pISSN 1598-2998, eISSN 2005-9256

Cancer Res Treat. 2020;52(4):1103-1111

https://doi.org/10.4143/crt.2020.337

Original Article

Open Access

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, PhD[1]
In Hye Song, MD, PhD[2]
Hyunna Lee, PhD[3]
Sungchul Kim, BS[1]
Dong Hyun Yang, MD, PhD[4]
Namkug Kim, PhD[5]
Dongho Shin, BS[6]
Yeonsoo Yoo, BS[6]
Kyowoon Lee, BS[7]
Dahye Kim, BBA[8]
Hwejin Jung, PhD[9]
Hyunbin Cho, BS[9]
Hyungyu Lee, PhD[9]
Taeu Kim, BA[10]
Jong Hyun Choi, BA[11]
Changwon Seo, MS[9]
Seong Il Han, BS[12]
Young Je Lee, BE[13]
Young Seo Lee, BA[14]
Hyung-Ryun Yoo, BS[15]
Yongju Lee, PhD[16]
Jeong Hwan Park, MD, PhD[17]
Sohee Oh, PhD[18]
Gyungyub Gong, MD, PhD[19]

**A list of author's affiliations appears at the end of the paper.*

Correspondence: Gyungyub Gong, MD, PhD
Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: 82-2-3010-4554
Fax: 82-2-472-7898
E-mail: gygong@amc.seoul.kr

Received April 20, 2020
Accepted June 29, 2020
Published Online June 30, 2020

*Young-Gon Kim and In Hye Song contributed equally to this work.

Purpose

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 SLNs 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.

Materials and Methods

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 6 weeks with two P40 GPUs. The algorithms were assessed in terms of the area under receiver operating characteristic curve (AUC).

Results

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.

Conclusion

In a challenge competition, accurate deep learning algorithms have been developed, which can be helpful in making frozen diagnosis of intraoperative SLN biopsy. Whether this approach has clinical utility will require evaluation in a clinical setting.

Key words

Breast neoplasms, Deep learning, Frozen sections, Neoplasm metastasis, Sentinel lymph node

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 (SLN) 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 SLN biopsy has been diminished in early breast cancer patients [10-13], SLN 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

Table 1. Clinicopathologic characteristics of the patients (resolution [width×height] of digital slide: 93,970×234,042)

	Training set (n=157)	Development set (n=40)	Validation set (n=100)	p-value[a]
Age (yr)	50 (28-80)	49 (30-68)	47 (34-75)	
Sex				
Female	157 (100)	40 (100)	100 (100)	> 0.99
Metastatic carcinoma				
Present, size > 2 mm	68 (43.3)	14 (35.0)	40 (40.0)	0.158
Present, size ≤ 2 mm	35 (22.3)	5 (12.5)	15 (15.0)	
Absent	54 (34.4)	21 (52.5)	45 (45.0)	
Neoadjuvant systemic therapy				
Not received	80 (51.0)	28 (70.0)	45 (45.0)	0.027
Received	77 (49.0)	12 (30.0)	55 (55.0)	
Histologic type				
IDC	149 (94.9)	32 (80.0)	86 (86.0)	0.005[b]
ILC	8 (5.1)	5 (12.5)	11 (11.0)	
MC	0	0	3 (3.0)	
Metaplastic carcinoma	0	3 (7.5)	0	
Histologic grade				
1 or 2	118 (75.2)	34 (85.0)	86 (86.0)	0.074
3	39 (24.8)	6 (15.0)	14 (14.0)	

Values are presented as median (range) or number (%). IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; MC, mucinous carcinoma. [a]p-values, calculated using the chi-square test, [b]For the histologic type, a chi-square test was conducted between IDC and non-IDC.

