pISSN 1598-2998, eISSN 2005-9256

https://doi.org/10.4143/crt.2022.055 Cancer Res Treat. 2023;55(2):513-522

Original Article

Diagnostic Assessment of Deep Learning Algorithms for Frozen Tissue Section Analysis in Women with Breast Cancer

Young-Gon Kim[1], **In Hye Song**[2], **Seung Yeon Cho**[3], **Sungchul Kim**[4], **Milim Kim**[5], **Soomin Ahn**[5], **Hyunna Lee**[6], **Dong Hyun Yang**[7], **Namkug Kim**[8], **Sungwan Kim**[1,9], **Taewoo Kim**[10], **Daeyoung Kim**[10], **Jonghyeon Choi**[11], **Ki-Sun Lee**[12], **Minuk Ma**[13], **Minki Jo**[13], **So Yeon Park**[5], **Gyungyub Gong**[14]

[1]Transdisciplinary Department of Medicine & Advanced Technology, Seoul National University Hospital, Seoul, [2]Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, [3]Interdisciplinary Program in Bioengineering, College of Engineering, Seoul National University, Seoul, [4]Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, [5]Department of Pathology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, [6]Health Innovation Big Data Center, Asan Institute of Life Science, Asan Medical Center, Seoul, [7]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, [8]Department of Convergence Medicine, Asan Institute of Life Science, Asan Medical Center, University of Ulsan College of Medicine, Seoul, [9]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, [10]Graduate School of AI, Korea Advanced Institute of Science and Technology, Daejeon, [11]Knowledge of AI Lab, NCSOFT, Seongnam, [12]Medical Science Research Center, Ansan Hospital, Korea University College of Medicine, Ansan, [13]School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, [14]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Purpose Assessing the metastasis status of the sentinel lymph nodes (SLNs) for hematoxylin and eosin–stained frozen tissue sections by pathologists is an essential but tedious and time-consuming task that contributes to accurate breast cancer staging. This study aimed to review a challenge competition (HeLP 2019) for the development of automated solutions for classifying the metastasis status of breast cancer patients.

Materials and Methods A total of 524 digital slides were obtained from frozen SLN sections: 297 (56.7%) from Asan Medical Center (AMC) and 227 (43.4%) from Seoul National University Bundang Hospital (SNUBH), South Korea. The slides were divided into training, development, and validation sets, where the development set comprised slides from both institutions and training and validation set included slides from only AMC and SNUBH, respectively. The algorithms were assessed for area under the receiver operating characteristic curve (AUC) and measurement of the longest metastatic tumor diameter. The final total scores were calculated as the mean of the two metrics, and the three teams with AUC values greater than 0.500 were selected for review and analysis in this study.

Results The top three teams showed AUC values of 0.891, 0.809, and 0.736 and major axis prediction scores of 0.525, 0.459, and 0.387 for the validation set. The major factor that lowered the diagnostic accuracy was micro-metastasis.

Conclusion In this challenge competition, accurate deep learning algorithms were developed that can be helpful for making a diagnosis on intraoperative SLN biopsy. The clinical utility of this approach was evaluated by including an external validation set from SNUBH.

Key words Breast neoplasms, Deep learning, Frozen sections, Neoplasm metastasis, Sentinel lymph node, Metastasis, Classification

Introduction

Breast cancer is the most common cancer among women. Digital pathology has contributed significantly to its primary and frozen section diagnosis, becoming a common procedure in multidisciplinary clinics [1]. While surgical removal of the primary tumor is necessary [2], it is also important to determine the metastatic status and surgical extent of regional lymph nodes. Sentinel lymph node (SLN) sampling or dissection is performed intraoperatively for this purpose [3-5]. When the tumor spreads beyond the primary location, it first drains into the sentinel nodes, making SLN biopsy a significant role in breast cancer cases [6]. Although evaluating frozen sections is more difficult than formalin-fixed paraffin-embedded (FFPE) sections because of inferior quality, the frozen section technique is recommended since it allows immediate consultation during surgery [7]. Recent advances in deep learning algorithms may not only aid in an accurate

Correspondence: Gyungyub Gong
Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: 82-2-2010-4554 Fax: 82-2-472-7898 E-mail: gygong@amc.seoul.kr

Received January 29, 2022 **Accepted** September 5, 2022
Published Online September 6, 2022

Co-correspondence: So Yeon Park
Department of Pathology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, 82 Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam 13620, Korea
Tel: 82-31-787-7712 Fax: 82-31-787-4012 E-mail: sypmd@snu.ac.kr

*Young-Gon Kim, In Hye Song, and Seung Yeon Cho contributed equally to this work.

diagnosis but minimize anesthesia time for patients and labor for pathologists [8,9].

