Challenge for Diagnostic Assessment of Deep Learning Algorithm for Metastases Classification in Sentinel Lymph Nodes on Frozen Tissue Section Digital Slides in Women with Breast Cancer

Young-Gon Kim
Asan Medical Center
In Hye Song
Seoul st. Mary's Hospital
Hyunna Lee
Asan Medical Center
Dong Hyun Yang
Asan Medical Center
Namkug Kim
University of Ulsan, College of Medicine, Asan Medical Center
Dongho Shin
Kakao brain
Yeonsoo Yoo
KaKao brain
Kyowoon Lee
Ulsan National Institute of Science and Technology
Dahye Kim
Chung-Ang University
Hwejin Jung
DoAI
Hyunbin Cho
DoAI
Hyungyu Lee
Lady Doak College
Taeu Kim
Sogang University
Jong Hyun Choi
Korea University
Changwon Seo
DoAI

Seong il Han
Korea Advanced Institute of Science and Technology
Young Je Lee
Yonsei University
Young Seo Lee
Ewha Womans University
Hyung-Ryun Yoo
University of Kwangwoon
Yongju Lee
Seoul National University
Jeong Hwan Park
Seoul National University
Gyungyub Gong (gygong@amc.seoul.kr)
Asan Medical Center
Sohee Oh
Seoul National University

Research article

Keywords: Breast Neoplasms, Deep Learning, Frozen Sections, Neoplasm Metastasis, Sentinel Lymph Node

Posted Date: February 10th, 2020

DOI: https://doi.org/10.21203/rs.2.23087/v1

Abstract

Assessing the status of metastasis in sentinel lymph nodes (SLNs) by pathologists is an essential task for the accurate staging of breast cancer. However, histopathological evaluation of sentinel lymph nodes by a pathologist is not easy and is a tedious and time-consuming task. The purpose of this study is to review a challenge competition (HeLP 2018) to develop automated solutions for the classification of metastases in hematoxylin and eosin–stained frozen tissue sections of SLNs in breast cancer patients. A total of 297 digital slides were obtained from frozen SLN sections, which include post–neoadjuvant cases (n = 144, 48.5%) in Asan Medical Center, South Korea. The slides were divided into training, development, and validation sets. All of the imaging datasets have been manually segmented by expert pathologists. A total of 10 participants were allowed to use the Kakao challenge platform for six weeks with two P40 GPUs. The algorithms were assessed in terms of the AUC (area under receiver operating characteristic curve). The top three teams showed 0.986, 0.985, and 0.945 AUCs for the development set and 0.805, 0.776, and 0.765 AUCs for the validation set. Micrometastatic tumors, neoadjuvant systemic therapy, invasive lobular carcinoma, and histologic grade 3 were associated with lower diagnostic accuracy. In a challenge competition, accurate deep learning algorithms have been developed, which can be helpful in making frozen diagnosis of intraoperative sentinel lymph node biopsy. Whether this approach has clinical utility will require evaluation in a clinical setting.

Introduction

Recently, implementation of digital pathology has been rising because of workforce crisis and increased need of consultation and collaboration. Digital pathology has many advantages in terms of time saving, slide storage, remote working, and second-opinion practice, and is becoming a part of routine procedure in diverse areas such as primary diagnosis, multidisciplinary clinic, and frozen section diagnosis[1]. Owing to rapid progress of technology, machine learning techniques using digital histopathological images have been investigated and showed satisfactory results in the detection of tumor areas and lymph node metastases in prostate, lung, and breast cancers[2 – 4].

Breast cancer is the most common cancer in women, accounting for approximately one-third of all cancers in women globally. For patients with localized breast cancer, the treatment of choice is surgical removal of the primary tumor[5]. In order to reduce disease recurrence or metastasis, lymph node sampling or dissection should be performed during surgery. Because axillary lymph node dissection may cause morbidity, such as arm-lymphedema and nerve injury, sentinel lymph node sampling is recommended in order to determine the nodal metastases status and if extensive lymph node dissection is required[6 – 9]. Although some recent studies suggested that the role of sentinel lymph node biopsy has been diminished in early breast cancer patients[10 – 13], sentinel lymph node sampling is still considered important due to its cost– and time– effectiveness and usually performed intraoperatively using the frozen section technique and which allows surgeons to make immediate decisions during surgery[14]. However, pathologists frequently experience problems while making diagnoses of frozen sections.

First, frozen section diagnosis should be made as quickly as possible in order to minimize the waiting time for surgeons which can cause surgical and anesthetic complications. The turnaround time of the frozen section diagnosis is usually kept less than 20 to 30 minutes, including the gross examination, tissue cutting and staining, and the microscopic examination[15]. Second, microscopic examination of a frozen section is more difficult than that of a conventional section because of inferior quality of the sections due to the frozen artifact. There are also components, such as capillaries, histiocytes, and germinal centers, in lymph nodes and which can be mistaken for metastatic carcinoma. Furthermore, frozen section diagnosis is extremely difficult in some patients who have underwent neoadjuvant systemic therapy before surgery. In order to overcome such difficulties, the deep learning algorithm might be helpful. For example, the 'CAncer MEtastases in LYmph nOdes challeNge' (CAMELYON16 and CAMELYON17) competitions disclosed that 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[4,16]. However, digital slides which were used in most of those previous studies had not been created from frozen tissue sections, but from formalin fixed paraffin embedded (FFPE) tissue sections. To our best knowledge, there has not been any reported study using frozen tissue section of SLNs until the present time. In addition, the previous studies did not include post–neoadjuvant cases, which has been increasing but difficult to histologically examine[17].

