International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 7 Issue 10 ǁ 2019 ǁ PP. 25-37

Detection of Lung Cancer from Whole-Slide Histopathologic Images Using Deep Learning Approach

Dr. Niranjan Mishra, Prakash Kumar Behera, Rudra Prasad Nanda,Manoranjan Sahoo

Department of Electronics and Communication Engineering, NM Institute of Engineering and Technology,Bhubaneswar , Odisha
Department of Electronics and Communication Engineering, Raajdhani Engineering College,Bhubaneswar,Odisha
Department of Electronics and Communication Engineering,Aryan Institute of Engineering and Technology Bhubnaeswar , Odisha
Department of Electronics and Communication Engineering,Capital Engineering College,Bhubaneswar,Odisha

The application of deep learning for the detection of lymph node metastases on histologic slides has attracted worldwide attention due to its potentially important role in patient treatment and prognosis. Despite this attention, false-positive predictions remain problematic, particularly in the case of reactive lymphoid follicles. In this study, a novel two-step deep learning algorithm was developed to address the issue of false-positive prediction while maintaining accurate cancer detection. Three-hundred and forty-nine whole-slide lung cancer lymph node images, including 233 slides for algorithm training, 10 slides for validation, and 106 slides for evaluation, were collected. In the first step, a deep learning algorithm was used to eliminate frequently misclassified noncancerous regions (lymphoid follicles). In the second step, a deep learning classifier was developed to detect cancer cells. Using this two-step approach, errors were reduced by 36.4% on average and up to 89% in slides with reactive lymphoid follicles. Furthermore, 100% sensitivity was reached in cases of macrometastases, micrometastases, and isolated tumor cells. To reduce the small number of remaining false positives, a receiver-operating characteristic curve was created using foci size thresholds of 0.6 mm and 0.7 mm, achieving sensitivity and specificity of 79.6% and 96.5%, and 75.5% and 98.2%, respectively. A two-step approach can be used to detect lung cancer metastases in lymph node tissue effectively and with few false positives.

Traditional microscope and glass slides have been used by pathologists to diagnose disease since the mid-19th century. The conventional workflow involves manual review of numerous glass slides and requires a significant amount of time and effort on the part of the pathologist.[1e3] The emergence of slide scanning machines capable of outputting high resolution digital slides has brought traditional pa-thology into the digital era, providing numerous advantages to the pathology workflow. One of these advantages is the ability to use computational techniques, including auto-mated image analysis, to aid pathologists in the examination and quantification of slides, thus reducing the time required for manual screening and improving the pathologist's ac-curacy, reproducibility, and workflow efficiency.

Recently, the application of deep learning techniques to assist diagnosis has attracted considerable interest in pa-thology. Convolutional neural networks (CNNs), in particular, have demonstrated enormous potential in medical image recognition tasks.[7e10] In pathology, CNNs have been used in several image recognition applications with valuable results, from tumor cell detection in primary breast cancer, to grading glioma and prostate cancer, to counting mitoses and segmentation of tumor-associated stroma, to building whole-slide image (WSI)-based prognostic data.[10e16]

A review of lymph nodes is critical for staging cancer and making appropriate therapeutic decisions.[17] Involvement of multiple lymph node levels is a key factor in determining prognosis, and careful assessment of lymph node status is

For this reason, the detection of metastases in lymph node tissue is an area that could stand to benefit from de-velopments in automated tissue classification using machine learning approaches. In 2016 and 2017 a series of compet-itive international challenges, Cancer Metastases in Lymph Nodes Challenge 2016 and 2017 (CAMELYON16 and CAMELYON17), were held to identify machine learning algorithms capable of detecting and staging breast cancer metastases.Some of the top-scoring entries in these challenges were able to demonstrate better performance in detecting micrometastases than a pathologist with time constraints,[20] as is usual with most pathologists working in busy practices. Similar to breast cancer, lymph node me-tastases in lung cancer play an important role in evaluating disease stage, selecting treatment options, and determining prognosis.[22,23] Among all cancers, lung cancer is the lead-ing cause of cancer-related death worldwide.[24]

Although lymph node metastases in lung cancer and breast cancer share some similar characteristics, they have certain distinct histologic features attributed to their respective cancer subtypes. In addition, mediastinal lymph nodes frequently show more prominently reactive histologic changes, including multiple hyperplastic lymphoid follicles and abundant anthracotic pigment-laden macrophages. These findings are not usual for extramediastinal locations, which may create difficulty when distinguishing tumor and non-tumor components in lymph node tissue using machine learning algorithms.

For tumor detection tasks, there is a tradeoff between achieving high sensitivity in detecting micrometastases and a high false-positive error rate, especially for the identifi-cation of isolated tumor cells (ITC).[21] Interestingly, errors made by deep learning algorithms do not strongly correlate with human errors, and are more often attributed to technical issues with the digital slide, such as out-of-focus areas or folds, misclassification of tumor confounding histologic patterns, or benign components of the lymph node that share morphological similarity with tumor tissue, including germinal centers, macrophages, and stroma.[6,25,26] Although technical errors can be prevented by more careful prepara-tion of slides, the histologic tumor mimics, especially by hyperplastic lymph nodes with reactive lymphoid follicles and enlarged germinal centers, remains an issue without an effective solution. Since these are common components of all lymph node sections, this limitation significantly restricts the clinical utility of algorithms used for metastatic tumor detection.

In this study, a deep learningebased software program with an integrated CNN algorithm was applied to the detection of lung cancer lymph node metastases in WSIs. A new method for metastatic tumor detection in lung cancer is proposed involving two steps of deep learning tissue classification in which the first step is used for exclusion of germinal centers and the second for tumor cell detection. The hypothesis is that this new approach can reduce false positives caused by tumor mimics and increase accuracy in the detection of lymph node metastases compared to using only one deep learning algorithm (one-step approach).

I. MATERIALS AND METHODS

This study was approved by the ethical board of Nagasaki University Hospital (19021824) and Kameda Medical Center (18-210).

Materials

A total of 349 lymph node slides from 101 lung cancer patients with various histologic tumor types and stages were enrolled. Slides were collected at Nagasaki University Hospital, Japan, from 2014 to 2018, and from Kameda General Hospital, Japan, from 2007 to 2018. Details on the WSI data used in this study are shown in Table 1. Of 349 slides, 233 slides were used for training algorithms, 10 slides were used for validation, and 106 slides were used for testing. The validation set, which was separate from the training and testing sets, was used in the first step of the study and for all parameter tuning and model design choices in the trials.

