD	.45	.37	.20	.70	.51±.17
	MDL	.66	**.61**	.04	.75	.56±.20
	Combination	**.68**	.48	**.27**	**.80**	**.64±.17**

Table 4: Comparison of $F1$ scores for the different methods applied to ER and CD8 images (NBx) of the test set.

3.1. Visual comparison

Figures 6 to 8 show detection results of lobular areas and comparisons with ground truth for the image with the best $F1$ score for each method. For all three methods, the best results were achieved in the NB34 image (ER). This case contains many compact lobules with dense epithelial tissue, which makes them stand out against the surrounding tissue quite clearly. In addition, most of its lobules

are positive for both ER and CD8 (i.e., many epithelial cells stained for ER are colored in brown and/or several brown colored CD8$^+$ T lymphocytes are in close contact to epithelial cells) further increasing their differences to the background.

Some characteristics of the different methods can be observed in these images. Correctly detected lobular structures from both MBU (Figure 6) and MTD (Figure 7) tend to have small stripes of false negative area around them, i.e., the detected area is usually smaller than the area that was manually outlined as ground truth. In the case of MBU, false negative areas within ground truth lobular objects occur due to the pixel–based classification instead of segmentation approaches using e.g., watershed or region growing. Intra–lobular stroma is usually not classified as lobular area due to missing contrasts by epithelial cells. This could be solved by additionally including some growing (and shrinking) procedures to reach smoother regions. The main advantage of MBU is that it is very fast (less than 2 min per image on a computer with a 3.6 GHz CPU, 32 GB RAM, using preprocessed nuclei detection) provided that the time–consuming nuclei detection has already been done, which is necessary anyway for cell population analysis. As mentioned before, we applied a robust method performing well for segmenting nuclei and classifying into different cell types by Definiens, which requires approximately 11 hours per image (single core usage in 3.4 GHz cluster environment with 384 GB RAM). In applications that do not quantify cells, it is also possible to use a rough and fast nuclei segmentation in MBU. For MTD, on average 38 min per image are needed on a computer with a 3.6 GHz CPU, 32 GB RAM. In contrast to MBU and MTD, there are practically no false negative regions in the result from MDL (Figure 8). The largest region labeled as false negative is actually a clearly stromal region inside a lobule, which is part of our ground truth.

False positive areas also occur for all methods; these are often located in the small space between adjacent lobular structures. Correctly detected lobular structures from MDL almost always have a small stripe of false positive area around them, i.e., the detected area is larger than the ground truth area. Although false positive or false negative borders around correctly detected lobular structures are quite narrow, they contribute to the measurements shown in Table 4, so that the detection quality needed for applications, as judged by visual inspection, may be better than these numbers suggest.

Some typical examples of detection problems for the different methods can be seen in Figures 9 to 11. In Figure 9, a subset of NB02, CD8 contains a false positive area from MBU, where the tissue has an altered structure due to the IHC processing. Such regions often cause false positive detections. Due to calculation of CDR, the method strongly relies on textures in which the color range is quite high, which could be caused for example by tissue processing artifacts due to antigene retrieval (bright pixels directly neighbored to dark pixels).

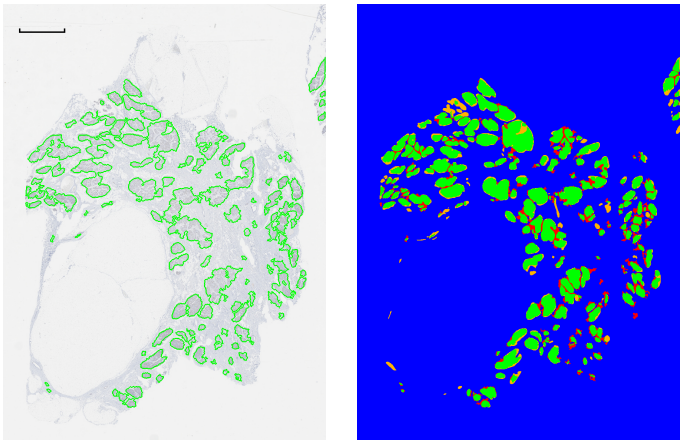

Figure 7: Image with best result for MTD. For the MTD, the best result was in NB34 ER, with a $F1$ score of 0.84. Left outlines of detected lobular areas in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Scale bar is 2 mm.

