From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge

Bandi, Peter; Geessink, Oscar; Manson, Quirine; van Dijk, Marcory; Balkenhol, Maschenka; Hermsen, Meyke; Bejnordi, Babak Ehteshami; Lee, Byungjae; Paeng, Kyunghyun; Zhong, Aoxiao

Total number of authors:
36

Published in:
I E E E Transactions on Medical Imaging

Link to article, DOI:
10.1109/TMI.2018.2867350

Publication date:
2018

Document Version
Peer reviewed version

From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens

***Abstract*—Automated detection of cancer metastases in lymph nodes has the potential to improve assessment of prognosis for patients. To enable fair comparison between the algorithms for this purpose, we set up the CAMELYON17 challenge in conjunction with the IEEE International Symposium on Biomedical Imaging 2017 conference in Melbourne.**

Over 300 participants registered on the challenge website, of which 23 teams submitted a total of 37 algorithms before the initial deadline. Participants were provided with 899 whole-slide images for developing their algorithms. The developed algorithms were evaluated based on the test set encompassing 100 patients and 500 whole-slide images. The evaluation metric used was a quadratic weighted Cohen's kappa.

We discuss the algorithmic details of the ten best pre-conference and two post-conference submissions. All these participants used convolutional neural networks in combination with pre- and postprocessing steps. Algorithms differed mostly in neural network architecture, training strategy and pre- and postprocessing methodology.

Overall, the kappa metric ranged from 0.89 to -0.13 across all submissions. The best results were obtained with pre-trained architectures such as ResNet. Confusion matrix analysis revealed that all participants struggled with reliably identifying isolated tumor cells, the smallest type of metastasis, with detection rates below 40%. Qualitative inspection of the results of the top participants showed categories of false positives, such as nerves or contamination, which could be targets for further optimization. Last, we show that simple combinations of the top algorithms result in higher kappa metric values than any algorithm individually, with 0.93 for the best combination.

***Index Terms*—breast cancer; sentinel lymph node; lymph node metastases; whole-slide images; grand challenge**

P. Bandi, O. Geessink, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, P.Bult, J. van der Laak, and G. Litjens were with the Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands. Q. Manson was with the Department of Pathology, University Medical Center Utrecht, Utrecht, the Netherlands. M. van Dijk was with the Department of Pathology, Rijnstate Hospital, Arnhem, Gelderland, the Netherlands. B. Lee, and K. Paeng were with Lunit Inc., Seoul, Gyeonggi-do, the Republic of Korea. A. Zhong, and Q. Li were with the Department of Radiology, Massachusetts General Hospital - Harvard Medical School, Boston, MA USA. F. Ghazvinian Zanjani, and S. Zinger were with the Department of Electrical Engineering, Technical University of Eindhoven, Eindhoven, the Netherlands. K. Fukuta was with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan. D. Komura was with the Department of Genomics Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. V. Ovtcharov was with Indica Labs, Corrales, NM USA. S. Cheng, and S. Zeng were with Huazhong University of Science and Technology, Wuhan, P. R. China. J. Thagaard, and A. B. Dahl were with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark. H. Lin, and H. Chen were with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, PRC. L. Jacobsson, and M. Hedlund were with ContextVision AB, Stockholm, Sweden. M. Çetin and E. Halıcı were with the Middle East Technical University, Ankara, Turkey. H. Jackson, and R. Chen were with Proscia Inc., Baltimore, MD USA. F. Both and J. Franke were with Karlsruhe Institute of Technology, Karlsruhe, Germany. H. Küsters-Vandevelde, and W. Vreuls were with the Department of Pathology, Canisius-Wilhelmina Hospital, Nijmegen, the Netherlands. Bram van Ginneken was with the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands.

I. Introduction

BREAST cancer is the most common cancer among women in the United States of America [1]. Within their lifetime, 12% of women are diagnosed with breast cancer. In 2017, an estimated 252,710 women were diagnosed with breast cancer, which accounts for 30% of all diagnosed cancer cases, and approximately 40,610 women died from the disease.

The prognosis of breast cancer patients is mainly determined by whether the cancer is organ-confined or has spread to other parts of the body [2]. An internationally accepted means to classify the extent of cancer is the tumor, (regional) lymph nodes, distant metastasis (TNM) staging system [3]. The TNM staging system is one of the most important tools for clinicians to select a suitable treatment for the patient. In breast cancer, TNM staging takes into account the size of the tumor (T-stage), whether the cancer has spread to the (regional) lymph nodes (N-stage), and whether the tumor has metastasized to other parts of the body (M-stage).

The axillary lymph nodes are typically the first location breast cancer metastasizes to. Currently, the status of these lymph nodes is almost always assessed by applying the sentinel lymph node procedure. This procedure tries to identify the nearest lymph nodes to which the tumor drains, which are then excised for pathologic examination [4], [5]. Typically, a blue dye and/or a radioactive tracer is injected in or near the tumor prior to surgery to identify these sentinel lymph nodes.

After formalin fixation and paraffin embedding, a couple of micrometers thin slices are cut from the excised nodes and placed on glass slides (typically 3-5 sections per lymph node). These slides are then stained with hematoxylin and

eosin (H&E) to highlight the cell nuclei and the general structural features of the tissue (Figure 1). Through microscopic assessment the pathologist screens the slides for tumor presence. If tumor cells are found, the pathologist measures their extent in order to determine the pathologic N stage (pN-stage) of the tumor. In case of unclear diagnosis on H&E, immunohistochemical (IHC) staining for cytokeratin can be used for clarification and is standard diagnostic practice in the Netherlands [6], [7].

The histopathological analysis of lymph nodes is time consuming, tedious and pathologists may miss small metastases [8]. The introduction of whole-slide imaging, which allows for the high-resolution digitization of glass slides, has paved the way for (partly) automating this work [9]. Automation can potentially improve the efficiency and accuracy of histopathological lymph node assessment.

In the medical image analysis research field, grand challenges have shown to be a very successful approach to quickly advance the state of the art. Typically, the challenge organizers define a clinically relevant task and release a sufficiently large and diverse training set to allow participants to build algorithms to solve a specific problem. Subsequently, algorithms are uniformly evaluated by the organizers to allow a fair performance comparison. There have been many successful challenges in recent years, in many medical imaging fields, for example: liver segmentation in CT (SLIVER07) [10], brain tumor segmentation in MRI (BRATS) [11], or lung nodule detection in CT (LUNA16) [12].