In the challenge competition originating from the HeLP (HEalthcare ai Learning Platform), several models have been developed. In this challenge setting, we aimed to evaluated the models' performances for classification of metastases per slide in hematoxylin and eosin–stained frozen tissue sections of SLNs of breast cancer patients.

Materials And Methods

Data Description

During routine surgical procedure for breast cancer in our institution, the excised sentinel lymph nodes were immediately submitted for frozen section. All of the sentinel lymph nodes were cut into 2-mm slices, entirely embedded in optimum cutting temperature (OCT) compound, and frozen in -20 to -30 For each lymph node, 5μm-thick frozen sections were cut and one or two sections were picked up on glass slides and stained with hematoxylin and eosin (H&E). In this study, a total of 297 digital slides of sentinel lymph nodes from 132 patients were retrospectively collected. Among those, 144 slides were made from sentinel lymph nodes of patients who had received neoadjuvant therapy (48.5%). The slides were divided into a training set, a development set and a validation set (157, 40, and 100 digital slides, respectively). Patient demographics are summarized in Table 1. The slides were scanned using a digital microscopy scanner (Pannoramic 250 FLASH; 3DHISTECH Ltd., Budapest, Hungary) in MIRAX format (.mrxs) and with a resolution of 0.221μm per pixel. The institutional review board for human investigations at Asan Medical Center (AMC) approved the study protocol with removal of all patient identifiers from the images and they waived the requirement for informed consent, in accordance with the retrospective design of this study.

Reference Standard

All the imaging datasets were segmented manually by one rater, and their annotations were confirmed by two clinically expert pathologists with six and 20 years' experience in breast pathology. Regions of metastatic carcinoma larger than 200μm in the greatest dimension were annotated as cancer with the in-house labeling tool, as shown in Figure 1.

Challenge Competition Environment

The challenge competition platform developed by Kakao was used to allocate two GPUs to each team. All of the competitors were allowed to access only paths of digital slides and corresponding mask images with Kakao platform. Docker image files that enables any of deep learning platform to run were used to train models and inference development and validation sets. Each team was given two P40 GPUs (NVIDIA, Santa Clara, CA, USA) resources for training models. Kakao platform used CUDA 9,0 and cuDNN 7.

During the first stage for four weeks, competitors were given 197 digital slides as the training and development set for four weeks. The training set (157 digital slides) with annotated masks was given for training the model, while the development set (40 digital slides) without masks was given for tuning the model. Model performance calculated by the evaluation matrix was listed on the leader board after inferencing the development set which was used for tuning the model. During the second stage for additional two weeks, the competitors were given 100 additional digital slides for final evaluation of their models with the optimal model derived from the development set.

Evaluation Metric

The algorithms were assessed for classifying between "metastasis" or "normal". AUC (area under receiver-operating characteristic curve) was evaluated by receiver-operating characteristic (ROC) analysis.

Competitors

Forty-five competitors who were interested in digital pathology or machine learning registered for this challenge within four weeks from the beginning of November 2018. Ten competitors were selected according to their inner commitments in accordance with the limited platform environment. Ten competitors were composed of students, researchers, and doctors experienced in medical image analysis using machine learning or deep learning. Only five competitors submitted their results on the leaderboard. The methodologic description is summarized in Table 2. All of the competitors selected only deep learning as the main architecture such as Inception v3[18] and Inception-ResNet[19] for classification of the tumor patch or U-Net[20] for segmentation of the tumor region. In one team which ranked high, random forest regression[21] was used to inference confidence by extracting high level features including the number of tumor regions, percentage of the tumor region over the entire tissue region, the area of the

largest tumor regions, etc., from the heat map generated using the deep learning method. Detailed descriptions of each algorithm are listed in Table 2.

Results

Model performances were sorted in descending order for the validation set as shown in Figure 2 and Table 3. Five teams submitted their results on the leader board in development and validation sets. For the development set, the top three algorithms showed 0.986, 0.985, and 0.945 AUC, while the others showed approximately 0.55 AUCs. For the validation set which consisted of 100 digital slides, the Fiffeb team showed the highest AUC 0.805 in the validation set compared with other teams such as the DoAI, GoldenPass, SOG, and Aron Research teams at AUC 0.776, 0.760, 0.540, and 0.470, respectively. Average times of the first three teams (Fiffeb, DoAI, and GoldenPass) in validation set were 10.8, 0.6, and 3.9 minutes, respectively.