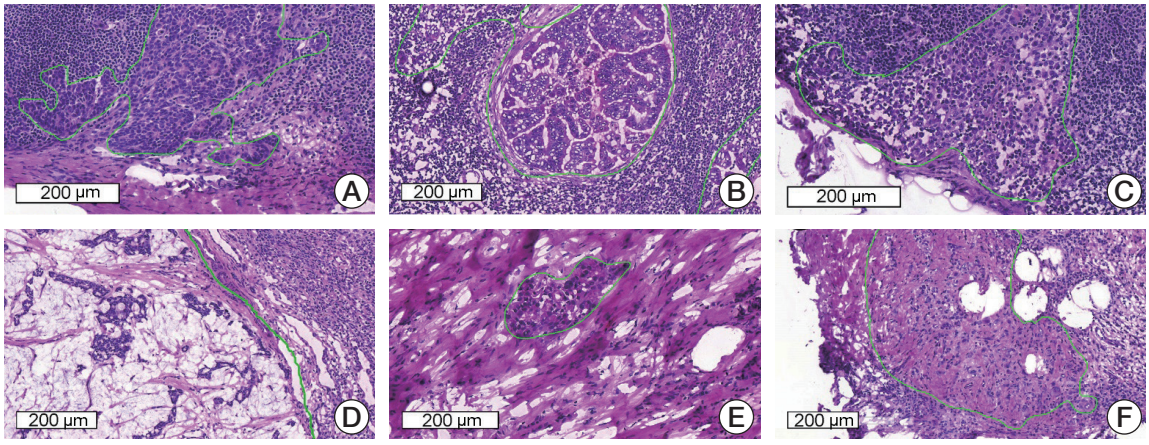

Fig. 1. Representative microscopic images of various metastatic carcinomas with annotation (H&E staining). (A) Invasive ductal carcinoma, histologic grade 2, consists of medium-sized tumor cells with moderate glandular formation. (B) Invasive ductal carcinoma, histologic grade 3, shows large-sized tumor 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.

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 evaluate 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

1. Data description

During routine surgical procedure for breast cancer in our institution, the excised SLNs were immediately submitted for frozen section. All of the SLNs were cut into 2-mm slices, entirely embedded in optimum cutting temperature compound, and frozen in –20°C to –30°C. For each lymph node, 5-μm-thick frozen sections were cut and one or two sections were picked up on glass slides and stained with hematoxylin and eosin. In this study, a total of 297 digital slides of SLNs from 132 patients were retrospectively collected. Among those, 144 slides were made from SLNs 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) without consideration of distribution of histologic type. Slides before a specific point in time were used as the training and development sets, and the other digital slides after that were used as the validation set. 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 0.221 μm per pixel.

2. Reference standard

All the imaging datasets were segmented manually by one rater, and their annotations were confirmed by two clinically expert pathologists with 6 and 20 years' experience in breast pathology. Regions of metastatic carcinoma larger than 200 μm in the greatest dimension were annotated as cancer with the in-house labeling tool, as shown in Fig. 1.

3. 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) 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 2 weeks, the

Table 2. Algorithm descriptions and hyper parameters

Team	Architecture	Input size (slide layer level)	Optimization (learning rate)	Augmentation real-time	Pre-processing	Post-processing; inference for confidence
Fiffeb	Inception v3, RFC	256×256×3 (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	512×512×3 (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	256×256×3 (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	300×300×3 (4) Slide	Adadelta (1e-3)	None	None	None; CNN output

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

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

Team	Development set AUC	Validation set AUC	Validation set					Time (min)
			ACC	TPR	TNR	PPV	NPV	
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	-

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

competitors were given 100 additional digital slides for final evaluation of their models with the optimal model derived from the development set.

4. Evaluation metric

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

5. Competitors

Forty-five competitors who were interested in digital pathology or machine learning registered for this challenge within 4 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 four competitors submitted their results on the leaderboard. The methodological description is summarized in Table 2. All of the competitors selected only deep learning as the main architecture such as Inception v3 [18] for classification of the tumor patch or U-Net [19] for segmentation of the tumor region. Instead of modifying their models, they focused on pre- and post-processing steps to achieve optimal results. In one team which ranked high, random forest regression [20] 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. Real time-based augmentation methods were adjusted while training models. Detailed descriptions of each algorithm are listed in Table 2.

6. Ethical statement

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.

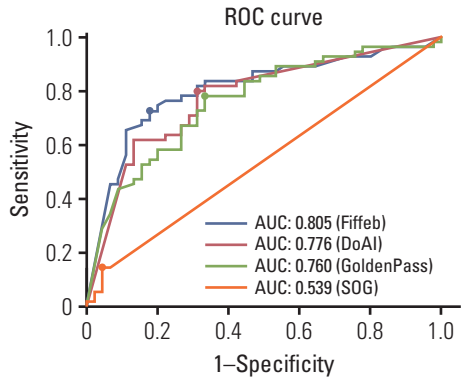

Fig. 2. Receiver operating characteristics (ROC) comparisons of models trained by four algorithms for the validation set and cut-off threshold value of each algorithm. The cutoff threshold value is dotted on each ROC curve. AUC, area under ROC.