In 2016, we organized the 'CAncer MEtastases in LYmph nOdes challeNge' (CAMELYON16) to improve automated breast cancer metastases detection in whole-slide images (WSIs) of sentinel lymph nodes [13]. As part of the challenge, we organized a reader study in which 11 pathologists under time constraint and 1 pathologist without time-constraint performed the same task as the algorithms in the challenge. We found that the best performing algorithms in the challenge perform at the level of the pathologist without time-constraint and perform significantly better than pathologists under time pressure. However, CAMELYON16 did not yet mimic clinical practice, limiting the conclusions that could be drawn from its results. We sought to amend these limitations with CAMELYON17. The following key changes were made to the setup of CAMELYON16:

- In CAMELYON16 we focused on classification of single WSIs whereas in CAMELYON17 we focus on patient-level pN-stage prediction including multiple WSIs per patient.
- Isolated tumor cells (ITC), the smallest type of metastasis, were excluded in CAMELYON16 and have now been included.
- Five centers providing cases were included instead of only two centers, allowing for a more accurate representation of preparation and staining diversity across laboratories and scanners.
- The challenge data set size increased from 399 to 1399 WSIs to get a better estimate of algorithm performance and allow participants to train better systems.

This paper discusses the results of the CAMELYON17 challenge, which were partly presented in a workshop during the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 in Melbourne, Australia. The next sections describe the data set, the challenge setup, and the algorithm evaluation strategy. Subsequently, we describe the methodology of the ten best pre-workshop and two post-workshop submissions and compare their results (ranking in Table I). Last, we discuss the results, the limitations of the study and recommendations for future work.

II. Materials

A. Whole-slide images

We included patients from five different medical centers from the Netherlands: slides from 130 lymph node resections from Radboud University Medical Center in Nijmegen (RUMC), 144 from Canisius-Wilhelmina Hospital in Nijmegen (CWZ), 129 from University Medical Center Utrecht (UMCU), 168 from Rijnstate Hospital in Arnhem (RST), and 140 from Laboratory of Pathology East-Netherlands in Hengelo (LPON). Of these patients we collected glass slides of H&E-stained sentinel lymph nodes. Whenever available, we also collected the corresponding IHC slides, stained for cytokeratin, to establish the reference standard. IHC slides were generally only available for more difficult cases for which in the H&E slides no tumor was detected on first reading. No consecutive H&E-slides from the same lymph node were included.

The glass slides were digitized with whole-slide scanners, resulting in WSIs. The slides from RUMC, CWZ and RST were scanned in the RUMC with an 3DHistech P250 whole-slide scanner with a pixel size of $0.24\,\mu\mathrm{m}$. The slides from LPON were scanned locally with their Philips IntelliSite Ultra Fast Scanner with a $0.25\,\mu\mathrm{m}$ pixel size. The UMCU used a Hamamatsu XR C12000 whole-slide scanner with a $0.23\,\mu\mathrm{m}$ pixel size.

The WSIs contained multiple resolution levels, with approximately 1×10^5 by 2×10^5 pixels at the highest resolution level. Each consecutive resolution level doubled the pixel size in both directions and halved the pixel count in each dimension. The typical file size of a WSI was about 4 GB, but it varied greatly depending on the scanner and tissue content of the image. The vendor-specific image formats were anonymized and converted to standard multi-resolution TIFF image files. For a description of the file format, see http://openslide.org/formats/generic-tiff/. The size of the complete data set was 3030.5 GB divided as 715.9 GB and 2314.6 GB between CAMELYON16 and CAMELYON17, respectively.

B. WSI labeling

Clinically, three types of metastases are distinguished, based on size: macro-metastases, micro-metastases and ITC (Table II). Although the clinical relevance of ITCs is debated, they have to be reported by pathologists and affect the pN-stage when no macro- or micro-metastases are present. When multiple metastases are present in a slide, the metastasis with the largest size determines the slide label.

TABLE I: CAMELYON17 combined leaderboard (pre- and post-workshop submissions)

Rank	Team	Affiliation	Kappa Score
1	Lunit	Lunit Inc.	0.8993
2	HMS-MGH-CCDS	Harvard Medical School, Mass. General Hospital, Center for Clinical Data Science	0.8806
3	VCA-TUe	Electrical Engineering Department, Eindhoven University of Technology	0.8729
4	MIL-GPAT	The University of Tokyo, Tokyo Medical and Dental University	0.8567
5	Indica Labs	Indica Labs	0.8554
6	chengshenghua	Huazhong University of Science and Technology, Britton Chance Center for Biomedical Photonics	0.8439
7	DTU	Technical University of Denmark	0.8098
8	IMT-CUHK	Imsight Medical Technology, Chinese University of Hong Kong, Xiamen University	0.7718
9	desuto	ContextVision	0.7640
10	METU-VISION	Middle East Technical University	0.7599
11	Proscia	Proscia Inc., Carnegie Mellon University, Moffitt Cancer Center	0.7594
12	ML-KA	Karlsruhe Institute of Technology	0.7330

TABLE II: Rules for assigning single cells or clusters of metastasized tumor cells to a metastasis category

Category	Size
Macro-metastasis	Larger than 2 mm
Micro-metastasis	Larger than 0.2 mm and/or containing more than 200 cells, but not larger than 2 mm
Isolated tumor cells	Single tumor cells or a cluster of tumor cells not larger than 0.2 mm or less than 200 cells

Every WSI was labeled with one of the ”macro”, ”micro”, or ”ITC” metastasis categories by a pathologist, based on the largest lesion present in the H&E stained slide using the corresponding cytokeratin-stained slide as a reference, if available. When no metastasis was present in the H&E stained slide, it was labeled ”negative”. Examples are shown in Figure 1 and 2.

C. Assigning pN-stage labels

The pN-stages are based on several slides per lymph node and, depending on the surgical procedure, several lymph nodes per patient. Furthermore, some pN-stages are based on lymph node locations or extra molecular tests. To keep the total data set size of CAMELYON17 within reasonable limits, the stages which require more than 5 lymph nodes per patient were excluded. Furthermore, as this is an image analysis challenge, we removed the stages that depend on non-imaging information. The final subset of pN-stages used in the challenge is indicated in Table III. For a full listing we refer the reader to the seventh edition of the TNM Classification of Malignant Tumors [3].

As it is almost impossible to find a roughly uniform distribution of patients across pN-stages at multiple institutions. For the purpose of this challenge we decided to create artificial patients. These artificial cases were constructed by grouping 5 WSIs from different patients from a single center as being from one individual, where each WSI resembled one lymph node. This facilitated a comparable pN-stage distribution between centers. We shared 40 of these artificial patients per medical center. The training set included 20 patients from each center with a disclosed pN-stage for each artificial patient and the metastasis label for each individual slide in the set.

The test set was composed of another 100 artificial patients (Table IV). The complete CAMELYON17 data set contained 1000 WSIs of H&E stained slides. The complete CAMELYON16 data set (training and test), was made available to give participants a good starting point for training algorithms. Altogether, 1399 WSIs were shared for the challenge (TableV).

