Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection

Insights Into the Black Box for Pathologists

Yun Liu, PhD; Timo Kohlberger, PhD; Mohammad Norouzi, PhD; George E. Dahl, PhD; Jenny L. Smith, MD; Arash Mohtashamian, MD; Niels Olson, MD; Lily H. Peng, MD, PhD; Jason D. Hipp, MD, PhD; Martin C. Stumpe, PhD

• ***Context.***—**Nodal metastasis of a primary tumor influences therapy decisions for a variety of cancers. Histologic identification of tumor cells in lymph nodes can be laborious and error-prone, especially for small tumor foci.**

Objective.—**To evaluate the application and clinical implementation of a state-of-the-art deep learning–based artificial intelligence algorithm (LYmph Node Assistant or LYNA) for detection of metastatic breast cancer in sentinel lymph node biopsies.**

Design.—**Whole slide images were obtained from hematoxylin-eosin–stained lymph nodes from 399 patients (publicly available Camelyon16 challenge dataset). LYNA was developed by using 270 slides and evaluated on the remaining 129 slides. We compared the findings to those obtained from an independent laboratory (108 slides from 20 patients/86 blocks) using a different scanner to measure reproducibility.**

Results.—**LYNA achieved a slide-level area under the receiver operating characteristic (AUC) of 99% and a tumor-level sensitivity of 91% at 1 false positive per patient on the Camelyon16 evaluation dataset. We also identified 2 "normal" slides that contained micrometastases. When applied to our second dataset, LYNA achieved an AUC of 99.6%. LYNA was not affected by common histology artifacts such as overfixation, poor staining, and air bubbles.**

Conclusions.—**Artificial intelligence algorithms can exhaustively evaluate every tissue patch on a slide, achieving higher tumor-level sensitivity than, and comparable slide-level performance to, pathologists. These techniques may improve the pathologist's productivity and reduce the number of false negatives associated with morphologic detection of tumor cells. We provide a framework to aid practicing pathologists in assessing such algorithms for adoption into their workflow (akin to how a pathologist assesses immunohistochemistry results).**

(*Arch Pathol Lab Med.* doi: 10.5858/arpa.2018-0147-OA)

Reviewing sentinel lymph node biopsies for evidence of metastasis is an important feature of breast cancer staging, concurrently impacting clinical staging and treatment decisions.[1] However, reviewing lymph nodes for the presence of tumor cells is a tedious, time-consuming, and potentially error-prone process. Although obtaining additional sections from the tissue block and performance of immunohistochemical (IHC) staining improve detection sensitivity, these techniques are associated with increased workload, costs, and reporting delays.

With the approval[2] and gradual implementation of whole slide imaging for primary diagnosis, utilization of computer-aided image analysis is becoming more feasible in routine diagnostic settings. In recent years, deep learning,[3] a kind of computer algorithm loosely inspired by biological neural networks, has significantly improved the ability of computers to identify objects in images.[4,5] In medicine, deep learning was used to diagnose referable diabetic retinopathy or diabetic macular edema and skin cancer with accuracy comparable to that of board-certified ophthalmologists or dermatologists.[6–8] Many other works using machine learning or deep learning for breast and other malignancies have been published.[9,10] Moreover, deep learning–based algorithms accurately detect metastatic breast cancer in lymph nodes, based on both slide-level and tumor-level receiver-operating-characteristic performance metrics.[11–13] The diagnostic accuracy of these algorithms is comparable to that of pathologists without time constraint, and significantly more accurate than pathologists in a simulated (1 minute per slide) environment.[14]

Accepted for publication August 27, 2018.

Supplemental digital content is available for this article. See text for hyperlink.

From Google AI Healthcare, Google Research, Mountain View, California (Drs Liu, Kohlberger, Norouzi, Dahl, Peng, Hipp, and Stumpe); and Laboratory Department, Naval Medical Center, San Diego, California (Drs Smith, Mohtashamian, and Olson).

Drs Liu, Kohlberger, Norouzi, Dahl, Peng, Hipp, and Stumpe are employees of Google Inc and own stock in the company. Drs Mohtashamian, Smith, and Olson have no relevant financial interest in the products or companies described in this article.

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the US Government.

