Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer

Babak Ehteshami Bejnordi, MS; Mitko Veta, PhD; Paul Johannes van Diest, MD, PhD; Bram van Ginneken, PhD; Nico Karssemeijer, PhD; Geert Litjens*, PhD; Jeroen AWM van der Laak*, PhD; The CAMELYON16 Consortium;

Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands (Ehteshami Bejnordi, van Ginneken, Karssemeijer). Medical Image Analysis Group, Eindhoven University of Technology, Eindhoven, The Netherlands (Veta). Department of Pathology, University Medical Center Utrecht, Utrecht, The Netherlands (van Diest). Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands (Litjens, van der Laak). Full lists of members and affiliations appear at the end of the paper.

Corresponding Author: Babak Ehteshami Bejnordi, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein 10, Nijmegen, The Netherlands, 6525 GA (Babak.EhteshamiBejnordi@radboudumc.nl).

*Authors contributed equally to the supervision of the study.

Revision data: 20 October 2017

Manuscript word count: 3967

Key points

Question: What is the discriminative accuracy of deep-learning-based algorithms compared with pathologists in detecting lymph node metastases in tissue sections of patients with breast cancer.

Finding: In cross sectional analyses that evaluated 32 algorithms submitted as part of a challenge, 7 deep learning algorithms showed greater discrimination than a panel of 11 pathologists, with an area under the curve of 0.994 (best algorithm) vs 0.884 (best pathologist).

Meaning: These findings suggest potential utility of deep learning algorithms for pathological diagnosis, but require assessment in a clinical setting.

IMPORTANCE: Application of deep learning techniques to whole-slide pathology images can potentially improve diagnostic accuracy and efficiency.

OBJECTIVE: Assess the performance of automated deep learning systems at detecting metastases in hematoxylin and eosin stained tissue sections of lymph nodes of patients with breast cancer and compare it to pathologists in a diagnostic setting.

DESIGN, SETTING AND PARTICIPANTS: Between November 2015 and November 2016, the CAMELYON16 challenge invited researchers to develop automated solutions for detecting lymph node metastases. A training set of whole-slide images (n=270) including 110 with and 160 without nodal metastases as verified by immunohistochemical staining, were provided to challenge participants to build their algorithms. Images were acquired at two different centers in the Netherlands using different staining protocols and whole-slide image scanners. Next, the performance of the algorithms was evaluated in an independent test set of 129 whole-slide images (49 with and 80 without tumor). The test set was also evaluated by 11 pathologists from the Netherlands, asked to ascertain the likelihood of nodal metastases for each image in a 2 hour session simulating routine pathology workflow and by 1 pathologist without time constraint.

EXPOSURE: Deep learning algorithms submitted as part of a challenge, a panel of 11 pathologists in a simulated diagnostic setting (working under restricted time) and one pathologist without time constraint.

MAIN OUTCOMES AND MEASURES: Outcomes were 1) the presence of specific metastatic foci; and 2) the absence versus presence of lymph node metastasis in a slide. Evaluation was performed using receiver operating characteristic curve analysis. The 11 pathologists participating in the simulation exercise rated their diagnostic confidence as: definitely normal, probably normal, equivocal, probably tumor or definitely tumor.

RESULTS: The AUC for the algorithms ranged from 0.556 to 0.994. The top-performing algorithm achieved a lesion-level true positive fraction comparable to that of the pathologist without time constraint (72.4%; 95% CI, 64.3%-80.4%) at a mean of 0.0125 false positives per normal whole-slide image. At the slide level, the best algorithm (AUC = 0.994; 95% CI, 0.983-0.999) performed significantly better than pathologists in a time constrained diagnostic simulation (mean AUC = 0.810; range, 0.738-0.884), p<0.001. The top 5 algorithms had mean AUC that was comparable to the pathologist interpreting the slides in the absence of time constraints (mean AUC = 0.960; range, 0.923-0.994 versus AUC = 0.966; 95% CI, 0.927-0.998).

CONCLUSIONS AND RELEVANCE: In the setting of a competition, some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; performance was comparable to an expert pathologist in the absence of time constraints. Whether this approach has clinical utility will require evaluation in a clinical setting.

INTRODUCTION

Full digitalization of the microscopic evaluation of stained tissue sections in histopathology has become feasible in recent years because of advances in slide scanning technology and cost reduction in digital storage. Advantages of digital pathology include: remote diagnostics, immediate availability of archival cases, and easier consultations with expert pathologists[1]. Also, the possibility for computer-aided diagnostics may be advantageous[2].

Computerized analysis based on deep learning (eText in the supplement) has shown potential benefits as a diagnostic strategy. Gulshan et al.[3] and Esteva et al.[4] demonstrated the potential of deep learning for diabetic retinopathy screening and skin lesion classification, respectively. Analysis of pathology slides is also an important application of deep learning, but requires evaluation for diagnostic performance.

Accurate breast cancer staging is an essential task performed by pathologists worldwide to inform clinical management. Assessing the extent of cancer spread by histopathological analysis of sentinel lymph nodes (SLN) is an important part of breast cancer staging. The sensitivity of SLN assessment by pathologists, however, is not optimal. A retrospective study showed that pathology review by experts changed the nodal status in 24% of patients[6]. Furthermore, SLN assessment is tedious and time-consuming. It has been shown that deep learning algorithms could identify metastases in SLN slides with 100% sensitivity while 40% of the slides without metastases could be identified as such[5]. This could result in a significant reduction in pathologists' workload.

The aim of this study was to investigate the potential of machine learning algorithms for detection of metastases in SLN slides and compare these to the performance of pathologists. To this end, the CAMELYON16 (CAncer MEtastases in LYmph nOdes challeNge 2016) competition was organized. Research groups around the world were invited to produce an automated solution for breast cancer metastases detection in SLN. Once developed, the performance of each algorithm was compared to the performance of a panel of 11 pathologists participating in a simulation exercise intended to mimic pathology workflow.

METHODS

Image datasets

To enable the development of diagnostic machine learning algorithms, we collected 399 whole-slide images of sentinel axillary lymph nodes (SLN) during the first half of 2015. SLNs were retrospectively sampled from 399 patients that underwent surgery for breast cancer at one of two hospitals in the Netherlands. The need for informed consent was waived by the institutional review board (research ethics committee of Radboudumc; file number 2016-2761). Whole-slide images where anonymized before making them available. All pathologists participating in this study were informed of, and agreed with the rationale and goals of this study. The participation of the pathologists was voluntarily and in accordance with the applicable Dutch rules concerning the review of research ethics committees and informed consent. In addition, the need to obtain informed consent from the panel of pathologists who participated in the study was waived by the research ethics committee. To enable assessment of the performance of the algorithms for slides containing micro and macrometastases as well as for negative slides, stratified random sampling was performed on the basis of the original pathology reports.

Whole-slide images were acquired at two different centers — Radboud University Medical Center (RUMC), and University Medical Center Utrecht (UMCU) — using two different scanners. RUMC whole-slide images were produced with a 3DHISTECH Pannoramic 250 Flash II digital slide scanner with a 20X objective lens (specimen level pixel size 0.243µm×0.243µm). UMCU whole-slide images were produced using the Hamamatsu XR C12000 digital slide scanner with a 40X objective lens (specimen level pixel size 0.226µm×0.226µm).

Reference standard

All metastases present in the slides were annotated under the supervision of expert pathologists. The annotations were first manually drawn by two students (one for each hospital) and then every slide was checked in detail by one of the two pathologists (PB and PvD for RUMC and UMCU, respectively; eFigure 1 in the supplement). In clinical practice, pathologists may opt to use immunohistochemistry (IHC) to resolve diagnostic uncertainty. In this study, obvious metastases were annotated without the use of IHC whereas, for all difficult cases and all cases appearing negative on H&E, IHC (anti-cytokeratin mouse monoclonal antibody, clone CAM 5.2, BD Biociences, San Jose, USA) was used (eFigure 2 in the supplement). This minimizes false negative interpretations. IHC is the most accurate method for metastasis evaluation and has little interpretation variability[7-9].

Pathologists differentiate between macrometastases (tumor cell cluster diameter>2 mm), micrometastases (tumor cell cluster diameter between 0.2 and 2 mm) and isolated tumor cells (solitary tumor cells or tumor cell clusters with diameter ≤ 0.2 mm and/or less than 200 cells). The largest available metastasis determines the slide-based diagnosis. Because the clinical value of having only isolated tumor cells in an SLN is disputed, we did not include such slides in our study and also did not penalize missing isolated tumor cells in slides containing micro- or macrometastases. Isolated tumor cells were, however, annotated in slides containing micro and macrometastases by the pathologists and included in the training whole-slide images. The set of images was randomly divided into a training (n = 270) and a test set (n = 129; details in **Table 1**). Both sets included slides with both micro and macrometastatic tumor foci as encountered in routine pathology practice.

Coding challenge

In the first (training) stage of the CAMELYON16 challenge, participants were given access to 270 whole-slide images of digitally scanned tissue sections. Each SLN metastasis in these images was annotated enabling participants to build their algorithms. In the evaluation stage, the performance of the participants' algorithms was tested on a second set of whole-slide images lacking annotation of SLN metastases. The output of each algorithm was sent to the challenge organizers by the participants for independent evaluation. Each team was allowed to make a maximum of three submissions. Multiple submissions were only allowed if the methodology of the new submission was distinct.

Tasks and evaluation metrics

Two tasks were defined: 1. identification of individual metastases in whole-slide images, and 2. classification of every whole-slide image as either containing or lacking SLN metastases. The tasks had different evaluation metrics and consequently resulted in two independent algorithm rankings. In task 1, algorithms were evaluated for their ability to identify specific metastatic foci in a whole-slide image. Challenge participants provided a list of potential metastasis locations with accompanying confidence scores in the range from 0 to 1. Algorithms were compared using a measure derived from the free-response receiver operator characteristic curve (FROC)[10]. The FROC curve shows the lesion-level true positive fraction versus the mean number of false positive detections in metastasis-free slides only. The FROC true positive fraction score that ranked teams in the first task was defined as the mean true positive fraction at 6 predefined false positive rates: 1/4, 1/2, 1, 2, 4, and 8 FPs per whole-slide image. Details on detection criteria for individual lesions can be found in the eMethods in the Supplement.

Task 2 evaluated the algorithms' ability to discriminate between 49 slides with and 80 without SLN metastases. In this case, identification of specific foci within images was not required. Participants provided a confidence, using the same rating schema as for task 1, indicating the probability that each whole-slide image contained any evidence of SLN metastasis from breast cancer. The area under the receiver operating characteristic curve (AUC) was used to compare the performance of the algorithms.

Performance of human experts

To establish a baseline for human expert performance, two experiments were conducted using the 129 slides in the test set, corresponding to the tasks defined above. In the first experiment, one expert pathologist (MCRFvD, >10 years of experience in pathology diagnostics, >2 years of experience in assessing digitized tissue sections) marked every single metastasis on a computer screen using high magnification. This task was performed without any time constraint. For comparison with the algorithms on task 2, the pathologist without time constraint indicated (during the same session) the locations of any (micro or macro) metastases per whole-slide image.

The setup without time constraint does not yield a fair measure of the accuracy of the routine diagnostic process. Preliminary experiments with four independent pathologists determined that two hours was a realistic amount of time for reviewing these 129 whole-slide images. We, therefore, asked 11 pathologists to independently assess the 129 glass slides in the test set in a simulation exercise designed to mimic routine diagnostic pathology workflow: a time limit of two hours was set but exceeding this limit was not penalized and every pathologist was allowed time to finish the entire set. The panel of the 11 pathologists (mean age 47.7 years; range 31-61 years) included one resident pathologist (third year of residency) and ten practicing pathologists (average number of years practicing 16.4; range 0-30; 0 is for one pathologist who just finished a 5 year residency program). Three of these pathologists had breast pathology as a special interest area.

The panel of 11 pathologists assessed the glass slides using a conventional light microscope and determined whether there was or was not any evidence of SLN metastasis in each image. This diagnostic task was identical to and used the same images as those evaluated by the algorithms in task 2. Pathologists indicated the level of confidence in their interpretation for each slide using five levels: definitely normal, probably normal, equivocal, probably tumor, definitely tumor. To obtain an empirical ROC curve, the threshold was varied to cover the entire range of possible ratings by the pathologists, and the sensitivity was plotted as a function of the false positive fraction (1 - specificity). To get estimates of sensitivity and specificity for each pathologist, the five levels of confidence were dichotomized by considering the confidence levels of ‘definitely normal’ and ‘probably normal’ as negative and all other levels as positive.

Statistical analysis. All statistical tests used in this study were two-sided considering a p-value < 0.05 as significant.

For task 1 and 2, confidence intervals of the FROC true positive fraction scores and areas under the ROC curve (AUC) were obtained using the percentile bootstrap method[11] for the algorithms, the panel of 11 pathologists, and the pathologist without time-constraint. The AUC values for the pathologists were calculated based on their provided five-point, ordinal scores.

For comparison of the AUC of the individual algorithms against the panel of 11 pathologists in task 2, we used multiple reader multiple case ROC (MRMC) analysis. The MRMC paradigm is frequently used for evaluating the performance of medical image interpretation and allows comparison of multiple readers reading the same cases while accounting for the different components of variance contributing to the interpretations[12,13]. Both the panel of readers and the algorithms as well as the cases were treated as random effects in this analysis. The panel of 11 pathologists represent the group of readers for modality one (diagnosing on glass slides) and an algorithm represents the reader for modality two (whole-slide images). Cases are the same set of slides/images seen by the panel and the algorithm. The AUC was the quantitative measure of performance in this analysis. The Dorfman-Berbaum-Metz significance testing, with Hillis improvements[14] was performed to test the null hypothesis that all effects are zero. The Bonferroni correction was used to adjust the p-values for multiple comparisons in the MRMC analysis (independent comparison of each of the 32 algorithms and the panel of pathologists).

Additionally, a permutation test[15] was performed to assess whether there was a statistically significant difference between the area under the ROC curve of the pathologists detecting macro-metastases compared to micro-metastases[16]. This test was also repeated for comparing the performance of pathologists for different histotypes: infiltrating ductal cancer versus all other histotypes. As the controls (slides not containing metastases) were the same in both groups, the permutation was only performed across the slides containing metastases. This test was performed for each individual pathologist and subsequently Bonferroni correction was applied to the obtained p-values.

No prior data was available for the performance of algorithms in this task. Therefore, no power analysis was used to predetermine the sample size.

iMRM 3.2 application developed by the Food and Drug Administration was used for MRMC analysis (available at https://github.com/DIDSR/iMRMC). An in-house developed script in Python 2.7 was used to obtain the percentile bootstrap CIs for the FROC and AUC scores (available at https://github.com/computationalpathologygroup/CAMELYON16). A custom script was written to perform the permutation tests and can be found at the same location.

RESULTS

The expert pathologist without time constraint required approximately 30 hours for assessing 129 whole-slide images. She did not produce any false positives in task 1 (i.e. non-tumorous tissue indicated as metastasis) but failed to identify 27.6% of individual metastases (lesion-level true positive fraction of 72.4% (95% CI, 64.3%-80.4%)) that were manifest when immunohistochemical staining was performed. At the slide level (task 2), the pathologist achieved a sensitivity of 93.8% (95% CI, 86.9%-100.0%), a specificity of 98.7% (95% CI, 96.0%-100.0%), and an AUC of 0.966 (95% CI, 0.927-0.998). The 11 pathologists in the simulation exercise spent a median of 120 minutes for 129 slides (range, 72-180 minutes). They achieved a mean sensitivity of 62.8% (95% CI, 58.9%-71.9%) with a mean specificity of 98.5% (95% CI, 97.9%-99.1%). The mean AUC was 0.810 (range, 0.738-0.884) (eTable 1 and eTable 2 in the Supplement show results for individual pathologists). eFigure 3 in the supplement shows the ROC curves for each of the 11 pathologists and the pathologist assessing the slides without time constraint.

The pathologists' results were further analyzed for their ability to detect micro versus macrometastases (see eResults in the supplement). It was found that the panel of 11 pathologists had a mean sensitivity of 92.9% (95% CI, 90.5%-95.8%) and mean AUC of 0.964 (range, 0.924-1.0) for detecting macrometastases compared to sensitivity of 38.3% (95% CI, 32.6%-52.9%) and AUC of 0.685 (range, 0.582-0.808) for micrometastases. Even the best performing pathologist in the panel missed 37.1% of the cases with only micrometastases.

Algorithm performance

Between November 2015 and November 2016, 390 research teams signed up for the challenge. Twenty-three teams submitted 32 methods for evaluation by the closing date (details in eTable 3 and eMethods in the Supplement). The majority of entries (25/32) were based on deep convolutional neural networks (eText in the Supplement). Besides deep learning, a variety of other approaches were attempted by CAMELYON16 participants. Different statistical and structural texture features were extracted (e.g. color SIFT features[17], local binary patterns[18], features based on gray-level co-occurrence matrix[19], etc.) combined with widely used supervised classifiers (e.g. support vector machines[20], random forest classifiers[21]). The performance and ranking of the entries for the two tasks are shown in **Table 2**. Overall, deep learning-based methods performed significantly better than other methods: the top-performing 19 entries in both tasks all used deep convolutional neural networks as the underlying methodology (see Table 2). Detailed method description for the participating teams can be found in eMethods in the Supplement.

Task1: Metastasis identification

The results of metastasis identification, as measured by the FROC true positive fraction score, are presented in Table 2 (eTable 4 in the supplement provides a more detailed summary of the results for the FROC analysis). The best algorithm, from team HMS & MIT (II), achieved an overall FROC true positive fraction score of 0.807 (95% CI, 0.732 - 0.889). The algorithm by team HMS & MGH (III) ranked 2nd in this task with an overall score of 0.760 (95% CI, 0.692 - 0.857). **Figure 1** presents the FROC curves for the top-five performing systems in the first task (FROC curves of all methods in eFigure 4 in the supplement). **Figure 2** shows several examples of metastases in the test set of CAMELYON16 and the probability maps produced by the top-three ranked algorithms (see eFigure 5 in the supplement).

Task 2: Whole-slide image classification

The results for all automated systems, sorted by their performance, are presented in **Table 2**. **Figures 3a** and **3b** show the ROC curves of the top-five ranked teams along with the operating points of the pathologists. eFigure 6 in the supplement shows the ROC curves for all methods. All 32 algorithms were compared to the panel of pathologists using MRMC analysis (see **Table 2**).

The top-performing system by team HMS & MIT (II) was a GoogLeNet model[22], which outperformed all other CAMELYON16 submissions with an AUC of 0.994 (95% CI, 0.983-0.999). This AUC exceeded the mean performance of the 11 pathologists (mean AUC of 0.810 (range, 0.738-0.884)) in the diagnostic simulation exercise (*p*<0.001, calculated using MRMC analysis[23], see Table 2). The top performing algorithm had an AUC comparable to that of the pathologist without time constraint (AUC = 0.966 (95% CI, 0.927-0.998)). Additionally, the operating points of all 11 pathologists were below the ROC curve of this method (**Figure 3a** and **3b**). The ROC curves for the two leading algorithms, the pathologist without time constraint, the mean ROC curve over the panel of 11 pathologists, and the panel participants with the highest and lowest AUCs are shown in **Figures 3c** and **3d**.

The second place algorithm by team HMS & MGH (III) used a fully convolutional ResNet-101[24] model. This system achieved an overall AUC of 0.976 (95% CI, 0.941-0.999), and yielded the highest AUC in detecting macrometastases (AUC 1.0). An earlier submission by this team, HMS & MGH (I), achieved an overall AUC of 0.964 (95% CI, 0.928-0.989) and ranked third. The fourth highest-ranked team was CULab (III) with a 16-layer VGG-net[25], followed by the first submission of the team HMS & MIT (I) with a 22-layer GoogLeNet. Overall, 7 of the 32 submitted algorithms had significantly higher AUCs than the panel of 11 pathologists (see **Table 2** for the individual p-values calculated using MRMC analysis).

The algorithms' results were further analyzed for comparing their performance in detecting micro and macrometastases (see eResults and eTable5 in the supplement). The top-performing algorithms performed similarly to the best performing pathologists in detecting macrometastases. Ten of the top-performing algorithms achieved better AUCs in detecting micrometastases than the best pathologist in the panel of 11 (mean algorithm AUC = 0.885 with range 0.812-0.997 versus best pathologist AUC = 0.808 (95% CI, 0.704-0.908)).

DISCUSSION

The CAMELYON16 challenge demonstrated that some deep learning algorithms were able to achieve better area under the ROC curve (AUC) than a panel of 11 pathologists participating in a simulation exercise for detection of lymph node metastases of breast cancer. To our knowledge, this is the first study that shows that interpretation of pathology images can be performed by deep learning based algorithms at an accuracy level that rivals human performance.