Results

Model performances were sorted in descending order for the validation set as shown in Table 3 and Fig. 2. Four teams submitted their results on the leader board in development and validation sets. For the development set, the Four algorithms showed 0.986, 0.985, 945, and 0.595 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, and SOG teams at AUC 0.776, 0.760, and 0.540 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

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

	Team			
	Fiffeb	DoAI	GoldenPass	SOG
Metastatic tumor size				
≤ 2 mm (n=33)				
TPR	0.600	0.667	0.667	0.067
FNR	0.400	0.333	0.333	0.933
> 2 mm (n=22)				
TPR	0.775	0.850	0.825	0.175
FNR	0.225	0.150	0.175	0.825
Neo-adjuvant therapy				
Not received (n=45)				
TPR	0.731	0.808	0.808	0.154
TNR	0.842	0.737	0.632	0.895
Received (n=55)				
TPR	0.724	0.793	0.759	0.138
TNR	0.808	0.654	0.692	1.000
Histologic type				
IDC (n=86)				
TPR	0.723	0.766	0.766	0.149
TNR	0.795	0.667	0.641	0.949
ILC (n=11)				
TPR	0.833	1.000	1.000	0.000
TNR	1.000	0.800	0.800	1.000
MC (n=3)				
TPR	0.500	1.000	0.500	0.500
TNR	1.000	1.000	1.000	1.000
Histologic grade				
1 or 2 (n=86)				
TPR	0.735	0.816	0.796	0.163
TNR	0.838	0.676	0.649	0.946
3 (n=14)				
TPR	0.667	0.667	0.667	0.000
TNR	0.750	0.750	0.750	1.000

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

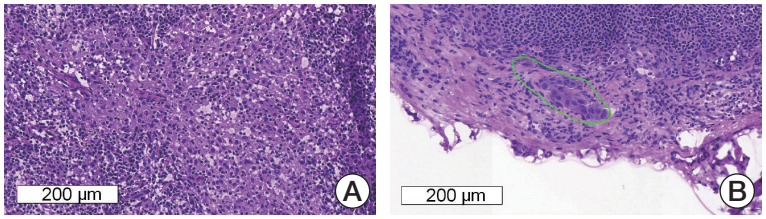

Fig. 3. Representative microscopic images of false-positive (A) and false-negative (B) cases. (A) Reactive histiocytes show abundant, eosinophilic cytoplasm and can be misinterpreted as metastatic carcinoma. (B) A very small focus of metastatic carcinoma (approximately 200 μm in the greatest dimension) is seen and which was missed by all four of the teams.

with the cutoff threshold determined by the Youden index [21] 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 2 mm 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 teams showed higher TPR and lower false-negative rate 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 Fig. 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 to 4.45 mm.

Discussion

In this current study, all of the competitors adopted convolutional neural network (CNN)–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 four teams, AUC was 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. Distribution of histologic type is different between training, development, and validation sets as shown in Table 1. Especially in the validation set, the number of slides obtained from patients after neoadjuvant systemic therapy was significantly higher than that in the development set. 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. Turnaround time between receiving samples and reporting in conventional frozen section diagnosis has been variably reported around 20-30 minutes, including gross examination, freezing, cutting, staining, and microscopic examination [22]. Time consumed for scanning can be varied upon the size of sections, type of scanning machine, magnification, and focus layering, but recent studies have reported that 3-9 minutes of median handling time for scanning [22,23]. 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., 448×448=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 SLN biopsy, false-negative results are more likely in micrometastases and favorable and/or lobular histology [24]. 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 lobular carcinoma patients revealed better TPR in the first three teams and better TNR in four teams than those from invasive ductal 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. Interestingly, TPR of mucinous carcinoma cases (0.5-1.0) was not lower than those of invasive ductal carcinoma (0.149-0.766) or invasive lobular carcinoma (0.000-1.000), although mucinous carcinoma was not included in training and validation sets. This might be due to some histologic similarities between mucinous carcinoma and other carcinomas, such as cluster formation, bigger cell size than lymphocytes, and nuclear size enlargement.

Our study has some strong significance compared to previously reported studies about possible usefulness of deep learning algorithm in diagnosis of SLN 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 SLNs 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 SLN 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 1 MB 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 two 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. Finally, the model performance can be influenced by various parameters including quality of tissue sections, staining quality and color differences, type of scanning machine, scanning environment, and accuracy of segmentation. Therefore, further studies for optimization of pre-processing of digital images might improve models' diagnostic performances.

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 [25]. 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.

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, SLN 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.

Conflicts of Interest

Conflicts of interest relevant to this article was not reported.

Acknowledgments

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), funded by the Ministry of Health & Welfare, Republic of Korea (HI18C0022).