For more detailed analysis, each algorithm was evaluated with the cutoff threshold determined by the Youden index[22] from the ROC curve in the validation set in terms of the accuracy (ACC), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV). The first placed team Fiffeb showed the highest AUC (0.805), ACC (0.770), TNR (0.822), and PPV (0.833), while the second placed team DoAI showed the highest TPR (0.800) and NPV (0.738).

In addition, model performance comparisons with clinical information for more detail, such as the metastatic tumor size (smaller or larger than 2mm in the greatest dimension), whether patients had received neoadjuvant systemic therapy, histologic type of tumor, and the histologic grade of the tumor was measured, as shown in Table 4. Four of the five teams showed higher TPR and lower FNR in lymph nodes with larger metastatic tumors. In lymph nodes obtained from patients who had received neoadjuvant systemic therapy, four teams showed lower TPR and two teams showed lower TNR. In terms of the histologic type, three teams showed higher TPR and four teams higher TNR in the invasive lobular carcinoma group than in the invasive ductal carcinoma group. When comparing performance between the histologic grades, four teams showed higher TPR, but only one team showed higher TNR in grade 1 or 2 than in grade 3.

Among the 100 slides in the validation set, 57 slides were correctly categorized by all top three teams (35 slides, true positive; 22 slides, true negative), four slides were incorrectly categorized as positive (false positive) by the top three teams, and six slides were incorrectly categorized as negative (false negative) by the top three teams, as shown in Figure 3. All of the four false positive slides were obtained from patients with invasive ductal carcinoma, histologic grade 2, and two slides were from neoadjuvant systemic therapy patients. Similarly, all of the six false negative slides were obtained from patients with invasive ductal carcinoma, i.e. five from histologic grade 2 patients and one from a histologic grade 3 patient, and three were from neoadjuvant systemic therapy patients. Four of the six false negative slides had micrometastases. The size range of metastatic carcinoma in the false negative slides was 0.13 mm to 4.45 mm

Discussion

In this current study, all of the competitors adopted CNN (convolutional neural network)-based deep learning methods as the main idea such as the classification or segmentation network, and which showed high performance at 0.805, 0.776, and 0.760 in terms of AUC for the top three teams.

Interestingly, in all five teams, AUC were lower in the validation set compared to that in the development set. This might be due to the difference in patient demographics, particularly with regard to neoadjuvant systemic therapy. In the validation set, the number of slides obtained from patients after neoadjuvant systemic therapy was significantly higher than that in the development set, as shown in Table 1. Neoadjuvant systemic therapy often causes fibrosis and macrophage infiltration in the tumor area and fragmentation and/or scattering of tumor clusters[17], and which can lead to difficulty in histologic examination. It might be suggested that this neoadjuvant systemic therapeutic effect caused a decrease of AUC in the validation set.

Inference time is also key point with this challenge so that methods can be adopted in routine clinical practice. Two different types of patch-based CNN methods, classification and segmentation network, have shown pros and cons. The number of outputs of the classification network in this challenge is same with the number of classes that the model classifies input patch into (i.e. 1 or 2) by encoding all input dimensions to compressed features for a precise decision. In case of segmentation network, the number of outputs is same with the number of input dimensions (i.e. 448448=200,704), which is approximately 100K or 200K times more than that of classification network. It is a factor reducing computational time. In our results, the first placed team using only classification network showed 0.3 higher AUC than that of the second placed team using only segmentation network, but too slow to deploy this into the real clinical routine while the computational time of the second placed team took 18.8 times faster than that of the first placed team. Ensemble of those different types of CNN networks should be considered to enhance model performance in routine clinical practice.

Next, we compared model performances according to the clinicopathologic factors of the patients. It is generally known that in manual examination of intraoperative sentinel lymph node biopsy, false negative results are more likely in micrometastases and favorable and/or lobular histology[23]. In the validation set, the top three teams showed better performances in lymph nodes with macrometastatic tumor, and which is consistent with manual examination and the CAMELYON16 study[4]. Lymph nodes which were obtained from non-neoadjuvant systemic therapy patients also revealed better performances, as discussed above. Lymph nodes from invasive ductal carcinoma patients revealed better TPR in all top three teams and better TNR in the top one team than those from invasive lobular carcinoma patients, although the number of slides from invasive lobular carcinoma patients is limited. This is in accordance with the general results in manual examination and the CAMELYON16 study. In the CAMELYON16 study, 29 among 32 teams showed higher AUC in the invasive ductal carcinoma set than in the non-invasive ductal carcinoma set. In addition, tumors of histologic grade 1 or 2 showed higher TPR in the top three teams, but lower TNR in two of the three teams than tumors of histologic grade 3, and which requires further studies.

We found that some cases were wrongly categorized by the first three teams. All of six false negative cases showed small-sized metastatic carcinoma, and which could result in false negativity. In contrast, four false positive cases did not reveal any common clinicopathologic feature. However, we assume that reactive histiocytic infiltration or prominent germinal centers in lymph nodes might cause false positivity. Manual confirmation is probably necessary, and so a screening tool that would expedite this process might have broad appeal.