To obtain an upper limit on what level of performance could be achieved by visual assessment of H&E-stained tissue sections, a single expert pathologist evaluated whole-slide images at high magnification in details and marked every single cluster of tumor cells. This took the pathologist 30 hours for 129 slides, which is infeasible in clinical practice. Although this pathologist was very good at differentiating metastases from false positives, 27.6% of metastases were missed compared to the reference standard obtained with the use of immunohistochemical staining to confirm the presence of tumor cells in cases where interpretation of slides was not clear-cut. This illustrates the relatively high probability of overlooking tumor cells in H&E-stained tissue sections. At the slide level, a high overall sensitivity and specificity for the expert pathologist analyzing each case without time constraint was observed.

To estimate the accuracy of pathologists in a routine diagnostic setting, 11 pathologists assessed the SLNs in a simulated exercise. The setting resembled diagnostic practice in The Netherlands, where use of IHC is mandatory for cases found negative on H&E reading. Compared to the expert pathologist interpreting the slides without time constraint, these pathologists were less accurate, especially on the slides which only contained micrometastases. Even the best performing pathologist on the panel missed over 37% of the cases with only micrometastases. Macrometastases were much less often missed. Specificity remained high, indicating that the task did not lead to a high rate of false positives.

The best algorithm achieved similar true positive fraction as the pathologist without time constraint when producing a mean of 1.25 false positive lesions in 100 whole-slide images and performed better when allowing for slightly more false positives. On the slide level, the leading algorithms performed better than the 11 pathologists in the simulation exercise.

All of the 32 algorithms submitted to CAMELYON16 used a discriminative learning approach to identify metastases in whole-slide images. The common denominator for the algorithms in the higher echelons of the ranking was that they used advanced convolutional neural networks. Algorithms based on manually engineered features performed less well.

Despite the use of advanced convolutional neural network architectures such as 16-layer VGG-Net[25], 22-layer GoogLeNet[22], and 101-layer ResNet[24], the ranking among teams using these techniques varied significantly, ranging from 1st to 29th. However, auxiliary strategies to improve system generalization and performance seemed more important. For example, team HMS & MIT improved their AUC in task 2 from 0.923 to 0.994 by adding a standardization technique[26] to help them deal with stain variations. Other strategies include exploiting invariances to augment training data (e.g. tissue specimens are rotation invariant), and addressing class imbalance (i.e. more normal tissue than metastases) by different training data sampling strategies (eDiscussion contains further examples of properties that distinguish the best-performing methods).

Previous studies on diagnostic imaging tasks in which deep learning reached human-level performance such as detection of diabetic retinopathy in retinal fundus photographs used reference standard based on the consensus of human experts[3]. This study, in comparison, generated reference standard using additional immunohistochemical staining, yielding an independent reference against which human pathologists could also be compared.

Limitations

This study has a number of limitations. The test dataset on which algorithms and pathologists were evaluated was enriched with cases containing metastases and, specifically, micro-metastases and thus is not directly comparable to the mix of cases pathologists encounter in clinical practice. Given the reality that most sentinel lymph nodes do not contain metastases, the dataset curation was needed to achieve a well-rounded representation of what is encountered in clinical practice without including an exorbitant number of slides. To validate the performance of machine learning algorithms, such as those developed in the Camelyon16 challenge, a prospective study is required. In addition, algorithms were specifically trained to discriminate between normal and cancerous tissue in the background of lymph node histological architecture, but might be unable to identify rare events such as co-occurring pathologies (e.g. lymphoma) or breast tissue. In addition, algorithm run-time was not included as a factor in the evaluation but might be relevant in, for example, frozen section analysis.

In this study, every pathologist was given a single H&E-stained slide per patient to determine the presence or absence of breast cancer metastasis. In a real clinical setting, multiple sections are evaluated for every lymph node, and typically multiple levels for each section are available. Also, in most hospitals pathologists request additional IHC staining in equivocal cases. Especially for slides containing only micro-metastases, this is a relevant factor affecting diagnostic performance. The detection of other pathologies in the SLN (e.g. lymphoma) which is relevant in routine diagnostics was not included in the present study.

In addition, the simulation exercise invited pathologists to review 129 H&E stained slides in about two hours to determine the presence of macroscopic or microscopic SLN metastasis. Although feasible in the context of this simulation, this may not represent the work pace in other settings. Less time constraint on task completion may increase the accuracy of SLN diagnostic review. In addition, pathologists may rely on IHC staining and the knowledge that all negative H&E slides will undergo review.

Conclusions

In the setting of a challenge competition, some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; performance was comparable to an expert pathologist in the absence of time constraints. Whether this approach has clinical utility will require evaluation in a clinical setting.

ACKNOWLEDGMENTS

The authors are grateful to the organizing committee of the 2016 IEEE International Symposium on Biomedical Imaging (ISBI) for hosting the workshop held as part of the study reported in this article. They also acknowledge their collaborators and financial support from their funding agencies.

Funding/Support: Data collection and annotation were funded by Stichting IT Projecten (STITPRO) Nijmegen, The Netherlands and by the Fonds Economische Structuurversterking (tEPIS/TRAIT project; LSH-FES Program 2009; DFES1029161 and FES1103JJT8U). The latter also supported (in kind) web-access to WSI. BEB was financially support by the European Union FP7 funded VPH-PRISM project under grant agreement 601040.

Reproducible Research Statement: The image data used for CAMELYON16 training and testing sets along with the lesion annotations are publicly available at (https://camelyon16.grand-challenge.org/download/). Because of the large size of the dataset, multiple options are provided for accessing/downloading the data. Python and Matlab codes used for performing evaluations of the performance of the algorithms are publicly available at (https://github.com/computationalpathologygroup/CAMELYON16).

Access to Data Statement: Babak Ehteshami Bejnordi had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Mr. Ehteshami Bejnordi reports personal fees from European Union, during the conduct of the study. Dr. Veta reports grants from Netherlands Organization for Scientific Research (NWO), outside the submitted work. Dr. van Ginneken reports other from Thirona, grants and other from Mevis Medical Solutions, outside the submitted work. Dr. Karssemeijer reports other from Volpara Solutions Ltd (Wellington, New Zealand), personal fees and other from QView Medical inc. (Los Altos, CA), personal fees and other from ScreenPoint Medical BV (Nijmegen, Netherlands), outside the submitted work. Dr. van der Laak reports grants from Stichting IT projecten Nijmegen, The Netherlands, non-financial support from Fonds Economische

Structuurversterking, grants from European Union, during the conduct of the study; personal fees from Philips, personal fees from ContextVision, personal fees from Diagnostic Services Manitoba, Winnipeg, Canada, outside the submitted work. Dr. Manson and Mr. Geessink report grants from Dutch Cancer Society, outside the submitted work. Dr. Beck, Dr. Beca, Dr. Khosla, and Dr. Wang report personal fees from PathAI inc., outside the submitted work. Dr. Beca also reports non-financial support from Nvidia Corp., other from Nvidia Corp., outside the submitted work. Dr. Ruusuvuori reports grants from Finnish Funding Agency for Innovation (269/31/2015), during the conduct of the study.

Role of the Funder/Sponsor: The funders and sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The CAMELYON16 Consortium:

Meyke Hermsen, BS; Quirine F Manson, MD, MS; Maschenka Balkenhol, MD, MS; Oscar Geessink, MS; Nikolaos Stathonikos, MS; Marcory CRF van Dijk, MD, PhD; Peter Bult, MD, PhD; Francisco Beca, MD, MS; Andrew H Beck, MD, PhD; Dayong Wang, PhD; Aditya Khosla, PhD; Rishab Gargeya; Humayun Irshad, PhD; Aoxiao Zhong, BS; Qi Dou, MS; Quanzheng Li, PhD; Hao Chen, PhD; Huang-Jing Lin, MS; Pheng-Ann Heng, PhD; Christian Haß, MS; Elia Bruni, PhD; Quincy Wong, BS, MBA; Ugur Halici, PhD; Mustafa Ümit Öner, MS; Rengul Cetin Atalay, MD; Matt Berseth, MS; Vitali Khvatkov, MS; Alexei Vylegzhanin, MS; Oren Kraus, MS; Muhammad Shaban, MS; Nasir Rajpoot, PhD; Ruqayya Awan, MS; Korsuk Sirinukunwattana, PhD; Talha Qaiser, BS; Yee-Wah Tsang, MD; David Tellez, MS; Jonas Annuscheit, BS; Peter Hufnagl, PhD; Mira Valkonen, MS; Kimmo Kartasalo, MS; Leena Latonen, PhD; Pekka Ruusuvuori, PhD; Kaisa Liimatainen, MS; Shadi Albarqouni, PhD; Bharti Mungal, MS; Ami George, MS; Stefanie Demirci, PhD; Nassir Navab, PhD; Seiryo Watanabe, MS; Shigeto Seno, PhD; Yoichi Takenaka, PhD; Hideo Matsuda, PhD; Hady Ahmady Phoulady, PhD; Vassili Kovalev, PhD; Alexander Kalinovsky, MS; Vitali Liauchuk, MS; Gloria Bueno, PhD; M. Milagro Fernandez-Carrobles, PhD; Ismael Serrano, PhD; Oscar Deniz, PhD; Daniel Racoceanu, PhD; Rui Venâncio, MS;

Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands (Hermsen, Balkenhol, Geessink, Bult, Tellez). Department of Pathology, University Medical Center Utrecht, Utrecht, The Netherlands (Manson, Stathonikos). Laboratorium Pathologie Oost Nederland, Hengelo, Netherlands (Geessink). PathAI inc., Cambridge, MA, USA (Beck, Wang, Khosla). BeckLab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA (Beca, Beck, Wang, Irshad). Rijnstate Hospital, Arnhem, The Netherlands (van Dijk). Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA (Khosla). The Harker School, San Jose, CA, USA (Gargeya). Center for Clinical Data Science, Gordon center for medical imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA (Zhong, Dou, Li). The Chinese University of Hong Kong, Hong Kong (Lin, Chen, Dou, Heng). ExB Research and Development GmbH, Germany (Haß, Bruni). Independent participant (Wong). Department of Electrical and Electronics Engineering, Middle East Technical University, Turkey (Halici, Öner). GSNAS Neuroscience and Neurotechnology, Middle East Technical University, Turkey (Halici). CanSyL, Graduate School of Informatics, Middle East Technical University, Turkey (Atalay). NLP LOGIX, Jacksonville, FL, USA (Berseth). Smart Imaging Technologies Co., Houston, TX, USA (Khvatkov, Vylegzhanin). Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada (Kraus). Tissue Image Analytics Lab, Department of Computer Science, University of Warwick, UK (Shaban, Rajpoot, Qaiser, Sirinukunwattana). Department of Pathology, University Hospitals Coventry and Warwickshire NHS Trust, UK (Rajpoot). Department of Computer Science and Engineering, Qatar University, Qatar (Awan). Department of Pathology, University Hospitals Coventry and Warwickshire NHS Trust, UK (Tsang). HTW Berlin, Berlin, Germany (Annuscheit, Hufnagl). BioMediTech Institute and Faculty of Medicine and Life Sciences, Tampere University of Technology, Tampere, Finland (Valkonen, Liimatainen, Kartasalo, Ruusuvuori). BioMediTech Institute and Faculty of Biomedical Science and Engineering, Tampere University of Technology, Tampere, Finland (Kartasalo). Prostate Cancer Research Center, Faculty of Medicine and Life Sciences and BioMediTech, University of Tampere, Tampere, Finland (Latonen). Faculty of Computing and Electrical Engineering, Tampere University of Technology, Pori, Finland (Ruusuvuori). Technical University of Munich, Munich, Germany (Albarqouni, Mungal, George, Demirci, Navab). Department of Bioinformatic Engineering, Osaka University (Watanabe, Seno, Takenaka, Matsuda). University of South Florida, FL, USA (Ahmady Phoulady). Biomedical Image Analysis Department, United Institute of Informatics Problems, Belarus National Academy of Sciences, Minsk, Belarus (Kovalev, Kalinovsky, Liauchuk). VISILAB, University of Castilla-La Mancha, Ciudad Real, Spain (Bueno, Fernandez-Carrobles, Serrano, Deniz). Sorbonne Universiteś, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale (LIB), Paris, France (Racoceanu). Pontifical Catholic University of Peru, San Miguel, Lima, Peru (Racoceanu). Sorbonne University, UPMC-Univ. Paris 6, Paris, France (Venâncio).

The CAMELYON16 Collaborators:
Ewout Schaafsma, MD, PhD; Benno Kusters, MD, PhD; Michiel vd Brand, MD; Lucia Rijstenberg, MD; Michiel Simons, MD; Carla Wauters, MD, PhD; Willem Vreuls, MD; Heidi Kusters, MD, PhD; Robert Jan van Suylen, MD, PhD; Hans van der Linden, MD, PhD; and Monique Koopmans, MD, PhD; Gijs van Leeuwen, MD, PhD; and Matthijs van Oosterhout, MD, PhD; Peter van Zwam, MD;

Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands (Schaafsma, Kusters, vd Brand, Rijstenberg, Simons). Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands (Wauters, Vreuls, Kusters). Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands (Suylen, van der Linden, Koopmans). St. Antonius Ziekenhuis, Nieuwegein, The Netherlands (van Leeuwen, van Oosterhout). Stichting PAMM, Eindhoven, The Netherlands (van Zwam)

References

1. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology.* 2017;70(1):134-145.
2. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* 2016;33:170-175.
3. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402-2410.
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 02/02/print 2017;542(7639):115-118.
5. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* 05/23/online 2016;6:26286.
6. Vestjens JHMJ, Pepels MJ, de Boer M, et al. Relevant impact of central pathology review on nodal classification in individual breast cancer patients. *Ann. Oncol.* 2012;23(10):2561-2566.
7. Reed J, Rosman M, Verbanac KM, Mannie A, Cheng Z, Tafra L. 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. *J. Am. Coll. Surg.* 2009/03/01 2009;208(3):333-340.
8. Chagpar A, Middleton LP, Sahin AA, et al. Clinical outcome of patients with lymph node-negative breast carcinoma who have sentinel lymph node micrometastases detected by immunohistochemistry. *Cancer.* 2005;103(8):1581-1586.
9. Pendas S, Dauway E, Cox CE, Giuliano R. Sentinel node biopsy and cytokeratin staining for the accurate staging of 478 breast cancer patients/Discussion. *The American surgeon.* 1999;65(6):500.
10. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. *Nucl. Instrum. Methods. Phys. Res. A: Accelerators, spectrometers, detectors and associated equipment.* 2011;648 Supplement 1:S297-S301.
11. Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* 1979/01 1979;7(1):1-26.
12. Gallas BD, Chan H-P, D'Orsi CJ, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad. Radiol.* 2012;19(4):463-477.
13. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis. *Acad. Radiol.* 2004;11(9):980-995.
14. Hillis SL, Obuchowski NA, Berbaum KS. Power Estimation for Multireader ROC Methods. *Acad. Radiol.* 2011/02/01/ 2011;18(2):129-142.
15. Upton G, Cook I. A dictionary of statistics 3e. Oxford university press; 2014.
16. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. Royal Meteorol. Soc.* 2002;128(584):2145-2166.
17. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 2004;60(2):91-110.
18. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002;24(7):971-987.
19. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 1973;SMC-3(6):610-621.
20. Cortes C, Vapnik V. Support-vector networks. *Mach. Learn.* 1995;20(3):273-297.
21. Breiman L. Random forests. *Mach. Learn.* 2001;45(1):5-32.
22. Szegedy C, Wei L, Yangqing J, et al. Going deeper with convolutions. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 7-12 June 2015, 2015.

23. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the Jackknife method. *Invest. Radiol.* 1992;27(9):723-731.

24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016, 2016.

25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.* 2014.

26. Ehteshami Bejnordi B, Litjens G, Timofeeva N, et al. Stain specific standardization of whole-slide histopathological images. *IEEE Trans. Med. Imaging.* 2016;35(2):404-415.

27. Kendall A, Badrinarayanan V, Cipolla R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680.* 2015.

28. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in neural information processing systems2012.

29. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention2015.

30. Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. Paper presented at: Proceedings of the IEEE International Conference on Computer Vision2015.

31. Viola P, Jones M. Fast and robust classification using asymmetric adaboost and a detector cascade. Paper presented at: Advances in Neural Information Processing Systems2002.

32. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging.* 2016;35(5):1313-1321.

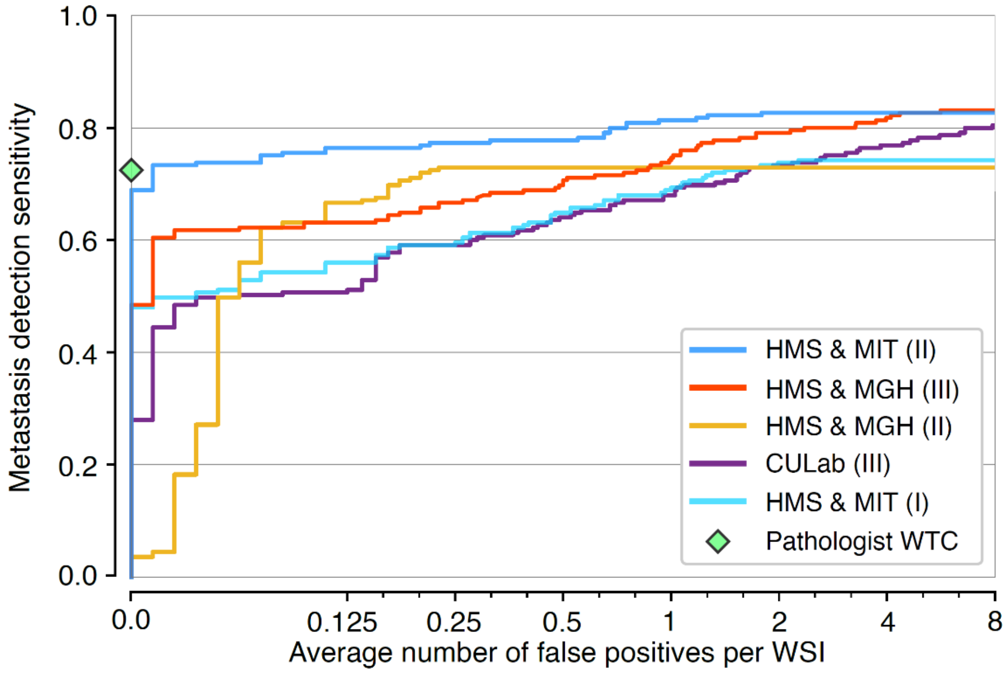


Figure 1. FROC curves of the top-five performing systems for the metastases identification task (task 1; measured on the 129 whole-slide images in the test set of which 49 contain metastatic regions). The range on the x-axis is linear between 0 and 0.125 and base-2 logarithmic scale between 0.125 and 8. Pathologist WTC refers to the pathologist who diagnosed the slides without time constraint. The pathologist did not produce any false positives and achieved a true positive fraction of 0.724 for detecting and localizing metastatic regions.