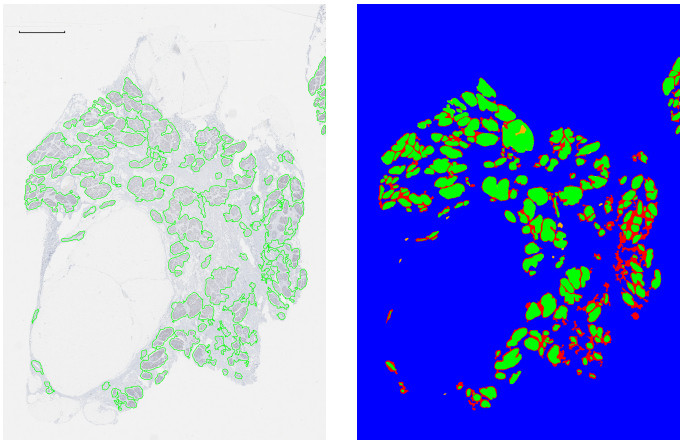

Figure 8: Image with best result for MDL. For MDL, the best result was in NB34 ER, with a $F1$ score of 0.80. Green lines show outlines of detected lobular structures. Left outlines of detected lobular areas in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Scale bar is 2 mm.

Figure 10 shows an example of false negative detection from MTD. This method relies on the measurement of nuclear density from detected nuclei. In the example, the lobular structure is relatively loose, meaning that epithelial tissue is sparse. In addition, the epithelial nuclei do not have large contrast to their surroundings, so that some are missed in the nuclei detection step. As a result, the nuclear density in the given area is below the required threshold, and the region was eliminated from the detection result. False positive detections (not shown) are often caused by vessels containing blood aggregations. These regions, though hardly stained, have a very similar texture

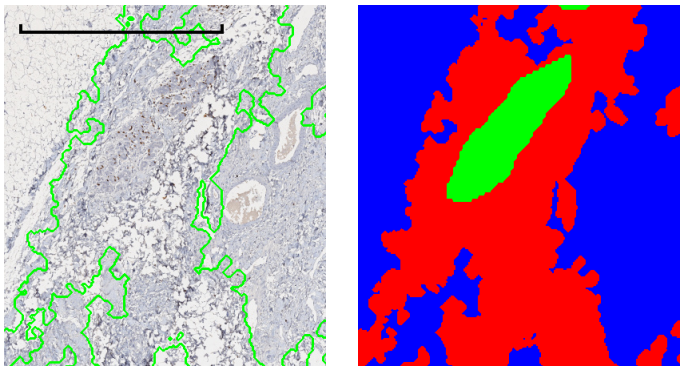

Figure 9: Example for false positive detection in MBU. Left outlines of detected lobular structures in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Example subset from NB02 CD8. False positive detections are caused by changes in tissue structure due to IHC processing. Scale bar is 1 mm.

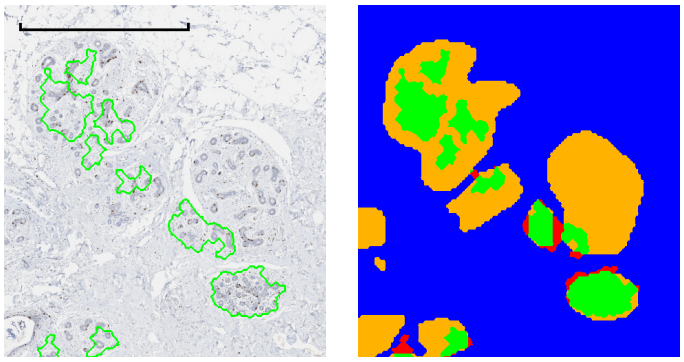

Figure 10: Example for false negative detection in MTD. Left outlines of detected lobular structures in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Example subset from NB16 CD8. Lobular structures are missed from detection because measured nuclear density was too low: epithelial tissue is sparse and nuclei have low contrast in image. Scale bar is 1 mm.

to that of lobules, which leads to their detection.

Figure 11 shows an image with a large false positive area from MDL. This method depends heavily on the quality of the training set: classes should be balanced, examples should be correctly labeled and they should also be representative of the variety that can be found in unseen data. This last criterion is the hardest to achieve and image NB28 CD8 exhibits a stromal texture that is not found in the others. This could be an artifact resulting from the physical process used to prepare the slide. While it is an infrequent event, it does happen more than once in the rest of our images.