Some deep learning algorithms showing better diagnostic performance than pathologists have been introduced in the CAMELYON 16 and 17 (Cancer Metastases in Lymph Nodes Challenge) competitions [10,11], in which FFPE tissue sections are used. For the validation of frozen sections in metastases classifications, we held the HeLP Challenge 2018 (HEalthcare ai Learning Platform), in which automated deep learning algorithms for detecting metastases in hematoxylin and eosin–stained frozen SLN tissue sections of breast cancer patients were developed [12]. The goal of this challenge was to discriminate between metastatic and normal tissues on digital pathology slides provided by Asan Medical Center (AMC). Four teams submitted their results to the leaderboard in the final stage, and three of them showed considerable area under the curve (AUC) values. However, major limitations of this competition included that all datasets were acquired from a single institution (AMC) and the clinicopathologic characteristics of tumors were randomly distributed among the training, development, and validation sets. The use of datasets from only one institution usually restricts the generalization of the model for multisite deployment owing to a lack of external validation. Different ratios of tumor characteristics in each training, development, and validation set likely cause overfitting to a particular ratio, leading to biased model tuning. Moreover, it is known that breast cancer patients with micro-metastasis (≤ 2 mm) in SLN do not require axillary node dissection [13]. Thus, determining metastatic tumor size in SLN is clinically meaningful.

In the second competition, HeLP Challenge 2019, we expanded our task to determine the presence of metastasis and also measure the longest diameter of the metastatic tumor, if one existed. Additional data were collected from Seoul National University Bundang Hospital (SNUBH) to allow for external validation. In addition, clinicopathologic characteristics of the tumor slides were distributed in the training, development, and validation sets as evenly as possible to balance the ratios among them. As the p-values are calculated, the p-value for each clinicopathologic factors was less than 0.001 except for one, which indicated that the dataset distribution was statistically significant, compared to the previous challenge setting [12]. Through this modified challenge setting, we aimed to evaluate the performance of deep learning models for classifying metastases per slide, measuring the largest metastatic tumor size, and ensuring the adaptability of the external dataset in hematoxylin and eosin–stained frozen SLN tissue sections of breast cancer patients.

Materials and Methods

1. Data description

We acquired 524 digital slides of SLNs from the two different institutions for routine frozen section surgical procedures [14]. At SNUBH, each excised SLN was immediately submitted, cut into 2-mm slices, entirely embedded in optimum cutting temperature compound, and frozen at –25°C. Each 5-μm-thick frozen section was cut, mounted on glass slides, and stained with hematoxylin and eosin (H&E). A total of 227 slides were scanned using a digital microscopy scanner (Pannoramic 250 Flash II, 3DHISTECH Ltd., Budapest, Hungary) in the MIRAX format (.mrxs) with a resolution of 0.389 μm per pixel (MPP) from SNUBH. As already introduced in our previous study [12], the data acquisition protocol was the same for AMC with negligible differences. At AMC, lymph nodes were cut into 2-3-mm slices and frozen at –20°C to –30°C. A total of 297 slides were scanned using a digital microscopy scanner (Pannoramic 250 Flash II, 3DHISTECH Ltd.) in .mrxs format with a resolution of 0.221 MPP. The most important and notable difference between the two institutes was the resolution (MPP) [15].

The dataset comprised 236 slides from AMC as the training set, 107 slides (61 from AMC and 46 from SNUBH) as the development set, and 181 slides from SNUBH as the validation set. The validation set consisted of primarily external institution data, and the purpose was to validate the adaptability of the deep learning models to generalize in external dataset. Each set involved sufficient consideration of the distribution of histologic type. Of the total dataset, 163 slides were obtained from patients who had received neoadjuvant therapy, which would be more challenging to histologically examine [16], prior to submission of the SLN samples for frozen sectioning. Table 1 summarizes the participants' demographic details.

2. Reference standard

For the AMC dataset, a single rater provided manual segmentation of all digital slides, and two clinically expert pathologists with 6 and 20 years of experience in breast pathology confirmed the annotations. A similar procedure was performed for the SNUBH dataset, where a single rater provided manual segmentation of the digital slides and an expert breast pathologist with 15 years of experience confirmed the annotations. Metastatic carcinomas with regions larger than 200 μm in the greatest digital slide dimension were annotated as the cancers.