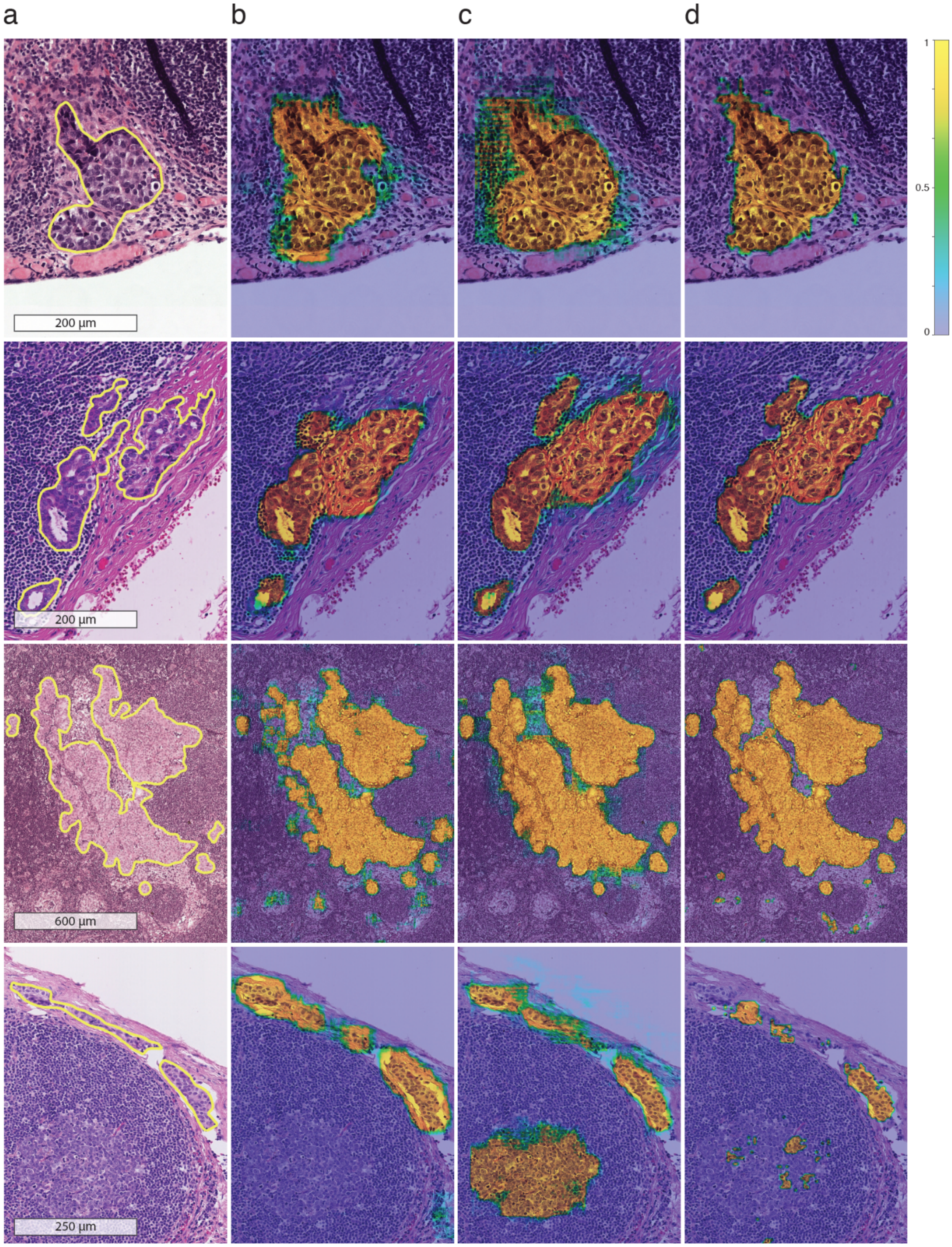


Figure 2. Example probability maps generated by the top-three performing systems for classification of microscopic images of H&E stained lymph node tissue sections. (a) Four annotated micrometastatic lesions in the test set of CAMELYON16. (b-d) Probability maps for teams HMS & MIT II, HMS & MGH III, and CULab III, respectively, overlaid on the original images. The color scale bar (top right) indicates the probability for each pixel to be part of a metastatic region. See eFigure 5 in the supplement for additional examples.

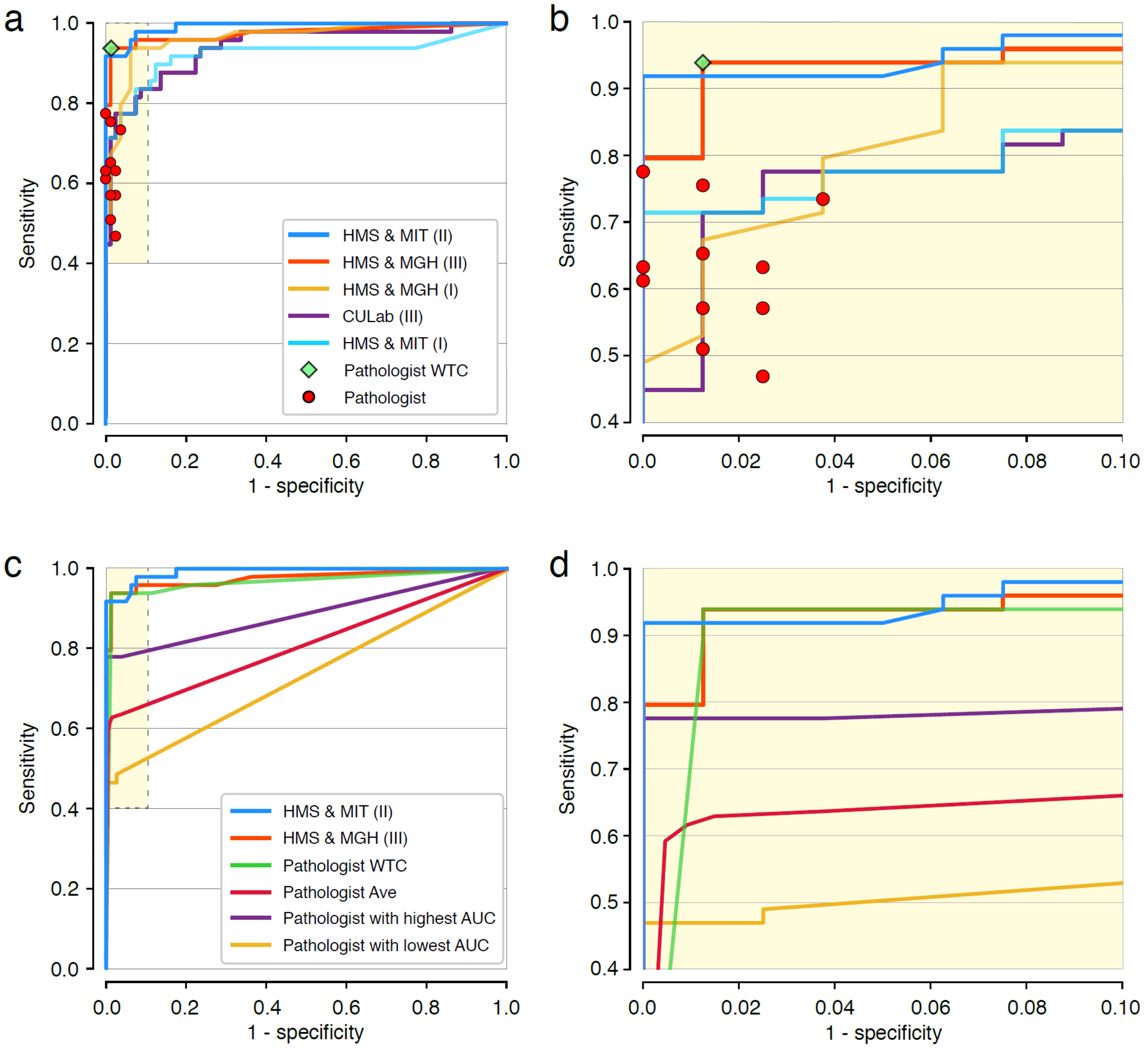
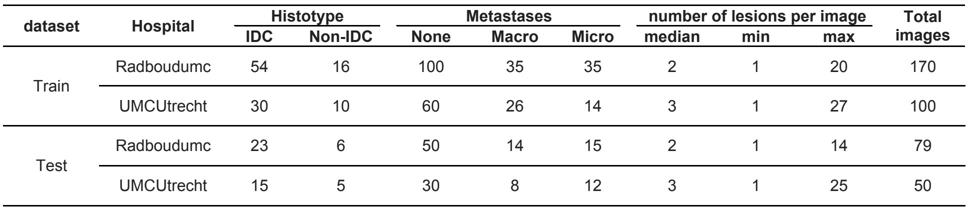


Figure 3. ROC curves of the top-performing systems (measured on the 129 whole-slide images in the test set of which 49 contain metastatic regions) for task 2. (**a**) ROC curves of the top-five performing systems and the operating points of the pathologists. A system achieves superior performance to a pathologist if the operating point of the pathologist lies below the system's ROC curve. The top-two deep learning based systems outperform all the 11 pathologists in our study. Pathologist WTC refers to the operating point of the pathologist who diagnosed the slides without time constraint. All the pathologists scored whole-slide images using five levels of confidence: definitely normal, probably normal, equivocal, probably tumor, definitely tumor. To generate estimates of sensitivity and specificity for each pathologist, negative was defined as confidence levels 'definitely normal' and 'probably normal' and all others as positive. (**b**) More detailed view of the highlighted area in *(a)*. (**c**) Comparison of the ROC curves of the top-two performing systems, the mean ROC over the panel of 11 pathologists (denoted as Pathologist Avg.) and the pathologist without time constraint. The mean ROC curve was computed using the pooled mean technique. This mean is obtained by joining all the diagnoses of the 11 pathologists and computing the resulting ROC curve as if it were one person analyzing 11×129=1419 cases. (**d**) More detailed view of the highlighted area in *(c)*.

Table 1. Summary of the number of whole-slide images in the training and testing sets including the number of cases with different primary tumor Histotypes (infiltrating ductal carcinoma (IDC) and other histotypes (non-IDC)) and the number of cases with different metastases size (micrometastases and macrometastases). The statistics of the number of lesions in the positive slides are also presented.



dataset	Hospital	Histotype		Metastases			number of lesions per image			Total images
		IDC	Non-IDC	None	Macro	Micro	median	min	max	
Train	Radboudumc	54	16	100	35	35	2	1	20	170
	UMCUtrecht	30	10	60	26	14	3	1	27	100
Test	Radboudumc	23	6	50	14	15	2	1	14	79
	UMCUtrecht	15	5	30	8	12	3	1	25	50

Table 2. Results of the submitted algorithms on the CAMELYON16 test set for the metastasis identification and the whole-slide image classification tasks. Algorithms are sorted based on their performance on the whole-slide classification task. The percentile bootstrap method was used to construct 95% confidence intervals for FROC true positive fraction scores (FROC scores) and AUCs. The results of the significant test with MRMC analysis for the comparison of each individual algorithm with the panel of 11 pathologist are provided. The p-values were adjusted for multiple comparisons using the Bonferroni correction in which the p-values are multiplied by the number of comparisons (32; comparison of the 32 submitted algorithms with the panel of pathologists). See eTable 3 and eMethods in the supplement for algorithms contact information and detailed description of each algorithm. See eText in the supplement for glossary of deep learning terminology.

Codename	Entry	metastasis identification		Whole-slide image Classification		Comparison to panel of 11 pathologists	Approach		Remarks
		FROC score	95% CI	AUC	95% CI		Deep learning	model	
HMS & MIT	II	0.807	0.732 - 0.889	0.9935	0.983 - 0.999	p<0.001	⍰	GoogLeNet[22]	Ensemble of two networks, stain standardization, extensive data augmentation, hard negative mining
HMS & MGH	III	0.760	0.692 - 0.857	0.9763	0.941 - 0.999	p<0.001	⍰	ResNet[24]	Fine-tuning pre-trained network, Fully convolutional network
HMS & MGH	I	0.596	0.578 - 0.734	0.9643	0.928 - 0.989	p<0.001	⍰	GoogLeNet[22]	Fine-tuning pre-trained network
CULab III	III	0.703	0.605 - 0.799	0.9403	0.888 - 0.980	p<0.001	⍰	VGG-16[25]	Fine-tuning pre-trained network, Fully convolutional network
HMS & MIT	I	0.693	0.600 – 0.819	0.9234	0.855 - 0.977	p=0.11	⍰	GoogLeNet[22]	Ensemble of two networks, Hard negative mining
ExB	I	0.511	0.363 - 0.620	0.9156	0.858 - 0.962	p=0.02	⍰	ResNet[24]	Varying class balance during training
CULab	I	0.544	0.467 - 0.629	0.9087	0.851 - 0.954	p=0.04	⍰	VGG-Net[25]	Fine-tuning pre-trained network
HMS & MGH	II	0.729	0.596 - 0.788	0.9082	0.846 - 0.961	p=0.04	⍰	ResNet[24]	Fine-tuning pre-trained network
CULab	II	0.527	0.335 - 0.627	0.9056	0.841 - 0.957	p=0.16	⍰	VGG-Net[25] & ResNet[24]	Fine-tuning pre-trained network, Cascade a VGG-Net that operates on low magnification images and a ResNet model that refines the results.
DeepCare	I	0.243	0.197 - 0.356	0.8833	0.806 - 0.943	p>0.99	⍰	GoogLeNet[22]	Fine-tuning pre-trained network
Quincy Wong	I	0.367	0.250 - 0.521	0.8654	0.789 - 0.924	p>0.99	⍰	SegNet[27]	Fine-tuning pre-trained network
METU	I	0.389	0.272 - 0.512	0.8642	0.786 - 0.927	p>0.99	⍰	4-layer CNN	Custom confidence filtering for post-processing
NLP LOGIX	I	0.386	0.255 - 0.511	0.8298	0.742 - 0.899	p>0.99	⍰	AlexNet[28]	Using a second stage random forest classifier to generate slide scores
Smart Imaging	II	0.339	0.239 - 0.420	0.8208	0.753 - 0.894	p>0.99	⍰	GoogLeNet[22]	Uses ensemble of the output from the team's first entry and GoogLeNet model
U of Toronto	I	0.382	0.286 - 0.515	0.8149	0.722 - 0.886	p>0.99	⍰	VGG-Net[25]	Combining the output of multiple CNNs trained on different magnifications by computing their mean
Warwick-QU	I	0.305	0.219 - 0.397	0.7958	0.711 - 0.871	p>0.99	⍰	U-Net[29]	Use of stain normalization
Radboudumc	I	0.575	0.446 - 0.659	0.7786	0.694 - 0.860	p>0.99	⍰	VGG-Net[25]	Extensive data augmentation, second stage CNN to generate slide level scores
HTW-Berlin	I	0.187	0.112 - 0.250	0.7676	0.665 - 0.853	p>0.99	⍰	CRFasRNN[30]	Fine-tuning pre-trained network
U of Toronto	II	0.352	0.292 - 0.511	0.7621	0.659 - 0.846	p>0.99	⍰	VGG-Net[25]	Combining the output of multiple CNNs trained on different magnifications by using an additional CNN
Tampere	I	0.257	0.171 - 0.376	0.7612	0.662 - 0.837	p>0.99	⍰	Random Forests[21]	Using a large set of intensity and texture features
Smart Imaging	I	0.208	0.119 - 0.306	0.7574	0.663 - 0.839	p>0.99	⍰	SVM[20] & Adaboost[31]	Cascade of SVM and Adaboost classifiers using texture features
Osaka University	I	0.347	0.234 - 0.463	0.7319	0.629 - 0.824	p>0.99	⍰	GoogLeNet[22]	
CAMP-TUM	II	0.273	0.194 - 0.379	0.7316	0.633 - 0.819	p>0.99	⍰	GoogLeNet[22]	Hard negative mining
USF	I	0.179	0.116 - 0.242	0.7270	0.611 - 0.823	p>0.99	⍰	Random Forests[21]	Using various intensity and texture features
NSS	I	0.165	0.116 - 0.195	0.7269	0.635 - 0.81	p>0.99	⍰	Rule-based	Multiple thresholds on several nucleus based features
Tampere	II	0.252	0.149 - 0.350	0.7133	0.612 - 0.801	p>0.99	⍰	7-layer CNN	Self-designed network architecture
CAMP-TUM	I	0.184	0.127 - 0.243	0.6911	0.580 - 0.787	p>0.99	⍰	Agg-Net[32]	Multi-scale approach for analyzing the images
Minsk Team	I	0.227	0.181 - 0.264	0.6890	0.568 - 0.804	p>0.99	⍰	GoogLeNet[22]	Separate models for different datasets, Hard negative mining
VISILAB	I	0.142	0.080 - 0.203	0.6532	0.551 - 0.748	p>0.99	⍰	Random Forests[21]	Using Haralick texture features[19]
VISILAB	II	0.116	0.063 - 0.177	0.6513	0.549 - 0.742	p>0.99	⍰	3-layer CNN	Self-designed network architecture
Anonymous	I	0.097	0.049 - 0.158	0.6277	0.530 - 0.717	p>0.99	⍰	Random Forests[21]	
LIB	I	0.120	0.079 - 0.182	0.5561	0.434 - 0.654	p>0.99	⍰	SVM[20]	Using various color and texture features
Pathologist WTC		0.724	0.643 - 0.804	0.966	0.927 - 0.998	—	—	—	The expert pathologist who scored without time a limit
Mean pathologist		—	—	0.810	0.750 – 0.869	—	—	—	The mean performance of 11 pathologists in a simulation exercise designed to mimic the routine workflow of diagnostic pathology

Supplementary Online Content

Ehteshami Bejnordi B, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*.

Supplementary information

eText

Deep Learning.

Application of traditional machine learning to medical image analysis typically involved human engineers collaborating with physicians to decide what kind of features are needed to recognize lesions or objects of interest in the images. These features were then extracted from medical images and fed to an algorithm which would assess whether the lesion or object was present in the image. Deep learning changes this in two important ways. The first refers to the 'deep' part in deep learning; whereas the traditional approach typically consisted of two steps, a deep network consists of many steps, or in deep learning terminology, layers. Each of these layers can perform feature extraction or classification, and because one layer feeds output into the next, features can hierarchically become more complex. The first layers, for example, can identify edges or circles and subsequent layers can combine these into more meaningful objects, eventually leading to complex structures such as faces in natural images. The second major difference with traditional machine learning is that features are no longer manually engineered, but learned automatically by the system. This is done by optimizing deep learning algorithms end-to-end, i.e. given an input, optimize the parameters across layers in such a way that the desired output is most likely. Typically this is done with an algorithm called backpropagation.

Most deep learning algorithms are based on artificial neural networks that are mathematical constructs that stack together 'nodes'. Nodes consist of simple multiplications and additions, combined with a non-linear transform and multiple nodes form a layer. By selecting how nodes between different layers are connected one can determine how features are extracted. Currently the most popular deep learning algorithm is the convolutional neural network (CNN). In a CNN nodes are connected in such a way that they model a convolution operation, which allows recognition of a single feature (a convolutional filter) across the entire image, making CNNs highly efficient for image processing. CNNs have revolutionized the field of computer visions, breaking records and attaining results that have eluded the community for years[1-4].

Glossary of deep learning and digital pathology terminology.

Model fine-tuning – using the weights of a model that was trained for one task as an initialization for training a model for a different task.

Model ensembling – combining the output from different models (e.g. by averaging the predictions) with the goal of improving the overall performance.

Hard-negative mining – discovering negative samples (in a detection problem) that are non-trivial to distinguish from positive samples.

Data augmentation – applying transformation to the training samples that create new, plausible training samples with the goal of increasing the training set size.

Staining normalization – modifying the color appearance of whole slide images such that it resembles some reference sample with the goal of reducing the appearance variability within a dataset.

Fully convolutional network – neural network consisting only of convolutional layers, or more generally, consisting only of layers that produce outputs for arbitrary input sizes (this enables a model to be trained on small images and then applied to larger images such as whole slide images).

AlexNet[1] – neural network architecture that was the winner of the ImageNet Large Scale Visual Recognition Challenge 2012 for the object detection, localization and classification tasks[5].

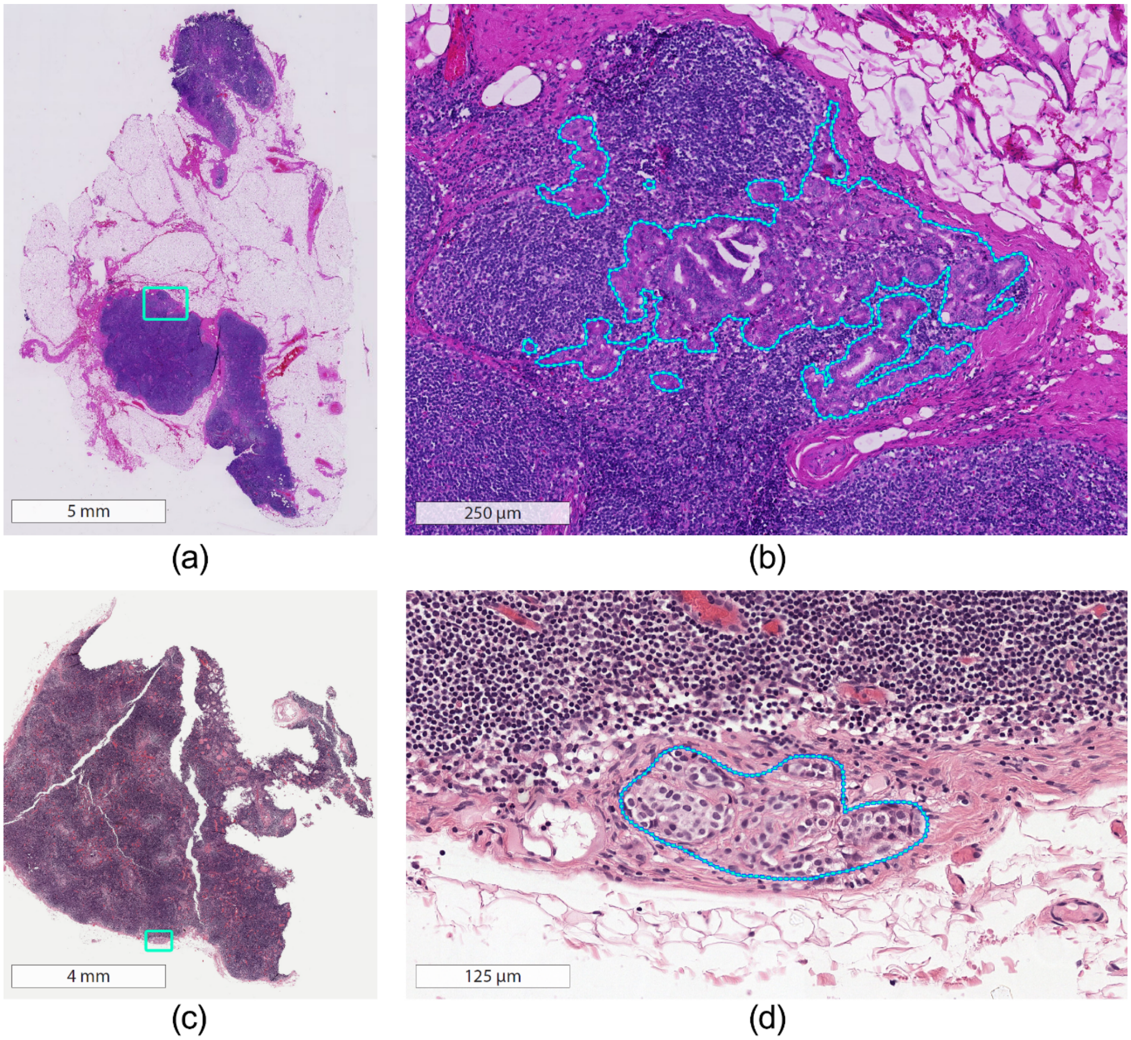
GoogLeNet[3] – neural network architecture that was the winner of the ImageNet Large Scale Visual Recognition Challenge 2014 for the object detection and classification tasks[5].