Alternatively, it is possible that the architecture or the fixed number of iterations are inadequate, and we may get a higher score by using more iterations or, better yet, by

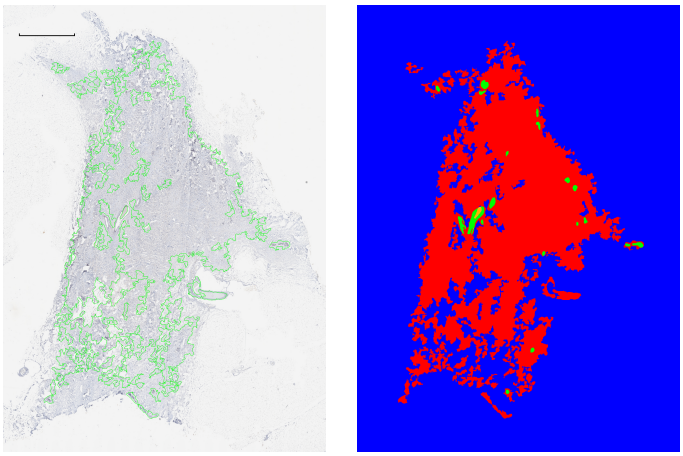

Figure 11: Example for false positive detection in MDL. Left outlines of detected lobular structures in green, right evaluation result (green: true positive areas, blue: true negative areas, red: false positive areas, orange: false negative areas). Example from NB28 CD8. Scale bar is 2 mm.

adding a validation step. Another disadvantage of MDL is that it requires a specific hardware, but the resulting speed is satisfying. Using a Nvidia Quadro K4200 GPU with 4 GB of memory, training for one fold takes less than two hours and evaluation for an image takes four minutes on average.

3.2. True positives in lobular structures

Histograms from per–lobular area analysis of true positive percentages (Figure 12) show the characteristics of the different methods in a more quantified manner.

In Figure 12 rows 1 and 3, histograms of percentage of ground truth area in detected lobular structures are shown, indicating how correct the detections are and how many false detections occurred. The first bin (0–5%) in the histograms of row 1 and 3 shows how many false positive objects appear in the results of the different methods. In all three methods, more false positive objects were detected in CD8 than in ER (2–3 times as many). Most false detections are observed in MBU, but on the other hand, the correctly found lobular areas are more cleanly detected (contain less non–lobular tissue) than in the other methods. For MTD, this picture is more heterogeneous, and in the results of MDL, most lobular structures contain some amount of non–lobular tissue (see also Figure 8).

Figure 12 rows 2 and 4 show histograms of percentage of detected area in ground truth regions, indicating how complete the detection is for existing ground truth regions and how many ground truth regions are missed from the detection. The first bin in the histograms of row 2 and 4 shows how many lobular structures are missed. It can be seen that MDL has the least total losses. Most ground truth regions are completely or almost completely