Further details on the histologic subtypes can be found in Supplemental Table S1. Metastases were classified following the clinical practice guidelines as macro-

metastases (the largest tumor deposit had a diameter 2

mm), micrometastases (0.2 to 2 mm), and ITC (<0.2 mm).[17,27]

Digitalization and Annotation

Glass slides were scanned into digital slides using an Aperio Scanscope CS2 digital slide scanner (Leica Biosystems, Buffalo Grove, IL) with a 40 objective (0.2517 mm/pixel). Digital slides were imported into HALO software version 2.2 (Indica Labs, Corrales, CA) for all subsequent steps, including annotation, training, and classification of digital slides. Tissue classification was performed using the HALO Tissue Classifier analysis module (random forest algorithm) and HALO AI (CNN, VGG network). Annotations used for training the tissue classification algorithms were drawn by one pathologist with 7 years of experience in pathology (H.H.N.P.), with supervision by an expert pulmonary pathologist (J.F.).

Trials of the New Strategy

First, a single classifier with two classes, tumor and non-tumor, was developed using a CNN algorithm. Around 4000 annotations, including various polygonal outlines, were provided for training of each class using high-resolution

Table 1 Data for the Whole-Slide Lymph Node Images Used for the Lung Cancer Metastasis Experiments

Category	Training	Validation	Testing	Total
Macrometastasis	100	4	24	128
Micrometastasis	7	0	23	30
ITC	0	0	2	2
Nonmetastasis	126	6	57	189
Total	233	10	106	349

Data are numbers of slides per category.

ITC, isolated tumor cells.

images (0.25 mm per pixel). The classifier was trained for

3 10^6 iterations. The results showed that, although all metastases were detected at the slide-based level (100% sensitivity), many false-positive foci were found in both metastatic and nonmetastatic slides (0% specificity). Lymphoid follicles were found to be a common tumor mimic and a frequent cause of false-positive foci using this classifier (Figure 1A).

In the next step, a second classifier was created with three classes: tumor, lymphoid follicle, and other tissue. The HALO AI CNN was trained with 2371 annotations labeled lymphoid follicles, 3902 annotations labeled tumor, and 3030 annotations for others. Training and classification were performed at high resolution (0.25 mm per pixel) for 7.4 10^5 training iterations. The hypothesis was that by separating lymphoid follicles into a distinct class, the tumor detection algorithm would better differentiate them from tumors. When it was applied in the validation set, there was similar sensitivity and specificity as the first classifier, due to misclassification of lymphoid follicles, especially reactive follicles (Figure 1B). Therefore, these two classifiers were not further used in this study.

A step-wise approach in which two separate classifiers were developed and linked was then tested. The first clas-sifier was designed to exclude lymphoid follicles from the rest of the tissue (including any tumor that might be pre-sent), and the second classifier was designed to detect tumor cells in the lymphoid-excluded tissue. In the first step, two different algorithms were tested to build lymphoid exclusion classifier, a random forest machine learning algorithm, and a deep learning CNN algorithm. In the next step, a deep learning algorithm was used to detect tumor cells. This method was based on the idea that a two-step algorithm could reduce most false-positive findings and produce a more accurate tumor detection tool in comparison to the one-step strategy (Figure 1, C and D).

Model Development

In the training step, annotations were divided into three separate component classes: tumor, lymphoid follicle, and other. Depending on the specific task of each machine learning model, different classes were chosen as the input data for training that model (Figure 2). Annotations of tumor were provided in various sizes, mimicking both macro and micrometastases, as well as ITC, to maximize the learning ability of the models.

Lymphoid Follicle Detection and Exclusion in the First Step

To determine the best model for lymphoid follicle detection, two different models were created. Model 1 was a random forest classifier [Lymphoid Follicle Random Forest Model (LFRFM)]. Random forest classifiers can obtain a good result with a small amount of training data.[28] In addition, based on the settings of the HALO software (analysis on texture and color of images) that prefers few and small training regions, 65 annotations were provided with two classes: lymphoid follicles (20 annotations) and others (45 annotations) including tumor area, for training the model. Training and classification were performed at a low reso-lution of 4.4 mm per pixel.

Model 2 was a deep learning classifier [Lymphoid Fol-licle CNN (LFCNN)]. Because a deep learning algorithm typically requires a considerable amount of training data for its multiple deep layers of structure to improve accuracy,[28] more annotations were provided with 2332 training regions in total, representing two classes: lymphoid follicles (1243 annotations) and others (1089 annotations) containing the tumor area. Training and classification were performed at a medium resolution of 1.04 mm per pixel with 3.35 10^4 training iterations.

Two models then were applied in the validation set to test the ability of lymphoid follicle detection. Among them, the model that could best identify lymphoid follicles on WSIs was chosen. All lymphoid follicles would then be excluded, and the layer without lymphoid follicles would be further analyzed in a second step using another deep learning model, model 3 [Tumor Detection CNN (TDCNN)], to detect cancer cells.

Tumor Cell Detection in the Second Step

The task of the second step was to detect cancer cells on lymph node slides using model 3 (TDCNN). In this step, the HALO AI CNN was trained with 10,155 total training an-notations representing two classes: tumor (4196 annota-tions) and others (5959 annotations). Training and classification were performed at a high resolution of 0.25 mm per pixel, which is equal to the high-power field of a microscope used in pathologic diagnoses. It was then trained for 1 10^7 iterations. Apart from its primary pur-pose, to analyze the layer without lymphoid follicles ach-ieved from the first step to detect metastatic tumors, the TDCNN model was also used to predict metastases in one step for purposes of comparison with the two-step deep learning algorithm.