VGG-net[6] – neural network architecture that was the winner of the ImageNet Large Scale Visual Recognition Challenge 2014 for the localization task[5].

ResNet[4] – neural network architecture that was the winner of the ImageNet Large Scale Visual Recognition Challenge 2015 for the object detection, localization and classification tasks[4,5].

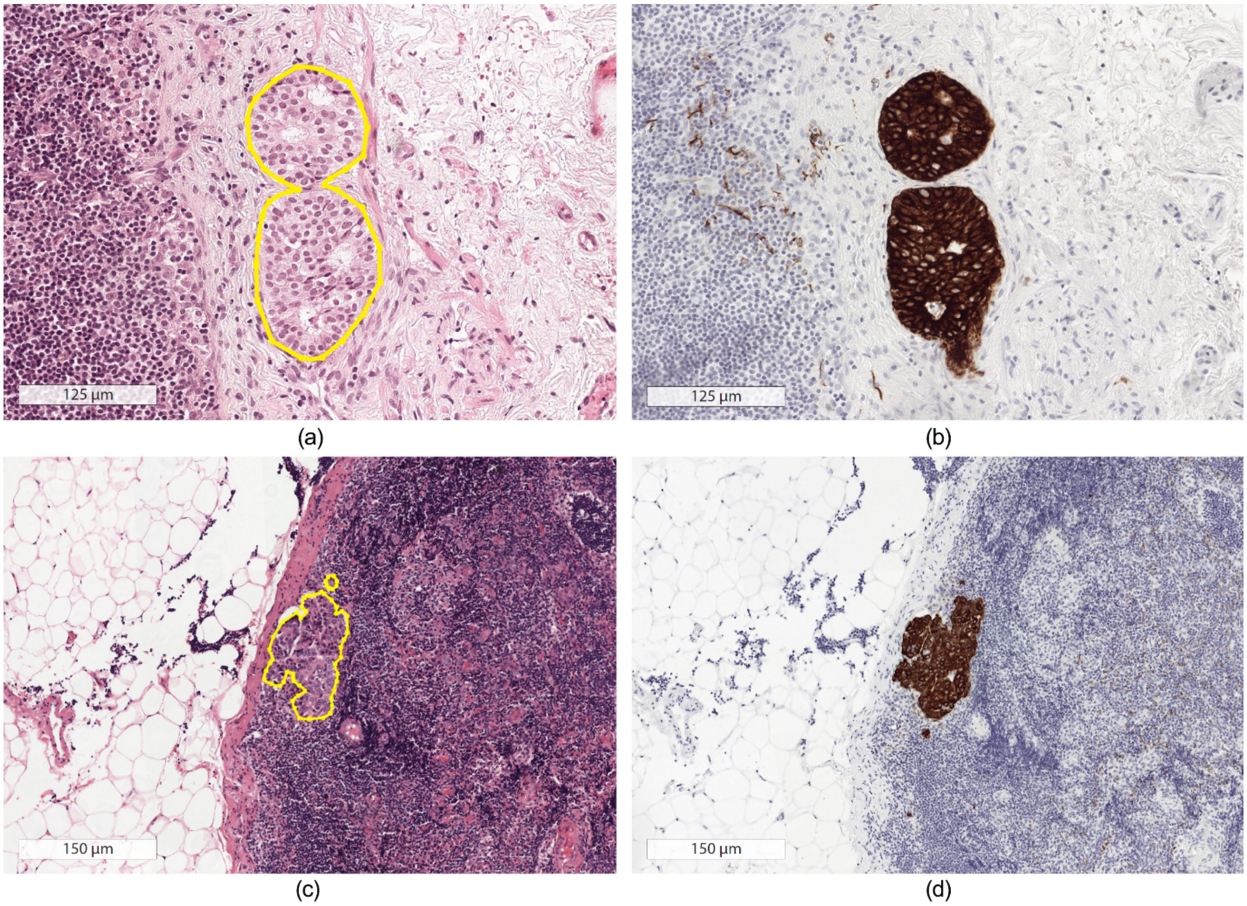
U-Net[7] and SegNet[8] – neural network architectures that were specifically designed for segmentation of biomedical images.

eFigure 1. Two example annotated areas of whole-slide images taken from the CAMELYON16 dataset



eFigure 1. Two example annotated areas of whole-slide images of hematoxylin and eosin stained lymph node tissue sections taken from the CAMELYON16 dataset. (a) and (c) show overviews of two examples of whole-slide images. (b) and (d) are magnified images, corresponding to rectangle areas in *(a)* and *(c)*, with detailed annotation of metastatic regions.

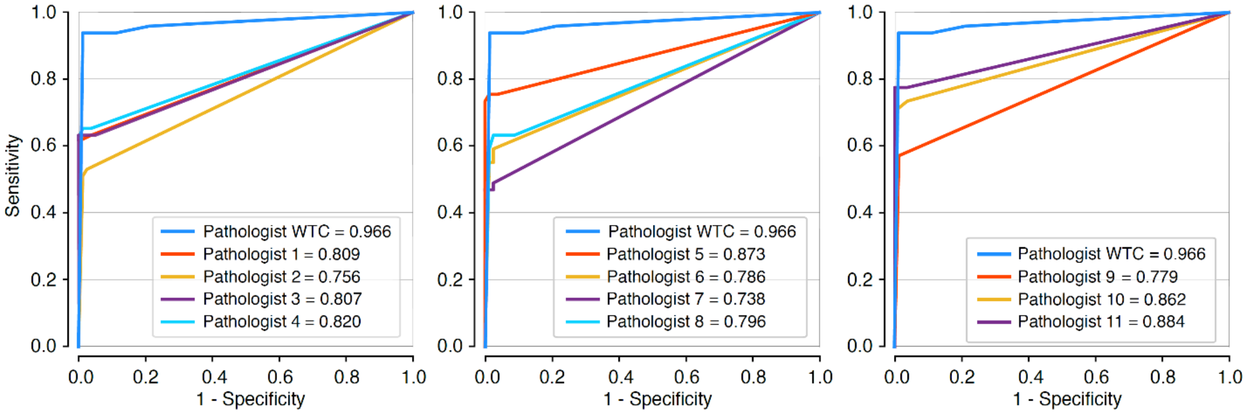
eFigure 2. Use of immunohistochemistry staining to generate reference standard



(a) (b) (c) (d)

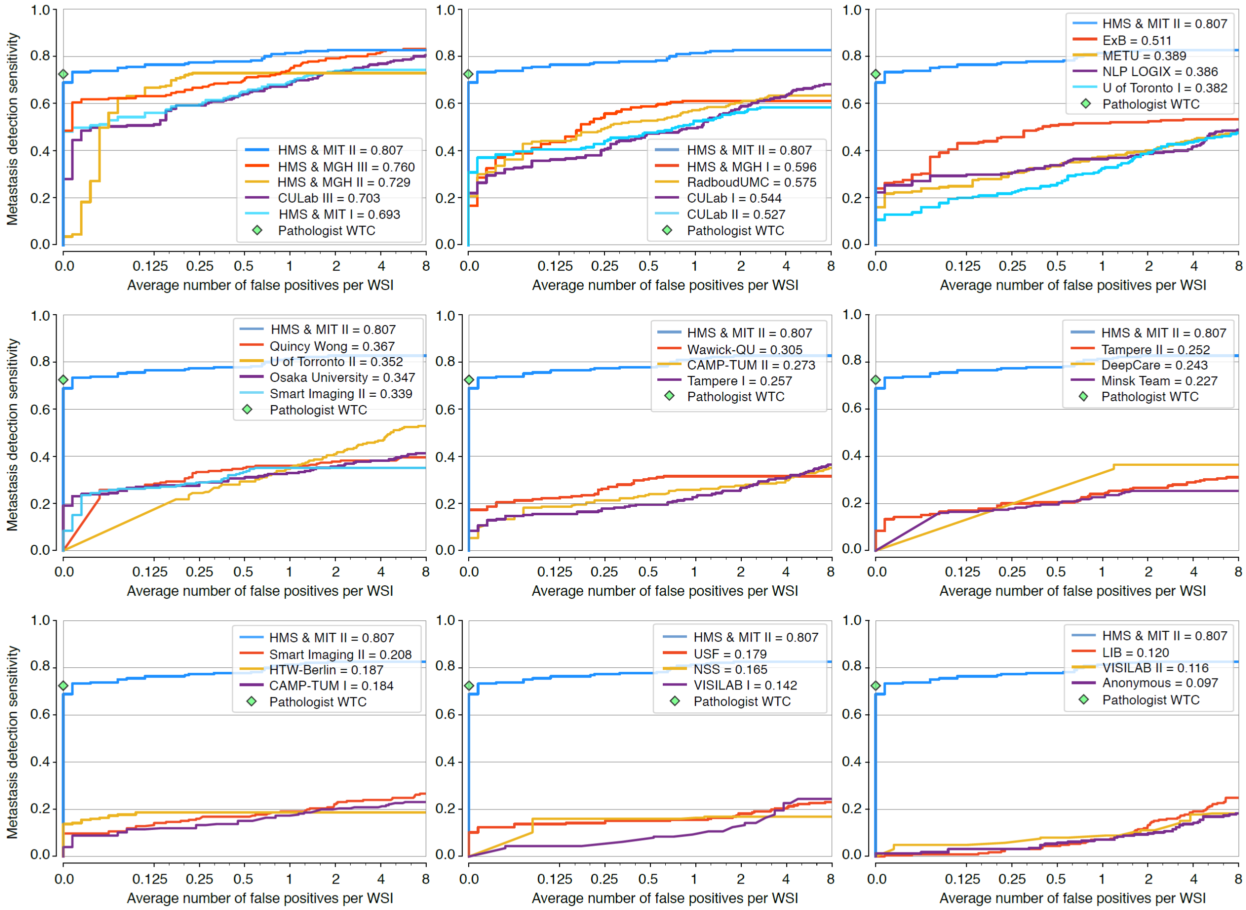
eFigure 2. Side by side visualization of hematoxylin and eosin (H&E) and immunohistochemistry (IHC) staining for generating reference standard. (a) and (c) show two example annotations made for two H&E stained images. (b) and (d) show corresponding tissue areas in *(a)* and *(c)*, stained with IHC. Note that IHC was only used for generating the reference standard in our challenge. Neither of the pathologists in our observer study nor participants of the challenge had access to this data. Immunohistochemical staining was performed with anti-CK8/18 (anti-cytokeratin mouse monoclonal antibody, clone CAM 5.2, BD Biociences, San Jose, USA). Binding of the antibody was visualized with a Brightvision® Poly-HRP-Anti Ms/Rb/Rt IgG biotin free detection system using BrightDAB® (Immunologic, Duiven, the Netherlands) as peroxidase-compatible chromogen and hematoxylin counterstaining.

eFigure 3. ROC curves of the panel of the 11 pathologists for task 2



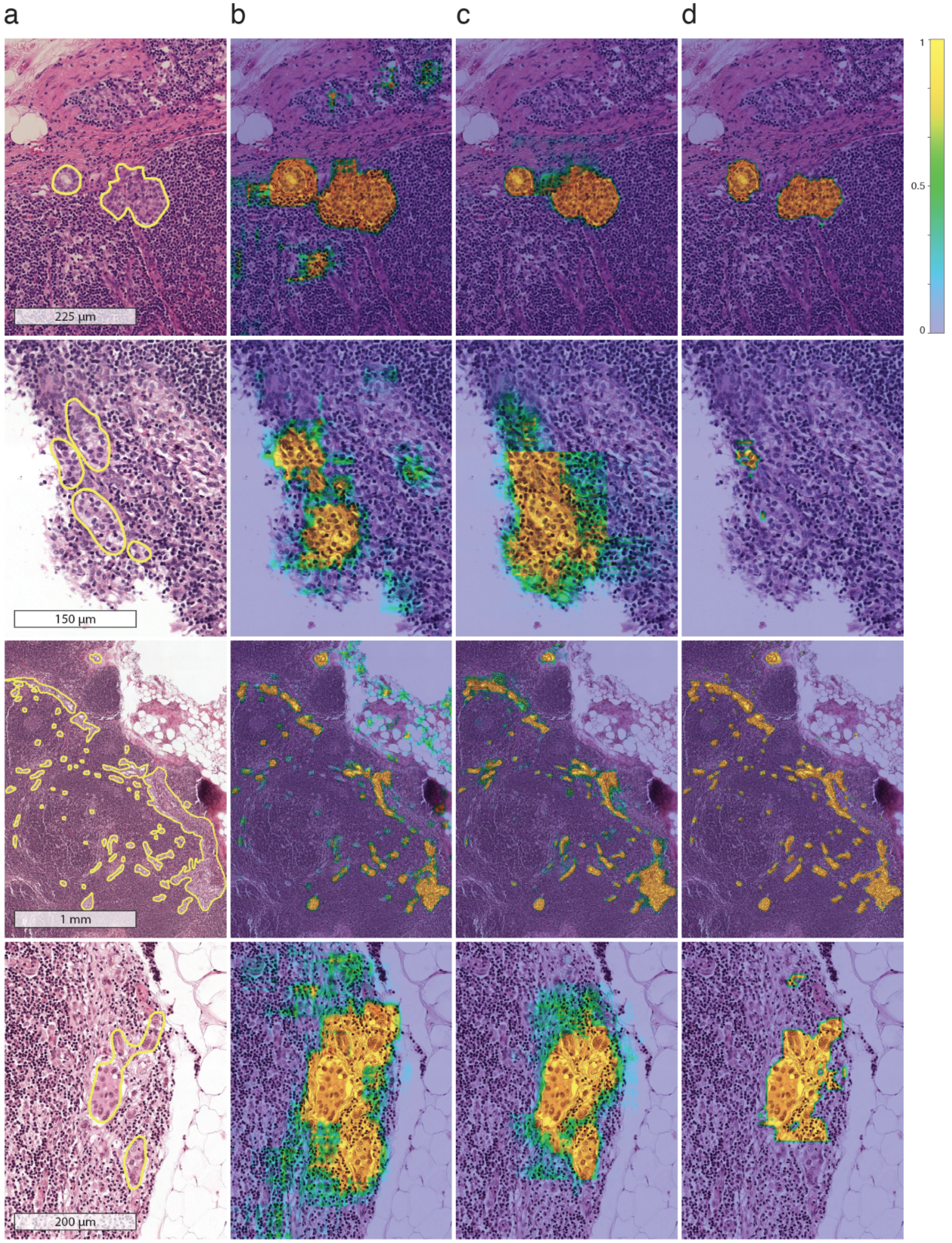
eFigure 3. ROC curves of the panel of 11 pathologists and their corresponding AUCs for task 2 (measured on the 129 whole-slide images in the test set of which 49 contain metastatic regions). All the pathologists scored whole-slide images using five levels of confidence: definitely normal, probably normal, equivocal, not confident, probably tumor, definitely tumor. To ease comparison, the ROC curve of the pathologist without time constraint (pathologist WTC) is shown in all subfigures.

eFigure 4. FROC curves of all participating teams for task 1



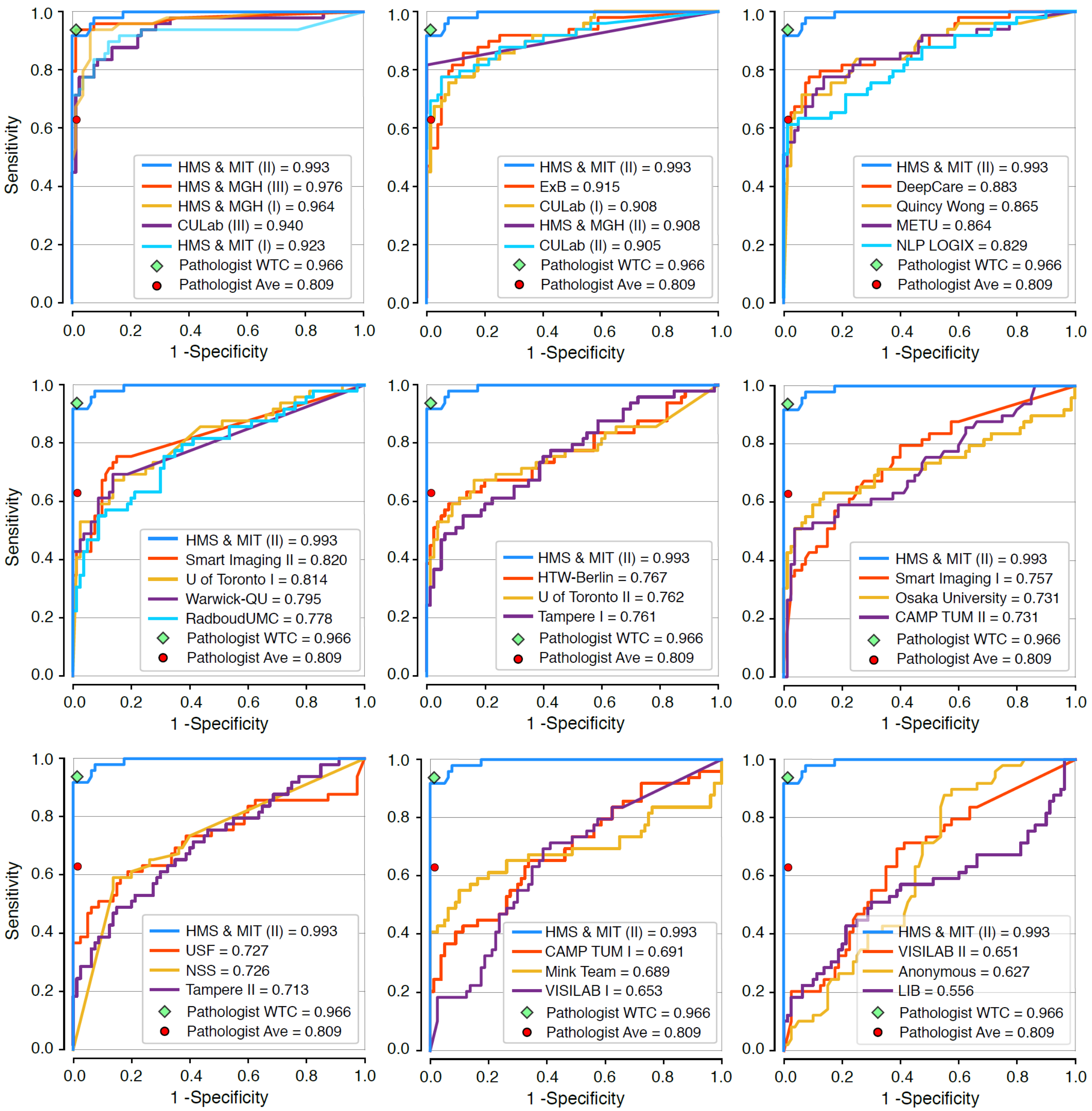
eFigure 4. FROC curves of all the 32 participating teams and their corresponding FROC true positive fraction scores for task 1 (measured on the 129 whole-slide images in the test set of which 49 contain metastatic regions). The operating point of the pathologist who scored the slides without time constraint (WTC) is shown as a green diamond. The range on the x-axis is linear between 0 and 0.125 and base-2 logarithmic scale between 0.125 and 8. The pathologist did not produce any false positives and achieved a true positive fraction of 0.724 for detecting and localizing metastatic regions. To ease comparison, the FROC curve of the best-performing system (HMS & MIT II) is shown in all subfigures.

eFigure 5. Example probability maps generated by the top-three performing systems



eFigure 5. Example probability maps generated by the top-three performing systems. (a) Four annotated metastatic lesions in the test set of CAMELYON16. (b-d) Probability maps for teams HMS & MIT II, HMS & MGH III, and CULab III, respectively, overlaid on the original images.

eFigure 6. ROC curves of all participating teams for task 2



eFigure 6. ROC curves of all the 32 participating teams and their corresponding AUCs for task 2 (measured on the 129 whole-slide images in the test set of which 49 contain metastatic regions). The operating point of the pathologist who scored the slides without time constraint (WTC) and the operating point of the mean of the panel of 11 pathologists are shown as green diamond and red circle, respectively. To ease comparison, ROC curve of the best-performing system (HMS & MIT II) is shown in all subfigures.

eTable1. Classification results by pathologists for the whole-slide image classification task (sensitivity and specificity)

eTable1. Classification results by the panel of 11 pathologists participating in the simulation exercise and the expert pathologist whiteout time constraint (WTC) on the CAMELYON16 test set for the whole-slide image classification task (task 2). The performances are measured in 129 whole-slide images in the test set of which 49 contain metastatic regions (comprising of 22 macro and 27 micrometastases, and 38 with primary tumor histotype of infiltrating ductal cancer (IDC) and 11 non-IDC). We report sensitivity and specificity for different scenarios: 1) differentiating all tumor slides from normal slides, 2) differentiating slides with macrometastases from normal slides while excluding micrometastases, 3) differentiating slides with micrometastases from normal slides while excluding macrometastases, 4) differentiating slides with primary tumor histotype of IDC from normal slides while excluding the rarer primary tumor histotypes (non-IDC), and 5) differentiating slides with non-IDC primary histotypes from normal slides while excluding slides with primary tumor histotype of IDC.