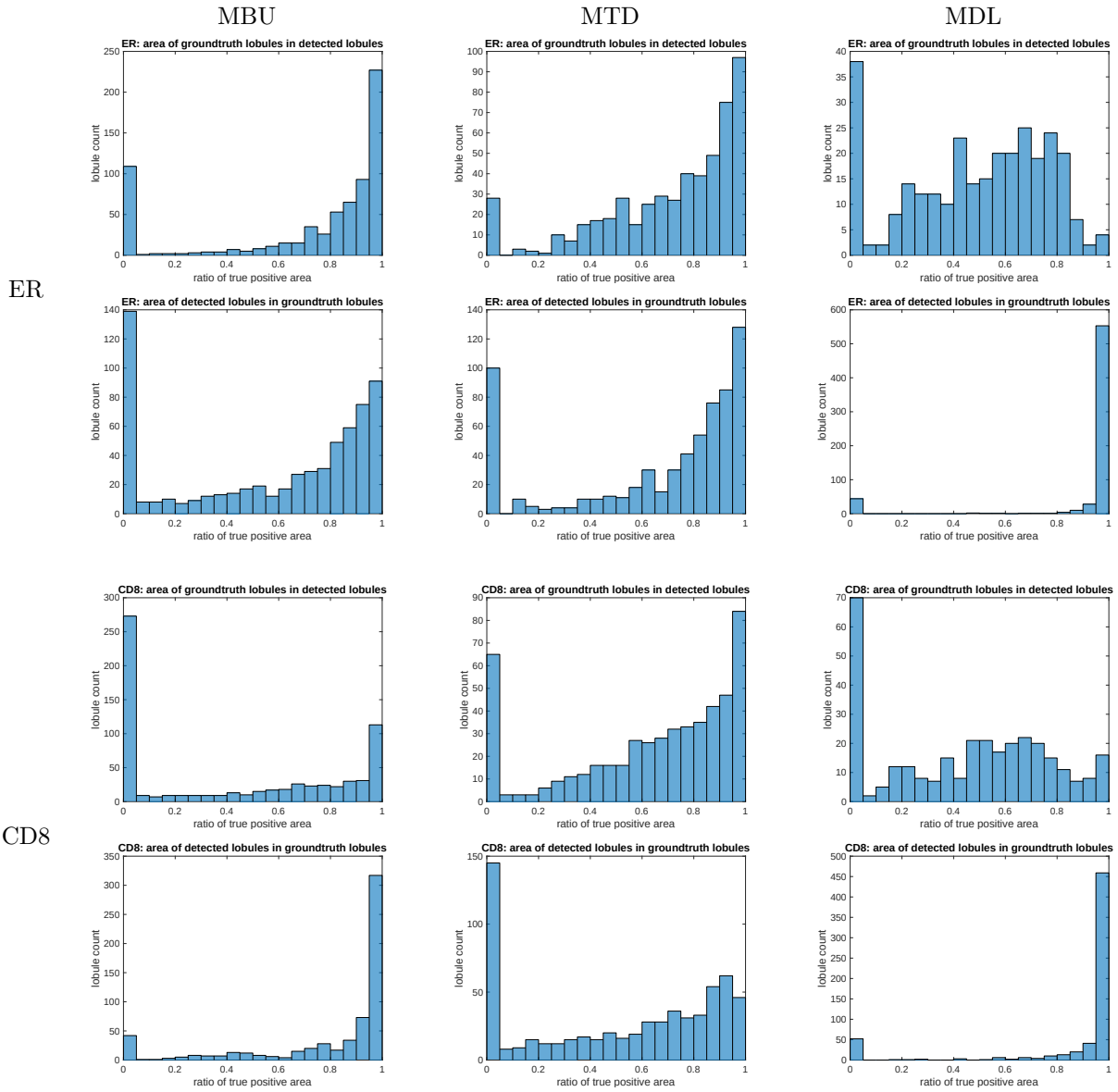

Figure 12: Histograms of true positive ratios for detected and ground truth areas. Left column: MBU, center column: MTD, right column: MDL. The upper two rows show histograms for ER stained images, the lower two rows for CD8 images.

detected, and there are very few where parts are missing. In the other two methods, with the exception of MBU at CD8, the total losses are higher and the detection is more heterogeneous: there are many lobular areas where only parts of the ground truth regions are detected. As described above, missing parts are often located along the borders of detected areas. While MDL shows no substantial differences between ER and CD8 images, MBU and MTD show some differences: for MBU, ER has more missed detections than CD8 and fewer complete detections. This is similar for MTD CD8. Both may be caused by method construction: in MBU the range r_X is higher for brown colored than blue colored cells compared to bright stroma, such that highly infiltrated CD8 tissue (NB16, NB34) is easier to detect. In contrast, MTD tends to miss such areas because the cell detection method is not adapted to CD8 staining. For both methods, it is possible to include a staining specific nuclei detection to avoid change in texture by immune cell infiltrations and utilize the highlighting of epithelial cells by ER.

In summary, MDL shows a high completeness in detection — hardly any structure is missed — but most of the detected lobular objects also contain non–lobular tissue. MBU has a number of missed as well as false detections, but correctly detected areas are relatively clean. MTD has more missed detections than MDL, but also more objects that contain mostly lobular tissue.