TABLE III: pN-stages used in the challenge

pN-Stage	Slide Labels
pN0	No micro-metastases or macro-metastases or ITC found.
pN0(i+)	Only ITC found.
pN1mi	Micro-metastases found, but no macro-metastases found.
pN1	Metastases found in 1 – 3 lymph nodes, of which at least 1 is a macro-metastasis.
pN2	Metastases found in 4 - 9 lymph nodes, of which at least 1 is a macro-metastasis.

TABLE IV: Patient-level characteristics for the CAMELYON17 data set

Center	Total Patients		Stages (Train)				
	Train	Test	pN0	pN0(i+)	pN1mi	pN1	pN2
CWZ	20	20	4	3	5	7	1
RST	20	20	4	2	5	6	3
UMCU	20	20	8	2	4	3	3
RUMC	20	20	3	2	4	8	3
LPON	20	20	5	2	3	6	4
Total	100	100	24	11	21	30	14

TABLE V: WSI-level characteristics for the complete data set

Center	Total WSIs		Metastases (Train)			
	Train	Test	Negative	ITC	Micro	Macro
CWZ	100	100	64	11	10	15
RST	100	100	58	7	23	12
UMCU	250	100	165	2	34	49
RUMC	349	100	210	8	64	67
LPON	100	100	61	8	5	26
Total	899	500	558	36	136	169

D. Detailed lesion annotations

In addition to the patient and slide level labels, a pathologist exhaustively annotated 10 WSIs from each of the 5 centers in the CAMELYON17 training set by carefully outlining each lesion in the WSIs with polygons (TableVI). The cytokeratin-stained slides were used as a reference, when available.

TABLE VI: Exhaustive annotations in the CAMELYON17 data set

Center	Total WSIs	Metastases (Train)		
		ITC	Micro	Macro
CWZ	10	3	3	4
RST	10	2	5	3
UMCU	10	2	4	4
RUMC	10	4	3	3
LPON	10	5	2	3
Total	50	16	17	17

Additionally, the detailed annotations of the 159 WSIs with metastases of the complete CAMELYON16 data set were made available. The annotation polygons were shared as a series of pixel coordinates on the highest resolution level in XML file format.

III. Methods

A. Challenge setup

We set up a website to share information about the challenge and to provide an interface for all challenge-related issues. The website was set up via https://www.grand-challenge.org, which has hosted over 155 biomedical image analysis challenges since 2007. The challenge website is accessible directly at https://camelyon17.grand-challenge.org.

On the website the participants could register and find a general overview of the challenge including the deadlines, a brief description of the biomedical background of the problem, a description of the data set, the rules of the challenge, the evaluation metrics, and Python code snippets for accessing the images and the annotations. Finally, through the website the participants could submit their results and access a forum to ask questions and provide comments.

Participants were granted access to the data set, forum and submission system after they registered and accepted the rules of the challenge. Anonymous participation was not allowed. The complete data set was made available under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. The license is available at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode. The complete CAMELYON16 and CAMELYON17 data sets were shared on Google Drive. Since the access to the services of Google are limited in the People's Republic of China (PRC) we mirrored the content of the shared Google Drive to Baidu Pan which is a local service in the PRC and can be accessed without restrictions.

The challenge aimed for a fair comparison of algorithms, therefore participants were not allowed to use other data sources. Making extra annotations on the training data set was only allowed if the annotations were subsequently submitted to the organizers along with the submission of the results so that these annotations can be made available to other participants.

The participants had to submit their results as CSV files through the challenge website. The deadline for pre-workshop submissions was April 6, 2017. Maximum 3 submissions were allowed per participant with a 4 page ISBI style paper accompanying each submission describing their methods. The 3 submissions had to be methodologically different. Resubmissions with simple hyper-parameter tuning were not allowed.

During the workshop at ISBI 2017 we presented the results of the challenge and invited the top 5 teams to present their methods. The results, presentations and participant's algorithms were shared via the challenge website after the workshop. Subsequently, the challenge was reopened for registration and submissions.

B. Metrics and evaluation

Within the challenge, participants were scored based on the ability of their algorithm to identify the pN-stages of the 100 test patients. To evaluate the performance of the algorithms, we used Cohen's kappa with 5 classes and quadratic weights [14] which is a statistic that measures inter-observer agreement for categorical variables.

Given n test patients and m categories (pN-stages), let n_{ij} denote the number of patients with the i^{th} pN-stage that were categorized to the j^{th} pN-stage. Let r_i denote the total number of patients with the i^{th} pN-stage and s_j the total number of patents categorized to the j^{th} pN-stage. Finally, let w_{ij} denote the disagreement weight associated with the i^{th} and the j^{th} pN-stages.

The weight matrix is

$$w_{ij} = (i-j)^2, \quad i,j \in 1..m \tag{1}$$

The mean observed degree of disagreement is

$$D_o = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} w_{ij} \tag{2}$$

The mean degree of disagreement expected by chance is

$$D_e = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{m} r_i s_j w_{ij} \tag{3}$$

Weighted kappa is then defined by

$$\kappa_w = \frac{D_e - D_o}{D_e} \tag{4}$$

The κ_w metric ranges from -1 to $+1$: a negative value indicates lower than chance agreement, zero indicates exact chance agreement, and a positive value indicates better than chance agreement. As pN-stages are ordinal, a quadratic weighted kappa was chosen to penalize misclassification which are more than one stage apart more severely.

In this paper we also use confusion matrices at the slide level for the top 4 teams to assess accuracies for specific types of metastases. This will allow us to identify the most promising areas of improvement for the algorithms. Furthermore, we qualitatively inspected the likelihood maps provided by the best two contestants to assess localization performance and identify common false positives and negatives.

Last, we assessed whether combining algorithms could lead to even better performance than each algorithm individually. We combined the submitted pN-stages and also the reported slide-level labels of the best 2 up till the best 12 teams by averaging the labels and by majority voting. The new slide-level labels were converted to pN-stages by applying the TNM-criteria. In case of a tie in majority voting, the highest pN-stage or slide-level label was selected from the votes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2867350, IEEE Transactions on Medical Imaging

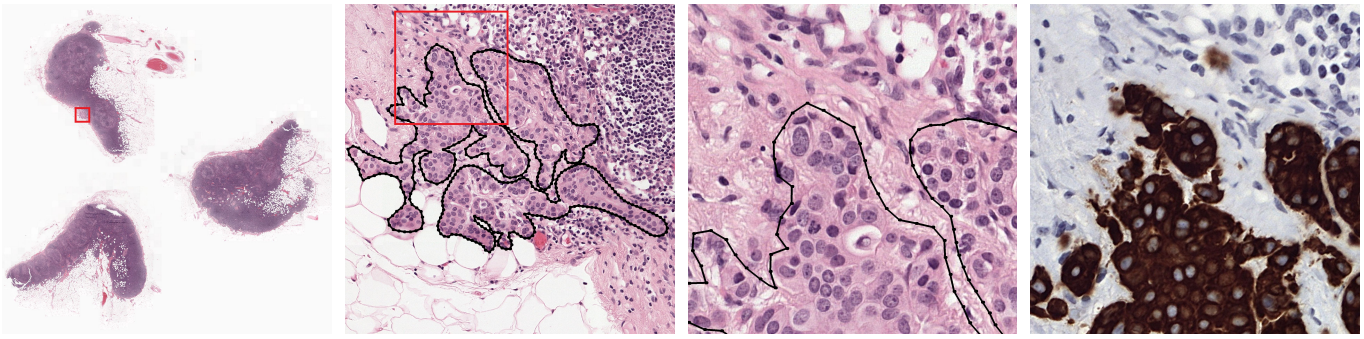

Fig. 1: Example of a WSI of a H&E stained section with a delineated micro-metastasis at increasing zoom levels, and the corresponding IHC (cytokeratin 8-18 stained) slide at the same location. The metastasis is outlined with black.