Codename	All cases		Metastases				Histotype			
			Macrometastases		Micrometastases		IDC		Non-IDC	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Pathologist 1	0.612	1	0.954	1	0.333	1	0.647	1	0.533	1
Pathologist 2	0.510	0.987	0.909	0.987	0.185	0.987	0.588	0.987	0.333	0.987
Pathologist 3	0.632	1	0.954	1	0.370	1	0.735	1	0.4	1
Pathologist 4	0.653	0.987	0.954	0.987	0.407	0.987	0.705	0.987	0.533	0.987
Pathologist 5	0.755	0.987	1	0.987	0.555	0.987	0.764	0.987	0.733	0.987
Pathologist 6	0.571	0.975	0.818	0.975	0.370	0.975	0.676	0.975	0.333	0.975
Pathologist 7	0.469	0.975	0.863	0.975	0.148	0.975	0.529	0.975	0.333	0.975
Pathologist 8	0.632	0.975	0.954	0.975	0.370	0.975	0.705	0.975	0.466	0.975
Pathologist 9	0.571	0.987	0.909	0.987	0.296	0.987	0.617	0.987	0.466	0.987
Pathologist 10	0.734	0.962	0.954	0.962	0.555	0.962	0.794	0.962	0.6	0.962
Pathologist 11	0.775	1	0.954	1	0.629	1	0.850	1	0.6	1
Mean pathologist	0.628	0.985	0.929	0.985	0.383	0.985	0.692	0.985	0.484	0.985
Pathologist WTC	0.938	0.987	1	0.987	0.888	0.9875	0.970	0.987	0.866	0.987

eTable2. Classification results by pathologists for the whole-slide image classification task (area under the ROC curve)

eTable2. Classification results by the panel of 11 pathologists participating in the simulation exercise and the expert pathologist without time constraint (WTC) on the CAMELYON16 test set for the whole-slide image classification task (task 2). The performances are measured in 129 whole-slide images in the test set of which 49 contain metastatic regions (comprising of 22 macro and 27 micrometastases, and 38 with primary tumor histotype of infiltrating ductal cancer (IDC) and 11 non-IDC). We report classification AUC for different scenarios: 1) differentiating all tumor slides from normal slides, 2) differentiating slides with macrometastases from normal slides while excluding micrometastases, 3) differentiating slides with micrometastases from normal slides while excluding macrometastases, 4) differentiating slides with primary tumor histotype of IDC from normal slides while excluding the rarer primary tumor histotypes (non-IDC), and 5) differentiating slides with non-IDC primary histotypes from normal slides while excluding slides with primary tumor histotype of IDC. We used percentile bootstrapping to construct 95% confidence interval. The results of the significance test for comparison of the performance of each pathologist for the detection of micro and macrometastases as well as for comparison of the performance for the detection of IDC and non-IDC metastases are presented (see the statistical analysis section). The p-values were adjusted for multiple comparisons using the Bonferroni correction.

Codename	All cases		Metastases					Histotype				
			Macrometastases		Micrometastases		Comparison of detection performance	IDC		Non-IDC		Comparison of detection performance
	AUC	95% CI	AUC	95% CI	AUC	95% CI		AUC	95% CI	AUC	95% CI	
Pathologist 1	0.809	0.732-0.876	0.976	0.918-1.0	0.673	0.577-0.777	p<0.001	0.817	0.729-0.899	0.791	0.665-0.916	p>0.99
Pathologist 2	0.756	0.679-0.82	0.948	0.874-1.0	0.599	0.510-0.672	p<0.001	0.785	0.696-0.858	0.689	0.569-0.831	p>0.99
Pathologist 3	0.807	0.738-0.876	0.976	0.916-1.0	0.669	0.562-0.757	p<0.001	0.861	0.779-0.937	0.685	0.566-0.825	p=0.34
Pathologist 4	0.820	0.744-0.885	0.976	0.915-1.0	0.692	0.590-0.787	p<0.001	0.847	0.762-0.922	0.758	0.623-0.891	p>0.99
Pathologist 5	0.873	0.802-0.926	1.0	1.0-1.0	0.769	0.659-0.859	p=0.01	0.878	0.797-0.949	0.862	0.737-0.969	p>0.99
Pathologist 6	0.786	0.711-0.854	0.924	0.838-0.993	0.674	0.577-0.76	p=0.03	0.844	0.758-0.921	0.656	0.543-0.778	p=0.15
Pathologist 7	0.738	0.663-0.805	0.930	0.843-1.0	0.582	0.502-0.65	p<0.001	0.773	0.683-0.854	0.658	0.548-0.791	p>0.99
Pathologist 8	0.796	0.715-0.866	0.969	0.904-1.0	0.654	0.549-0.739	p<0.001	0.835	0.743-0.91	0.707	0.576-0.854	p>0.99
Pathologist 9	0.779	0.707-0.845	0.948	0.869-1.0	0.642	0.545-0.72	p<0.001	0.803	0.710-0.884	0.727	0.599-0.857	p>0.99
Pathologist 10	0.862	0.796-0.927	0.976	0.917-1.0	0.769	0.651-0.859	p=0.01	0.893	0.815-0.957	0.793	0.670-0.919	p>0.99
Pathologist 11	0.884	0.816-0.941	0.976	0.917-1.0	0.808	0.704-0.908	p=0.03	0.924	0.845-0.983	0.793	0.660-0.919	p=0.25
Mean pathologist	0.810	0.750-0.869	0.964	0.930-0.997	0.685	0.619-0.746	—	0.842	0.775-0.907	0.738	0.630-0.846	—
Pathologist WTC	0.966	0.927-0.998	0.994	0.977-1.0	0.943	0.868-0.995	p=0.87	0.976	0.932-1.0	0.943	0.848-1.0	p>0.99

eTable3. Participating teams in CAMELYON16

eTable3. Teams participating in CAMELYON16. Each method is identified with a codename used in the text. See **eMethods** for details about each method.

Codename	Contributors	Institutions	Training Model
HMS & MIT (I & II)	Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H Beck	Harvard Medical School and Massachusetts institute of Technology	(Models I & II) 22 layer GoogLeNet[3]
HMS & MGH (I, II & III)	Aoxiao Zhong, Quanzheng Li	Harvard Medical School and Massachusetts General Hospital	(Model I) 22 layer GoogLeNet[3], (Model II) 101 ResNet[4] (Model III) 101 fully convolutional ResNet[4]
ExB	Christian Hass, Urko Sanchez, Ivan Vasilev, Tony Mey, and Elia Bruni	ExB Research and Development GmbH	34 layer ResNet[4]
CULab (I, II & III)	Hao Chen, Huang-Jing Lin, Qi Dou, and Pheng-Ann Heng	The Chinese Univ. of Hong Kong	(Model I) VGG-16[6], (Model II) cascade of VGG-16[6] and ResNet[4]-152 (Model III) VGG-16[6]
Quincy Wong	Quincy Wong	Independent participant	37 layer SegNet[8]
METU	Ugur HALICI, Mustafa Ümit ÖNER, and Rengül Çetin Atalay	Middle East Technical Univ.	4 layer CNN
NLP LOGIX	Matt Berseth	NLP LOGIX	7 layer AlexNet[1]
Smart Imaging (I & II)	Vitali Khvatkov, Alexei Vylegzhanin	Smart Imaging Technologies Co.	(Model I) SVM[9] and Adaboost[10], (Model II) Combination of model I and a 22 layer GoogLeNet[3]
U of Toronto (I & II)	Oren Kraus	Univ. of Toronto	(Models I & II) 10 layer VGG-like[6] network
Warwick-QU	Muhammad Shaban, Talha Qaiser, Ruqayya Awan, Korsuk Sirinukunwattana, Yee-Wah Tsang, and Nasir Rajpoot	University of Warwick	15 layer U-Net[7]
Radboudumc	David Tellez	Radboud Univ. Medical Center	VGG-13[6]
HTW-Berlin	Jonas Annuscheit, Peter Hufnagl	HTW-BERLIN	CRFasRNN[11]
Tampere I	Mira Valkonen, Kimmo Kartasalo, Kaisa Liimatainen, Leena Latonen, Pekka Ruusuvuori	Univ. of Tampere	Random Forests[12]
Osaka University	Seiryo Watanabe , Shigeto Seno, Yoichi Takenaka, Hideo Matsuda	Osaka Univ.	22 layer GoogLeNet[3]
USF	Hady Ahmady Phoulady	Univ. of South Florida	Random Forests[12]
NSS	Nandakumar P, Sarath PC, Vishnu Prasad M, Yadukrishnan M, and Sreejith Valsan M	NSS college of Engineering	Multiple thresholds
Tampere II	Kaisa Liimatainen, Kimmo Kartasalo, Mira Valkonen, Leena Latonen, Pekka Ruusuvuori	Univ. of Tampere	7 layer CNN
CAMP-TUM (I & II)	Bharti Munjal, Amil George, Shadi Albarqouni, Stefanie Demirci, Nassir Navab	Technical Univ. of Munich	(Model I) 5 layer Agg-Net[13], (Model II) 22 layer GoogLeNet[3]
Minsk Team	Vassili Kovalev, Alexander Kalinovsky, and Vitali Liauchuk	United Institute of Informatics Problems	22 layer GoogLeNet[3]
VISILAB (I & II)	M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Gloria Bueno	Univ. of Castilla-La Mancha	(Model I) Random Forests[12], (Model II) 3 layer CNN
Anonymous	Anonymous	Anonymous	Random Forests[12]
LIB	R. Venâncio, B. Ben Cheikh, A. Coron, and D. Racoceanu	Sorbonne Univ.	SVM[9]
DeepCare	Tong Xu	DeepCare Inc.	22 layer GoogLeNet[3]

eTable4. Summary of results for the metastasis identification task (task1)

eTable4. Results of the submitted algorithms on the CAMELYON16 test set for the metastasis identification task. We report the overall FROC scores and the true positive fraction at several values for the mean number of false positives per whole-slide image (FPs/WSI). The final FROC true positive fraction score (FROC score) that ranked teams in the this task was defined as the mean true positive fraction at 6 predefined false positive rates: 1/4, 1/2, 1, 2, 4, and 8 FPs per whole-slide image. Note that the pathologist scoring without time constraint had an overall true positive fraction of 72.4% without any false positives.

Codename	FROC score	True positive fraction at the different false positive values					
		$^{1}/_{4}$ FPs/WSI	$^{1}/_{2}$ FPs/WSI	1 FPs/WSI	2 FPs/WSI	4 FPs/WSI	8 FPs/WSI
HMS & MIT II	0.807	0.773	0.778	0.813	0.827	0.827	0.827
HMS & MGH III	0.760	0.667	0.707	0.747	0.791	0.818	0.831
HMS & MGH II	0.729	0.729	0.729	0.729	0.729	0.729	0.729
CULab III	0.703	0.591	0.640	0.680	0.733	0.769	0.804
HMS & MIT I	0.693	0.596	0.649	0.693	0.738	0.742	0.742
HMS & MGH I	0.596	0.556	0.587	0.609	0.609	0.609	0.609
RadboudUMC	0.575	0.493	0.524	0.569	0.600	0.631	0.631
CULab I	0.544	0.404	0.471	0.493	0.582	0.631	0.684
CULab II	0.527	0.440	0.476	0.524	0.560	0.582	0.582
ExB	0.511	0.458	0.507	0.516	0.520	0.533	0.533
METU	0.389	0.307	0.333	0.373	0.400	0.444	0.476
NLP LOGIX	0.386	0.307	0.338	0.364	0.387	0.418	0.502
U of Toronto I	0.382	0.244	0.293	0.351	0.409	0.467	0.529
Quincy Wong	0.367	0.333	0.351	0.360	0.378	0.382	0.396
U of Toronto II	0.352	0.222	0.262	0.324	0.391	0.436	0.476
Osaka University	0.347	0.289	0.311	0.329	0.356	0.382	0.413
Smart Imaging II	0.339	0.289	0.338	0.351	0.351	0.351	0.351
Warwick-QU	0.305	0.262	0.307	0.316	0.316	0.316	0.316
CAMP-TUM II	0.273	0.213	0.240	0.258	0.276	0.298	0.356
Tampere I	0.257	0.178	0.196	0.227	0.267	0.311	0.364
Tampere II	0.252	0.200	0.204	0.240	0.267	0.289	0.311
DeepCare	0.243	0.000	0.000	0.364	0.364	0.364	0.364
Minsk Team	0.227	0.178	0.196	0.227	0.253	0.253	0.253
Smart Imaging I	0.208	0.160	0.169	0.191	0.222	0.240	0.267
HTW-Berlin	0.187	0.187	0.187	0.187	0.187	0.187	0.187
CAMP-TUM I	0.184	0.133	0.151	0.173	0.200	0.213	0.231
USF	0.179	0.151	0.151	0.156	0.182	0.204	0.231
NSS	0.165	0.160	0.160	0.164	0.169	0.169	0.169
VISILAB I	0.142	0.062	0.084	0.093	0.142	0.227	0.244
LIB	0.120	0.031	0.044	0.071	0.133	0.191	0.249
VISILAB II	0.116	0.058	0.080	0.089	0.111	0.178	0.182
Anonymous	0.097	0.031	0.058	0.071	0.098	0.142	0.182

eTable5. Summary of results for the whole-slide image classification task (task2)

eTable 5. Results of the submitted algorithms on the CAMELYON16 test set for the whole-slide image classification task. We report classification AUC for different scenarios: 1) differentiating all tumor slides from normal slides, 2) differentiating slides with macrometastases from normal slides while excluding micrometastases, 3) differentiating slides with micrometastases from normal slides while excluding macrometastases, 4) differentiating slides with primary tumor histotype of infiltrating ductal cancer (IDC) from normal slides while excluding the rarer primary tumor histotypes (non-IDC), and 5) differentiating slides with non-IDC primary histotypes from normal slides while excluding slides with primary tumor histotype of IDC.

Codename	**All cases**		**Macrometastases**		**Micrometastases**		**IDC**		**Non-IDC**	
	AUC	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**
HMS & MIT II	0.9935	0.983-0.999	0.9905	0.973-1.0	0.9972	0.989-1.0	0.9926	0.979-1.0	0.9954	0.983-1.0
HMS & MGH III	0.9763	0.941-0.999	1.0	1.0-1.0	0.9569	0.893 – 0.999	0.9785	0.928-1.0	0.9712	0.920-1.0
HMS & MGH I	0.9643	0.928-0.989	0.9932	0.983-1.0	0.9407	0.876-0.987	0.9724	0.946-0.993	0.9458	0.857-0.997
CULab III	0.9403	0.888-0.980	0.9875	0.961-1.0	0.9019	0.812-0.962	0.9529	0.909-0.983	0.9117	0.785-0.991
HMS & MIT I	0.9234	0.855-0.977	0.9596	0.862-1.0	0.8939	0.794-0.971	0.9055	0.807-0.978	0.9642	0.915-0.996
ExB	0.9156	0.858-0.962	0.9948	0.985-1.0	0.8509	0.749-0.932	0.9276	0.855-0.981	0.8883	0.777-0.973
CULab I	0.9087	0.851-0.954	0.9966	0.989-1.0	0.8370	0.742-0.913	0.9290	0.868-0.974	0.8625	0.750-0.960
HMS & MGH II	0.9082	0.846-0.961	1.0	1.0-1.0	0.8333	0.738-0.917	0.9118	0.833-0.968	0.90	0.795-1.0
CULab II	0.9056	0.841-0.957	0.9926	0.972-1.0	0.8347	0.722-0.925	0.9311	0.852-0.983	0.8479	0.720-0.953
DeepCare	0.8833	0.806-0.943	0.9705	0.903-1.0	0.8123	0.704-0.895	0.8932	0.808-0.954	0.8608	0.756-0.973
Quincy Wong	0.8654	0.789-0.924	0.9821	0.952-1.0	0.7703	0.634-0.874	0.8888	0.805-0.953	0.8125	0.657-0.940
METU	0.8642	0.786-0.927	0.9897	0.982-1.0	0.7618	0.630-0.867	0.8877	0.802-0.958	0.8108	0.655-0.941
NLP LOGIX	0.8298	0.742-0.899	0.9863	0.951-1.0	0.7023	0.564-0.812	0.8838	0.796-0.947	0.7075	0.538-0.864
Smart Imaging II	0.8208	0.753-0.894	0.9818	0.962-0.997	0.6895	0.566-0.791	0.8289	0.732-0.913	0.8025	0.664-0.917
U of Toronto I	0.8149	0.722-0.886	0.9514	0.866-0.996	0.7037	0.563-0.804	0.8673	0.779-0.931	0.6963	0.517-0.846
Warwick-QU	0.7958	0.711-0.871	0.9909	0.971-1.0	0.6368	0.513-0.733	0.8393	0.742-0.915	0.6971	0.547-0.836
Radboudumc	0.7786	0.694-0.860	0.9318	0.866-0.992	0.6537	0.536-0.779	0.7923	0.690-0.89	0.7475	0.591-0.88
HTW-Berlin	0.7676	0.665-0.853	0.9591	0.872-0.999	0.6115	0.459-0.736	0.7610	0.617-0.872	0.7825	0.627-0.911
U of Toronto II	0.7621	0.659-0.846	0.9698	0.923-0.996	0.5928	0.442-0.71	0.8294	0.719-0.91	0.6096	0.394-0.805
Tampere I	0.7612	0.662-0.837	0.9687	0.926-0.994	0.5921	0.472-0.703	0.7772	0.647-0.875	0.7250	0.589-0.843
Smart Imaging I	0.7574	0.663-0.839	0.9386	0.880-0.977	0.6097	0.473-0.719	0.7706	0.639-0.860	0.7275	0.597-0.845
Osaka University	0.7319	0.629-0.824	0.9852	0.964-0.998	0.5254	0.361-0.662	0.8051	0.686-0.899	0.5658	0.364-0.762
CAMP-TUM II	0.7316	0.633-0.819	0.9585	0.906-0.995	0.5468	0.409-0.660	0.7596	0.640-0.855	0.6683	0.485-0.827
USF	0.7270	0.611-0.823	0.9380	0.840-0.995	0.5551	0.401-0.674	0.7706	0.636-0.869	0.6283	0.427-0.820
NSS	0.7269	0.635-0.81	0.8562	0.756-0.928	0.6215	0.511-0.749	0.7925	0.686-0.877	0.6783	0.430-0.739
Tampere II	0.7133	0.612-0.801	0.8909	0.782-0.964	0.5685	0.427-0.67	0.7765	0.669-0.861	0.5700	0.398-0.734
CAMP-TUM I	0.6911	0.580-0.787	0.8863	0.779-0.959	0.5319	0.407-0.67	0.7540	0.649-0.846	0.5483	0.364-0.742
Minsk Team	0.6890	0.568-0.804	0.7693	0.568-0.804	0.6236	0.507-0.783	0.7423	0.604-0.855	0.5683	0.348-0.768
VISILAB I	0.6532	0.551-0.748	0.7756	0.671-0.878	0.5535	0.412-0.673	0.6807	0.572-0.775	0.5908	0.428-0.776
VISILAB II	0.6513	0.549-0.742	0.7696	0.662-0.873	0.5549	0.413-0.674	0.6765	0.564-0.766	0.5942	0.432-0.779
Anonymous	0.6277	0.530-0.717	0.7420	0.629-0.838	0.5344	0.421-0.631	0.6364	0.531-0.734	0.6079	0.472-0.732
LIB	0.5561	0.434-0.654	0.8153	0.687-0.91	0.3449	0.219-0.49	0.6051	0.467-0.724	0.4450	0.258-0.650

eMethods

CAMELYON16 evaluation metrics.