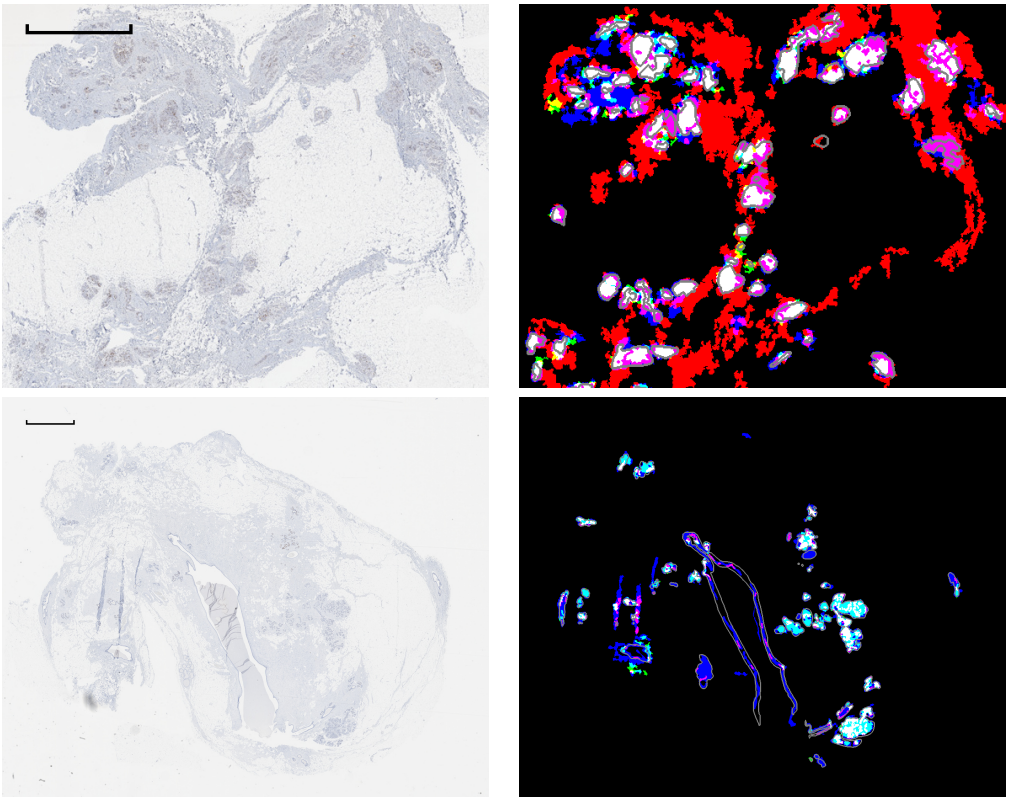

Figure 13: Combination of methods. White: labeled as lobular by all three methods; yellow: labeled as lobular by MBU and MTD; magenta: labeled as lobular by MBU and MDL; cyan: labeled as lobular by MTD and MDL; red: labeled as lobular by MBU only; green: labeled as lobular by MTD only; blue: labeled as lobular by MDL only. Gray lines: border lines of ground truth. Top: A subset of NB02 CD8 shows the effect of the combination. Detections by one method only (red, green, blue) often lie outside of the ground truth. Ground truth regions are almost completely covered, in this example often by all three methods or by combinations of MBU and MDL (magenta), sometimes by combinations of MTD and MDL (cyan) or combinations of MBU and MTD (yellow). Bottom: In NB25 ER, large ductal areas are only detected by MDL (blue). Scale bars are 2 mm.

3.3. Combination

The top part of Figure 13 visualizes the effect of the combination of results. The $F1$ score results can be found in Table 4 (row "Combination" for both stainings). The results show that in most of the cases the $F1$ score of the combination is higher than or at least as high as that of every single method. The exceptions are NB12 (ER) and NB25 (both stainings) from the test set, where the combination result is still second best but with a larger distance to the best single method.

NB25 has a large amount of ductal tissue, which is incorporated in the ground truth but for the most part only detected by MDL (Figure 13, bottom). If the target of an analysis are primarily the lobules but not the ducts, the combination seems to be the better choice also in this case. Further, using both the combination results and the results of MDL might be helpful in distinguishing lobular and ductal tissue.

4. Conclusion

In order to automate the detection of lobular structures in digital WSIs of normal breast tissue, we developed and compared several image analysis methods: two methods combining manually designed features and cell detection and a featureless machine learning–based image analysis.