3. Challenge competition environment

The platform for the contest was developed by Kakao Brain, and the competitors were allowed to access the data

Table 1. Clinicopathologic characteristics of the patients in the AMC and SNUBH datasets (resolution [width×height] of the digital slide: 93,615×232,948 pixels [AMC] and 56,462×132,956 pixels [SNUBH])

	Training set	Development set		Validation set	p-value[a]
	AMC (n=236)	AMC (n=61)	SNUBH (n=46)	SNUBH (n=181)	
Age (yr)	50 (30-72)	49 (28-80)	51 (38-73)	52 (25-87)	
Female sex	236 (100)	61 (100)	46 (100)	181 (100)	> 0.99
Metastatic carcinoma					
Size > 2 mm	113 (47.9)	30 (49.2)	18 (39.1)	65 (35.9)	0.114
Size ≤ 2 mm	28 (11.9)	6 (9.8)	8 (17.4)	36 (19.9)	
Absent	95 (40.2)	25 (41.0)	20 (43.5)	80 (44.2)	
Neo-adjuvant therapy					
Not received	122 (51.7)	30 (49.2)	42 (91.3)	167 (92.3)	< 0.001
Received	114 (48.3)	31 (50.8)	4 (8.7)	14 (7.7)	
Histologic type					
IDC	201 (85.2)	50 (82.0)	46 (100)	177 (97.8)	< 0.001[b]
ILC	18 (7.6)	4 (6.5)	0	2 (1.1)	
Others	17 (7.2)	7 (11.5)	0	2 (1.1)	
Histologic grade					
1 or 2	188 (79.7)	49 (80.3)	28 (60.9)	112 (61.9)	< 0.001
3	48 (20.3)	12 (19.7)	18 (39.1)	69 (38.1)	

Values are presented as median (range) or number (%). AMC, Asan Medical Center; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; SNUBH, Seoul National University Bundang Hospital. [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.

only through the given paths using Docker image files. More details of the challenge platform and environment are introduced in our review of the previous challenge [12]. The competitors were informed about the details of the challenge environment and the dataset two days prior to the start of the competition. They were also notified of the difference between the two datasets from the two institutions, such as MPP, magnification, and staining intensity. However, more details involving the organization of the slides in the dataset were kept undisclosed to ensure fairness of the contest. For the first five weeks of the challenge, 343 digital slides were provided as training and development sets. Annotated masked images were provided in addition to the 236 digital slides of the training set to train the model. For the next two weeks, an additional 181 digital slides consisting of only SNUBH data were opened for the competitors to use as the final validation of their best-tuned models. During this period, the digital slides from the development set were no longer available, as the competitors were not allowed to additionally tune the model based on the development set once the validation set is open. The model's final performance was submitted to the leaderboard, and the scores and ranks were displayed in real time. Details of the algorithms for each team are presented in S1 Table.

4. Evaluation metric

The algorithms were evaluated for their ability to classify the digital frozen tissue section slides as "metastasis slides" or "normal slides" and measure the size of the longest diameter of the metastatic tumor. For the statistical analysis of the classification task, receiver operating characteristic (ROC) analysis at the slide level was performed, and the AUC was measured to compare the algorithms. As for the size measurement task, the assessment was made in terms of accuracy regarding the size of the largest metastasis. The error range for the size measurement evaluation was ±5%. Positive labels were given for predictions of size within the given error range, while negative labels were given for any other predictions outside this range. These binary labels, either positive or negative, were compared with the labels of metastasis in each digital slide and evaluated in terms of accuracy. This accuracy score was named "Scores of Major Axis" throughout the challenge.

5. Competitors

Registration for this challenge began in mid-November 2019 and lasted for 3 weeks. Ten teams were selected for participation from among the total registered teams. Toward the end of the contest, nine teams submitted their results to the leaderboard for the development set; finally, only four teams submitted their results for the validation set. The top three

Table 2. Final scores of performances of classification of tumor slides and prediction of major axes

Team	Phase 1. Development set (AMC+SNUBH)			Phase 2. Validation set (SNUBH)		
	AUC of slides	Scores of major axis	Total score	AUC of slides	Scores of major axis	Total score
GoldenPass	0.901	0.523	0.712	0.891	0.525	0.708
MediTrain	0.838	0.411	0.624	0.809	0.459	0.634
DRM	0.542	0.402	0.472	0.736	0.387	0.561

AMC, Asan Medical Center; AUC, area under the curve; DRM, DeepRunningMachine; SNUBH, Seoul National University Bundang Hospital.

Table 3. Performance comparison of classification task of tumor slides

Team	Phase 2. Validation set (SNUBH)				
	ACC	TPR	TNR	PPV	NPV
GoldenPass	0.845	0.772	0.938	0.940	0.765
MediTrain	0.790	0.644	0.975	0.970	0.684
DRM	0.724	0.634	0.838	0.831	0.644

ACC, accuracy; DRM, DeepRunningMachine; NPV, negative predictive value; PPV, positive predictive value; SNUBH, Seoul National University Bundang Hospital; TNR, true-negative rate; TPR, true-positive rate.

teams were the GoldenPass, MediTrain, and DeepRunningMachine (DRM) teams, and their methodological descriptions are shown in S1 Table. The results of only these three teams, who demonstrated meaningful outputs, were used for the review and analysis of this challenge.