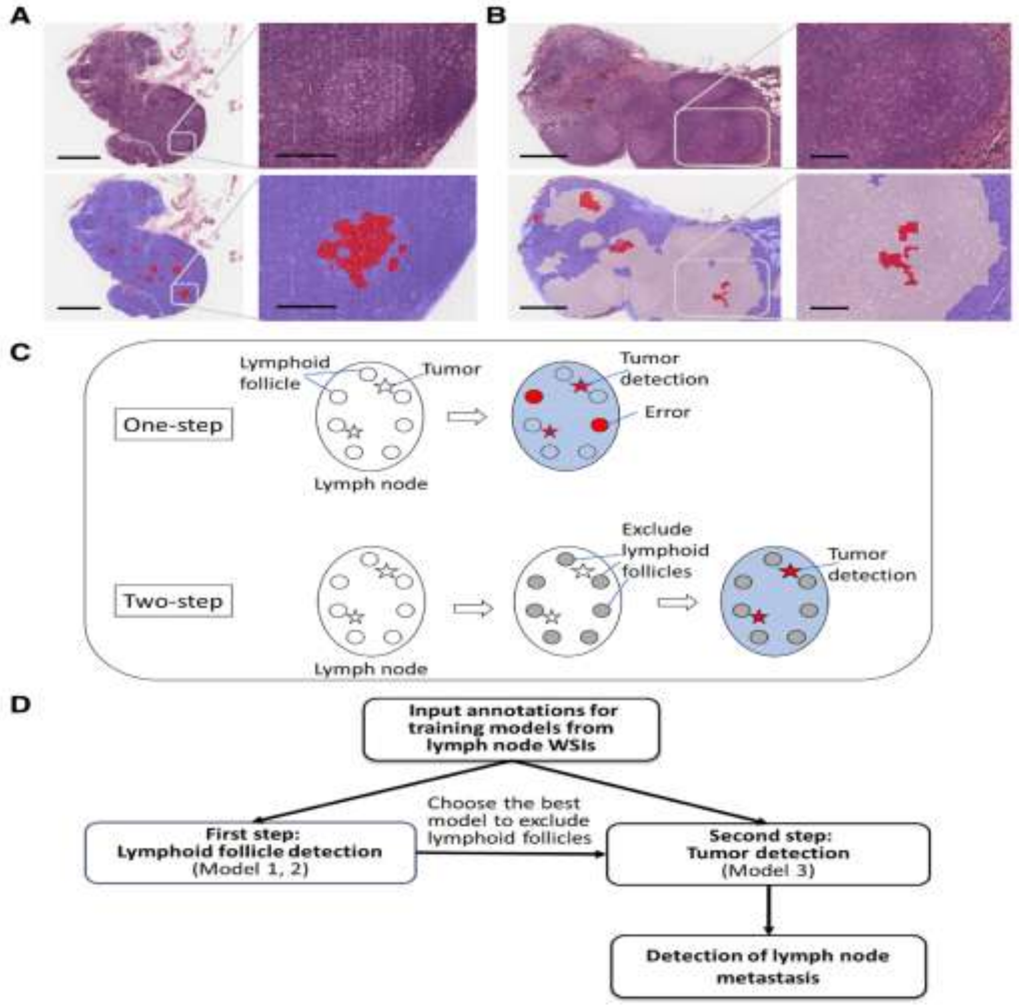

A
B
C
Lymphoid follicle
Tumor
One-step
Lymph node
Tumor detection
Error
Two-step
Lymph node
Exclude lymphoid follicles
Tumor detection
D
Input annotations for training models from lymph node WSIs
Choose the best model to exclude lymphoid follicles
First step: Lymphoid follicle detection (Model 1, 2)
Second step: Tumor detection (Model 3)
Detection of lymph node metastasis

Figure 1 A and B: Multiple large false-positive areas are represented as cancer foci in germinal centers of lymphoid follicles detected by the first and second trial set of deep learning classifiers. Upper row: hematoxylin and eosin staining of original images. Lower row: red: tumor, blue: other, and white: lymphoid follicle. Boxed areas in left panels of A and B are shown in higher magnification in right panels. C: The hypothesis was that the two-step deep learning algorithm can achieve better results compared with the one-step method. D: Strategy for two-step deep learning algorithms in detail. Scale bars: 2.0 mm (A, upper and lower left panels); 300 mm (A and B, upper and lower right panels); 500 mm (B, upper and lower left panels). WSI, whole-slide image.

Convolutional Neural Network Training and Application

In this study, HALO AI settings were fixed at a probability threshold of 50% on the tumor heatmap for the outcome, indicating that only pixels with more than 50% possibility of displaying a tumor were labeled as positive for the cancer class on the WSIs. The GeForce GTX TITAN X graphics card (NVIDIA, Santa Clara, CA) provided the required GPU for HALO AI. HALO AI utilizes the Caffe engine and a fully convolutional version of the VGG architecture[29] with all padding removed. Training was conducted on patches of 435 at the defined resolution. The patches were generated by picking a random class (with equal probability for each class), a random image containing annotation for the chosen class, and a random point inside a region of the chosen class and image. The patches were cropped around the chosen point and further augmented with random rota-tions and random shifts in hue, saturation, contrast, and brightness. The model was pre-trained on ImageNet and then trained for the defined number of iterations using RMSProp[30] (delta of 0.9) with a learning rate of 1e-3, a reduction in the learning rate by 10% every 10,000 itera-tions, and an L2 regularization of 5e-4. Because there was no padding in the model during analysis, the tile size was increased to 1867 1867, increasing the performance without changing the output.

Evaluation of Results

In the first step (lymphoid follicle exclusion), two models were tested in the described trial slide set to evaluate their ability to identify lymphoid follicles. Each image was divided into small patches (100 100 mm) and compared with the annotation of the pathologist (ground truth for lymphoid follicle detection). The patch would be considered as i) a true positive if the predicted area overlapped by more than 50% with ground truth, ii) a false positive if there was less than 50% overlap with ground truth, iii) a true negative if there was no positive prediction outside of ground truth areas, or iv) a false negative if there was no positive pre-diction inside a ground truth area. To evaluate the outcome of each model, accuracy was calculated based on the sum of the accuracy of all patches for all images.

In the second step, evaluation was performed at the slide level. After analysis, slides were labeled as metastasis or no metastasis based on the presence or absence of a tumor classification on the slide and as macrometastases, micro-metastases, or ITC based on the largest diameter of positive area measured manually with a ruler in HALO software. The maximum positive area identified by the algorithm was chosen if multiple metastatic foci were identified on a single slide. This result was compared against the ground truth, which was provided by the recorded diagnosis of the expert pathologist.

For optimal assessment of false-positive reduction on slides, the testing data set was split into two groups: with and without lymphoid follicles, due to the fact that not all lymph node slides contain lymphoid folli-cles. In this study, lymphoid follicles were defined as reactive lymphoid follicles with enlargement in shape and size, a prominent germinal center, a mantle zone, and numerous tingible body macrophages with mixed centroblasts and centrocytes. The slides in the group without lymphoid follicles had only small lymphocyte aggregations or a few small-to-normalesized lymphoid follicles. The two groups were then analyzed using both the one-step and the two-step methods, to evaluate the amount of error reduction achieved by the two-step method. The formula for false-positive area reduction was calculated as:

$$\text{Reduction of false positive ðFPÞ} \;\; z \;\; \frac{\text{FP area of one step FP area of two step}}{\text{FP area of one step}}$$

Statistical significance was determined using a one-sided t-test, with $P < 0.05$ considered statistically significant. Statistical analysis was performed using the Stata statistical software package version 14.2 (StataCorp LP, College Station, TX).