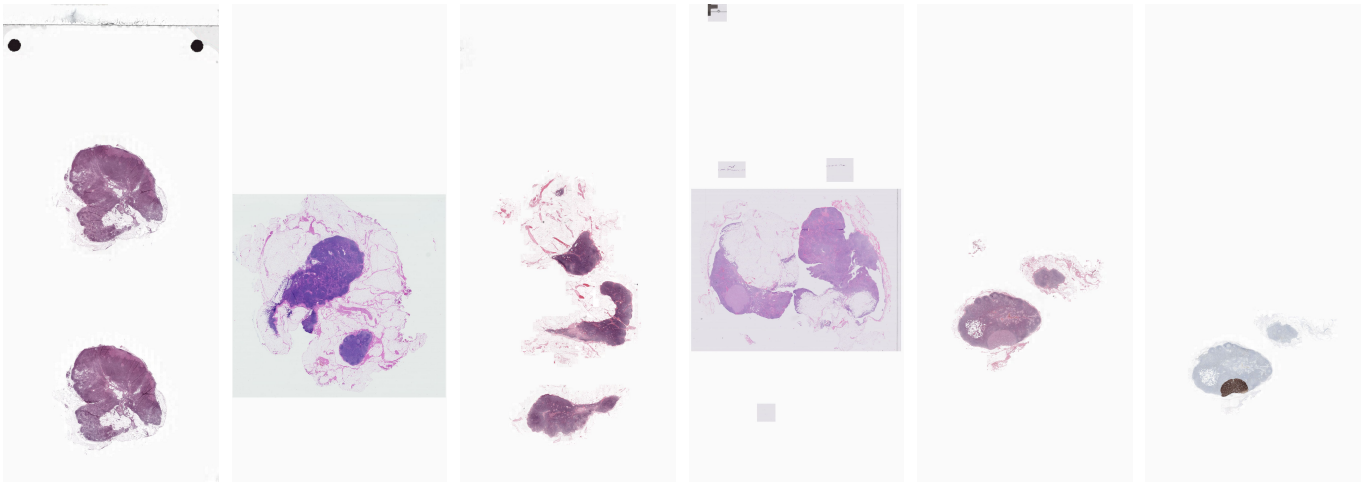

Fig. 2: Low-resolution examples of WSIs. One H&E stained slide from each medical center in the training set and the corresponding IHC (cytokeratin 8-18 stained) slide for the last H&E stained slide.

C. Summary of submitted algorithms

We had 300 registered participants before March 1, 2017 when the test data set was released and over a 1000 by the time of writing this article.

Altogether 23 teams submitted their results before the workshop deadline. To keep the paper concise we only present the methodology and results of the ten best performing algorithms. We also received four submissions after the challenge was reopened (but before 31st December 2017), of which one was a resubmission and one was excluded for not providing sufficient algorithmic detail. The other two post-workshop submissions were included. This resulted in a total of twelve algorithms which are presented in this paper.

All the twelve teams followed the same basic algorithmic steps: preprocessing, slide-level classification, slide-level postprocessing, and patient-level classification. We first give a brief summary and then cover each of the four steps in more detail.

In the preprocessing step all teams started with identifying the tissue regions on the WSIs. Typically, large parts of the slide do not contain tissue (Figure 2), and do not need to be processed. Therefore, the preprocessing step is essential for developing efficient algorithms. Subsequently, to perform metastases detection in each slide, all twelve teams trained convolutional neural networks architectures (CNN) with image tiles extracted from the identified tissue regions (normal and metastatic areas). The trained networks were then applied to the test images to obtain metastasis-likelihood maps. Within the postprocessing step, most participants thresholded the likelihood map and collected several features from the identified cancerous areas and used a separate classifier (e.g. random forest) to determine the class of the WSI: negative, ITC, micro, or macro. Last, the participants typically followed the pN-stage definitions to combine their slide-level findings into a patient-level pN-stage.

1) Preprocessing: All participants used a preprocessing step to identify tissue regions in the WSIs. All participants used simple filtering and thresholding algorithms, mostly Otsu's adaptive threshold at a low resolution level [15]. Differences between the methods were mainly found in which color space the thresholds were applied, for example RGB (red-green-blue), HSV (hue-saturation-value), or HSI (hue-saturation-intensity), and the type of morphological operations that were used to refine the thresholded image. For example, team 4 and 11 used a median filter to remove small regions, team 5 used connected component analysis and size filtering, and team 6 used morphological hole-filling. For a full listing we refer to Table VII.

TABLE VII: Differences in preprocessing and augmentation. G: Grayscale (mean value of RGB channels), CCA: Connected component analysis, Max/Min Δ: Threshold on the difference between maximum and minimum value across RGB channels.

Rank	Resolution Level	Color Space	Threshold Type	Morphological Operations
1	5	G	Value of G	-
2	5	G	Otsu's	-
3	6	G	Otsu's	-
4	6	HSV	Otsu's	Median filter,
5	8	G	Otsu's	Size filter using CCA
6	7	RGB	Max/Min Δ	Hole filling
7	6	HSI, H&E	I in HSI, E in H&E	Remove small obj., Closing
8	4	G	Otsu's	-
9	6	G	Otsu's	-
10	2	G	Value of G	-
11	0	HSV	Otsu's	Median filter, Size filter using CCA
12	4	G	Yen's [16]	Variance filter, Mean filter

2) Slide-level classification: Almost all participants used the CAMELYON16 WSIs, the 50 exhaustively annotated CAMELYON17 WSIs, and all the negative WSIs from the CAMELYON17 data set to develop their algorithms. Team 4, 7 and 12 used only the CAMELYON16 data set.