Results

The model performances were sorted in descending order for the development set and validation set as shown in Tables 2 and 3. Nine teams submitted their results to the leaderboard for the development set, while five teams submitted their results for the validation set. Among them, the results of only the top three teams were considered meaningful because the lower-ranked teams showed AUC values below 0.500, which is too low to be accountable. For the development set, the three teams showed AUC values of 0.901, 0.838, and 0.542 for the slides and 0.523, 0.411, and 0.402 for the major axis. For the validation set, which consisted of 181 digital slides from SNUBH, the GoldenPass team showed the highest AUC (0.891) for the validation set (vs. those of the MediTrain and DRM teams of 0.809 and 0.736, respectively). All teams showed a decrease in AUC when the slides from AMC were eliminated in the validation set except for the DRM team, which demonstrated a large increase in performance for the external dataset only. For the major axis measurement, all teams showed an increase, although small, while the DRM team had a decreased score. A comparison of ROC curves for calculating the AUC values for each team is illustrated in Fig. 1. The first-place team, GoldenPass, showed confidence scores of 0-1 for each inference of their results, whereas the other two teams showed only a binary form of the prediction results with a 0 or 1. This difference is shown in Fig. 1, where the ROC curve of the GoldenPass team shows the staircase phenomenon, while the curves of the MediTrain and DRM teams were drawn from only three points. From the ROC curves, the optimal cut-off threshold was determined by the Youden's Index to evaluate each algorithm.

While the curves in Fig. 1 demonstrate the model performances of classifying normal and metastasis slides, the AUC values and ROC curves, as shown in Fig. 2, were additionally computed for performances of classifying micro-metastasis (≤ 2 mm) and macro-metastasis (> 2 mm). The slides with metastasis smaller than 2 mm were counted as the same label as normal slides in this case, and comparison of AUC values are shown in Table 4, along with the corresponding ROC curves in Fig. 2. When micro-metastases were considered as normal, the top two teams showed higher AUC values, and the values between the teams showed larger gaps. The performance comparison for both evaluations is visually summarized in confusion matrix representation in S2 Fig.

Model performance was additionally evaluated by comparing performance according to clinicopathologic characteristics. This clinical information (Table 5) includes the size of the metastatic tumor (whether its greatest dimension is smaller or larger than 2 mm), neoadjuvant therapy status, histologic type, and histologic grade. The top two teams

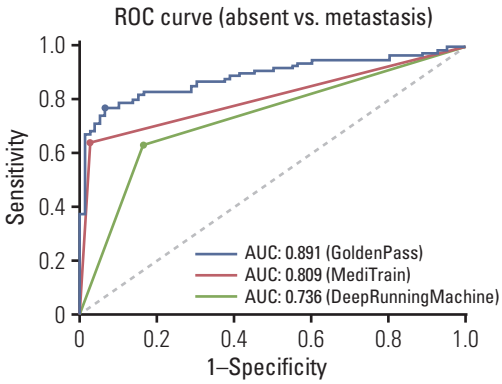

Fig. 1. Receiver operating characteristic (ROC) curve comparisons of models trained by the three algorithms for the validation set. AUC, area under the curve.

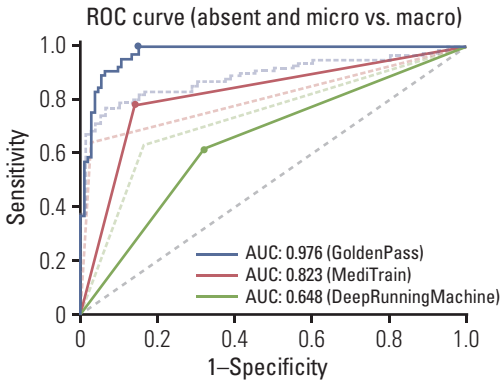

Fig. 2. Receiver operating characteristic (ROC) curve comparisons of models for classifying micro-metastasis as normal (The original ROC curves from Fig. 1 are shown with dotted lines.). AUC, area under the curve.