Our study has some strong significance compared to previously reported studies about possible usefulness of deep learning algorithm in diagnosis of sentinel lymph node metastasis[4,16]. First, we used digital slides from frozen sections which were made intraoperatively, while previous studies used FFPE sections. Since frozen sections have lower quality due to tissue artifact compared with FFPE sections, it is more difficult to examine frozen sections than FFPE sections. However, what is used to determine the surgical extent intraoperatively in the real world is frozen sections, not FFPE sections. Therefore, we suggest that studies of the deep learning algorithm with sentinel lymph nodes would be more practical if frozen sections are used. Second, our dataset includes a high proportion (48.5%) of post–neoadjuvant patients. The role of neoadjuvant therapy in breast cancer treatment has been increasing these days, but it is much more difficult to histologically diagnose sentinel lymph node metastasis after neoadjuvant therapy[17]. During case selection, we included more post–neoadjuvant cases than clinical setting with an intention of making our dataset unique and more useful. To reduce false positive or false negative issues technically, the deep learning models should be re-trained with those regions and different hyper-parameters such as class weights or loss weights. Those regions with different hyper-parameters have deep learning models intensively trained as strong positive regions with this strategy. Applications using these methods can be adopted in routine clinical practice by showing attention map with augmented reality and training itself robustly with false positive cases selected by pathologists with on-line learning.

Our contest has several limitations. First, only paths to access the training, development, and validation sets were given to competitors, which means that they had no way to check the heat map generated by their models as all dataset contests provided were not available in public. Competitors were not allowed to check processing in the middle of training for the same reason. Only less than 1MB log data could be saved and given to competitors for the purpose of debugging after training processing to check if and how the training is going well. It was also not available how much time was spent for training and analyses. This might be one of key reasons of the models with relatively low accuracies. Second, only 2 GPUs were given to each competitor, and it could be limited resource, although this constraint makes competitors fair. Third, we did not perform immunohistochemistry to confirm metastatic carcinoma on frozen section slides. On the contrary to FFPE sections, multiple frozen sections which were made from the same tissue fragment showed quite different shapes due to the tissue artifact. Therefore, immunohistochemistry is not as helpful in frozen sections as in FFPE sections to annotate tumor cells. In addition, it is impossible to retrospectively perform immunohistochemistry on frozen sections. Instead, when we annotate tumor cells in frozen sections, we review matched FFPE sections with cytokeratin immunohistochemistry in order to minimize annotation error. Finally, the high proportion of post–

neoadjuvant cases or cases with micrometastases could have negatively affected the diagnostic accuracy of algorithms in this study. It would have been nicer if we could divide the dataset into multiple groups and develop different algorithms based on patients' information, such as neoadjuvant status, histologic type, or histologic grade of tumor. However, it was impossible due to the limited number of digital slides. We hope to expand our dataset and include such analysis in our further study.

Possibly because of the characteristics of our dataset and the above limitations, even the top three algorithms in this study showed relatively lower performance than the other first prized in CAMELYON16, and lower diagnostic accuracy than average of pathologists[24]. However, we believe that it is worth holding a digital pathology challenge competition using frozen tissue sections in open innovation manner. For adjusting algorithms into routine clinical practice, HeLP is preparing another challenge competition to handle other problems such as localization of micro-metastasis and processing time.

Recognition abilities of deep learning and human could be complement each other. In addition, algorithms with deep learning can be used as computer aided system to help doctors diagnose. For example, virtual reality technology can help making quack accurate decision or alert a doctor who misses critical parts.

Conclusion

We held a challenge competition during six weeks to resolve the problem for classification of digital pathology slides with metastases in hematoxylin and eosin–stained frozen tissue sections of SLNs of breast cancer patients. The top three competitor teams achieved very high AUCs in the development set while they performed slightly lower AUC in the validation set. In this open innovation manner, the deep learning algorithms could be developed and evaluated, which might be helpful in the frozen diagnosis of intraoperative, sentinel lymph node biopsy. Further studies are required in order to increase the accuracy and decrease the time consuming required to apply the deep learning algorithm in the clinical setting.

References

1 Williams, B. J., Bottoms, D. & Treanor, D. Future-proofing pathology: the case for clinical adoption of digital pathology. *Journal of clinical pathology* **70**, 1010-1018 (2017).

2 Wang, S. *et al.* Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific reports* **8**, 10393 (2018).

3 Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports* **6**, 26286 (2016).

4 Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318**, 2199-2210 (2017).

5 Harrison, T. R., Kasper, D. L. & Fauci, A. S. *Harrison's Principles of Internal Medicine 19th Ed*. (McGraw-Hill AccessMedicine, 2015).

6 Hayes, S. C., Janda, M., Cornish, B., Battistutta, D. & Newman, B. Lymphedema after breast cancer: incidence, risk factors, and effect on upper body function. *Journal of clinical oncology* **26**, 3536-3542 (2008).

7 Fleissig, A. *et al.* Post-operative arm morbidity and quality of life. Results of the ALMANAC randomised trial comparing sentinel node biopsy with standard axillary treatment in the management of patients with early breast cancer. *Breast cancer research and treatment* **95**, 279-293 (2006).