We showed that despite a moderate variance for any given staining, each gives good results based on visual check of accuracy, and that the output could be used as basis for further image analysis (e.g., cell populations identification and quantification). We also showed that combining all the methods by pixel–level majority voting improves precision and might help with subclassification of lobular tissue into lobules and ducts. We will leverage the knowledge gained in this study to tackle the issue of lobular structure detection in cancerous breast tissue and expand the concept of ROI–targeted immune cell detection in oncoimmunology.

5. Acknowledgements

The authors gratefully acknowledge funding by German Federal Ministry of Education and Research (BMBF) for the eMed project SYSIMIT (grant 01ZX1308A). The authors thank Nicole Krönke for experimental work.

Hardware for MDL was partially funded by French National Center for Scientific Research (CNRS).

6. Author contributions

GA developed MDL, AG MTD, and NSS MBU. GA, AG, and NSS designed the evaluation and combination and prepared the manuscript. RS applied the nuclei detection used in MBU and established exchange of image analysis data. FF provided the expert annotations and pathological input. GF, CW, BN, and FF helped in study design and edited the manuscript. All authors read and approved the final manuscript.

7. Conflict of interest statement

RS is a full–time employee of Definiens AG. GA, NSS, BN, GF, FF, CW, and AG did not report any conflict of interest.

References

[1] Elenbaas, B., Spirio, L., Koerner, F., Fleming, M.D., Zimonjic, D.B., Donaher, J.L., et al. Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Development 2001;15:50–65.

[2] Douglas-Jones, A.G.. Lymphocytic lobulitis in breast core biopsy: a peritumoral phenomenon. Histopathol 2006;48:209–212.

[3] Hermsen, B.B.J., von Mensdorff-Pouilly, S., Fabry, H.F.J., Winters, H.A.H., Kenemans, P., Verheijen, R.H.M., et al. Lobulitis is a frequent finding in prophylactically removed breast tissue from women at hereditary high risk of breast cancer. J Pathol 2005;206:220–223. doi:10.1002/path.1774.

[4] Daniel, C., Rojo, M.G., Klossa, J., Della Mea, V., Booker, D., Beckwith, B.A., et al. Standardizing the use of whole slide images in digital pathology. Comput Med Imaging Graph 2011;35:496–505. doi:10.1016/j.compmedimag.2010.12.004.

[5] Fuchs, T.J., Buhmann, J.M.. Computational pathology: Challenges and promises for tissue analysis. Comput Med Imaging Graph 2011;35:515–530. doi:10.1016/j.compmedimag.2011.02.006.

[6] Alilou, M., Kovalev, V., Taimouri, V.. Segmentation of cell nuclei in heterogeneous microscopy images: A reshapable templates approach. Comput Med Imaging Graph 2013;37:488–499. doi:10.1016/j.compmedimag.2013.07.004.

[7] Zhang, L., Kong, H., Chin, C.T., Liu, S., Chen, Z., Wang, T., et al. Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts. Comput Med Imaging Graph 2014;38:369–380. doi:10.1016/j.compmedimag.2014.02.001.

[8] Zhang, X., Xing, F., Su, H., Yang, L., Zhang, S.. High-throughput histopathological image analysis via robust cell segmentation and hashing. Med image analysis 2015;26:306–315.

[9] Krüger, J.M., Wemmert, C., Sternberger, L., Bonnas, C., Dietmann, G., Gançarski, P., et al. Combat or surveillance? evaluation of the heterogeneous inflammatory breast cancer microenvironment. J Pathol 2013;229:569–578. doi:10.1002/path.4150.

[10] Li, C., Anderson, B., Daling, J., Moe, R.. Trends in incidence rates of invasive lobular and ductal breast carcinoma. Jama 2003;289:1421–1424.

[11] Gulbahce, H.E., Vanderwerf, S., Blair, C., Sweeney, C.. Lobulitis in nonneoplastic breast tissue from breast cancer patients: association with phenotypes that are common in hereditary breast cancer. Hum Pathol 2014;45:78–84. doi:10.1016/j.humpath.2013.08.008.

[12] Degnim, A.C., Brahmbhatt, R.D., Radisky, D.C., Hoskin, T.L., Stallings-Mann, M., Laudenschlager, M., et al. Immune cell quantitation in normal breast tissue lobules with and without lobulitis. Breast Cancer Res Treat 2014;144:539–549. doi:10.1007/s10549-014-2896-8.