Author Details

[1]Department of Biomedical Engineering, Asan Institute of Life Science, Asan Medical Center, University of Ulsan College of Medicine, Seoul, [2]Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, [3]Health Innovation Big Data Center, Asan Institute for Life Science, Asan Medical Center, Seoul, Departments of [4]Radiology and [5]Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, [6]KakaoBrain-BrainCloud Team, Seongnam, [7]Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan, [8]Image Laboratory, School of Computer Science and Engineering, Chung-Ang University, Seoul, [9]DoAI Inc., Seoul, [10]Department of Business Management and Convergence Software, Sogang University, Seoul, [11]Data Science & Business Analytics Lab, School of Industrial Management Engineering, College of Engineering, Korea University, Seoul, [12]Software Graduate Program, School of Computing, College of Engineering, Korea Advanced Institute of Science and Technology, Seoul, [13]Department of Biomedical Engineering, Yonsei University, Seoul, [14]Department of Social Studies Education, College of Education, Ewha Womans University, Seoul, [15]Department of Math, University of Kwangwoon, Seoul, [16]Department of Electrical and Computer Engineering, Seoul National University, Seoul, Departments of [17]Pathology and [18]Biostatistics, Seoul National University College of Medicine and SMG-SNU Boramae Medical Center, Seoul, [19]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

References

1. Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: the case for clinical adoption of digital pathology. J Clin Pathol. 2017;70:1010-8.
2. Wang S, Chen A, Yang L, Cai L, Xie Y, Fujimoto J, et al. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Sci Rep. 2018;8:10393.
3. Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6:26286.
4. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318:2199-210.
5. Kasper DL, Fauci AS, Hauser SL, Longo DL, Jameson JL, Loscalzo J. Harrison's principles of internal medicine. 19th ed.

New York: McGraw-Hill; 2015.
6. Hayes SC, Janda M, Cornish B, Battistutta D, Newman B. Lymphedema after breast cancer: incidence, risk factors, and effect on upper body function. J Clin Oncol. 2008;26:3536-42.
7. Fleissig A, Fallowfield LJ, Langridge CI, Johnson L, Newcombe RG, Dixon JM, 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 Res Treat. 2006;95:279-93.
8. Lyman GH, Temin S, Edge SB, Newman LA, Turner RR, Weaver DL, 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. 2014;32:1365-83.
9. Manca G, Rubello D, Tardelli E, Giammarile F, Mazzarri S, Boni G, et al. Sentinel lymph node biopsy in breast cancer: indications, contraindications, and controversies. Clin Nucl Med. 2016;41:126-33.
10. Galimberti V, Cole BF, Viale G, Veronesi P, Vicini E, Intra M, 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. Lancet Oncol. 2018;19:1385-93.
11. Giuliano AE, Ballman KV, McCall L, Beitsch PD, Brennan MB, Kelemen PR, 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. 2017;318:918-26.
12. Wang J, Tang H, Li X, Song C, Xiong Z, Wang X, et al. Is surgical axillary staging necessary in women with T1 breast cancer who are treated with breast-conserving therapy? Cancer Commun (Lond). 2019;39:25.
13. Donker M, van Tienhoven G, Straver ME, Meijnen P, van de Velde CJ, Mansel RE, 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. Lancet Oncol. 2014;15:1303-10.
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. Eur J Cancer. 2006;42:617-20.
15. Chen Y, Anderson KR, Xu J, Goldsmith JD, Heher YK. Frozen-section checklist implementation improves quality and patient safety. Am J Clin Pathol. 2019;151:607-12.
16. Bandi P, Geessink O, Manson Q, Van Dijk M, Balkenhol M, Hermsen M, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans Med Imaging. 2019;38:550-60.
17. Honkoop AH, Pinedo HM, De Jong JS, Verheul HM, Linn SC, Hoekman K, et al. Effects of chemotherapy on pathologic and biologic characteristics of locally advanced breast cancer. Am J Clin Pathol. 1997;107:211-8.
18. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: 2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27-30; Las Vegas, NV, USA.
19. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2015 Oct 5-9; Munich, Germany. Cham: Springer; 2015. p. 234-41.
20. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002;2:18-22.
21. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32-5.
22. Laurent-Bellue A, Poullier E, Pomerol JF, Adnet E, Redon MJ, Posseme K, et al. Four-year experience of digital slide telepathology for intraoperative frozen section consultations in a two-site French academic department of pathology. Am J Clin Pathol. 2020;154:414-23.
23. Menter T, Nicolet S, Baumhoer D, Tolnay M, Tzankov A. Intraoperative frozen section consultation by remote whole-slide imaging analysis: validation and comparison to robotic remote microscopy. J Clin Pathol. 2020;73:350-2.
24. Akay CL, Albarracin C, Torstenson T, Bassett R, Mittendorf EA, Yi M, et al. Factors impacting the accuracy of intra-operative evaluation of sentinel lymph nodes in breast cancer. Breast J. 2018;24:28-34.
25. Houpu Y, Fei X, Yang Y, Fuzhong T, Peng L, Bo Z, et al. Use of Memorial Sloan Kettering Cancer Center nomogram to guide intraoperative sentinel lymph node frozen sections in patients with early breast cancer. J Surg Oncol. 2019;120:587-92.