Table 4. Performance comparison of classifying micro-metastasis as tumor versus micro-metastasis as normal

Team	AUC (validation set)	
	Absent vs. Tumor (including ≤ 2 mm) (Fig. 1)	Absent (including ≤ 2 mm) vs. Tumor (Fig. 2)
GoldenPass	0.891	0.976
MediTrain	0.809	0.823
DRM	0.736	0.648

AUC, area under the curve; DRM, DeepRunningMachine.

showed a higher true-positive rate (TPR) and a lower false-negative rate (FNR) in slides with metastatic tumors larger than 2 mm, while the third-place team showed the opposite with a higher TPR and lower FNR for smaller metastatic tumor slides. Two teams showed a lower TPR for slides obtained from patients who had not received neoadjuvant therapy, while the other team showed lower TPR for slides of samples from patients with a history of neoadjuvant therapy. Two teams showed a lower true-negative rate (TNR) for slides with a neoadjuvant therapy history, while the first-place team (GoldenPass) showed an especially significant drop in TNR. For cases in which the metastatic carcinoma was invasive lobular carcinoma (ILC), all of the top three teams showed higher TPR and TNR values in contrast to cases of invasive ductal carcinoma (IDC). In terms of comparing performance according to histologic grade, the GoldenPass team showed better performance, although there was a very small difference in the classification of SLN with a histologic grade of 1 or 2, while higher values in both TPR and TNR were obtained for histologic grade 3 for the other two teams with the exception of the MediTrain team, which showed a higher TNR for histologic grade 1 or 2 samples.

The top three teams correctly classified 100 slides, including 39 true-positive and 61 true-negative, and all three incorrectly classified nine slides as negative (false-negative) among the 181 slides in the validation set. The first-place team had five false-positive slides that the other two teams correctly classified. These wrongly categorized slides are represented in Fig. 3. The second- and third-place teams incorrectly classified one slide as positive (false-positive), while the DRM team incorrectly classified 12 slides as positive. One false-positive slide obtained by the two teams was IDC histologic grade 2 without a history of neoadjuvant therapy. All nine false-negative slides were obtained from patients with the IDC histologic type who did not receive neoadjuvant systemic therapy: six were from patients with histologic grade 1 or 2 cancer, while the other three were from patients with histologic grade 3 cancer. Among those false-negative slides, all nine had micro-metastases (size range, 0.15 to 1.91 mm).

Among the 65 lymph nodes with a metastasis greater than 2 mm, the GoldenPass team predicted 44 of them as being larger than 2 mm; of them, 15 were within the allowed error range. The MediTrain and DRM teams predicted 37 and 40 cases, respectively, as being larger than 2 mm, and four and three of them, respectively, were within the allowed error range. For the 36 SLN samples with micro-metastasis,

Table 5. Performance comparison of determining clinicopathologic characteristics of tumors

	Team		
	GoldenPass	MediTrain	DeepRunningMachine
Metastatic tumor size			
Absent (n=80)			
TPR	0.938	0.975	0.838
FNR	0.063	0.025	0.163
≤ 2 mm (n=36)			
TPR	0.361	0.389	0.667
FNR	0.639	0.611	0.333
> 2 mm (n=65)			
TPR	1.000	0.785	0.615
FNR	0.000	0.215	0.385
Neoadjuvant therapy			
Not received (n=167)			
TPR	0.766	0.649	0.628
TNR	0.959	0.986	0.836
Received (n=14)			
TPR	0.857	0.571	0.714
TNR	0.714	0.857	0.857
Histologic type			
IDC (n=177)			
TPR	0.765	0.643	0.633
TNR	0.937	0.975	0.835
ILC+mixed (n=4)			
TPR	1.000	0.667	0.667
TNR	1.000	1.000	1.000
Histologic grade			
1 or 2 (n=112)			
TPR	0.794	0.619	0.619
TNR	0.939	0.980	0.796
3 (n=69)			
TPR	0.737	0.684	0.658
TNR	0.935	0.968	0.903

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

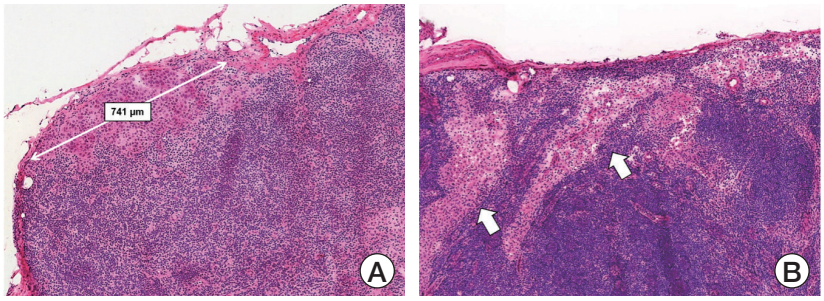

Fig. 3. Representative examples of false-negative and false-positive cases in hematoxylin and eosin stains. (A) A case with micro-metastasis (741 μm in diameter), which was predicted as negative by the three teams (visual field: 9.6×). (B) A case with sinus histiocytosis (shown with arrows) mimicking metastasis, which was predicted as positive by GoldenPass team (visual field: 8.9×).

whose size was less than or equal to 2 mm, the GoldenPass and DRM teams did not have predictions smaller than 2 mm, while the MediTrain team predicted 10 of them as smaller than 2 mm, with one being within the given error range.