In the lesion-based evaluation, a lesion was deemed to be identified if the location of the identified region was within the annotated reference standard lesion. If there were multiple findings for a single reference standard region, only the detection with the highest likelihood was considered while the lower likelihood findings were not considered false positives. All detections that were not within a specific distance (75 µm) from the reference standard annotations were counted as false positives.

In practice, there can be multiple small tumor regions that lie in the proximity of each other. Pathologists, however, consider all of these clusters as a single region. Therefore, it is important to consider them as a single lesion for the evaluation. We followed the guideline described by Cserni et al.[14] for merging these regions. Regions that were two or five cells apart (~75µm) were considered as a single entity. Subsequently, we used the following steps to obtain the evaluation masks: (1) Applying distance transform on the inverse binary mask of reference standard, (2) Thresholding the distance transformed image (T=154), (3) Labeling the connected components in the binary image. The resulting evaluation mask was a labeled image in which different tumor regions received different unique labels. This evaluation mask was used for the computation of the FROC curve.

Method descriptions.

This section contains the descriptions of all methods that were submitted to the CAMELYON16 challenge, excluding two teams (Anonymous and NSS) that did not submit sufficient details to be included in this section (The scores and ranking of all teams including these two teams are provided in Table 2. For brevity and improved readability, each method is presented in a standardized and formatted fashion. All methods follow a similar workflow: 1) The whole-slide images are preprocessed, 2) A machine learning model for detection of tumor regions is trained, 3) The machine learning model is used to produce a tumor probability map for the slide and 4) The probability map is post-processed to produce tumor lesion locations and scores, and a score for the entire slide. This general workflow is reflected in the structure of the method description. The **Introduction** section highlights key aspects of the method. The **Preprocessing** section contains the description of the steps that were taken to separate the tissue regions in the slides from the non-relevant background and standardize the tissue appearance (e.g. by performing staining normalization). The **Deep learning framework** section, which is relevant only for methods that use deep learning as the underlying methodology, contains details regarding the neural network architecture, data sampling policy and optimization procedure that was used to train the models. The methods that are based on conventional machine learning approaches have an analogous **Classification framework** section that describes the classification and feature extraction techniques that were used. The **Metastasis identification task** section describes the steps that were taken to compute the locations of the lesions in the whole-slide images along with corresponding probability scores. Finally, the **Whole-slide image classification task** section describes the steps that were taken to compute the probability score for the whole-slide image.

METHOD 1

Team name: Minsk Team

Authors: Vassili Kovalev, Alexander Kalinovsky, and Vitali Liauchuk

Affiliation: Department of Biomedical Image Analysis, United Institute of Informatics Problems, Belarus National Academy of Sciences, Surganova St., 6, 220012 Minsk, Belarus

Email: vassili.kovalev@gmail.com

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: two separate CNNs for different scanner types and two iterations of hard-negative mining.

Preprocessing

- Tissue detection: Color thresholding and morphological operations
- Preprocessing magnification: Image level 7 (pixel size = 31.1×31.1 μm^2)
- Staining normalization: None, separate systems were trained for images from different labs

Deep learning framework

Architecture:

- 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: 256×256
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: 150,000 positive and 150,000 negative
- Patch sampling strategy: Two iterations of hard-negative mining were performed. The training set was expanded with patches from regions of non-tumor tissue that the system was initially misclassifying as metastasis.
- Data augmentation: None

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Random sampling from a uniform distribution
- Batch size: 32
- Batch normalization[15]: Yes
- Regularization: 50% dropout[16] in final layers
- Learning rate: Initialized at 0.01 and decreased to 1.0e-5 with exponential decay
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 280,000 iterations

Metastasis identification task

1. The probability map, generated at level 7, was thresholded at 0.99 and post-processed with morphological filtering.
2. Connected components were extracted.
3. Components smaller than 5 pixels were removed.
4. The morphological skeletons of the remaining connected components were extracted.

5. The center of gravity of the component was calculated.
6. The point on the skeleton closest to the center of gravity was selected as the lesion coordinate.
7. The lesion score was calculated as: $\min(\frac{\text{region size}}{30}, 1)$.

Whole-slide image classification task

A histogram of the probability map was calculated. Subsequently, a logistic regression model was trained to map this histogram to a probability value for the entire image.

Results

This method achieved an FROC true positive fraction score of 0.227 for task 1 and an AUC of 0.689 (95% CI, 0.568 - 0.804) for task 2. The method ranked 23rd and 28th in the first and the second leaderboards, respectively.

METHOD 2

Team name: Radboudumc

Authors: David Tellez

Affiliation: Radboud University Medical Center Nijmegen, Geert Grootteplein-Zuid 10, 6525GA Nijmegen, The Netherlands

Email: David.TellezMartin@radboudumc.nl

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: augmentation with Gaussian blurring and mapping of the tumor probability maps to slide level scores with a second-stage CNN model.

Preprocessing

- Tissue detection: Color thresholding
- Preprocessing magnification: Image level 2 (pixel size = 0.97×0.97 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 15-layer VGG-like[6] network

Patch sampling:

- Patch size: 256×256
- Level: 2 (pixel size = 0.97×0.97 μm^2)
- Number of training samples: 150,000 positive and 150,000 negative
- Patch sampling strategy: Patches were sampled uniformly from positive and negative regions. Normal patches were sampled from negative slides as well as non-metastatic regions in tumor slides.
- Data augmentation: Rotation, vertical and horizontal mirroring and random Gaussian blurring

Parameters:

- Optimization method: ADAM[18]
- Weight initialization: Xavier's method[19]
- Batch size: 16
- Batch normalization[15]: Yes
- Regularization: L_2-regularization (1.0e-6) and 50% dropout[16]
- Learning rate: Exponential learning rate decay when the validation accuracy plateaued for 2,000 iterations
- Activation function: Leaky ReLu[20]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 20,000 iterations

Metastasis identification task

1. The probability maps were eroded and subsequently thresholded.
2. Connected components were extracted from the thresholded probability map.
3. Multiple points were uniformly sampled per region as lesion detection points.
4. The lesion probability was calculated as the mean probability of the pixels inside the connected component.

Whole-slide image classification task

A separate CNN, with the same architecture as the one trained for localizing metastases, was trained taking as input the probability map at low resolution to directly predict whether the slide contains metastasis or not.

Results

This method achieved an FROC true positive fraction score of 0.575 for task 1 and an AUC of 0.779 (95% CI, 0.694 - 0.860) for task 2. The method ranked 7th and 17th in the first and the second leaderboards, respectively.

METHOD 3 & 4
Team name: HMS & MIT (I & II)

Authors: Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew Beck

Affiliation: Harvard Medical School (BIDMC) and Massachusetts Institute of Technology (CSAIL), USA

Email: dwang5@bidmc.harvard.edu

Introduction

Two methods were submitted. Both methods are based on deep convolutional neural networks (CNNs). Key aspects include: feature-based post-processing to compute lesion and slide scores and a separately trained model with hard-negative samples. The main difference between the first and second methods are the use of a whole-slide image stain standardization algorithm[21] and more comprehensive data augmentation strategy in the second method.

Preprocessing

- Tissue detection: Conversion to the HSV color space[22] and subsequent Otsu thresholding[23] for each channel. Final tissue mask is achieved by combining the individual channel masks.
- Preprocessing magnification: Image level 5 (pixel size = 7.8×7.8 μm^2)
- (Method I) Staining normalization: None
- (Method II) Staining normalization: Whole-slide image color standardizer (WSICS)[21]

Deep learning framework

Architecture:

- 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: 224×224
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: Two million for each class
- Patch sampling strategy: Patches were sampled uniformly from positive and negative regions. Hard-negative mining was performed after initial classification to augment the training set.
- (Method I) Data augmentation: Rotation, random cropping
- (Method II) Data augmentation: Rotation, random cropping and addition of color noise

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Random sampling from a Gaussian distribution
- Batch size: 32
- Batch normalization[15]: No
- Regularization: L_2-regularization (0.0005) and 50% dropout[16]
- Learning rate: 0.01, multiplied by 0.5 every 50,000 iterations
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 300,000 iterations

Metastasis identification task

1. Obtain probability maps from the initial model (the model without hard-negative mining) and the model with hard-negative mining.

2. Threshold the probability map of the initial model at 0.9.
3. Extract connected components.
4. Take the center point of each connected component as the lesion location.
5. The lesion probability score is calculated as the sum the values in that region in both probability maps.
6. (Method II only) Each lesion score is additionally weighted by the slide-based score (obtained from the whole-slide image classification task).

Whole-slide image classification task

A set of global and local features were calculated for the entire slide. The global features are:

- The ratio between the area of metastatic regions and the tissue area
- The sum of all cancer metastases probabilities detected in the metastasis identification task, divided by the tissue area

These global features were calculated at 5 different thresholds (0.5, 0.6, 0.7, 0.8 and 0.9) resulting in 10 global features.

Local features were calculated based on the two largest metastatic candidate regions at a threshold of 0.5. In total 9 features per region were calculated resulting in a total of 18 features. The local features are:

- Area: The area of connected region
- Eccentricity: The eccentricity of the ellipse that has the same second-moments as the region
- Extend: The ratio of region area over the total bounding box area
- Bounding box area
- Major axis length: The length of the major axis of the ellipse that has the same normalized second central moments as the region
- Max/mean/min intensity: The maximum/mean/minimum probability value in the region
- Aspect ratio of the bounding box
- Solidity: Ratio of region area over the surrounding convex area

Using the 28-length feature vectors a random forest classifier[12] was trained to assign the slide level score.

Results

The first method (HMS & MIT I) achieved an FROC true positive fraction score of 0.693 for task 1 and an AUC of 0.923 (95% CI, 0.855 - 0.977) for task 2. This method ranked 5th on both leaderboards. The second method (HMS & MIT II) achieved an FROC true positive fraction score of 0.807 for task 1 and an AUC of 0.993 (95% CI, 0.983 - 0.999) for task 2. This method ranked 1st on both leaderboards.

METHOD 5

Team name: ExB

Authors: Christian Hass, Elia Bruni

Affiliation: ExB Research and Development

Email: bruni@exb.de

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of the ResNet[4] architecture and varying class balance during training.

Preprocessing

- Tissue detection: The image was divided into 5×5 tiles. For each tile, if the mean color difference between different RGB channels was lower than a threshold, the tile was considered as background.
- Preprocessing magnification: Image level 6 (pixel size = 15.6×15.6 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 34-layer ResNet[4]

Patch sampling:

- Patch size: 256×256
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: 1.6 million
- Patch sampling strategy: Training was started with a balanced sampling between the positive and negative class. As the training proceeded the distribution of positive/negative samples was slowly changed to match the original distribution in the images.
- Data augmentation: Rotation and mirroring

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: MSRA initialization[24]
- Batch size: 16
- Batch normalization[15]: Yes
- Regularization: L_2 regularization (0.0001)
- Learning rate: Initial learning rate of 0.01, which was reduced to 0.001 after 40,000 iterations, and to 0.0001 after 60,000 iterations
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 100,000 iterations

Metastasis identification task

1. Threshold the probability map at level 0 and remove small positive areas (< 834 pixels at a threshold of 0.4) from the probability map
2. Perform regional non-maxima suppression
3. Extract the center of gravity of the remaining regions

4. The lesion score for each region is computed as the maximum probability within the region

Whole-slide image classification task

The slide score was computed as the maximum score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.511 for task 1 and an AUC of 0.916 (95% CI, 0.858 - 0.962) for task 2. The method ranked 10th and 6th in the first and the second leaderboards, respectively.

METHOD 6

Team name: HTW-Berlin

Authors: Jonas Annuscheit, Peter Hufnagl

Affiliation: HTW Berlin, Berlin, Germany

Email: Jonas.Annuscheit@student.htw-berlin.de

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of a conditional random field as recurrent neural network[11] on top of a fully convolutional network[25] and the use of a pre-trained network for initialization of weights.

Preprocessing

- Tissue detection: The difference between the red and green channels from the RGB color space was thresholded to identify tissue regions.
- Preprocessing magnification: Image level 4 (pixel size = 3.9×3.9 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- CRFasRNN[11]

Patch sampling:

- Patch size: 440×440
- Level: 4 (pixel size = 3.9×3.9 μm^2)
- Number of training samples: 400,000
- Patch sampling strategy: All patches were uniformly sampled from positive slides.
- Data augmentation: Rotation and mirroring

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Pre-trained model trained on the Pascal VOC12 dataset[26]
- Batch size: 1
- Batch normalization[15]: Yes
- Regularization: 50% dropout[16]
- Learning rate: Fine-tuning using a learning rate of $6.0e^{-13}$
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 3 epochs

Metastasis identification task

1. The probability map was thresholded
2. The center of gravity of each region was considered as the lesion location and the probability value at that location was taken as the lesion score.
3. For low probability regions the surrounding area was scanned and the highest probability was assigned as likelihood for that region to be a metastasis.

Whole-slide image classification task

The slide score was computed as the maximum lesion score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.187 for task 1 and an AUC of 0.768 (95% CI, 0.665 - 0.853) for task 2. The method ranked 25th and 18th in the first and the second leaderboards, respectively.

METHOD 7

Team name: NLP LOGIX

Authors: Matt Berseth

Affiliation: NLP LOGIX, LLC.

Email: matt.berseth@nlplogix.com

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of ADAM optimization[18], computation of lesion and slide scores with second-stage random forest classifiers[12] and use of GrabCut[27] and watershed transform[28] for lesion segmentation.

Preprocessing

- Tissue detection: The image was divided into non-overlapping 256×256 tiles. Patches that had fewer than 500 unique colors or where the most frequently occurring RGB color code made up more than 90% of the patches pixels were considered background.
- Preprocessing magnification: Image level 0 (pixel size = 0.24×0.24 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 7-layer AlexNet[1]

Patch sampling:

- Patch size: 256×256
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: 250,000
- Patch sampling strategy: 15% positive patches, 85% negative patches
- Data augmentation: Rotation and mirroring

Parameters:

- Optimization method: ADAM[18]
- Weight initialization: Random sampling from a truncated normal distribution
- Batch size: 50
- Batch normalization[15]: Yes
- Regularization: 50% dropout[16]
- Learning rate: 0.0001
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 50,000 iterations (the training was stopped when the validation loss stopped decreasing)

Metastasis identification task

1. Probability map was segmented with GrabCut[27] and watershed segmentation[28].
2. The center of gravity of each region was considered as the location of the lesion candidate.
3. Summary statistics on cluster size and probability distribution were fed to a random forest classifier[12] to determine the lesion score.

Whole-slide image classification task

Summary metrics from all lesion candidate clusters were calculated and fed to another random forest classifier[12] to determine the slide score.

Results

This method achieved an FROC true positive fraction score of 0.386 for task 1 and an AUC of 0.830 (95% CI, 0.742 - 0.899) for task 2. The method ranked 12th and 13th in the first and the second leaderboards, respectively.

METHOD 8

Team name: Quincy Wong

Authors: Quincy Wong

Affiliation: Independent participant

Email: qwong77@yahoo.ca

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of SegNet[8] architecture (encoder-decoder network) pre-trained with weights from VGG-16[6] and good results with only very limited additional training data.

Preprocessing

- Tissue detection: Tiles containing tissue were selected based on overall intensity value. If the value was too high the tile was considered background.
- Preprocessing magnification: Image level 1 (pixel size = 0.49×0.49 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 37-layer SegNet[8] (encoder-decoder network)

Patch sampling:

- Patch size: 480×360
- Level: 1 (pixel size = 0.49×0.49 μm^2)
- Number of training samples: Less than 1000 per class
- Patch sampling strategy: Roughly balanced sampling, manual addition of visually interesting patches
- Data augmentation: Mirroring of only manually selected visually interesting regions

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Pre-trained weights of VGG-16[6]
- Batch size: 2
- Batch normalization[15]: Yes
- Regularization: 50% drop out[16] on selected deeper/middle layers
- Learning rate: 0.001
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 50,000 epochs

Metastasis identification task

1. Candidate lesions were determined by thresholding of the probability map (threshold value of 0.98) and morphologic operations.
2. The regions were downsampled to the resolution of level 4 (pixel size = 3.9×3.9 μm^2).

3. Centroids of remaining regions were calculated. Lesions with an area below 50 pixels received a probability penalty of 0.15 for each 4 pixels below 50. Larger centroids were given a bonus but never exceeded 1.0.

Whole-slide image classification task

The slide score were computed as the maximum lesion score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.367 for task 1 and an AUC of 0.865 (95% CI, 0.789 - 0.924) for task 2. The method ranked 14th and 11th in the first and the second leaderboards, respectively.

METHOD 9

Team name: Osaka University

Authors: Seiryo Watanabe , Shigeto Seno, Yoichi Takenaka, Hideo Matsuda

Affiliation: Department of biomedical engineering, Osaka University, Japan

Email: s-wtnb@ist.osaka-u.ac.jp

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of the GoogLeNet[3] architecture and use of an averaging filter in the post-processing stage.

Preprocessing

- Tissue detection: 300×300 tiles were extracted from the image and saved to disk. If the file size (JPEG compressed) was smaller than 18KB the tile was considered background and removed. For the remaining tiles, a threshold of 200 was used on the green and blue color channels to identify background pixels.
- Preprocessing magnification: Image level 0 (pixel size = 0.24×0.24 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: 300×300
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: One million
- Patch sampling strategy: Balanced sampling from all training slides
- Data augmentation: None

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Random sampling from a Gaussian distribution
- Batch size: 24
- Batch normalization[15]: No
- Regularization: None
- Learning rate: 0.01
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 10 million iterations

Metastasis identification task

1. Probability maps were generated by first dividing the image at level 0 into non-overlapping patches of size 300×300 and classifying each patch.
2. Pixels in the probability map with a value lower than 0.1 were suppressed.
3. The probability maps were filtered with a local 3×3 averaging filter.
4. The resulting probability map was thresholded (threshold value of 0.5) and the center points of the resulting regions were considered candidate lesions.

5. The lesion scores were computed as the maximum probability value within the regions.

Whole-slide image classification task

The slide score was computed as the maximum lesion score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.347 for task 1 and an AUC of 0.732 (95% CI, 0.629 - 0.824) for task 2. The method ranked 16th and 23rd in the first and the second leaderboards, respectively.

METHOD 10
Team name: METU

Authors: Ugur Halici, Mustafa Ümit Öner

Affiliation: Departments of Electrical and Electronics Engineering, GSNAS Neuroscience and Neurotechnology, and Graduate School of Informatics, Middle East Technical University, Turkey

Email: halici@metu.edu.tr

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of a custom CNN architecture with relatively few layers yet good performance, and custom confidence filtering for post-processing.

Preprocessing

- Tissue detection: Otsu thresholding
- Preprocessing magnification: Image level 7 (pixel size = 31.1×31.1 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- Custom CNN – 2 convolutional layers and 2 fully connected layers

Patch sampling:

- Patch size: 64×64
- Level: 2 (pixel size = 0.97×0.97 μm^2)
- Number of training samples: 240,000 samples per class
- Patch sampling strategy: Negative samples were sampled only from negative slides
- Data augmentation: 48×48 random cropping from 64×64 patches

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Xavier's method[19]
- Batch size: 128
- Batch normalization[15]: No
- Regularization: L_2 regularization (0.0018)
- Learning rate: Initial learning rate was set to 0.1 and updated at 750,000 iterations to 0.01
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 1.125 million iterations

Metastasis identification task

1. The probability map was filtered with Gaussian filters and thresholded.
2. Connected components were extracted. Each connected component was considered a candidate region.
3. For each candidate lesion, the point farthest to the boundaries among points that have probability values in the interval of $[\max(P) - 0.2, \max(P)]$ was selected as representative.
4. The probability value at the representative location was taken as the lesion score.