II. RESULTS

Lymphoid Follicle Exclusion in the First Step

The two models performed differently with respect to lymphoid follicle prediction. The accuracy of model 1 (LFRFM) and model 2 (LFCNN) was 51.7% and 94.5%, respectively. The LFCNN model showed a well-fitting shape with the original lymphoid follicles seen in hema-toxylin and eosin images, whereas the LFRFM showed many false positives in which tumor cells were mis-classified as lymphoid follicles (Figure 3). On the basis of these results, the LFCNN model was chosen to eliminate all lymphoid follicles from the slides prior to the second step.

Tumor Cell Detection in the Second Step

To evaluate, in detail, the tumor detection results at the slide level, a confusion matrix was created showing accuracy in percentage terms for tumor prediction in different sized metastases (Table 2). The two-step deep learning algorithm performed well in identifying positive slides, including all macrometastases, micrometastases, and ITC with 100% accuracy. Examples of metastasis prediction are displayed in Figure 4. By contrast, the algorithm worked poorly in identifying negative slides. All negative slides still retained some tiny foci of false positivity, which were considered by the algorithm as either ITC (31.6%) or micrometastasis (68.4%). Therefore, the sensitivity and specificity at the slide-based level of all slides were calculated as 100% and 0%, respectively.

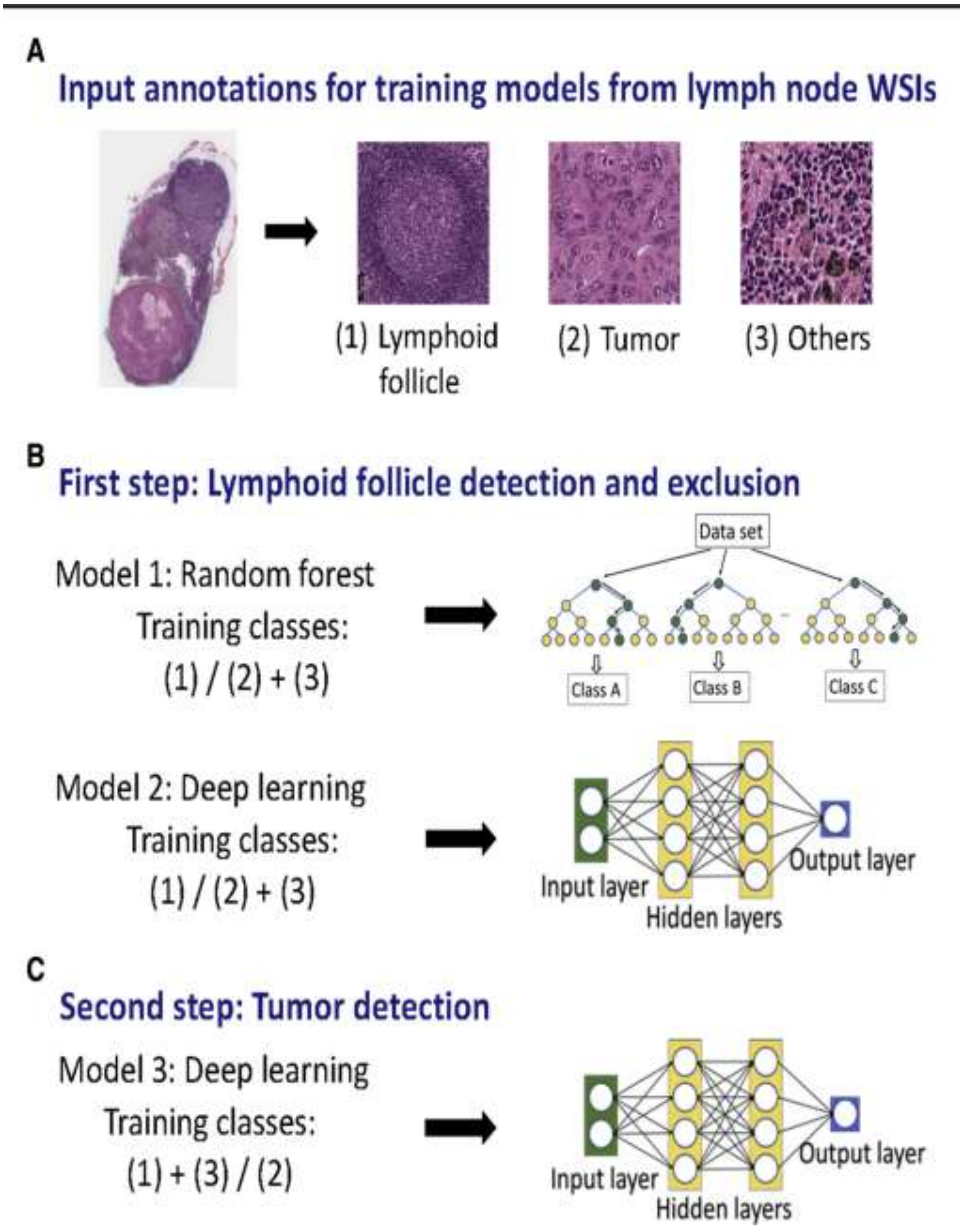


Figure 2 Input data information with detailed features of the three models used in the two-step method. A: Input data with three different types of annotation based on hematoxylin and eosin slides [numbered from (1) to (3)]. B: Two separated models (Lymphoid Follicle Random Forest Model and Lymphoid Follicle Convolutional Neural Network) were created in the first step to identify lymphoid follicles, with two classes of lymphoid follicles [slide (1)] and others including tumor [slides (2) þ (3)] used for training. C: A deep learning algorithm (Tumor Detection Convolutional Neural Network) was trained with two classes of tumor [slide (2)] and others including lymphoid follicle [slides (1) þ (3)] to detect tumor cells in the second step. WSI, whole-slide image.