Discussion

Recent advances in technology and equipment have led to the expansion of digital pathology in many countries. Digital pathology includes primary diagnosis based on whole slide imaging, telepathology, and computer-aided diagnosis using image analysis software [17]. A computer-aided diagnosis is defined as the interpretation of digitized histological images using a computational diagnostic system [18]. Currently, deep learning is generally considered the most promising computer-aided diagnosis method. Computer-aided diagnosis using deep learning methods showed good performance for classification, prognostication, and prediction of breast cancer, prostate cancer, gastrointestinal cancer, skin cancer, etc. [19-24].

Digital pathology has also been implemented and validated for intraoperative frozen section diagnosis [25-27]. For primary diagnosis, most frozen section slides were successfully scanned, and findings of glass and digitalized slides showed excellent agreement. In addition, digital pathology has apparent advantages for consultation since pathologists can save a considerable amount of time and effort if they simply use telepathology instead of actually moving to see glass slides or show them to other pathologists. However, the application of computer-aided diagnosis in frozen section pathology is still in its infancy. There have been several studies on the quantification of steatosis using deep learning for frozen liver biopsy sections [28,29] but few studies on computer-aided diagnosis in frozen section pathology of cancer surgery. Our group previously held HeLP Challenge 2018 to develop a deep learning algorithm for the diagnosis of SLN sections in breast cancer surgery as summarized in the introduction section. We then held HeLP Challenge 2019, which aimed to expand the dataset, measure the metastatic tumor sizes, and improve the overall algorithm performance.

In this study, all of the participants of the top three teams included convolutional neural network–based deep learning methods for classification or segmentation networks, which resulted in adequately high performance with AUC values of 0.891, 0.809, and 0.736. Notably, the performances of the top three teams were better than those of HeLP Challenge 2018, which were AUC values of 0.805, 0.776, and 0.760. We believe that this enhancement could be due to dataset expansion and algorithm improvement. Further data collection and training might enable the implementation of computer-aided diagnosis in frozen section pathology.

The model performances were compared and evaluated according to the clinicopathologic characteristics of the patients. Although the top two teams showed a lower TPR in micro-metastasis than in the macro-metastasis, the third-place team showed a paradoxical result with a higher TPR in micro-metastasis. The models of the top two teams were generally trained well to distinguish metastatic tumor slides, and they showed similar aspects that smaller the tumor sizes, the more difficult it was to classify. On the other hand, while the model of DRM team was not trained generally enough to classify metastatic tumors, the result was in consistence with the previous study that revealed Inception-v1, also known as GoogLeNet, as the best performing network in micro-metastasis [10]. Although it is a previous version of Inception-v4 employed by DRM team, they both share the same inception modules, which may have contributed in robustness in micro-metastasis. Such aspect may have seen amplified since the number of slides for the micro-metastasis was the smallest of the three categories in metastatic tumor size.

A main modification in this second competition was the addition of the dataset from SNUBH to enable the evaluation of deep learning models for adaptability in an external dataset. Interestingly, two of the teams showed higher total scores for the validation set than for the development set. The first-place team, GoldenPass, had a decreased total score in the validation set, but the absolute value of the difference was the smallest among the three teams. In other words, this can be interpreted as the GoldenPass team showing the most similar performance in the development and validation sets. Since the purpose of external validation is to assess the model's adaptability in a dataset from another domain, such results may be an indication of deep learning model robustness. This might be due to the difference in pre-processing methods, particularly with regard to the handling of input data acquired from two different institutes. As already mentioned in the previous section, the primary difference between the AMC and SNUBH datasets is related to the definite size of each pixel, referred to as MPP, which is determined at the point of slide scanning. If input patches are extracted from the same slide layer level, the patch resolution in the AMC data would be approximately 1.7 times that of the SNUBH data patch. To minimize the influence of this domain gap, the GoldenPass team extracted patches from level 4 for the AMC slides and level 3 for the SNUBH slides and rescaled them. They also applied stain normalization, which can reduce the variations in color and intensity in H&E-stained images obtained at different time points and in different laboratories (S1 Table). This suggests that consideration of the domain gap during the training led to the maintenance of a small change in performance between the

Table 6. Performance differences in major axis prediction according to variations in error range

Team	Scores of major axis prediction error range				
	±5%	±10%	±15%	±20%	±30%
GoldenPass	0.525	0.552	0.597	0.613	0.630
MediTrain	0.459	0.492	0.508	0.525	0.580
DRM	0.387	0.475	0.492	0.503	0.514

DRM, DeepRunningMachine.

development and validation sets.