Whole-slide image classification task

The slide score was computed as the maximum lesion score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.389 for task 1 and an AUC of 0.864 (95% CI, 0.786 - 0.927) for task 2. The method ranked 11th and 12th in the first and the second leaderboards, respectively.

METHOD 11

Team name: Warwick-QU

Authors: Muhammad Shaban[1], Talha Qaiser[2], Ruqayya Awan[1], Korsuk Sirinukunwattana[2], Yee-Wah Tsang[2], and Nasir Rajpoot[2]

Affiliation: [1]Department of Computer Science and Engineering, College of Engineering, Qatar University

[2]Department of Computer Science, University of Warwick, England

Email: muhammad.shaban@qu.edu.qa

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of a CNN model in the preprocessing stage to segment the tissue regions and use of a U-NET-like[7] architecture for lesion segmentation.

Preprocessing

- Tissue detection: Fully convolutional CNN
- Preprocessing magnification: Image level 2 (pixel size = 0.97×0.97 μm^2)
- Staining normalization: Reinhard staining normalization[29]

Deep learning framework

Architecture:

- 15-layer U-NET[7]

Patch sampling:

- Patch size: 428×428
- Level: 2 (pixel size = 0.97×0.97 μm^2)
- Number of training samples: 8,000 positive and 12,000 negative
- Patch sampling strategy: Positive patches were extracted from all metastasis annotations. Negative patches were extracted with random sampling. Spectral clustering was applied to find visually distinct patches for training for both classes.
- Data augmentation: None

Parameters:

- Optimization method: Adadelta[30]
- Weight initialization: Random initialization
- Batch size: 10
- Batch normalization[15]: No
- Regularization: 50% dropout[16]
- Learning rate: initially set to 0.001
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 120,000 iterations

Metastasis identification task

1. Two binary lesion masks were computed using two different thresholds.
2. Lesion regions with an area ratio of less than 0.2 (computed as the area ratio of the same lesion in the two thresholded masks) were removed from further consideration.
3. The lesion centroid was extracted as the lesion location.

4. The lesion score was extracted as the minimum probability within the lesion weighted by its area.

Whole-slide image classification task

The probability of the largest tumor region was used as slide probability score.

Results

This method achieved an FROC true positive fraction score of 0.305 for task 1 and an AUC of 0.796 (95% CI, 0.711 - 0.871) for task 2. The method ranked 18th and 16th in the first and the second leaderboards, respectively.

METHOD 12 & 13
Team name: CAMP-TUM (I & II)

Authors: Bharti Munjal, Amil George, Shadi Albarqouni, Stefanie Demirci, Nassir Navab

Affiliation: Technische Universitat Munchen, Computer Aided Medical Procedure (CAMP), Munich, Germany

Email: shadi.albarqouni@tum.de

Introduction

This method is based on deep convolutional neural networks (CNNs). Two submissions were made based on two different network architectures. Key aspects of the method with better performance include: use of the GoogLeNet[3] architecture, hard-negative mining and postprocessing with a random forest classifier[12] trained with region-level features.

Preprocessing

- Tissue detection: Otsu thresholding[23]
- Preprocessing magnification: Image level 3 (pixel size = 1.94×1.94 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- Method (I): 5-layer AggNet[13] (multi-scale network)
- Method (II): 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: Patches of size 33×33 (method I) and 224×224 (method II) were extracted from level 3 (pixel size = 1.94×1.94 μm^2) and level 6 (pixel size = 15.6×15.6 μm^2), respectively.
- Number of training samples: 2 million patches for method I and 240,000 patches for method II.
- Patch sampling strategy: Initially a CNN model was trained with uniformly sampled patches from level 6. Subsequently, a new CNN model was trained with patches sampled from level 3 including false positives of the first model. Normal patches were sampled from both positive and negative slides.
- Data augmentation: Rotation and flipping.

Parameters:

- Optimization method: Adaptive gradient descent (AdaGrad)[31]
- Weight initialization: Xavier's method[19]
- Batch size: 32
- Regularization: 50% dropout[16]
- Learning rate: 0.0001
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 1.1 million iterations

Metastasis identification task

For each slide, a probability map was produced using the CNN model trained with patches from level 3. Candidate metastatic regions were detected by smoothing the probability maps with a Gaussian filter and thresholding with a threshold value optimized on a validation set. For method II, further postprocessing was performed. For each candidate in the resulting probability map, the area, orientation, major/minor axis length ratio, and probability map statistics (max, min, and mean) were extracted. The final probability score for each candidate was produced with a random forest classifier[12] trained with these features.

Whole-slide image classification task

For each slide, a probability score was produced by averaging the probability values of the top three candidate metastases.

Results

The first method (CAMP-TUM I) achieved an FROC true positive fraction score of 0.184 for task 1 and an AUC of 0.691 (95% CI, 0.580 - 0.787) for task 2. This method ranked 26th and 27th on the first and second leaderboards, respectively. The second method (CAMP-TUM II) achieved an FROC true positive fraction score of 0.273 for task 1 and an AUC of 0.737 (95% CI, 0.633 - 0.819) for task 2. This method ranked 19th and 22nd on the first and second leaderboards, respectively.

METHOD 14

Team name: TAMPERE II

Authors: Kaisa Liimatainen, Kimmo Kartasalo, Mira Valkonen, Leena Latonen, Pekka Ruusuvuori

Affiliation: BioMediTech, University of Tampere, Finland

Email: kaisa.liimatainen@tut.fi

Introduction

This method is based on deep convolutional neural networks (CNNs). A key aspect of this method is the use of a VGG-like[6] CNN model.

Preprocessing

- Tissue detection: Otsu thresholding[23] applied to the S component of the HSV color space[22] and morphological operations to remove spurious regions.
- Preprocessing magnification: Image level 5 (pixel size = 7.8×7.8 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 7-layer VGG-like[6] architecture with five convolutional and two fully connected layers

Patch sampling:

- Patch size: 32×32
- Level: 5 (pixel size = 7.8×7.8 μm^2)
- Number of training samples: 32,000 from both classes
- Patch sampling strategy: Normal patches were uniformly sampled from both negative and positive slides.
- Data augmentation: Translation and flipping

Parameters:

- Optimization method: Stochastic gradient descent with momentum[32]
- Weight initialization: Random sampling from a uniform distribution
- Batch size: 16
- Regularization: None
- Learning rate: Initial learning rate of 0.01 with momentum of 0.9
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 20,000 iterations

Metastasis identification task

For each slide, a probability map at image level 6 was produced with the trained CNN model. Candidate metastatic regions were detected by max-filtering of the probability map with 3×3 kernel, thresholding, and removal of small connected components. The remaining connected components in the thresholded probability map were considered candidate regions with probability scores equal to the maximum probability value within the region.

Whole-slide image classification task

The slide score was computed as the maximum lesion score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.252 for task 1 and an AUC of 0.713 (95% CI, 0.612 - 0.801) for task 2. The method ranked 21st and 26th in the first and the second leaderboards, respectively.

METHOD 15 & 16
Team name: VISILAB (I & II)

Authors: M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Gloria Bueno

Affiliation: VISILAB, E.T.S.I.I, University of Castilla-La Mancha, Ciudad Real, Spain

Email: Gloria.Bueno@uclm.es

Introduction

This method is based on a random forest classifier[12] using texture features. The authors performed a comparative analysis with a CNN-based method (method II).

Preprocessing

- Tissue detection: Color thresholding
- Preprocessing magnification: Level 6 (pixel size = 15.55×15.55 μm^2)
- Staining normalization: None

Classification framework:

Classifier:

- Method I: Random forest classifiers[12] with 50 decision trees
- Method II: 3-layer CNN with two convolutional layers

Features for metastasis identification (method I):

- Haralick texture features[33]

Features for whole-slide image classification (method I):

- Morphometric features: Area, convex area, convex hull, Euler number, extent, fill area, major axis length, minor axis length, perimeter, solidity.
- Geometric features: Bounding box, centroid, eccentricity, equivalent diameter, orientation, extrema.

Patch sampling:

- Patch size: For method I, patches of size 400×400 from level 0 were extracted. These patches were resampled to a size of 40×40 for the CNN used in method II.
- Number of training samples: 90,000 positive samples and 8.5 million negative samples
- Patch sampling strategy: Uniform sampling. Normal patches were sampled from both negative and positive slides.

Parameters (method II):

- Optimization method: Stochastic gradient descent with momentum[32]
- Weight initialization: Xavier's method[19]
- Batch size: 64
- Regularization: L_2 regularization (0.0005)
- Learning rate: 0.0005 with momentum 0.9
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 480 iterations

Metastasis identification task

For method I, for each slide, all non-overlapping regions of size 400×400 from level 0 were classified as metastasis or non-metastasis using the random forest classifier and Haralick texture features. For method II, the probability maps were generated using the trained CNN. In both methods, the resulting probability map was thresholded with a threshold value of 0.7, and post-processed with morphological operators to connect neighboring regions using a dilation operation with a disk-shaped structuring element with radius 10.

Whole-slide image classification task

For each region, several morphometric and geometric features (in the probability map) were extracted such as: area, bounding box, centroid, convex area, convex hull, eccentricity, equivalent diameter, Euler number, extent, filled area, major axis length, minor axis length, orientation, perimeter, extrema and solidity (features from the MATLAB *regionprops* function). Subsequently, for each slide, these region-based features were summarized by calculating the mean, standard deviation, sum, minimum, maximum, median, mode, variance, covariance, kurtosis and skewness of each feature. Finally an SVM classifier was used to compute a score for each slide.

Results

The first method (VISILAB I) achieved an FROC true positive fraction score of 0.142 for task 1 and an AUC of 0.653 (95% CI, 0.551 - 0.748) for task 2. This method ranked 29th on both leaderboards. The second method (VISILAB II) achieved an FROC true positive fraction score of 0.116 for task 1 and an AUC of 0.651 (95% CI, 0.549 - 0.742) for task 2. This method ranked 31st and 30th on the first and second leaderboards, respectively.

METHOD 17 & 18

Team name: U of Toronto (I & II)

Authors: Oren Kraus

Affiliation: University of Toronto, Electrical and Computer Engineering, Canada

Email: oren.kraus@mail.utoronto.ca

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: use of multiple CNN models trained at different magnification levels and use of learned deconvolutional layers for upsampling. Two different approaches for merging the results from multiple CNNs were investigated, which resulted in two submissions.

Preprocessing

- Tissue detection: Otsu thresholding[23]
- Preprocessing magnification: Image level 5 (pixel size = 7.8×7.8 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- 10-layer VGG-like[6] fully convolutional neural network[25]

Patch sampling

- Patch size: 300×300 tiles from levels 2 (pixel size = 0.97×0.97 μm^2), 3 (pixel size = 1.9×1.9 μm^2), 4 (pixel size = 3.9×3.9 μm^2) and 5 (pixel size = 7.8×7.8 μm^2).
- Number of training samples: 18,432
- Patch sampling strategy: For positive slides, one third of the patches were sampled from positive regions, one third from metastasis border regions, and one third from negative regions. The number of patches sampled from each slide was proportional to the tissue area.
- Data augmentation: Rotation, translation, and flipping

Parameters:

- Optimization method: Adam[18]
- Weight initialization: Xavier's method[19]
- Batch size: 16
- Regularization: 50% dropout[16]
- Learning rate: 0.0003
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 3.6 million iterations

Metastasis identification task

Two different approaches were used for computing probability maps. The first method took the mean across different scales as the final probability map. In the second method, the outputs of the CNNs trained at different magnification were used as inputs to another CNN model that produced a merged probability map. The final probability map was thresholded and postprocessed with morphological operators to identify positive regions. The probability score of each region was defined as the maximum value of the probability map within the region.

Whole-slide image classification task

The probability score for each whole-slide image was produced with a logistic regression classifier trained with features describing the detected metastatic regions: mean, min., max. probability score and size features.

Results

The first method (U of Toronto I) achieved an FROC true positive fraction score of 0.352 for task 1 and an AUC of 0.815 (95% CI, 0. 0.722 - 0.886) for task 2. This method ranked 15th on both leaderboards. The second method (U of Toronto II) achieved an FROC true positive fraction score of 0.382 for task 1 and an AUC of 0.762 (95% CI, 0.659 - 0.846) for task 2. This method ranked 13th and 19th on the first and second leaderboards, respectively.

METHOD 19

Team name: USF

Authors: Hady Ahmady Phoulady

Affiliation: University of South Florida, Tampa, USA

Email: parham.ap@gmail.com

Introduction

This method is based on a random forest classifier[12] using color and texture features. A key aspect of this method is the use of a lymphocyte probability map in the preprocessing step to exclude non-tumor regions.

Preprocessing

- Tissue detection: Thresholding of a lymphocyte probability map using a hierarchical multilevel thresholding method[34] to exclude non-tumor regions
- Preprocessing magnification: Level 4 (pixel size = 3.9×3.9 μm^2)
- Staining normalization: None

Classification framework:

Classifier:

- Random forest classifier[12]

Features:

- Grayscale intensity histogram
- Gray-level concurrence features[33]
- Local binary patterns[35]

Patch sampling:

- Patch size: 101×101
- Level: 1 (pixel size = $0.49\times$ 0.49 μm^2)
- Number of training samples: 500,000 positive samples and one million negative samples
- Patch sampling strategy: Patches were sampled with higher frequency in metastatic regions. Normal patches were sampled from both negative and positive slides.

Metastasis identification task

The random forest classifier was used to produce a probability map that was post-processed with Gaussian filtering and thresholded to obtain metastatic regions. Each region was assigned a probability score equal to the mean of the probability of the region.

Whole-slide image classification task

The probability score for each whole-slide image was computed as the weighted arithmetic mean (with weights 3 and 1) of the two metastatic regions with the highest probability scores.

Results

This method achieved an FROC true positive fraction score of 0.179 for task 1 and an AUC of 0.727 (95% CI, 0.611 - 0.823) for task 2. The method ranked 27th and 24th in the first and the second leaderboards, respectively.

METHOD 20
Team name: TAMPERE I

Authors: Mira Valkonen, Kimmo Kartasalo, Kaisa Liimatainen, Leena Latonen, Pekka Ruusuvuori

Affiliation: BioMediTech, University of Tampere, Finland

Email: valkonen.mira@gmail.com;

Introduction

This method is based on a random forest classifier[12] using texture features. A key aspect of this method is the use of nuclei density features.

Preprocessing

- Tissue detection: Otsu thresholding[23] applied to the S component from the HSV color space[22] and morphological operations to remove spurious regions.
- Preprocessing magnification: Image level 5 (pixel size = 7.8×7.8 μm^2)
- Staining normalization: Histogram matching

Classification framework:

Classifier:

- Random forest classifier[12] with 50 classification trees

Features:

- Gray-level concurrence features[33]
- SIFT descriptors[36]
- Local binary patterns[35]
- Histogram of oriented gradients (HOG)[37]
- The independent elements of the co-variance matrix of the ellipses fitted to the extracted maximally stable extremal regions (MSER)[38] and number of MSER regions.
- All texture features were extracted from both the hematoxylin and eosin channels obtained with color deconvolution[39].
- Nuclei density descriptors (mean inter-nuclei distance and number of nuclei) from watershed-based nuclei segmentation.

Patch sampling:

- Patch size: 200×200
- Level: 5 (pixel size = 7.8×7.8 μm^2)
- Number of training samples: 200,000 positive samples and 200,000 negative samples
- Patch sampling strategy: Normal patches were randomly sampled from both negative and positive slides, including metastatic region borders.

Metastasis identification task

For each slide, a probability map was produced with the trained random forest classifier. Candidate metastatic regions were detected by max. filtering of the probability map, thresholding, and connected component analysis. The connected components in the thresholded probability map were considered candidate regions with probability scores equal to the mean probability value within the region.

Whole-slide image classification task

The slide score was computed as the maximum score within the slide.

Results

This method achieved an FROC true positive fraction score of 0.257 for task 1 and an AUC of 0.761 (95% CI, 0.662 - 0.837) for task 2. The method ranked 20th on both leaderboards.

METHOD 21 & 22

Team name: SMART IMAGING (I & II)

Authors: Vitali Khvatkov, Alexei Vylegzhanin

Affiliation: Smart Imaging Technologies Co., US

Email: vitali.khvatkov@simagis.us

Introduction

This team submitted two methods for evaluation. The first method uses a conventional machine learning approach, while the second method is based on a combination of deep learning and conventional machine learning using handcrafted features. Key aspects include: multiscale analysis, use of nuclei density features and use of the GoogLeNet[3] architecture. The proposed solution is available on the Simagis Live platform (http://web-pathology.net).

Preprocessing

- Tissue detection: Ensemble of SVM classifiers with 27 color and texture features
- Preprocessing magnification: Image level 1 (pixel size = 0.49×0.49 μm^2)
- Staining normalization: Transform color coordinates to modified HSV color space[22] as follows, (1) transform color to HSV space; (2) shift Hue by 160 by subtracting/adding 160 from/to H value , (3) trim white color by removing pixels with V (value) above 0.9 threshold, (4) cluster to 3 phase system by K-mean clustering of pixel colors in modified HSV color system. Normalized images have been used in all detection/classification steps of the algorithm.

Classification framework (method I):

Classifier:

- Ensemble of SVM classifiers[9] used to classify patches at 3 different resolutions.
- Candidate regions produced by the SVM classifiers were further processed using multiscale cascade of AdaBoost[10] models.

Features:

- Combination of rotation-invariant local binary patterns[35] and color features (features selection and optimization was done using the *caret* package in R).

Patch sampling:

- Patch sampling strategy: Normal patches were sampled from both positive and negative slides. The patch sizes for the different resolutions are given in the table below.

Level	Pixel Size (μm)	Patch Size (pixels)	Patches from "tumor" class	Patches from "negative" class
Level 1	0.24×0.24 μm^2	16x16	12630	11449
Level 2	0.49×0.49 μm^2	64x64	6296	7907
Level 3	1.90×1.90 μm^2	128x128	6296	7907

Deep learning framework (method II)

Architecture:

- 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: 128×128 from
- Level: 1 (pixel size = 0.49×0.49 μm^2)
- Number of training samples: 14,000
- Patch sampling strategy: Equal number of positive and negative patches were randomly sampled. Normal patches were sampled from both negative and positive slides.
- Data augmentation: Translation, rotation, and flipping

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Xavier's method[19]
- Batch size: 128
- Regularization: 50% dropout[16]
- Learning rate: 0.01
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 420,000 iterations

Metastasis identification task

The probability maps for the first method were computed using the Adaboost classifier. The probability maps for the second method were produced with a combination of the Adaboost and CNN classifiers. Candidate lesions were localized with the geographic clustering algorithm[40]. Geographic clustering algorithm identified geographic clusters of tiles and center of cluster on slide. The composite probability for each candidate was computed as weighted mean of probability of tiles in the cluster. The weights for each cluster member (tile) was computed using the measure of "compactness".

Whole-slide image classification task

The slide scores were produced with a SVM classifier that uses features that summarize the distribution of the candidate lesions in the slide.

Results

The first method (SMART IMAGING I) achieved an FROC true positive fraction score of 0.208 for task 1 and an AUC of 0.757 (95% CI, 0. 0.663 - 0.839) for task 2. This method ranked 24th and 21st on the first and second leaderboards, respectively. The second method (SMART IMAGING II) achieved an FROC true positive fraction score of 0.339 for task 1 and an AUC of 0.821 (95% CI, 0.753 - 0.894) for task 2. This method ranked 17th and 14th on the first and second leaderboards, respectively.

METHOD 23, 24 & 25

Team name: CULab (I, II & III)

Authors: Hao Chen, Huang-Jing Lin, Qi Dou, and Pheng-Ann Heng

Affiliation: Department of Computer Science and Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong

Email: jackie.haochen@gmail.com

Introduction

This method is based on deep convolutional neural networks (CNNs). There are three submissions by this team, each employing a different CNN architecture. A key aspect of the best performing method is the use of a fully convolutional architecture for dense predictions.