With the above-described two-step approach, there was a significant reduction in error achieved by the deep learning tissue classification method compared with that of the one-step method. In addition, accurate identification of true cancer cells was retained (Figure 5A). In groups with and without lymphoid follicles, a remarkable elimination of false-positive areas was noted with the two-step approach. In slides that contained prominent reactive lymphoid folli-cles, an 89% reduction in false-positive area was reached by lymphoid follicle exclusion, with large false-positive foci

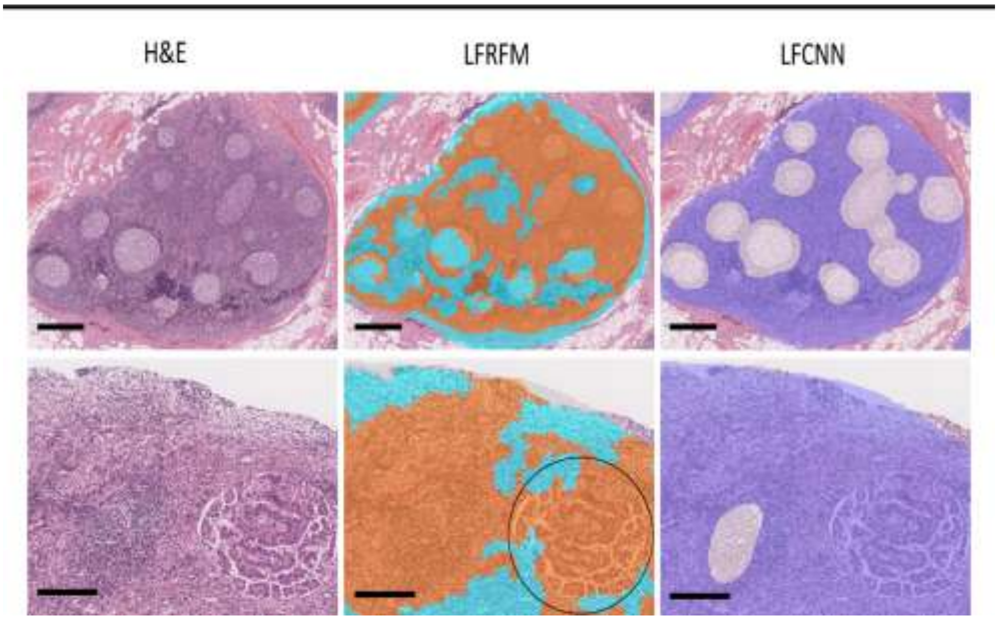


Figure 3 Lymphoid follicle detection task in the first step. The top row shows different shapes and sizes of lymphoid follicle prediction from two models. Lymphoid Follicle Convolutional Neural Network (LFCNN) shows better results (LFCNN, top row), concordant with the lymphoid follicles of the original image [hematoxylin and eosin (H&E), top row]. The bottom row shows false positives, in which the Lymphoid Follicle Random Forest Model (LFRFM) labels lymphoid follicles as tumors (LFRFM, bottom row). By contrast, there is no error in the same area of the LFCNN result (LFCNN, bottom row). LFRFM orange: lymphoid follicles, cyan: others. LFCNN white: lymphoid follicles, blue: others. Black circle indicates incorrect recognition. Scale bars: 500 mm (top row); 200 mm (bottom row).

accounting for the majority of this reduction. Including all slides, the results showed 36.5% and 5.4% reductions in false-positive area using the two-step algorithm in groups with and without lymphoid follicles, respectively;

$P < 0.0001$ (Figure 5B).

Some small false-positive foci that were unrelated to lymphoid follicles remained, causing a 0% specificity. Applying the solution of removing positive detection area by size, different levels of specificity and sensitivity were obtained. A receiver-operating characteristic (ROC) curve was plotted based on the largest diameter of the deleted positive area, to evaluate metastasis classification, which reached an area under the curve of 0.922 (Figure 6). The two best filters were 0.6 mm, which achieved a sensitivity and specificity of

79.6% and 96.5%, respectively, and 0.7 mm, which achieved a sensitivity and specificity of 75.5% and 98.2%, respectively. The data underlying the ROC curve are provided in Supplemental Table S2.

Table 2 Slide-Level Confusion Matrices Indicating Accuracy (in Percentages) of the Two-Step Deep Learning Algorithm Prediction with Variously Sized Metastases

	Prediction			
Ground-truth	Negative	ITC	Micro	Macro
Negative	0	31.6	68.4	0
ITC	0	100	0	0
Micro	0	0	100	0
Macro	0	0	0	100

ITC, isolated tumor cells; Macro, macrometastasis; Micro, micrometastasis.

The training and testing set was largely (95%) composed of the most common histologic types of lung cancer, namely adenocarcinoma and squamous cell carcinoma. Interest-ingly, the two-step deep learning algorithm could also detect lymph node metastases in much rarer types of lung cancer, including pleomorphic carcinoma, spindle cell carcinoma, large cell neuroendocrine carcinoma, and signet ring cell carcinoma (Supplemental Table S1 and Supplemental Figure S1).

III. DISCUSSION

This study describes the successful detection of lymph node metastases using a novel two-step deep learning approach within a commercially available deep learning platform. In addition, it is the first study on the detection of lung cancer lymph node metastases from WSIs. The two-step approach is a useful protocol, which can serve as a tool to assist with the work of the pathologist.

The subject of the CAMELYON challenge was breast cancer metastases, and although this is histologically different from lung cancer, metastatic diseases share the same overall characteristics: atypical cancer cells with hyperchromatic enlarged nuclei arranged in various patterns such as glands, clusters, etc, on the background of lymph node tissue. On that basis, the result of this study is com-parable with that of the CAMELYON challenge. Notably, our method achieved an accuracy that is relatively better than what has been achieved thus far in the CAMELYON challenges, particularly with regard to the detection of ITC and micrometastases.[21] Although ITC prediction was an issue for many CAMELYON teams, it was not in our case. ITC prediction rates for the CAMELYON teams were low, and their instruments performed poorly in terms of accuracy (0% to 34.3%), with the current best algorithm reaching only 11.4% accuracy. Only the top two teams could identify micrometastases well, with rates of 75.9% and 83.1%, whereas all other algorithms detected only two-thirds of all cases or less.[21] In this study, the sensitivity for detection of micrometastases and ITC (100% for both) was significantly higher than those of previous teams. In addition, various histologic subtypes were included for both training and testing steps, and the algorithm successfully detected these tumors in the testing step (Supplemental Table S1 and Supplemental Figure S1).