Corresponding author: Jason D. Hipp, MD, PhD, Google AI Healthcare, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (email: hipp@google.com).

Computer algorithms may significantly improve a pathologist's workflow. However, from a clinical perspective, they have not achieved wide-scale acceptance, and from a regulatory perspective they have not yet been fully examined. Despite good performance metrics, the safety and quality[15] profile of these algorithms have not been completely addressed. The practicing pathologist has little technical understanding of the underlying algorithms, diagnostic accuracy and error rates, as well as the utility of these programs in clinical practice.

In this study, we applied the current state-of-the-art algorithm (LYmph Node Assistant, or LYNA) to the Cancer Metastases in Lymph Nodes 2016 challenge dataset (Camelyon16),[14] as well as a set of 108 images from a different source. We analyze LYNA's performance by dividing our analysis into 2 sections: one at the whole-slide level and the other at the level of tissue patches, or regions of interest (ROIs). We show that LYNA is unaffected by common histology artifacts such as poor fixation, overstaining, chatter, and coverslip bubbles. The technology, as such, can be used by various institutions with different histopathology laboratory procedures and digital infrastructure. LYNA, however, at times did falsely identify giant cells, germinal centers, and histiocytes as tumor foci. While these misidentified cells are readily diagnosed by an experienced pathologist, we hypothesize that the image distortion may have confused the algorithm. We will discuss these findings and show preliminary data on a small cohort of images that indicate that even without further modification of the LYNA algorithm, it is capable of identifying other metastatic tumors in lymph nodes. Lastly, we anticipate that our analysis into the quantitative performance, examination of how the algorithm works, and comprehensive tissue-level error analysis can be used as a framework for evaluation of future algorithms and their implementation into the clinical workflow.

MATERIALS AND METHODS

Study Design and Image Acquisition

We obtained data from 2 sources: the Camelyon16[14] challenge containing 399 slides, and a separate dataset ("DS2") that we digitized, containing 108 slides from 20 patients (86 tissue blocks). In the Camelyon16 dataset, 244 slides were digitized with a 3DHISTECH (Budapest, Hungary) Pannoramic 250 Flash II digital slide scanner and 155 slides were digitized with a Hamamatsu (Hamamatsu, Japan) XR C12000 digital slide scanner. DS2 slides were digitized with an Aperio (Leica Biosystems, Buffalo Grove, Illinois) AT2 whole-slide scanner. All slides were scanned at a resolution of 0.24 ± 1 μm per pixel. The number of slides containing metastasis and the purpose of each dataset are detailed in Table 1. Institutional review board waived the need for informed consent for the use of Camelyon16 slides. All work related to the DS2 dataset was approved by their Institutional review board.

Ground Truth Determination for Training and Validation

The Camelyon16 ground truth diagnoses were provided by the organizers and are described in detail elsewhere.[14] Briefly, the slides were reviewed by 1 of 2 pathologists, and cases that were interpreted as negative on hematoxylin-eosin (H&E) alone, or were otherwise considered diagnostically challenging, were subjected to IHC confirmation (anti-cytokeratin, CAM 5.2 [BD Biosciences, San Jose, California]).

For DS2, two US board-certified pathologists (with at least 10 years of experience) independently graded each slide as macrometastasis, micrometastasis, isolated tumor cells (ITCs), or negative. The pathologists also provided a brief description of the tumor location within micrometastasis- and ITC-containing slides for ease of adjudication. A third pathologist reviewed the slides when the 2 pathologists rendered conflicting diagnoses. The third pathologist referred to additional sections from the tissue block or reviewed the corresponding cytokeratin IHC stain for challenging cases. To remain consistent with the Camelyon16 challenge evaluation and College of American Pathologists guidelines (since ITCs are considered N0), we excluded two DS2 slides that contained only ITCs. For transparency, a pathologic analysis of LYNA's performance on these 2 slides is included in the Results section.