8 Lyman, G. H. *et al.* Sentinel lymph node biopsy for patients with early-stage breast cancer: American Society of Clinical Oncology clinical practice guideline update. *J Clin Oncol* **32**, 1365-1383 (2014).

9 Manca, G. *et al.* Sentinel lymph node biopsy in breast cancer: indications, contraindications, and controversies. *Clinical nuclear medicine* **41**, 126-133 (2016).

10 Galimberti, V. *et al.* Axillary dissection versus no axillary dissection in patients with breast cancer and sentinel-node micrometastases (IBCSG 23-01): 10-year follow-up of a randomised, controlled phase 3 trial. *The Lancet Oncology* **19**, 1385-1393 (2018).

11 Giuliano, A. E. *et al.* Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis: the ACOSOG Z0011 (Alliance) randomized clinical trial. *Jama* **318**, 918-926 (2017).

12 Wang, J. *et al.* Is surgical axillary staging necessary in women with T1 breast cancer who are treated with breast-conserving therapy? *Cancer Communications* **39**, 25 (2019).

13 Donker, M. *et al.* Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. *The lancet oncology* **15**, 1303-1310 (2014).

14 Celebioglu, F., Sylvan, M., Perbeck, L., Bergkvist, L. & Frisell, J. Intraoperative sentinel lymph node examination by frozen section, immunohistochemistry and imprint cytology during breast surgery–a prospective study. *European journal of cancer* **42**, 617-620 (2006).

15 Chen, Y., Anderson, K. R., Xu, J., Goldsmith, J. D. & Heher, Y. K. Frozen-Section Checklist Implementation Improves Quality and Patient Safety. *American journal of clinical pathology* (2019).

16 Bandi, P. *et al.* From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE transactions on medical imaging* **38**, 550-560 (2019).

17 Honkoop, A. H. *et al.* Effects of chemotherapy on pathologic and biologic characteristics of locally advanced breast cancer. *American journal of clinical pathology* **107**, 211-218 (1997).

18 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818-2826.

19 Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. in *Thirty-First AAAI Conference on Artificial Intelligence.*

20 Ronneberger, O., Fischer, P. & Brox, T. in *International Conference on Medical image computing and computer-assisted intervention.* 234-241 (Springer).

21 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22 (2002).

22 Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32-35 (1950).

23 Akay, C. L. *et al.* Factors impacting the accuracy of intra-operative evaluation of sentinel lymph nodes in breast cancer. *The breast journal* **24**, 28-34 (2018).

24 Houpu, Y. *et al.* Use of Memorial Sloan Kettering Cancer Center nomogram to guide intraoperative sentinel lymph node frozen sections in patients with early breast cancer. *Journal of surgical oncology* **120**, 587-592 (2019).

Declarations

Acknowledgements

This work was supported by Kakao and Kakao Brain corporations and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), and was funded by the Ministry of Health & Welfare, Republic of Korea (HI18C0022).

Author Contributions

Y.-G.K. and I.H.S. analyzed data, searched literature, generated Figures, and interpreted data. G.G., D.H.Y, H.L. and N.K. designed and supervised the study. D.S. and Y.Y. provided platform. K.L., D.K., H.J., H.C., H.L., T.K., J.H.C., C.S., S.I.H., Y.J.L., Y.S.L., H.-R.Y., Y.L., J.W.P., and S.O. experimented with algorithms. All of the authors were involved in writing the paper and had final approval of the submitted and published versions.

Additional Information

The authors declare no competing interests.

Tables

Table 1. Clinicopathologic characteristics of the patients

		Training set (n = 157)	Development set (n = 40)	Validation set (n = 100)	*P*-value*
Age (median and range)		50 (28 – 80)	49 (30 – 68)	47 (34 – 75)	
Sex	Female	157 (100%)	40 (100%)	100 (100%)	1
Metastatic carcinoma	Present, size > 2mm	68 (43.3%)	14 (35%)	40 (40%)	0.158
	Present, size ≤ 2mm	35 (22.3%)	5 (12.5%)	15 (15%)	
	Absent	54 (34.4%)	21 (52.5%)	45 (45%)	
Neoadjuvant systemic therapy	Not received	80 (51.0%)	28 (70%)	45 (45%)	**0.027**
	Received	77 (49.0%)	12 (30%)	55 (55%)	
Histologic type	IDC	149 (94.9%)	32 (80%)	86 (86%)	**0.005****
	ILC	8 (5.1%)	5 (12.5%)	11 (11%)	
	MC	0 (0%)	0 (0%)	3 (3%)	
	metaplastic carcinoma	0 (0%)	3 (7.5%)	0 (0%)	
Histologic grade	1 or 2	118 (75.2%)	34 (85%)	86 (86%)	0.074
	3	39 (24.8%)	6 (15%)	14 (14%)	

Abbreviations: IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; MC, mucinous carcinoma.

* *P*-values, calculated using the c^2 test.