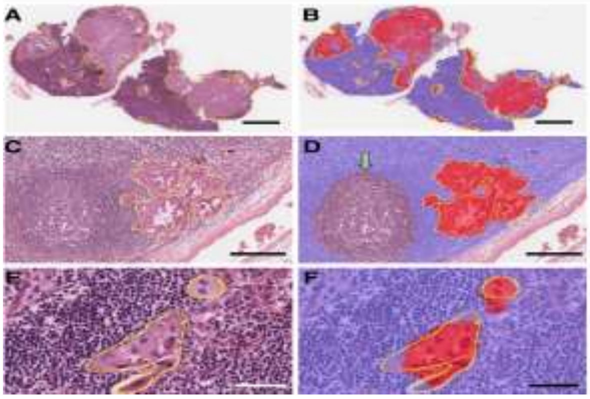
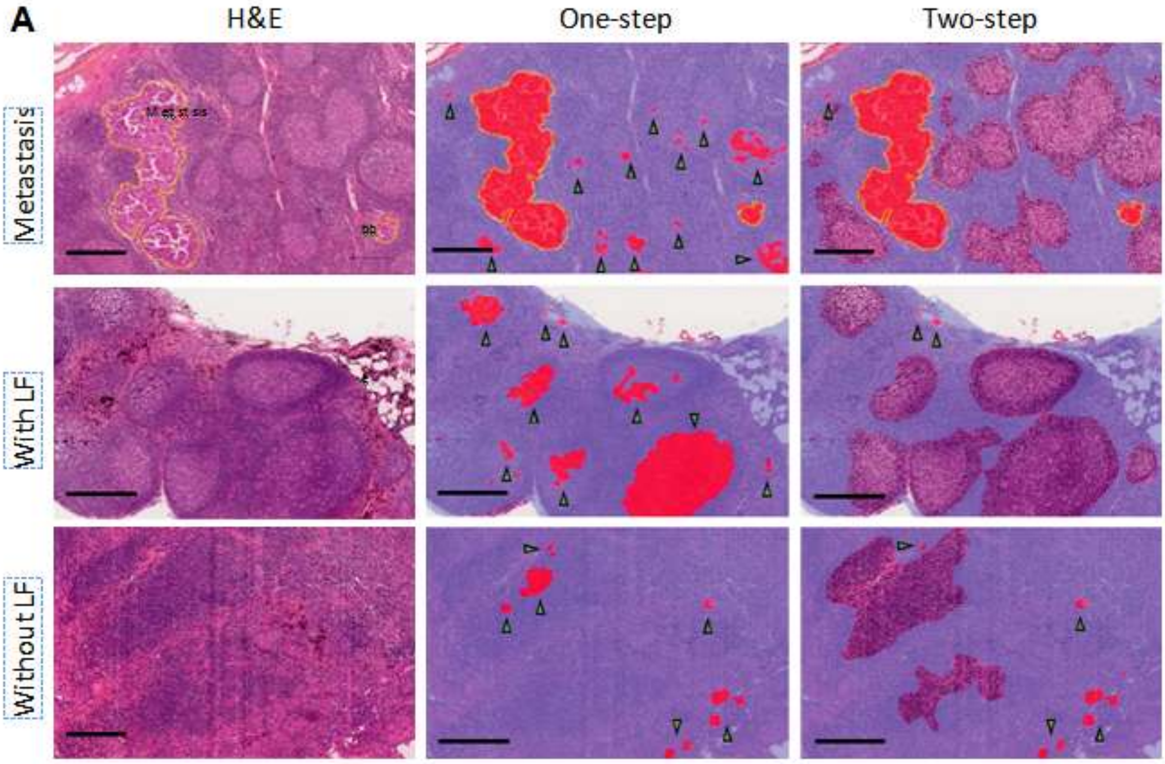


Figure 4 Detection of variously sized metas-tases using the two-step deep learning algorithm [Lymphoid Follicle Convolutional Neural Network (LFCNN) þ Tumor Detection Convolutional Neural Network (TDCNN)]. Macro-metastasis (A and B), micrometastasis (C and D), and isolated tumor cells (E and F) were detected by this method. The original hematoxylin and eosin staining in the lymphoid follicle area in the image (D, green arrow) suggests that it is totally deleted by LFCNN and cannot be further analyzed in the second step by TDCNN. Red: tumor; blue: others; yellow outline: annotation of tumor as ground-truth by pathologist. Scale bars: 2.0 mm (A and B, top row); 200 mm (C and D, middle row); 50 mm (E and F, bottom row).

Another strong point of this work is the development of a unique two-step approach to remove false-positive errors in metastatic tumor cell detection. In CAMELYON16, many of the submissions suffered from high false-positive error rates.[20] In CAMELYON17, teams tried, with limited suc-cess, to reduce this error by including hard-negative mining steps. In fact, their overall errors were worse, and the ac-curacy at the slide level decreased, with 67 of the 500 slides misclassified by the best-ranked team.[21] By observing the error exhibited in the results of CAMELYON16 and in our own experience, the germinal center area of lymphoid fol-licles were identified as a frequent source of false-positive error. Most of the false-positive foci that were success-fully eliminated by the exclusion step in this study were within large, active lymphoid follicles. The false-positive foci that remained after the exclusion step were outside of lymphoid follicles and mostly composed of small clusters of sinus histiocytes or fibroblasts (Figure 5A). Slides with reactive lymphoid follicles have the potential to have larger false-positive areas due to cancer-mimicking germinal cen-ter areas. As such, slides without reactive lymphoid follicles showed fewer prediction errors.

The effective exclusion step (first step) in this study was based on the better LFCNN algorithm chosen between two models: LFCNN and LFRFM. A small number of anno-tations were provided in the training set (65 annotations) for LFRFM owing to the fixed setting of the random forest classifier structure in the software. Although the random forest classifier achieved satisfactory results only with a small training data set in some studies,[28] it showed diffi-culty in separating the germinal center and marginal zone, two major components of the lymphoid follicle; this indi-cated inadequate accuracy of classifying heterogeneous subjects. Conversely, LFCNN is a deep learning algorithm that can benefit from a large amount of data to improve performance. A total of 2332 annotations with 3.35 10^4 iterations was considered to be a reasonable amount of data for training a deep learning algorithm. The regularity in shape and size of lymphoid follicles may explain the highly accurate predictions for a deep learningebased model such as LFCNN.