The model architecture employed by the top two teams involved a feature pyramid network (FPN), known as a multi-scale feature extractor, while the third-place team employed Inception-v4 and support vector machine (SVM). Based on the model architectures, the use of FPN may expect to minimize the influence of MPP difference between the datasets, since the network makes use of feature maps extracted from various scales. This may have contributed in increasing the performance of major axis measurement task. On the other hand, implementation of Inception-v4 and SVM to train the geometric features extracted by the model may have optimized the output for the classification task only. Although the top two teams both equally employed FPN architectures, consideration of MPP in the patch extraction stage by the first-place team may have further contributed to the enhancement in the final performance.

Model performance in the classification task was notably low in slides with micro-metastatic tumors and high in slides with ILC. Pathologists' manual examination of intraoperative SLN biopsy is generally difficult in cases of micrometastases and lobular histology [30]; hence, poor performance for discriminating micro-metastatic tumor slides is probable. However, the peculiarly high TNR in slides with ILC may be due to the amount of data in the validation set, which included an extremely low proportion (1.1%) of cases of lobular histology. In addition, the model performance in the major axis measurement task was generally low for all teams. This might be the reason for the strict error range allowed in the contest. An error range ±5% was used to compute the participants' scores and ranks, but in fact, error ranges of approximately ±15%-20% are acceptable for determining the sizes of metastases in actual clinical examinations. Relatively low major axis prediction scores could be complemented by increasing the allowed error range (Table 6).

For additional analysis, the slides with no metastatic tumor or micro-metastatic tumors only (≤ 2 mm) were considered as negative, and the slides with macro-metastasis (> 2 mm) were considered as positive. There are two reasons for this. Firstly, if frozen biopsy reveals micro-metastases only, then axillary lymph node dissection is not required. Therefore, clinical significance of micro-metastasis is much less than macro-metastasis. Secondly, when annotating tumor areas for this challenge, the pathologist did not annotate metastatic tumor clusters smaller than 2 mm because that was too labor-intensive. This classification could possibly affect the learning ability of tumor detecting algorithms. In such setting, the top two teams showed larger AUC values and the GoldenPass team showed especially large increase, which can be interpreted as that their model was better fit for discriminating the macro-metastasis.

Although the current breast cancer treatment guidelines do not recommend axillary lymph node dissection in the micro-metastasis, some surgeons still prefer to do additional lymph node sampling in the setting of micro-metastasis, or just. Therefore, it might be helpful for pathologists if deep learning algorithms can sensitively detect very small foci of metastatic tumor cells, including micro-metastasis or even isolated tumor cells. We suggest that further studies including more delicate annotation and intense learning process can improve tumor detecting ability of the algorithms.

We held a 7-week-long challenge competition to develop deep learning algorithms for the analysis of digital pathology slides with H&E-stained frozen tissue sections of SLN samples from breast cancer patients. In contrast to the previous challenge we held, here we tried to develop more helpful and practical models for the diagnosis of frozen intraoperative SLN biopsy samples by adding the major axis measurement task and external dataset. The measurement task was to help determine whether the size of a metastasis requires its resection, and an external dataset was used to evaluate the models' robustness and adaptability to data from another institution. The top three ranked teams achieved high AUC values and acceptably high scores for major axis prediction despite a strictly limited error range in the evaluation. The deep learning models proposed in this challenge may be used for clinical trials in the future to compare the performances between the computer-aided diagnosis versus the pathologist's examination. Moreover, follow-up studies could be conducted with the expansion cohort to adjust the proposed algorithms into routine clinical practice, which our

future works will focus on. Yet, further studies are required to increase the micro-metastases detection accuracy and implement concise and time-saving models for application in routine clinical settings.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (https://www.e-crt.org).

Ethical Statement

The study protocols were approved by the Institutional Review Board Committees of AMC (IRB number: 2018-0583), University of Ulsan College of Medicine, Seoul, Korea, and SNUBH (IRB number: B-1806-472-106), Seoul National University College of Medicine, Gyeonggi, Korea, which waived the need for informed patient consent.

Author Contributions

Conceived and designed the analysis: Kim S (Sungchul Kim), Kim M, Ahn S, Lee H, Yang DH, Kim N, Kim S (Sungwan Kim), Park SY, Gong G.
Collected the data: Gong G, Park SY.
Contributed data or analysis tools: Kim M, Ahn S, Lee H.
Performed the analysis: Kim YG, Song IH, Cho SY.
Wrote the paper: Kim YG, Song IH, Cho SY, Kim S (Sungchul Kim), Kim M, Ahn S, Lee H, Yang DH, Kim N, Kim S (Sungwan Kim), Kim T, Kim D, Choi J, Lee KS, Ma M, Jo M, Park SY, Gong G.
Searched literature: Kim YG, Song IH, Cho SY.
Supervision: Kim S (Sungchul Kim), Kim M, Ahn S, Lee H, Yang DH, Kim N, Kim S (Sungwan Kim), Park SY, Gong G.
Experimented with algorithms: Kim T, Kim D, Choi J, Lee KS, Ma M, Jo M.