Computer Image Analysis: LYNA (Algorithm) Development

Our deep learning–based image analysis workflow is divided into 2 stages: algorithm development and algorithm application (Figure 1). To develop the algorithm, we randomly sampled square image patches of size 128 pixels at high power (≈32 μm on a side). This patch size was selected to encompass several cells and was also used by Litjens et al.[11] The algorithm takes as input a larger square patch of size 299 pixels (≈75 μm) to provide additional context akin to how a pathologist reviews a slide. We used 299 pixels as the input size because it is the default input size of the deep learning architecture used, Inception (V3).[16]

A deep learning architecture is a series of mathematical operations arranged in a hierarchy of layers. Earlier layers tend to produce low-level image features (such as edges), and later layers use the low-level features to construct more abstract features (such as shapes).[4] A simple tool for visual exploration of deep learning features is available at playground.tensorflow.org. While the operations and their order are predetermined by the architecture, the parameters of the operations are automatically learned, a process called *training* in the machine learning literature. The correct predictions are termed *labels*, which are determined and annotated by pathologists by outlining tumors at the pixel level. Specifically, for each slide in the Camelyon16 dataset, tumors (if any) are outlined at the pixel level. When extracting image patches, we also extract the corresponding label of the tissue patch (benign: 0 or tumor: 1) and train the algorithm by repeatedly adjusting the weights of the algorithm to reduce the error on the image patches seen by the algorithm.

We improve upon a previously published algorithm[12] by increasing the ratio of normal to tumor patches seen by the algorithm to 4:1 to reduce false positives. We further enhanced the computational efficiency of the training process, which improved the diversity of tissues "seen" by the algorithm during its training phase. These changes substantially increased both tumor-level sensitivity and slide-level area under the receiver operating characteristic (AUC, see Results). A detailed explanation of our deep learning algorithm is included in the Supplemental Digital Content.

Computer Image Analysis: Algorithm Usage

After training, LYNA was used by exhaustive application across the slide. This creates a 2-dimensional table of numbers, where each number indicates the predicted tumor likelihood of the corresponding ≈32-μm square tissue patch. In practice, we only applied LYNA to patches that contained tissue by using a conservative threshold similar to that of Janowczyk and Madabhushi.[9] These predictions were visualized as a heatmap, where blue indicates low and red indicates high likelihood, or with regions of high tumor likelihood highlighted (Figure 1). To obtain a slide-level prediction, we used the maximum predicted value across all 100,000 high-magnification (×40) patches in each slide.

Computer Image Analysis: Color Normalization

Hematoxylin-eosin slides typically vary dramatically in appearance across institutions because of factors such as tissue preparation, staining protocols, and oxidation in the laboratory. When these slides are digitized by a whole slide scanner, additional variation can be introduced (eg, by digital white balance). We found that normalizing for these variations[12] improved algorithm

Table 1. Characteristics of Datasets Used in This Study					
Dataset	Purpose	No. of Slides	No. of Patients	No. of Slides With Macrometastasis	No. of Slides With Micrometastasis
Camelyon16	Algorithm development	270 (160 normal, 110 tumor)	270	49 (18%)[a]	61 (23%)[a]
Camelyon16	Evaluation	129 (80 normal, 49 tumor)	129	21 (16%)	28 (22%)
Dataset2 ("DS2")	Further evaluation	108 (52 normal, 54 tumor, 2 ITC-only slides removed)	20 (86 blocks)	50 (47%)	4 (4%)

Abbreviation: ITC, isolated tumor cells.

[a] Estimated from the pixel-level tumor annotations provided by the Camelyon16 organizers.

performance when used in conjunction with the other improvements described above. Briefly, we transformed the colors into a hue-saturation-density space[17] that accounts for the nonlinear relationship between stain (such as H&E) amount and pixel intensity values. We then apply a slide-specific transformation[18] to change each slide's color statistics to a reference slide. This is a simplified version of the approach described by Bejnordi et al.[19] For this work, we used the median color statistics across the training set as the reference; these statistics were not optimized for pathologist review. Figure 2, A through F, shows sample image patches before and after normalization by our approach. This color normalization was applied only for LYNA review and not pathologists' review for the images in this study.