With respect to the different types of algorithms, all participants used CNNs. Specifically, they used variants of common network architectures: ResNet [17], GoogLeNet/Inception [18], VGG-Net [19], U-Net [20], and one team used DenseNet [21]. In contrast to CAMELYON16, none of the included twelve algorithms used a custom architecture. Team 2 and 4 used significantly adapted versions of the common architectures. Team 2 used a variant of ResNet-101 called DeepLab [22]. DeepLab employs convolution with dilated filters instead of downsampling (e.g. max-pooling) to increase the spatial resolution of the network when applied in a fully-convolutional fashion. Furthermore, in order to combat reduced localization accuracy due to inherent translational invariance in CNNs the architecture also uses conditional random fields (CRF). Team 4 used GoogLeNet in their ensemble to create texture representation by taking the location-wise outer product of the feature maps at the 'inc4d' layer. Subsequently, these are averaged across location to obtain a single feature vector. This vector is then fed into a softmax classifier. This approach is similar to that of the bilinear CNNs [23].

Five of the teams used model ensembles but only 2 teams, team 4 and 12 used fundamentally different networks in their ensembles. For example, team 4 used a combination of 2 GoogLeNets with different input patch sizes and a Resnet-50 architecture. The rest of the teams used instances of the same architecture with different initialization, parameters or patch augmentation settings. Eight of the teams used pre-trained networks for the challenge. They all used networks that were pre-trained on the ImageNet challenge [24], except team 2 who used a network that has been pre-trained on Microsoft COCO challenge [25].

All participants extracted small image patches of metastases and normal areas from the WSIs to train their CNNs, although the exact patch size and pixel resolution differed substantially. For the complete details of the network architectures and training parameters we refer to Table VIII.

In addition, almost all teams performed extensive data augmentation to increase the variation in the training set; only team 6 did not use any data augmentation. Random mirroring and rotations of 90°, 180°and 270°were the most popular augmentation strategies. Two teams applied rotations with angles sampled from the continuous [0°, 360°] interval instead. Other strategies included random cropping of patches, and applying affine transformations (e.g. scaling).

In addition, to make their CNNs robust to color variation caused by differences between scanners and/or staining protocols, most participants used patch color augmentation in the HSV, RGB or H&E color spaces by adding noise to the individual color channels. Some of the teams used additional brightness, contrast and gamma adjustments. Two teams took a completely different approach and tried to use stain normalization algorithms [26] to ensure a uniform color distribution across the images. For the complete details of the augmentation strategies we refer to Table IX.

3) Slide-level postprocessing: All participants used the trained networks to generate metastasis-likelihood maps for the WSIs. Team 3 used test time augmentations to generate the likelihood map. Test time augmentation refers to the practice of applying training augmentations to patches at test time to get multiple metastasis likelihoods per patch. Often these are then averaged to obtain the final likelihood for that patch, but team 3 used the most certain likelihood (i.e. closest to 1.0). To obtain the actual metastasis candidates most teams thresholded the likelihood maps and post-processed the resultant binary masks. A typical strategy, used for example by team 1 and 6 is to remove small detections to reduce the amount of false positives. Instead of thresholding, team 3 and team 12 used conditional random fields to assign pixel labels [27].

Assigning a slide-level label is trivial in case of perfect pixel level classification: a metastasis class can be assigned by measuring the largest detected area (Table II). Only two teams used this approach in CAMELYON17. As we already learned in the CAMELYON16 challenge, many algorithms submitted by participants suffer from high false positive rates [13]. The winner in CAMELYON16 solved this by extracting features from the binary detection mask and the likelihood map and feeding these features to a random forest classifier. This approach was replicated by several participants in CAMELYON17. Features that were typically used are, for example, the number of detected metastases, mean detection size and standard deviation, mean detection likelihood and standard deviation. Team 3 used a different approach by applying a more extensive rule-based system. To better separate between micro-metastases and ITC, they tried to calculate the number of cells via color deconvolution and thresholding on the hematoxylin channel. Subsequently, the DBSCAN algorithm was used to group together small metastases areas which were in close proximity [28].

Most teams determined the patient-level pN-stages by applying the rules according to the definition of pN-stages, except team 9 and 12. Team 9 combined the extracted features

TABLE VIII: Network architecture and training details. DO: Dropout, HNM: Hard Negative Mining

Rank	Architecture	Pre-trained	Ensemble (Size)	Batch Norm	Patch Sizes	Image Level	Batch Size	Iterations	Training Set Size	L2 Loss	DO	HNM
1	ResNet-101	ImageNet	x (3)	x	256×256	0	32	5×10^5	4.5×10^7	-	-	-
2	ResNet-101	COCO	-	x	960×960	1	10	2×10^4	1.6×10^6	x	-	x
3	GoogLeNet	ImageNet	-	x	299×299	1	32	9.5×10^4	3×10^6	x	x	x
4	GoogLeNet, ResNet-50	ImageNet	x (3)	-	256×256, 512×512	0	128	1×10^5	5×10^5	x	-	x
5	VGG	ImageNet	-	-	435×435	0	1	1.3×10^6	1.3×10^6	x	x	-
6	GoogLeNet	ImageNet	-	x	299×299	0	32	1×10^5	1.2×10^7	-	x	x
7	GoogLeNet	-	-	x	128×128	1	32	3.2×10^4	1×10^6	-	x	x
8	VGG	ImageNet	-	-	244×244	0	75	3×10^5	1.3×10^7	x	x	x
9	U-Net	-	-	-	512×512	2	20	1.5×10^6	1.3×10^6	-	-	x
10	U-Net	-	x (2)	-	256×256, 512×512	2	16	5.3×10^5	1.1×10^6	-	-	-
11	GoogLeNet	ImageNet	-	-	256×256	0	32	2.5×10^5	5×10^6	x	-	-
12	Dense U-Net, Densenet	-	x (3)	x	416×416	2	4 - 75	2.5×10^5	1.8×10^5	x	x	-

TABLE IX: Augmentation methods. AT: Affine transformations, AGN: Additive Gaussian noise, M90: multiples of 90°

Rank	Mirroring	Rotation	Color	Other
1	x	[0°, 360°]	HSV	contrast
2	x	-	RGB	brightness and contrast
3	x	M90	HSV	-
4	x	M90	-	cropping
5	x	M90	HSV	-
6	-	-	-	-
7	x	M90	H&E	gamma adjustment
8	x	[0°, 360°]	RGB	cropping
9	x	M90	stain norm.	-
10	x	M90	-	-
11	x	M90	stain norm.	brightness, zoom, AT, and AGN
12	x	M90	HSV	contrast

TABLE X: Likelihood map postprocessing, slide-level classification and pN-stage assignment. TH: Threshold, CRF: Conditional Random Field, RFC: Random Forest Classifier, RBS: Rule-based System, GBT: Gradient Boosted Trees, SVM: Support Vector Machine.