** For the histologic type, a c^2 test was conducted between IDC and non-IDC.

Table 2. Algorithm descriptions and hyper parameters

Team	Architecture	Input size (Slide layer level)	Optimization (Learning rate)	Augmentation	Pre-processing	Post-processing; Inference for confidence
Fiffeb	Inception v3, RFC	2562563(6) Patch	SGD (0.9)	Color augmentation, horizontal flip, random rotation	Otsu thresholding, Tumor (>90%) and non-tumor (0% and >20%)	Generation of heat map with image level 7 and feeding morphological information into FRC; RFC output.
DoAI	U-Net	5125123(0) Patch	SGD (1e-1, decay 0.1 each 2 epochs)	Rotation, horizontal and vertical flip,	None	De-noising for false positive reduction; CNN output.
GoldenPass	U-Net, Inception v3	2562563(4) Patch	Adam (1e-3, 5e-4)	Rotation, horizontal and vertical flip, brightness (0.5~1)	Otsu thresholding, Tumor (>100%)	None; Max value for heat-map
SOG	Simple CNN	3003003(4) Slide	Adadelta (1e-3)	None	None	None; CNN output.
Aron Research	Inception-ResNet	2992993(8) Patch	Adam (1e-3, decay 0.1 each 10 epochs)	Rotation, adding noise for saturation, hue, and contrast	While pixel thresholding	Gaussian smoothing; Mean value for heat-map

Abbreviations: RFC, random forest classifier, CNN, convolutional neural network; SGD, stochastic gradient descent.

Table 3. Performance and average time (minute) comparison for classification of tumor slide

Team	AUC		ACC	TPR	TNR	PPV	NPV	Time (min.)
	Development set	Validation Set	Validation Set					
Fiffeb	**0.986**	**0.805**	**0.770**	0.727	**0.822**	**0.833**	0.712	10.8
DoAI	0.985	0.776	0.750	**0.800**	0.689	0.759	**0.738**	0.6
GoldenPass	0.945	0.760	0.730	0.782	0.667	0.741	0.714	3.9
SOG	0.595	0.540	0.510	0.145	0.956	0.800	0.478	-
Aron Research	0.525	0.470	0.450	0.000	1.000	0.000	0.450	-

Abbreviations: AUC, area under the curve; ACC, accuracy; TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value.

Table 4. Performance comparison for determining the clinicopathologic characteristics of tumors

Team	Metastatic tumor size				Neo-adjuvant therapy			
	2mm (n=33)		>2mm (n=22)		Not received (n=45)		Received (n=55)	
	TPR	FNR	TPR	FNR	TPR	TNR	TPR	TNR
Fiffeb	0.600	0.400	0.775	0.225	0.731	0.842	0.724	0.808
DoAI	0.667	0.333	0.850	0.150	0.808	0.737	0.793	0.654
GoldenPass	0.667	0.333	0.825	0.175	0.808	0.632	0.759	0.692
SOG	0.067	0.933	0.175	0.825	0.154	0.895	0.138	1.000
Aron Research	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000

Team	Histologic type						Histologic grade			
	IDC (n=86)		ILC (n=11)		MC (n=3)		1 or 2 (n=86)		3 (n=14)	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Fiffeb	0.723	0.795	0.833	1.000	0.500	1.000	0.735	0.838	0.667	0.750
DoAI	0.766	0.667	1.000	0.800	1.000	1.000	0.816	0.676	0.667	0.750
GoldenPass	0.766	0.641	1.000	0.800	0.500	1.000	0.796	0.649	0.667	0.750
SOG	0.149	0.949	0.000	1.000	0.500	1.000	0.163	0.946	0.000	1.000
Aron Research	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000

Abbreviations: MC, mucinous carcinoma; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; TPR, true positive rate; FNR, false negative rate; TNR, true negative rate