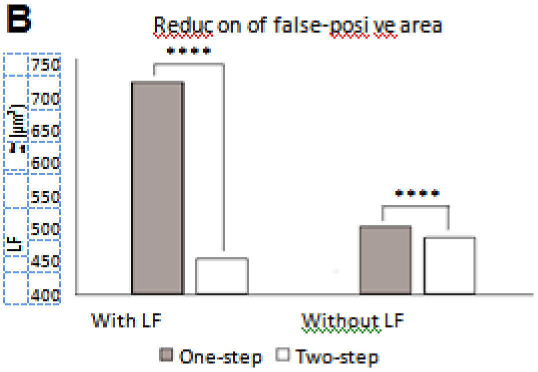


Figure 5 Tumor detection with reduction of the false positive area by two-step deep learning algorithms. A: One-step deep learning presents multiple false-positive foci (middle column), whereas two-step reduces almost all of them (right column). Note that correct tumor prediction still remains after the second step (A, two-step, top row). B: The average of false-positive areas shows statistically significant dif-ferences in error reduction between the one-step and two-step approaches in groups with and without lymphoid follicles. Red: tumor; blue: others; yellow outline: ground-truth of tumor; green arrowheads: false-positive foci. ****P < 0.0001. Scale bars: 500 mm (A, top and middle rows); 300 mm (A, bottom row). H&E, hematoxylin and eosin; LF, lymphoid follicles.

The low percentage of error reduction obtained in the group without lymphoid follicles and when combining all slide sets was due to the accumulation of multiple small pseudopositive areas on the WSI. Size filtering was a common method used in CAMELYON challenges to reduce false-positive rates.[20,21] In this study, an ROC curve was created using different sizes of false-positive foci as the filter and by comparing the sensitivity and specificity of metas-tasis detection. The size of removal foci could be up to 1.2 mm, which caused some missing cases of micrometastases. In this study, on the basis of the ROC curve, the two best size filters were 0.6 mm and 0.7 mm. Although these filters excluded some foci of true micrometastases, it is notable that the role of small metastases and ITCs in the prognosis of lung cancer is less certain when compared with macro-metastases. Some studies have shown no association be-tween small-sized metastases and patient survival rate, raising the suggestion that lung cancer patients should not be upstaged based on the presence of only small metasta-ses.[31,32] Depending on the purpose of the deep learning algorithm, used as a screening tool or as an assistant for confirmation of the pathologist’s diagnosis, different cutoff points can be used to achieve different sensitivities and specificities.

False-posive rate

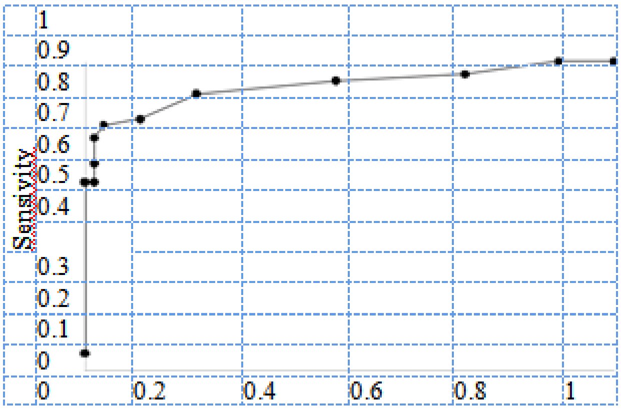


Figure 6 A receiver-operating characteristic (ROC) curve for slide-level metastasis detection with removal of different sizes of small positive-predictive foci. Adjusting the filter from 0 to 1.2 mm, an ROC curve of lymph node metastasis detection was generated by the two-step deep learning algorithm, which achieved an area under the curve of 0.922. The two best filters were 0.6 mm and 0.7 mm, indicating removal of small detections with sizes 0.6 mm and 0.7 mm, respectively. For the 0.6-mm filter, the sensitivity and specificity were 79.6% and 96.5%, respectively, and for the 0.7 mm filter, the sensitivity and specificity were 75.5% and 98.2%, respectively.

The requirement to exhaustively evaluate numerous lymph node slides in routine practice makes the pathol-ogist prone to fatigue and at risk of missing metasta-ses.[18,19] In previous studies, pathologists had more false negatives in metastatic detection (reduced sensitivity), and although deep learning algorithms could exceed the pa-thologist's sensitivity, they frequently paid a cost of decreased specificity (increased false positives).[21,33] Therefore, it is recommended to combine the strength of assistive algorithms with the specificity and expertise of pathologists in clinical practice, rather than relying on algorithms alone.[6,21] In other words, deep learning algo-rithms like the one described in this paper could be used to highlight areas of the tissue that should be evaluated more rigorously by the pathologist. Importantly, the concept of carcinoma detection on the background of reactive lymph nodes is not limited to lung cancer and can be applied to metastatic lymph nodes of different solid cancers; therefore, this approach has the potential to contribute to the healthcare workflow on a wider scale, where automation through deep learning is becoming an increasingly important component.[34e36]

There are several limitations in this study that require further exploration. First, the version of HALO AI software used in this study has a fixed probability threshold of 50% for the tumor class, causing inherently high sensitivity and low specificity. Using both steps of deep learning and filtering to remove small false positives, this limitation could be minimized. The next version of HALO AI will provide flexible settings for users to change the probability threshold, which may improve classification such that size filtering is no longer required. Another limitation is the relatively small number of cases with micrometastases (23 cases) and ITC (2 cases), as well as the number of cases in the validation data set that was used for the trials and the first-step models (10 cases).

Despite these limitations, this two-step deep learning method has potential to be applied in multiple research di-rections and could be incorporated into the pathology workflow in the future. To improve the existing algorithm's accuracy in metastasis detection, the number of slides collected in the daily workflow and the histologic spectrum of lung cancer types used for training will be expanded in future work. Using the next version of HALO AI, it will be possible to increase the tumor probability threshold to test whether this can be used to improve the balance between sensitivity and specificity in tumor prediction. The two-step approach will continue to be used to evaluate its usefulness in different lymph node slides of lung cancer and tumors of various other origins. On the basis of these results, the most promising algorithms and methods with the highest sensi-tivity and fair specificity will be used in a clinical trial, comparing the accuracy of the pathologist alone to the pathologist aided by algorithms, to investigate whether deep learning is additive to the existing clinical workflow. In the current condition of pathologist shortages worldwide, this method could be effectively integrated into the pathology workflow, potentially improving the quality of diagnoses and reducing the pathologist's workload.

In summary, using deep learning software with a two-step classification approach, it is possible to detect lung cancer metastases in lymph node tissue with high sensitivity, regardless of histologic type. An initial classification step can be used to effectively remove false positive predictions caused by lymphoid follicles. In light of these results, this method may be a useful addition to the clinical workflow.

IV. ACKNOWLEDGMENTS

H.H.N.P., M.F., T.F., K.K., and J.F. conceived and designed the study; H.H.N.P. wrote the manuscript; H.H.N.P. and A.B. acquired data; H.H.N.P. performed sta-tistical analysis; A.B. and J.F. revised the manuscript; and J.F. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Supplemental Data
Supplemental material for this article can be found at http://doi.org/10.1016/j.ajpath.2019.08.014.