Rank	Likelihood Map Filtering	Binary Mask Generation	Slide-level Classifier	pN-Stage Assignment
1	-	1 TH	RFC	rule-based
2	-	2 THs	RFC	rule-based
3	upsampling	CRF	RBS	rule-based
4	-	3 THs	RFC	rule-based
5	-	3 THs	-	rule-based
6	-	1 TH	RFC	-
7	-	1 TH	RFC	rule-based
8	-	5 THs	RFC	rule-based
9	-	-	GBT	model-based
10	downsampling	1 TH	-	rule-based
11	Gaussian filtering	Otsu's	RFC	rule-based
12	morphological smoothing	CRF	SVM	regression

of all 5 slides per patient and used gradient boosted trees to determine the pN-stage of the patient directly. Team 12 on the other hand used a regression on slide-level prediction instead of direct rule based method to construct pN-stage from slide classifications.

Team 10 built a two stage binary decision tree to determine the metastases category on the individual slides. First they differentiated between negative and ITC; or micro and macro categories. Then they further divided the two sets into negative or ITC; and micro- or macro-metastases accordingly. At each step they used a different combination of the outputs of the 2 networks. For the complete details of the slide-level postprocessing we refer to Table X.

IV. Results

The metric used to rank the algorithms, the quadratic-weighted κ score, ranged from 0.8993 to -0.1341 for all 23 participating teams and from 0.8993 to 0.7330 for the methods included in this paper. As such, in terms of agreement, performance ranged from near-perfect agreement to worse-than-chance when including all participants. For a complete listing of the top 12 teams and their κ scores we refer to Table I.

Confusion matrices at the slide-level were also generated for the best 4 teams to inspect the quantitative results in more detail (Table XI). We can see that all teams performed well in identifying negative slides and slides containing macro-metastases. All teams performed poorly in identifying ITC, although the range in accuracy is quite large (0 – 34.3% correct). Teams 1 and 2 additionally performed well on slides containing micro-metastases, whereas team 3 and 4 performed significantly worse.

When combining the submissions of multiple teams a best κ score of 0.9261 was obtained by combining the slide-level classification of teams 1, 2 and 3 by averaging slide-level labels. This is 0.0268 higher than the single best team. The κ scores of the 5 best combinations are shown in Table XII. Focusing on the pN-stage classification specifically, the best single team assigned 76 out of 100 patients to the correct pN-stage, whereas the best combination got 77 out of 100 correct. Furthermore, the largest difference between the predicted pN-stage and the reference pN-stage was 3 stages for team 1 and only 2 stages for the best combination. Miss with larger than

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2867350, IEEE Transactions on Medical Imaging

TABLE XI: Slide-level confusion matrices of the best four teams with the accuracy indicated in percentages. BC is the best combination of algorithms. The cell colors range from white, representing low error rate to red, representing high error rate.

		Submitted																			
		Negative					ITC					Micro					Macro				
	Team	1	2	3	4	BC	1	2	3	4	BC	1	2	3	4	BC	1	2	3	4	BC
Reference	Negative	95.7	95.4	84.2	97.7	92.7	3.1	0.4	12.7	0.0	6.1	0.8	4.2	1.9	1.5	1.2	0.4	0.0	1.2	0.8	0.0
	ITC	51.4	68.6	62.8	94.3	51.4	11.4	14.3	34.3	0.0	40.0	37.2	17.1	2.9	5.7	8.6	0.0	0.0	0.0	0.0	0.0
	Micro	12.1	15.7	12.0	36.2	13.3	2.4	1.2	19.3	0.0	4.8	83.1	75.9	66.3	53.0	80.7	2.4	7.2	2.4	10.8	1.2
	Macro	3.3	1.6	0.8	4.1	0.0	0.0	0.0	2.5	0.0	4.1	5.7	11.5	6.5	7.4	4.9	91.0	86.9	90.2	88.5	91.0

1 difference occurred 5 times for the single best algorithm and only twice for the best combination. At slide-level the best combination performed equally on negative and macro-metastasis class slides, a few percentage points worse on micro-metastasis class slides but was almost 30 percentage points better in identifying ITC (Table XI).

TABLE XII: Kappa scores of different algorithm combination outputs

Teams	Type	Combination	Kappa Score
1 – 3	slide-level	mean	0.9261
1 – 3	slide-level	majority	0.9236
1 – 5	slide-level	majority	0.9226
1 – 3	patient-level	mean	0.9209
1 – 2	patient-level	mean	0.9175

Evaluation of the likelihood maps of team 1 and 2 provided insight in the performance of their algorithms, and clarified some of the false positives and false negatives. Examples of the likelihood maps are depicted in Figure 3. On the first row, a nerve is depicted that was identified by both teams as a metastasis. On the second row of Figure 3 an example of contamination is shown. The tissue sample was contaminated with a small piece of breast tissue during glass slide preparation. This contamination is not a metastasis but was picked up as such by both systems. Rows three and four show a macro- and micro-metastasis, respectively. The macro-metastasis was missed by team 1 and misclassified as a micro-metastasis by team 2. The micro-metastasis was missed by both teams. Both these metastases showed very diffuse infiltration of the healthy tissue, making it challenging to identify them. Last, the fifth row shows a micro-metastasis nicely segmented by both team 1 and 2. The detection of team 1 was a bit more precise since they correctly identified the extending arms on the top left and right side.

V. Discussion

Given that the participation requirements for CAMELYON17 were very high in terms of amounts of data that had to be processed within a limited time frame, both the quality and quantity of submissions was high. With 37 submissions, it was even slightly higher than for CAMELYON16, which had 32 submissions at the initial deadline.

The submitted algorithms were not only able to detect the presence of metastases but also measure their extent to derive the metastasis category, including ITC, and to determine the pN-stage that is used in clinical practice. Therefore, the outcome of CAMELYON17 more directly relates to clinical practice and the submitted algorithms can more readily be evaluated in that context.

A key observation is that the best performing algorithms do well on slides containing macro-metastases and metastasis-free slides. However, even the current best algorithm still performs very poorly on identifying ITC with only 11.4% accuracy. It has to be noted that ITC only play a very limited role in the pN-staging system and often are also missed by pathologists on H&E-stained slides [8]. These very small metastases can subsequently be picked up by using additional IHC staining.

The data set contained only 36 whole-slide images with ITCs of which only 16 were annotated. This could limit the performance of the algorithms detecting ITC. However, we think another reason might be also be important: to achieve high sensitivity on the small ITC lesions, one most likely needs to allow more small false positives in normal cases (i.e. it is harder to get rid of spurious detections automatically). For example, team 3, which used a rule-based system to obtain slide level classifications, was the best in ITC detection but at a cost of the highest false positive ratio in normal images.

The most important aspect of a well-performing system in terms of pN-staging and clinical relevance is its ability to detect macro- and micro-metastases. There are only relatively minor differences in the ability of the best systems to pick-up macro-metastases as the accuracies are within 5%. As such, most of the difference in the ranking is caused by the ability of the top two to identify micro-metastases much better than all other algorithms.