Figures

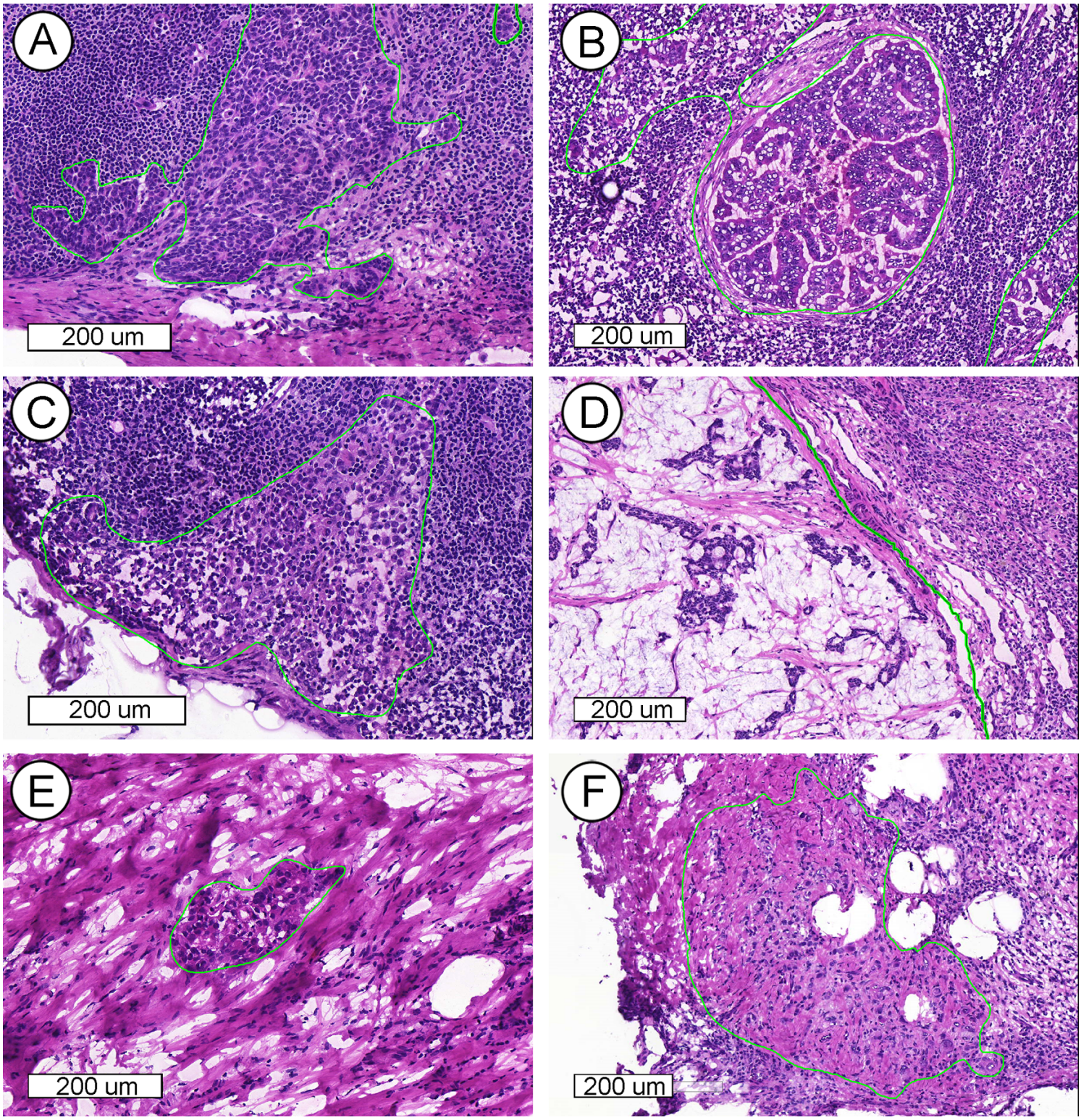

Figure 1

Representative microscopic images of various metastatic carcinomas with annotation. (A) Invasive ductal carcinoma, histologic grade 2, consists of medium-sized tumor cells with moderate glandular formation. (B) Invasive ductal carcinoma, histologic grade 3, shows large-sized tumor cells with poor glandular formation. (C) Tumor cells are small- to medium-sized and poorly cohesive in invasive lobular carcinoma. (D) Mucinous carcinoma contains abundant extracellular mucin. (E) & (F) Invasive ductal

carcinoma after neoadjuvant systemic therapy shows fragmented clusters of tumor cells (E) or singly scattered, atypical tumor cells (F) in the fibrotic background (Hematoxylin and eosin).