Unpacking the Black Box: Examining Mechanisms of LYNA's Predictions

We examined how LYNA made its predictions by computing the amount that each pixel affects LYNA's output prediction, in other words, the gradient of the output with respect to each input pixel. Next, we smoothed these gradients by averaging across several versions of the input image that have some noise artificially added to the pixels, a technique called *SmoothGrad*.[20] In each input patch, we selected the 5% most important pixels. Using the percentile instead of a fixed value allowed this threshold to be invariant to the absolute gradient magnitudes in each image. Finally, we visualized the pixels in the original image that were within 1 μm of these "important pixels." From our experience, because truly important pixels tended to cluster around a few cells, this final "importance dilation step" typically highlighted whole cells, making the final image more easily interpretable (Results).

Statistical Analysis

LYNA was evaluated by using metrics based on receiver-operating-characteristic curves (ROCs) at the slide level and tumor level as in the Camelyon16 challenge.[14] The slide-level area under ROC (AUC) was used to assess the ability of LYNA to discriminate between benign and metastasis-containing slides. The tumor-level ROC (called free-response ROC, or FROC) was used to assess the sensitivity of LYNA for individual tumor foci at various numbers of false positives flagged per slide. We report the sensitivities at several false-positive rates: 0, 0.25, 1, and 8. The last 3 values were chosen to match the official Camelyon16 evaluation metrics, while "0" allows comparisons with the reported performance of a pathologist in the challenge, who did not make any false-positive diagnosis.

RESULTS

Quantitative Evaluation of LYNA

LYNA operates by exhaustively generating a prediction for each tissue patch on every slide. Some examples of these predictions are shown in Figure 3, A through P. The maximum value (corresponding to the most suspicious tissue patch) across the 100,000 predictions in each slide is the slide-level prediction. Thus we evaluate both slide-level predictions and the patch-level predictions. Using the Camelyon16 evaluation set, LYNA achieved a slide-level AUC (nodal metastasis: present or absent) of 99.3% (95% CI, 98.1%–100%), similar to the previous best of 99.4%. The second- and third-ranked teams achieved 97.6% and 96.4%, respectively.[14] For comparison, the Camelyon16 challenge also tasked a practicing pathologist with evaluating the same slides digitally, where they achieved an AUC of 96.6% without any time constraint. Under time constraint of a minute per slide, the average of 11 pathologists' AUC was significantly lower at 81.0%. On the Camelyon16 dataset, setting LYNA's threshold to capture all of the positive cases (ie, 100% sensitivity and negative predictive value), which might be used, for example, to prioritize review of positive

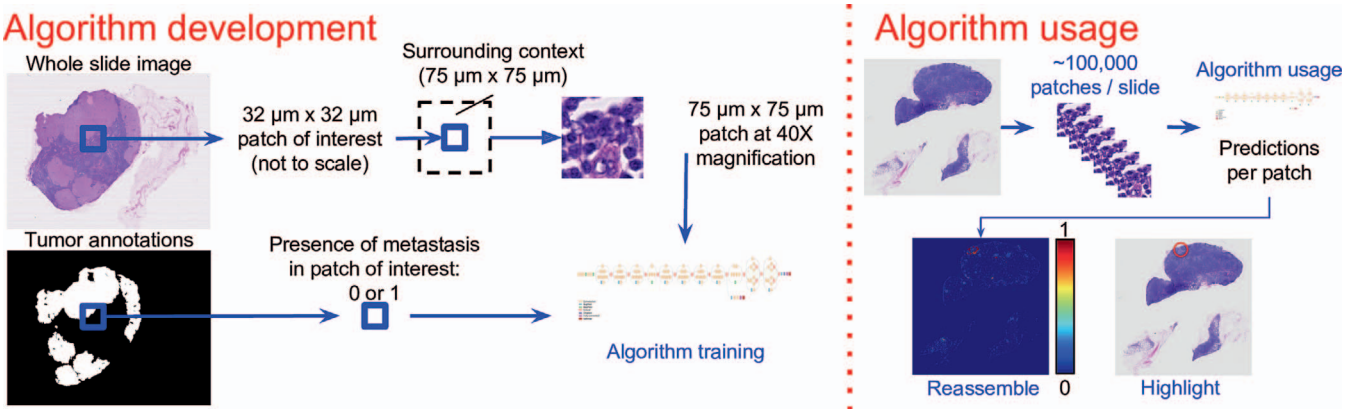

Figure 1. *Overview of patch-based methodology for hematoxylin-eosin–stained images. Each pixel corresponds to approximately 0.25 μm (32 μm = 128 pixels; 75 μm = 299 pixels). The algorithmic usage section for a slide can be completed in under a minute on average on a cloud computer platform.*