Preprocessing

- Tissue detection: Color thresholding
- Preprocessing magnification: Image level 5 (pixel size = 7.8×7.8 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- Method I: VGG-16[6]
- Method II: Cascade of two CNNs[41]. The first CNN (VGG-16) works with lower magnification images (level 1), has very high sensitivity and quickly eliminates many negative regions. The second CNN, a 152-layer ResNet architecture[4], refines the results from the first model.
- Method III: Fully convolutional network adapted from VGG-16[6] for dense predictions

Patch sampling:

- Patch size: 224×224 for second CNN of method II; 244x244 for other networks.
- Level: 0 (pixel size =0.24×0.24 μm^2) used for method III and the second CNN of method II. Level 1 (pixel size = 0.49×0.49 μm^2) used for method I and the first CNN of method II.
- Number of training samples: 15 million (5% positive)
- Patch sampling strategy: Uniform sampling
- Data augmentation: Translation and flipping

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Pre-trained with the ImageNet dataset[5]
- Batch size: 10 for ResNet-152, and 50 for the other architectures
- Regularization: L_2 regularization (0.0005)
- Learning rate: Initially set at 0.001 and decreased by a factor of 10 every 100,000 iterations.
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 300,000 iterations

Metastasis identification task

For each slide, a probability map was produced using the trained CNN model (at level 0 with a stride of 32). Candidate metastatic regions were detected by filtering the probability map with a median filter (kernel size of 3×3) and thresholding. Each connected component in the resulting probability map was considered a candidate detection with a probability score equal to the maximum probability value within the region. This procedure was used for all three methods.

Whole-slide image classification task

The slide score was computed as the maximum score within the slide.

Results

The first method (CULab I) achieved an FROC true positive fraction score of 0.544 for task 1 and an AUC of 0.909 (95% CI, 0.851 - 0.954) for task 2. This method ranked 8th and 7st on the first and second leaderboards, respectively. The second method (CULab II) achieved an FROC true positive fraction score of 0.527 for task 1 and an AUC of 0.906 (95% CI, 0.841 - 0.957) for task 2. This method ranked 9th on both leaderboards. The third method (CULab III) achieved an FROC true positive fraction score of 0.703 for task 1 and an AUC of 0.942 (95% CI, 0.888 - 0.980) for task 2. This method ranked 4th on both leaderboards.

METHOD 26

Team name: DeepCare

Authors: Tong Xu

Affiliation: DeepCare Inc.

Email: txu@deepcare.com

Introduction

This method is based on deep convolutional neural networks (CNNs). Key aspects include: the use of the pre-trained GoogLeNet[3] architecture and a second-stage SVM classifier for computing slide scores.

Preprocessing

- Tissue detection: Multi-thresholding in the HSV[22] color space.
- Preprocessing magnification: Image level 3 (pixel size $= 1.94\times1.94\ \mu m^2$)
- Color normalization: None

Deep learning framework

Architecture:

- 22-layer GoogLeNet[3]

Patch sampling:

- Patch size: 256×256
- Level : 0 (pixel size = $0.24\times0.24\ \mu\text{m}^2$)
- Number of training samples: 700,000
- Patch sampling strategy: Patches were uniformly sampled from positive and negative regions. Negative samples were taken from both positive and negative slides. For positive slides, additional negative samples were taken from regions bordering metastatic regions.
- Data augmentation: Mirroring and rotation of the positive samples

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Pretrained GoogLeNet model with the ImageNet dataset[5]
- Batch size: 64
- Batch normalization[15]: Yes
- Regularization: L_2 regularization (0.0005)
- Learning rate: Initialized at 0.01 and decreased every 100,000 iterations by a factor of 0.1
- Activation function: ReLu[17]
- Loss function: Cross-entropy
- Number of training epochs/iterations: 120,000 iterations

Metastasis identification task

Using the trained GoogLeNet model, a probability map was generated for each slide. Candidate regions were produced with connected component analysis. Regions with an area smaller than 20 pixels were rejected as false positives. The lesion scores were computed as the mean of the probability values within the region. The center of gravity and the probability score of the lesions with a probability higher than 0.85 were reported.

Whole-slide image classification task

For each whole-slide image, five binary masks containing metastatic connected components were generated by applying multiple thresholds of 0.5, 0.6, 0.7, 0.8 and 0.9 on the probability map. Subsequently, two types of features including 5 shape and 3 statistics-based probability features were extracted from the five multi-thresholded regions. These features include:

- Area
- Eccentricity
- Major and minor axis length of the ellipse that has the same normalized second central moments as the region
- Ratio of pixels in the region to the pixels in the total bounding box
- The mean, maximum, and variance of the probability values inside the multi-thresholded regions of each candidate

Overall, a 40-dimensional feature vector was extracted from each candidate region. The normalized 40-dimensional feature vectors were then fed into a SVM classifier[9] to discriminate between tumor and non-tumor regions. The trained SVM classifier was used to discriminate annotated metastases in positive slides from candidate findings in negative whole-slide images. The probability score for the whole-slide images were computed as the weighted mean of the detected tumor regions present in the whole-slide images.

Results

This method achieved an FROC true positive fraction score of 0.243 for task 1 and an AUC of 0.883 (95% CI, 0.806 - 0.943) for task 2. The method ranked 22nd and 10th on the first and second leaderboards, respectively.

METHOD 27
Team name: LIB

Authors: R. Venâncio, B. Ben Cheikh, A. Coron, and D. Racoceanu

Affiliation: Sorbonne Universiteś, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale (LIB), Paris, France

Email: rui.venancio.t@gmail.com

Introduction

This method is based on a SVM[9] classifier using color and texture features for automated detection of metastatic cancer from whole-slide images of sentinel lymph nodes.

Preprocessing

- Tissue detection: *K*-means clustering
- Preprocessing magnification: Level 4 (pixel size = 3.9×3.9 μm^2)
- Staining normalization: Reinhard staining normalization[29]

Classification framework:

Classifier:

- SVM[9]

Features:

- Haralick texture features[33]
- Law's texture energy measures[42]
- Features were selected with sequential forward selection

Patch sampling:

- Patch size: 800×800
- Level: 0 (pixel size = 0.24×0.24 μm^2)
- Number of training samples: 1,100 positive and 1,130 negative
- Patch sampling strategy: Patches were sampled uniformly from positive and negative regions. Negative samples were taken from both positive and negative slides.

Metastasis identification task

The slides were divided in rectangular patches. Each patch was classified as positive or negative. The centroids of regions larger than four connected positive patches were selected as candidate lesion locations. The lesion scores were calculated as the mean of the probability values of all patches within the region.

Whole-slide image classification task

If the number of positive patches in a whole-slide image was larger than 11, the mean of the probabilities of all positive patches was calculated and reported as the whole-slide score.

Results

This method achieved an FROC true positive fraction score of 0.120 for task 1 and an AUC of 0.556 (95% CI, 0.434 - 0.654) for task 2. The method ranked 30th and 32nd on the first and second leaderboards, respectively.

METHOD 28, 29 & 30

Team name: HMS-MGH (I, II & III)

Authors: Aoxiao Zhong, Quanzheng Li

Affiliation: Gordon Center for Medical Imaging, Clinical Data Science Center, Harvard Medical School, Massachusetts General Hospital

Email: zhongaoxiao@gmail.com

Introduction

Three methods were submitted. The first two submissions are similar to the methods of the Harvard & MIT team, based on patch-wise classification using GoogLeNet[3] and ResNet-101[4], respectively. The third submission is based on dense prediction using fully convolutional ResNet-101 architecture with atrous convolution and atrous spatial pyramid pooling[43].

Preprocessing

- Tissue detection: Otsu thresholding[23]
- Preprocessing magnification: Image level 5 (pixel size = 0.78×0.78 μm^2)
- Staining normalization: None

Deep learning framework

Architecture:

- Method I: GoogLeNet[3]
- Method II: ResNet-101[4]
- Method III: Fully convolutional ResNet-101 architecture with atrous convolution and atrous spatial pyramid pooling (deeplab v2[43])

Patch sampling:

- Patch size: 224×224 for methods I and II, and 512×512 for method III
- Level: 0 (pixel size = 0.24×0.24 μm^2) for methods I and II, and level 1 (pixel size = 0.49×0.49 μm^2) for method III
- Number of training samples: 400,000 with 25% of positive patches for methods I and II. On-line sampling with approximately 25% positive samples for method III.
- Patch sampling strategy: Negative patches were sampled from both negative slides and normal regions in positive slides.
- Data augmentation: Mirroring and random cropping for all methods

Parameters:

- Optimization method: Stochastic gradient descent
- Weight initialization: Pre-trained model with the ImageNet dataset[5] for methods I and II, and pre-trained model with the MS-COCO dataset[44] for method III.
- Batch size: 64 for method I, 128 for method II, 10 for method III
- Regularization: L_2 regularization was used for all methods. The regularization coefficients were 0.0002, 0.0001, and 0.0005 for methods I, II and III, respectively.
- Learning rate: The learning rate was initialized at 0.001 and divided by 10 when the error plateaued for method I and II. The learning rate was initialized at 2.5e-4 and multiplied by 0.9 every 40,000 iterations for method III.
- Activation function: ReLu[17]
- Loss function: Cross-entropy

- Number of training epochs/iterations: 150,000 iterations for method I, 180,000 iterations for method II and 40,000 iterations for method III

Metastasis identification task

1. Perform connected component analysis of the thresholded probability map (the threshold was set to 0.9 for methods I and II and 0.95 for method III).
2. The centroids of the connected components were used as candidate location.
3. The mean probability values of the connected components were used as the lesion scores.
4. Regions with major-axis length smaller than 200 µm were removed as false positives.

Whole-slide image classification task

Higher level features were extracted from the tumor heatmaps (computed using the *regionprops* function in *skimage*[45]) with thresholds of 0.5 and 0.9 for methods I and II, and thresholds of 0.5 and 0.95 for method III. All these features are computed for the largest detected candidate in the whole-slide image:

- The major axis length
- The ratio between the area of the candidate region and the total bounding box area
- Eccentricity of the ellipse that has the same second-order moments as the region
- Total area
- Mean intensity

A random forest classifier[12] was trained with these features and subsequently used to produce the probability score for each slide.

Results

The first method (HMS-MGH I) achieved an FROC true positive fraction score of 0.596 for task 1 and an AUC of 0.965 (95% CI, 0.928 - 0.989) for task 2. This method ranked 6th and 3rd on the first and second leaderboards, respectively. The second method (HMS-MGH II) achieved an FROC true positive fraction score of 0.729 for task 1 and an AUC of 0.908 (95% CI, 0.846 - 0.961) for task 2. This method ranked 3rd and 8th on the first and second leaderboards, respectively. The third method (HMS-MGH III) achieved an FROC true positive fraction score of 0.760 for task 1 and an AUC of 0.976 (95% CI, 0.941 - 0.999) for task 2. This method ranked 2nd on both leaderboards.

eResults

Stratification according to metastasis size and primary tumor histotype in task 2

The pathologists' results were further analyzed in two subcategories: analysis according to metastasis size and primary tumor histotype (eTable 1 and eTable 2 in the Supplement). Pathologist without time constraint achieved a better sensitivity and AUC for detecting macrometastases (sensitivity of 100% and AUC of 0.994 (95% CI, 0.977-1.0)) and metastases originating from infiltrating ductal carcinoma (IDC) (sensitivity of 97.0% (95% CI, 89.7%-100%) and AUC of 0.976 (95% CI, 0.932-1.0)) compared to micrometastases (sensitivity of 88.8% (95% CI, 75.0%-100%) and AUC of 0.943 (95% CI, 0.868-0.995)) and non-IDC cases (sensitivity of 86.6% (95% CI, 66.7%-100%) and AUC of 0.943 (95% CI, 0.848-1.0)), respectively (no statistically significant difference for comparison of AUCs, p=0.87 (Bonferroni corrected) for comparison of the performance for the detection of micro and macrometastases, and p>0.99 for comparison of the performance for the detection of IDC and non-IDC cases). For all 11 pathologists in the simulated routine diagnostic setting, the performance was significantly higher (See eTable 2 for individual p-values) for detection of macrometastases (mean sensitivity of 92.9% (95% CI, 90.5%-95.8%) and mean AUC of 0.964 (range, 0.924-1.0)) compared to micrometastases (mean sensitivity of 38.3% (95% CI, 32.6%-52.9%) and mean AUC of 0.685 (range, 0.582-0.808)). We also observed that metastases originating from IDC (mean sensitivity of 69.2% (95% CI, 65.4%-77.4%) and mean AUC of 0.842 (range, 0.773-0.924)) were more often detected compared to non-IDC cases (mean sensitivity of 48.4% (95% CI, 43.2%-59.7%) and mean AUC of 0.738 (range, 0.656-0.862)) (but not significantly, see eTable2 for the p-values for each pathologist).

The top-ranking systems performed similarly to the best performing pathologists in detecting macrometastases. The performance of the algorithms in detecting micrometastases, however, was considerably more variable. Many of the top-ranked algorithms achieved better AUCs than the best pathologist in the panel of 11 (best pathologist AUC = 0.808 (95% CI, 0.704-0.908) versus best algorithm AUC = 0.997 (95% CI, 0.989-1.0)) in detecting micrometastases. The AUC of the two leading algorithms (AUC = 0.997 (95% CI, 0.989-1.0) and 0.957 (95% CI, 0.893-0.999), respectively) even surpassed that of the pathologist without time constraint (AUC = 0.9430 (95% CI, 0.868-0.995)).

With regard to the primary tumor histotype, the majority of the algorithms had higher AUCs for detecting IDC metastases than metastases of other types. The top-four performing algorithms achieved higher AUCs than the panel of 11 pathologists in detecting metastases of both IDC and non-IDC histotypes (see eTables 2 and eTable5).

eDiscussion

Potential reasons for large variability in CNN performance

The modest performance of some of the algorithms based on convolutional neural networks (CNN), in many cases, could be attributed to choosing a low magnification to process the slide, or selecting a very small patch size for training. Consequently, the system either lacks the detailed information present in the higher magnifications or loses the contextual information that could be captured by a larger patch size. Despite using the right magnification, patch size and state-of-the-art CNN architectures, achieving satisfactory results can be challenging. Training deep learning models can involve many hyperparameter settings (e.g. learning rate, regularization strength, mini-batch size, etc.). Successful and efficient training and debugging of large scale CNNs requires careful selection and adjustment on these hyperparameters, and finding out the relation between hyperparameters and validation errors.

Properties of the top-performing algorithms

We can summarize the main properties of the high-ranked teams based on 4 main characteristics: network architecture, patch-sampling strategy, preprocessing and data augmentation, and network ensemble.

One common property of the leading teams is that they all used very deep state-of-the-art CNN architectures such as GoogLeNet[3], VGG-Net[6], and ResNet[4]. The leading team, HMS & MIT (II), trained a 22-layer GoogLeNet model and enriched the training data by adding false positive findings produced by an initial model. By doing this, the network becomes more knowledgeable on recognizing the more difficult normal regions. The CNNs used in systems HMS & MGH (III), HMS & MGH (II), and CULab (III) were ResNet-101, GoogLeNet-22 and VGG-Net-16, respectively, all initialized by weights from pre-trained networks and fine-tuned with the challenge data. ResNet-101 was pre-trained on the MS-COCO dataset[44] and the other two models were pre-trained on the large scale 1000-class ImageNet dataset[5]. The high performance of these methods is in accordance with previous studies which have validated the efficacy of transfer learning strategies[46-48]. Some of the key factors contributing to the outstanding performance of the HMS & MIT (II) system were the use of the whole-slide image color standardizer (WSICS) algorithm[21] to normalize the appearance of whole-slide images, and the incorporation of a more rigorous data augmentation strategy including rotation, flipping, random cropping, and the addition of random offsets to each RGB color channel. The ResNet-101 model used in the system of HMS & MGH (III) used very large image patches of size 512×512 that were more than double the input size of all the other systems used in this challenge. On top of that, the use of atrous convolution (dilated convolution) and spatial pyramid pooling[43] enabled the system to capture objects as well as image context at multiple scales.

Another factor contributing to the success of some of the top-ranking algorithms is the use of network ensembles. The winning team used an ensemble of a network trained on standardized whole-slide images and a network trained on original whole-slide images to report the probabilities for each finding. The first submission of this team, HMS & MIT (I), ranking fifth, used an ensemble of two networks (networks trained before and after hard-negative mining).

To generate a slide based score for the second task, the majority of the teams assigned the maximum probability among the detected lesions in the whole-slide image as the confidence score for that slide. Prior to this assignment, they mostly removed small areas of positive findings, and/or applied Gaussian/median filtering. Although this approach worked well for many of the teams, including CU-Lab (III) and ExB research that were ranked fourth and sixth in the image classification task, it may not take into account metastases characteristics (e.g. slides containing multiple high-score findings or slides containing larger metastases could have increased chance of containing metastases). In contrast, the systems HMS & MIT (I & II) used a random forest classifier employing a variety of geometrical and morphological features extracted from each probability map. Details of these features can be found in **eMethods**. The use of a learning-based algorithm to produce a confidence score from a whole-slide image probability map is likely the centerpiece of this algorithm that makes it the top-performing system for the first task.

Finally, one interesting property of the top-performing system HMS & MIT (II) in the metastasis identification task is that it uses the output of the discriminative classifier that produces a slide-based confidence score, to weigh the score of each finding in the second task. This top-down analysis reduces the number of false-positives, particularly in normal slides.

eReference

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in neural information processing systems2012.
2. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 2013;35(8):1915-1929.
3. Szegedy C, Wei L, Yangqing J, et al. Going deeper with convolutions. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 7-12 June 2015, 2015.
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016, 2016.
5. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 2015;115(3):211-252.
6. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.* 2014.
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention2015.
8. Kendall A, Badrinarayanan V, Cipolla R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680.* 2015.
9. Cortes C, Vapnik V. Support-vector networks. *Mach. Learn.* 1995;20(3):273-297.
10. Viola P, Jones M. Fast and robust classification using asymmetric adaboost and a detector cascade. Paper presented at: Advances in Neural Information Processing Systems2002.
11. Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. Paper presented at: Proceedings of the IEEE International Conference on Computer Vision2015.
12. Breiman L. Random forests. *Mach. Learn.* 2001;45(1):5-32.
13. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging.* 2016;35(5):1313-1321.
14. Cserni G, Bianchi S, Boecker W, et al. Improving the reproducibility of diagnosing micrometastases and isolated tumor cells. *Cancer.* 2005;103(2):358-367.
15. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167.* 2015.
16. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 2014;15(1):1929-1958.
17. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Paper presented at: Proceedings of the 27th International Conference on Machine Learning2010.
18. Kingma D, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 2014.
19. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Paper presented at: aistats2010.
20. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. Paper presented at: ICML Workshop on Deep Learning for Audio, Speech and Language Processing2013.
21. Ehteshami Bejnordi B, Litjens G, Timofeeva N, et al. Stain specific standardization of whole-slide histopathological images. *IEEE Trans. Med. Imaging.* 2016;35(2):404-415.

22. Joblove GH, Greenberg D. Color spaces for computer graphics. *SIGGRAPH Comput. Graph.* 1978;12(3):20-25.
23. Otsu N. A threshold selection method from gray-level histograms. *Automatica.* 1975;11(285-296):23-27.
24. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Paper presented at: Proceedings of the IEEE International Conference on Computer Vision2015.
25. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2015.
26. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* 2015;111(1):98-136.
27. Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts. Paper presented at: ACM transactions on graphics (TOG)2004.
28. Meyer F. Topographic distance and watershed lines. *Signal processing.* 1994;38(1):113-125.
29. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Computer Graphics and Applications.* 2001;21(5):34-41.
30. Zeiler MD. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701.* 2012.
31. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 2011;12(Jul):2121-2159.
32. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Cognitive modeling.*5(3):1.
33. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 1973;SMC-3(6):610-621.
34. Ahmady Phoulady H, Goldgof DB, Hall LO, Mouton PR. Nucleus segmentation in histology images with hierarchical multilevel thresholding. Paper presented at: Proceedings of SPIE Medical Imaging2016.
35. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002;24(7):971-987.
36. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 2004;60(2):91-110.
37. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition2005.
38. Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing.* 2004;22(10):761-767.
39. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* 2001;23(4):291-299.
40. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining1996.
41. Chen H, Dou Q, Wang X, Qin J, Heng PA. Mitosis detection in breast cancer histology images via deep cascaded networks. Paper presented at: Proceedings of the thirtieth AAAI Conference on Artificial Intelligence2016.
42. Laws KI. Rapid Texture Identification. Paper presented at: Proceedings of SPIE Medical Imaging1980.

43. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915.* 2016.
44. Lin T, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. Paper presented at: European Conference on Computer Vision2014.
45. Van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ.* 2014;2:e453.
46. Bengio Y. Deep learning of representations for unsupervised and transfer learning. Paper presented at: Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop2011.
47. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Paper presented at: Advances in neural information processing systems2014.
48. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops2014.