ORCID iDs

Young-Gon Kim : https://orcid.org/0000-0003-2148-1299
In Hye Song : https://orcid.org/0000-0001-6325-3548
Seung Yeon Cho : https://orcid.org/0000-0002-8756-1331
So Yeon Park : https://orcid.org/0000-0002-0299-7268
Gyungyub Gong : https://orcid.org/0000-0001-5743-0712

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Acknowledgments

This research was supported by 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).

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. Kasper D, Fauci A, Hauser S, Longo D, Jameson JL, Loscalzo J. Harrison's principles of internal medicine. 19th ed. New York: McGraw Hill; 2015.
3. 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.
4. 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.
5. 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.
6. Zahoor S, Haji A, Battoo A, Qurieshi M, Mir W, Shah M. Sentinel lymph node biopsy in breast cancer: a clinical review and update. J Breast Cancer. 2017;20:217-27.
7. 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.
8. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301-9.
9. Kim YG, Choi G, Go H, Cho Y, Lee H, Lee AR, et al. A fully automated system using a convolutional neural network to predict renal allograft rejection: extra-validation with gigapixel immunostained slides. Sci Rep. 2019;9:5123.
10. 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.
11. 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.
12. Kim YG, Song IH, Lee H, Kim S, Yang DH, Kim N, et al. 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. Cancer Res Treat. 2020;52:1103-11.

13. 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.
14. 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.
15. Kim YG, Kim S, Cho CE, Song IH, Lee HJ, Ahn S, et al. Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections. Sci Rep. 2020;10:21899.
16. 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.
17. Chong Y, Kim DC, Jung CK, Kim DC, Song SY, Joo HJ, et al. Recommendations for pathologic practice using digital pathology: consensus report of the Korean Society of Pathologists. J Pathol Transl Med. 2020;54:437-52.
18. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019;249:286-94.
19. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. Comput Biol Med. 2020;127:104065.
20. Lino-Silva LS, Xinaxtle DL. Artificial intelligence technology applications in the pathologic diagnosis of the gastrointestinal tract. Future Oncol. 2020;16:2845-51.
21. Li F, Yang Y, Wei Y, He P, Chen J, Zheng Z, et al. Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. J Transl Med. 2021;19:348.
22. Fitzgerald J, Higgins D, Mazo Vargas C, Watson W, Mooney C, Rahman A, et al. Future of biomarker evaluation in the realm of artificial intelligence algorithms: application in improved therapeutic stratification of patients with breast and prostate cancer. J Clin Pathol. 2021;74:429-34.
23. Gupta P, Huang Y, Sahoo PK, You JF, Chiang SF, Onthoni DD, et al. Colon tissues classification and localization in whole slide images using deep learning. Diagnostics (Basel). 2021; 11:1398.
24. Kuntz S, Krieghoff-Henning E, Kather JN, Jutzi T, Hohn J, Kiehl L, et al. Gastrointestinal cancer classification and prognostication from histology using deep learning: systematic review. Eur J Cancer. 2021;155:200-15.
25. Cima L, Brunelli M, Parwani A, Girolami I, Ciangherotti A, Riva G, et al. Validation of remote digital frozen sections for cancer and transplant intraoperative services. J Pathol Inform. 2018;9:34.
26. Ramaswamy V, Tejaswini BN, Uthaiah SB. Remote reporting during a pandemic using digital pathology solution: experience from a tertiary care cancer center. J Pathol Inform. 2021;12:20.
27. Evans AJ, Brown RW, Bui MM, Chlipala EA, Lacchetti C, Milner DA, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. Arch Pathol Lab Med. 2022;146:440-50.
28. Sun L, Marsh JN, Matlock MK, Chen L, Gaut JP, Brunt EM, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine. 2020;60:103029.
29. Perez-Sanz F, Riquelme-Perez M, Martinez-Barba E, de la Pena -Moral J, Salazar Nicolas A, Carpes-Ruiz M, et al. Efficiency of machine learning algorithms for the determination of macrovesicular steatosis in frozen sections stained with sudan to evaluate the quality of the graft in liver transplantation. Sensors (Basel). 2021;21:1993.
30. Akay CL, Albarracin C, Torstenson T, Bassett R, Mittendorf EA, Yi M, et al. Factors impacting the accuracy of intraoperative evaluation of sentinel lymph nodes in breast cancer. Breast J. 2018;24:28-34.