[13] Grote, A., Abbas, M., Linder, N., Kreipe, H.H., Lundin, J., Feuerhake, F.. Exploring the spatial dimension of estrogen and progesterone signaling: detection of nuclear labeling in lobular epithelial cells in normal mammary glands adjacent to breast cancer. Diagnostic Pathol 2014;9:S11. doi:10.1186/1746-1596-9-S1-S11.

[14] Eramian, M., Daley, M., Neilson, D., Daley, T.. Segmentation of epithelium in H&E stained odontogenic cysts. J Microscopy 2011;244:273–292. doi:10.1111/j.1365-2818.2011.03535.x.

[15] Di Cataldo, S., Ficarra, E., Acquaviva, A., Macii, E.. Automated segmentation of tissue images for computerized IHC analysis. Comp Meth Progr Biomed 2010;100:1–15. doi:10.1016/j.cmpb.2010.02.002.

[16] Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordlin, S., et al. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic Pathol 2012;7. doi:10.1186/1746-1596-7-22.

[17] Roullier, V., Lezoray, O., Ta, V.T., Elmoataz, A.. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. Comput Med Imaging Graph 2011;35:603–615.

[18] Lomenie, N., Racoceanu, D.. Point set morphological filtering and semantic spatial configuration modeling: Application to microscopic image and bio-structure analysis. Pattern Recog 2012;45:2894–2911.

[19] Wernick, M.N., Yang, Y., Brankov, J.G., Yourganov, G., Strother, S.C.. Machine learning in medical imaging. Signal Processing Magazine, IEEE 2010;27:25–38. doi:10.1109/MSP.2010.936730.

[20] Bengio, Y.. Learning deep architectures for AI. Foundations Trends Mach Learn 2009;2:1–127. doi:10.1561/2200000006.

[21] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.. Gradient-based learning applied to document recognition. In: Proc. IEEE; vol. 86. 1998,.

[22] Krizhevsky, A., Sutskever, I., Hinton, G.E.. Imagenet classification with deep convolutional neural networks. In: NIPS. Curran Associates, Inc.; 2012, p. 1097–1105.

[23] Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: SPIE Med. Imag.; vol. 9041. 2014, p. 3–15. doi:10.1117/12.2043872.

[24] Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II; chap. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-40763-5; 2013, p. 411–418. URL: http://dx.doi.org/10.1007/978-3-642-40763-5_51. doi:10.1007/978-3-642-40763-5_51.

[25] Feuerhake, F., Sigg, W., Höfter, E.A., Unterberger, P., Welsch, U.. Cell proliferation, apoptosis, and expression of bcl–2 and bax in non–lactating human breast epithelium in relation to the menstrual cycle and reproductive history. Breast Cancer Res Treat 2003;77:37–48.

[26] Harvey, J., Clark, G., Osborne, C.K., Allred, D.C.. Estrogen receptor status by immunohistochemistry is superior to

the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol 1999;17:1474–1481.

[27] Hammond, M.E., Hayes, D., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., et al. Asco-cap guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol 2010;28:2784–2795.

[28] Denkert, C., Von Minckwitz, G., Brase, J., Sinn, B., Gade, S., Kronenwett, R., et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2–positive and triple-negative primary breast cancers. J Clin Oncol 2015;33:983–991.

[29] Dice, L.R.. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

[30] Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B.. Histopathological image analysis: A review. IEEE Rev Biomed Eng 2009;2:147–171. doi:10.1109/RBME.2009.2034865.

[31] Brieu, N., Pauly, O., Zimmermann, J., Binnig, G., Schmidt, G.. Slide specific models for segmentation of differently stained digital histopathology whole slide images. Proc SPIE 2016;9784:978410–7.

[32] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:14085093 2014;.

[33] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. Imagenet large scale visual recognition challenge. Internat J Comput Vision 2015;115:211–252. doi:10.1007/s11263-015-0816-y.

[34] Duchi, J., Hazan, E., Singer, Y.. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011;12:2121–2159.

[35] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20:226–239. doi:10.1109/34.667881.

[36] Jain, A.K., Duin, R.P.W., Mao, J.. Statistical pattern recognition: A review. IEEE Trans Pattern Anal Mach Intell 2000;22:4–37.