REFERENCES

[1]. Raab SS, Grzybicki DM, Janosky JE, Zarbo RJ, Meier FA, Jensen C, Geyer SJ: Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer 2005, 104:2205e2213

[2]. Nakhleh RE: Error reduction in surgical pathology. Arch Pathol Lab Med 2006, 130:630e632

[3]. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson ANA, Nelson HD, Pepe MS, Allison KH, Schnitt SJ, O'Malley FP, Weaver DL: Diagnostic concordance among patholo-gists interpreting breast biopsy specimens. JAMA 2015, 313: 1122e1132

[4]. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B: Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009, 2:147e171

[5]. Holten-Rossing H, Talman M-LM, Jylling AMB, Lænkholm A-V, Kristensson M, Vainer B: Application of automated image analysis reduces the workload of manual screening of sentinel lymph node biopsies in breast cancer. Histopathology 2017, 71:866e873

[6]. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, Thng F, Peng L, Stumpe MC: Impact of deep learning assistance on the histopathologic review of lymph nodes for meta-static breast cancer. Am J Surg Pathol 2018, 42:1636e1646

[7]. Ghaznavi F, Evans A, Madabhushi A, Feldman M: Digital imaging in pathology: whole-slide imaging and beyond. Annu Rev Pathol 2013, 8:331e359

[8]. Krizhevsky A, Sutskever I, Hinton GE: ImageNet classification with deep convolutional neural networks. Commun ACM 2017, 60: 84e90

[9]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: Going deeper with convolutions. Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition: Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE), 2015. pp. 1e9

[10]. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen-van de Kaa C, Bult P, van Ginneken B, van der Laak J: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016, 6:26286

[11]. Cires¸an DC, Giusti A, Gambardella LM, Schmidhuber J: Mitosis detection in breast cancer histology images with deep neural net-works. Med Image Comput Comput Assist Interv 2013, 16:411e418

[12]. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A: A deep con-volutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016, 191:214e223

[13]. Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A: Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Edited by Gurcan MN, Madabhushi A. In SPIE Proceedings Vol. 9041: Medical Imaging 2014: Digital Pathology. Bellingham WA: International Society for Optics and Photonics, 2014. pp. 904103

[14]. Ertosun MG, Rubin DL: Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. AMIA Annu Symp Proc 2015, 2015:1899e1908

[15]. Ehteshami Bejnordi B, Mullooly M, Pfeiffer RM, Fan S, Vacek PM, Weaver DL, Herschorn S, Brinton LA, van Ginneken B, Karssemeijer N, Beck AH, Gierach GL, van der Laak JAWM, Sherman ME: Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. Mod Pathol 2018, 31:1502e1512

[16]. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C, Lundin J: Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018, 8

[17]. Gress DM, Edge SB, Greene FL, Washington MK, Asare EA, Brierley JD, Byrd DR, Compton CC, Jessup JM, Winchester DP, Amin MB, Gershenwald JE: Principles of cancer staging. Edited by Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, Sullivan DC, Jessup JM, Brierley JD, Gaspar LE, Schilsky RL, Balch CM, Winchester DP, Asare EA, Madera M, Gress DM, Meyer LR. In AJCC Cancer Staging Manual. Cham, Switzerland: Springer International Publishing, 2017. pp. 3e30

[18]. van Diest PJ: Histopathological workup of sentinel lymph nodes: how much is enough? J Clin Pathol 1999, 52:871e873

[19]. Weaver DL, Krag DN, Manna EA, Ashikaga T, Harlow SP, Bauer KD: Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. Mod Pathol 2003, 16: 1159e1163

[20]. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al: Diagnostic assess-ment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017, 318: 2199e2210

[21]. 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: 550e560

[22]. Andre F, Grunenwald D, Pignon J-P, Dujon A, Pujol JL, Brichon PY, Brouchet L, Quoix E, Westeel V, Le Chevalier T: Survival of patients with resected N2 nonesmall-cell lung cancer: evidence for a subclassification and implications. J Clin Oncol 2000, 18:2981e2989

[23]. Betticher DC, Hsu Schmitz S-F, Tötsch M, Hansen E, Joss C, von Briel C, Schmid RA, Pless M, Habicht J, Roth AD, Spiliopoulos A, Stahel R, Weder W, Stupp R, Egli F, Furrer M, Honegger H, Wernli M, Cerny T, Ris H-B: Mediastinal lymph node clearance after docetaxel-cisplatin neoadjuvant chemotherapy is prognostic of survival in patients with stage IIIA pN2 none small-cell lung cancer: a multicenter phase II trial. J Clin Oncol 2003, 21:1752e1759

[24]. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018, 68:394e424

[25]. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH: Deep Learning for Identifying Metastatic Breast Cancer. arXiv 2016. arXiv:1606.05718v1

[26]. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado GS, Hipp JD, Peng L, Stumpe MC: Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv 2017. arXiv:1703.02442

[27]. James DB, Mary KG, Christian W: TNM Classification of Malignant Tumours. ed 8. Hoboken, NJ: John Wiley & Sons, 2018

[28]. Han T, Jiang D, Zhao Q, Wang L, Yin K: Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Trans Inst Meas Control 2018, 40:2681e2693

[29]. Long J, Shelhamer E, Darrell T: Fully Convolutional Networks for Semantic Segmentation. arXiv 2014. arXiv:1411.4038.

[30]. Tieleman T, Hinton G: Lecture 6.5: RMSProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning 2012, 4:26e31

[31]. Marchevsky AM, Gupta R, Kusuanco D, Mirocha J, McKenna RJ: The presence of isolated tumor cells and micrometastases in the intratho-racic lymph nodes of patients with lung cancer is not associated with decreased survival. Hum Pathol 2010, 41:1536e1543

[32]. Marchevsky AM, Qiao J-H, Krajisnik S, Mirocha JM, McKenna RJ: The prognostic significance of intranodal isolated tumor cells and micrometastases in patients with nonesmall cell carcinoma of the lung. J Thorac Cardiovasc Surg 2003, 126:551e557

[33]. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al: Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018, 15:20170387

[34]. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI: A survey on deep learning in medical image analysis. Med Image Anal 2017, 42:60e88

[35]. Baskin II, Winkler D, Tetko IV: A renaissance of neural networks in drug discovery. Expert Opin Drug Discov 2016, 11:785e795

[36]. Miotto R, Wang F, Wang S, Jiang X, Dudley JT: Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018, 19:1236e1246