With regard to false positive detections, all algorithms still struggle with benign areas that occur rarely in the training set, for example the nerve shown in Figure 3 or contamination caused by tissue processing in the lab. Several teams tried to circumvent this by including hard-negative mining steps, but with limited success. Most likely this is caused by the fact that these benign areas are so rare that it is impossible to learn an accurate representation, even with the three terabytes of data in the CAMELYON17 data set. A potential avenue

to address this issue is by incorporating model uncertainty, for example via test-time dropout [29]. Another type of false positive which is hard to address is the contamination shown in row 2 of Figure 3. This can only be identified as a false positive detection when the global context of the slide is taken into consideration. As all competing algorithms use mostly local information (i.e. patches) to train their models, this can not be incorporated. An efficient strategy to add this global context to deep networks is interesting for further research.

We tried to identify the key characteristics in terms of methodology for the top performing algorithms. One important observation is that it is not possible to achieve competitive results using only a pre-trained GoogLeNet. Many groups tried this approach, modeled after the winner of CAMELYON16, but their results vary substantially. We also know from CAMELYON16 that pre-training in itself does not improve performance, but does offer the benefit of much faster convergence [30]. Especially in the context of a time-limited challenge, the reduced training time is beneficial. The fact that results vary substantially, even when using the same, pre-trained architecture indicates that the way the networks are trained or fine-tuned is more important than the architecture itself.

Observing the training processes used by the teams that are included in this paper, it can be concluded that the data being fed to the system is inherently important. All the participating teams extracted high-resolution patches from the WSIs. The best eight algorithms used either level 0 ($0.25\,\mu\mathrm{m}$ pixel size) or level 1 ($0.5\,\mu\mathrm{m}$ pixel size). The details that are available on high resolution levels are likely necessary for achieving good performance for this task. The amount of context included in the patches did vary greatly between teams. The smallest spatial area was 256×256 pixels at the highest resolution level, while the the largest spatial area corresponds to 1920×1920 pixels at that same level. Given the results, the context provided by 256×256 pixels at level 0 was enough for achieving good performance; larger contexts were not needed.

An interesting characteristic of the best performing algorithm is that it was trained for up to a magnitude more iterations than most of the other contenders. Only team 5, 9 and team 10 trained longer but they were either using a VGG architecture, which has roughly three times as many parameters as the ResNet-101 used by team 1, or U-net, which has not been pre-trained. Team 1 also used the largest number of patches in their training set of all contenders. These observations together might indicate that this network has learned from a more varied set of patches, which could explain why it generalizes best on the test set.

The majority of the teams focused on the most challenging patches using hard negative mining. This seemed to benefit performance overall as it was used by seven out of ten best teams, even though the best performing team did not use it. In CAMELYON17, the hard negative mining was used more widely than in CAMELYON16, where only two of the top-performing algorithms used it.

One of the other lessons learned from CAMELYON16 was that proper handling of stain variation between centers is key to good performance. In CAMELYON17 this is even more important as we now included 5 centers instead of 2. In CAMELYON16 the best performing team used color normalization to pre-process all the slides, whereas in this challenge most of the teams relied on heavy color augmentation to force their networks to be robust to color variation. In clinical practice such networks would be more desirable since they do not rely on a preprocessing step that could potentially fail.

A common question after every Grand Challenge in medical image analysis is whether the problem, in this case automatic identification of breast cancer metastases in sentinel lymph nodes, has been adequately solved. Up till now we can confidently state that this is not yet the case. Despite the excellent results of the participating teams the fact that a straightforward combination of the 3 top teams yields a 0.0268 better kappa score than the current best of 0.8993 shows that there is still room for improvement for the individual algorithms. Even more so when we take into account that even the best combination only classifies 77 out of 100 patients correctly. The errors are even worse at the slide level. The best ranked team misclassified 67 of the 500 slides in the test set. Overall 10 slides containing micro-metastases and 4 slides containing macro-metastases were classified as negative. That would be an unacceptable error in clinical practice. Although the poor performance on ITC is not immediately relevant from a clinical perspective, it could undermine the trust clinicians have in such algorithms. Improving the algorithm performance in this aspect is thus still worthwhile.

In terms of future work, the CAMELYON17 challenge will remain open for new submissions to allow improved algorithms to obtain better results. With respect to extending the scope of the challenge, adding the IHC stains as an extra layer of information is an option which would bring the challenge even closer to clinical practice. Alternatively, lymph nodes with metastases from other tumor entities, such as melanoma or colon cancer could be added. From a practical perspective, sharing 3 terabytes of data with participants all around the world has been challenging. Increasing the data size even further could render an expanded challenge impractical or impossible for many to participate in. A possible alternative would be to host the data at a single location and provide an environment to the participants that they could access remotely and where they could use the data to develop their algorithms without having to download it.

Summarizing, the algorithms competing in CAMELYON17 have proven that it is possible to automatically analyze histopathological WSIs in a clinically relevant setting, but can not yet be implemented without some form of supervision by a clinical expert. In their current state the algorithms could potentially effectively aid clinicians by pre-screening the WSIs. The pre-screened images could steer the attention of the pathologist to the relevant areas and ease the pN-staging by outlining metastases in advance.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017." *CA: a cancer journal for clinicians*, vol. 67, pp. 7–30, Jan. 2017.

[2] N. Howlader, A. M. Noone, M. Krapcho, D. Miller, K. Bishop, C. L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin, "SEER Cancer Statistics Review, 1975-2014, National Cancer Institute. Bethesda, MD,

http://seer.cancer.gov/csr/1975_2014/," based on November 2016 SEER data submission, posted to the SEER web site, April 2017. [Online]. Available: http://seer.cancer.gov/csr/1975_2014/

[3] L. H. Sobin, M. K. Gospodarowicz, and C. Wittekind, *TNM Classification of Malignant Tumours*, 7th ed. Wiley-Blackwell, 2011.

[4] A. E. Giuliano, K. K. Hunt, K. V. Ballman, P. D. Beitsch, P. W. Whitworth, P. W. Blumencranz, A. M. Leitch, S. Saha, L. M. McCall, and M. Morrow, "Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial." *JAMA*, vol. 305, pp. 569–575, Feb. 2011.

[5] A. E. Giuliano, K. V. Ballman, L. McCall, P. D. Beitsch, M. B. Brennan, P. R. Kelemen, D. W. Ollila, N. M. Hansen, P. W. Whitworth, P. W. Blumencranz, A. M. Leitch, S. Saha, K. K. Hunt, and M. Morrow, "Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis: The ACOSOG Z0011 (alliance) randomized clinical trial." *JAMA*, vol. 318, pp. 918–926, Sep. 2017.

[6] A. Chagpar, L. P. Middleton, A. A. Sahin, F. Meric-Bernstam, H. M. Kuerer, B. W. Feig, M. I. Ross, F. C. Ames, S. E. Singletary, T. A. Buchholz, V. Valero, and K. K. Hunt, "Clinical outcome of patients with lymph node-negative breast carcinoma who have sentinel lymph node micrometastases detected by immunohistochemistry," *Cancer*, vol. 103, pp. 1581–1586, 2005.