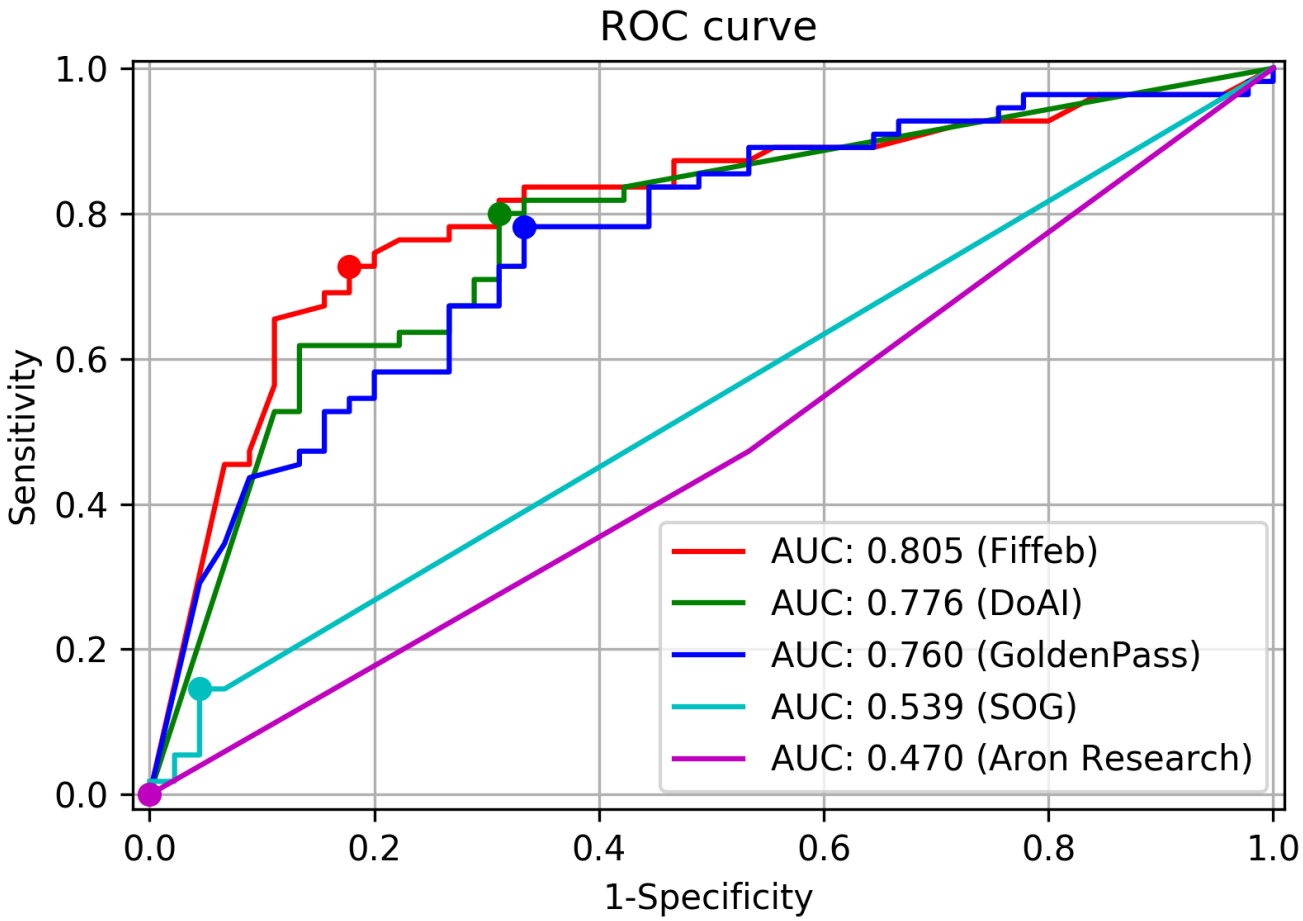

Figure 2

Receiver-operating characteristics (ROC) comparisons of models trained by five algorithms for the validation set and cutoff threshold value of each algorithm. The cutoff threshold value is dotted on each ROC curve.

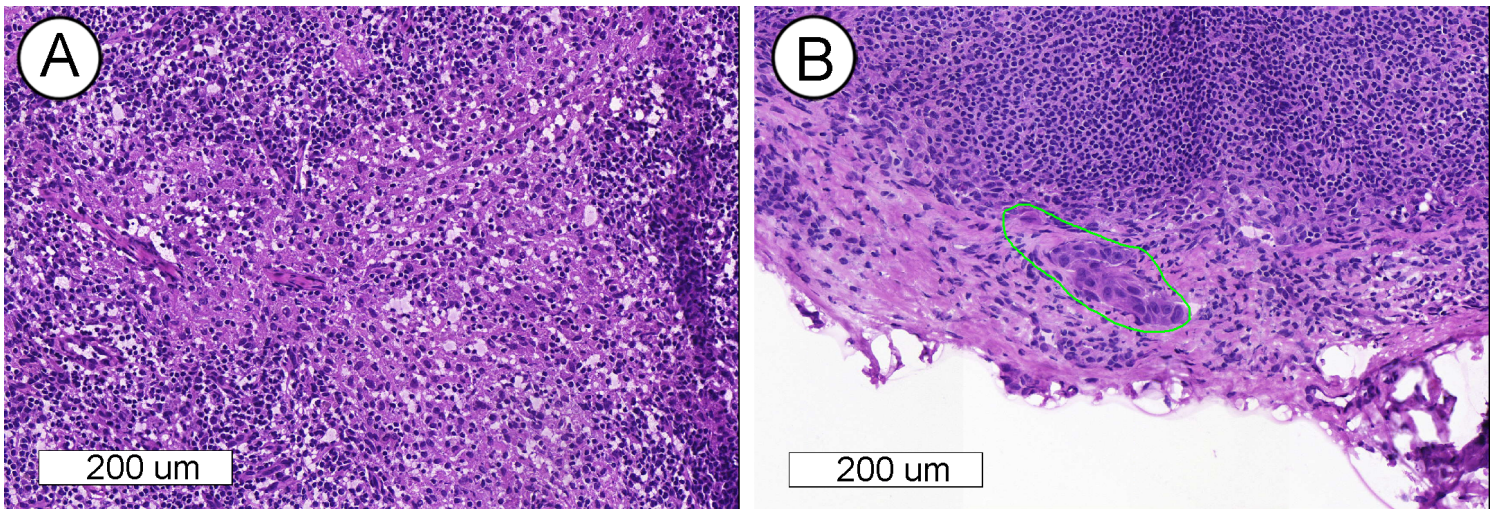

Figure 3

Representative microscopic images of false-positive (A) and false-negative (B) cases. (A) Reactive histiocytes show abundant, eosinophilic cytoplasm and can be misinterpreted as metastatic carcinoma. (B) A very small focus of metastatic carcinoma (approximately 200 μm in the greatest dimension) is seen and which was missed by all five of the teams.