[7] J. Reed, M. Rosman, K. M. Verbanac, A. Mannie, Z. Cheng, and L. Tafra, "Prognostic implications of isolated tumor cells and micrometastases in sentinel nodes of patients with invasive breast cancer: 10-year analysis of patients enrolled in the prospective east carolina university/anne arundel medical center sentinel node multicenter study," *Journal of the American College of Surgeons*, vol. 208, pp. 333–340, 2009.

[8] P. J. van Diest, C. H. M. van Deurzen, and G. Cserni, "Pathology issues related to sn procedures and increased detection of micrometastases and isolated tumor cells." *Breast disease*, vol. 31, pp. 65–81, 2010.

[9] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Nature Scientific Reports*, vol. 6, p. 26286, 2016. [Online]. Available: http://dx.doi.org/10.1038/srep26286

[10] B. van Ginneken, T. Heimann, and M. Styner, "3D Segmentation in the Clinic: A Grand Challenge," in *3D Segmentation in the Clinic: A Grand Challenge*, 2007, pp. 7–15.

[11] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, agatay Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, pp. 1993–2024, 2015.

[12] A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. v. d. Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. v. d. Gugten, P. A. Heng, B. Jansen, M. M. J. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. M. C. Manders, A. Sora-Mengana, J. C. Garca-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. A. Zuidhof, B. v. Ginneken, and C. Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017. [Online]. Available: https://arxiv.org/abs/1612.08012

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

[14] J. L. e. a. Fleiss, "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619., 1973.

[15] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, pp. 62–66, 1979.

[16] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Trans. Image Process.*, vol. 4, pp. 370 – 378, 1995.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385*, 2015.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *arXiv:1512.00567*, 2015.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, 2015, pp. 234–241.

[21] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 640–651, Apr. 2017.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2017.

[23] T.-Y. Lin and S. Maji, "Visualizing and Understanding Deep Texture Representations," in *Computer Vision and Pattern Recognition*, 2016.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 1–42, 2014.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, 2014.

[26] B. Ehteshami Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Holler, A. Homeyer, N. Karssemeijer, and J. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 404–415, Sep 2016. [Online]. Available: http://dx.doi.org/10.1109/TMI.2015.2476509

[27] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *arXiv:1210.5644*, 2012.

[28] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 96, no. 34, 1996, pp. 226–231.

[29] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016, pp. 1050–1059.

[30] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe, "Detecting Cancer Metastases on Gigapixel Pathology Images," *arXiv:1703.02442*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2867350, IEEE Transactions on Medical Imaging

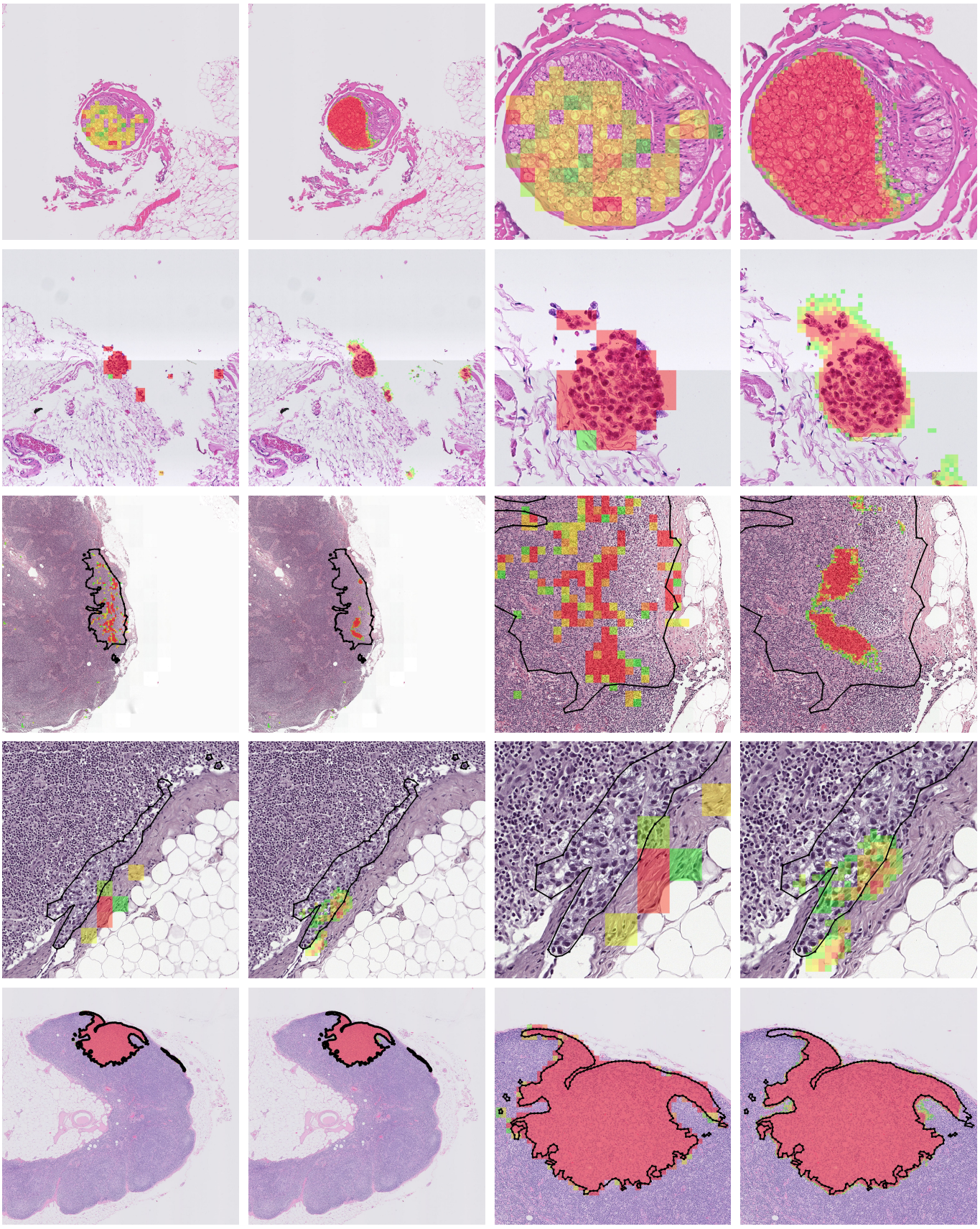

Fig. 3: Examples of likelihood maps. Columns: maps of teams 1 and 2 on low and high magnifications. Rows: nerve, contamination, missed macro-metastasis, missed micro-metastasis, identified micro-metastasis. The colors range from green to red, representing low to high probability respectively. The reference is annotated in black.