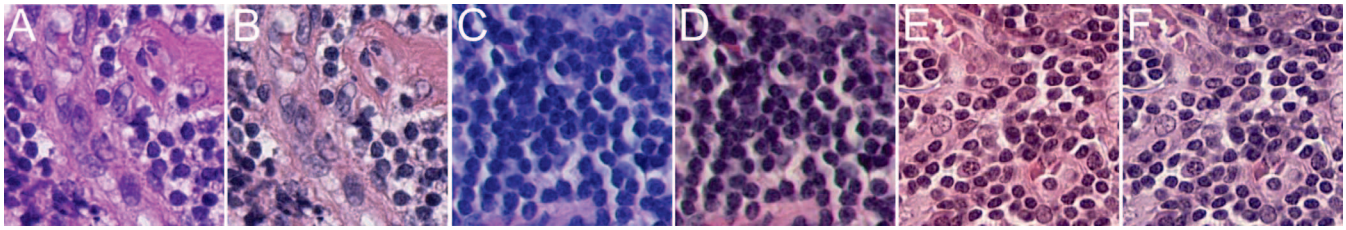

Figure 2. *Color normalization to reduce histology and scanner variations, applied to hematoxylin-eosin–stained images. Representative patches at ×40 magnification (75 μm) were selected from 3 representative cases (in terms of stain appearance). Colors were normalized to the median color statistics of the training set slides and were not optimized for pathologist review. A and B, Slide Normal_102 at ×40 magnification: (A) eosinophilic appearance and (B) after color normalization. C and D, Slide Test_122 at ×40 magnification: (C) basophilic appearance and (D) after color normalization. E and F, Slide Test_007 at ×40 magnification: (E) original appearance and (F) after color normalization.*

slides, results in a positive predictive value of 79% (49 of 62 slides). A threshold that captured all of the negative cases (ie, 100% specificity and positive predictive value), used for example to order IHC stains in advance (if consistent with institutional practice for reviewing negative cases), results in a negative predictive value of 96% (80 of 83 slides).

However, when evaluated for its ability to detect all the tumors on every slide, LYNA performed significantly better than the winning algorithm in Camelyon16. LYNA detected all 40 macrometastases without any false positives, and achieved 69% sensitivity (161 of 225 tumor foci; 95% CI, 63%–78%) for all the metastases before the first false-positive patch-level prediction. When allowed 1 patch-level false-positive prediction per tumor-negative slide, LYNA achieved a sensitivity of 91% (205 of 225 tumor foci), halving the false-negative rate relative to the previous best result of 81%[14] (183 of 225 tumor foci). The second- and third-ranked teams achieved 75% (168 of 225 tumor foci) and 73% (164 of 225 tumor foci), respectively.[14] For comparison, the Camelyon16 practicing pathologist achieved a sensitivity of 72% (163 of 225 tumor foci) after spending 30 hours reviewing 129 slides. Results on other false-positive thresholds are reported in Table 2.

LYNA also detected micrometastases in 2 "normal" slides (Normal_086 and Normal_144; Figure 3, A though F), which were not detected by other participants in Camelyon16 before LYNA. These were brought to the attention of and confirmed by the Camelyon16 challenge organizers, who confirmed that tumor was present in the normal slides. Fortunately, these were data processing errors, and the patients' diagnoses were correct.

Of the 108 slides in the DS2 dataset, 2 slides contained only ITCs and were excluded from the performance metrics as well to match diagnostic guidelines[21] and the Camelyon16 protocol,[14] where no slides contained only ITCs. LYNA confidently detected the ITCs in 1 of these 2 slides but not the other (Figure 3, M through P). On the remaining 106 slides in DS2, LYNA achieved a similar slide-level AUC of 99.6% (95% CI, 98.8%–100%). On the DS2 dataset, a threshold that captures all of the positive cases results in a positive predictive value of 87% (54 of 62 slides), and a threshold that captures all of the negative cases results in a

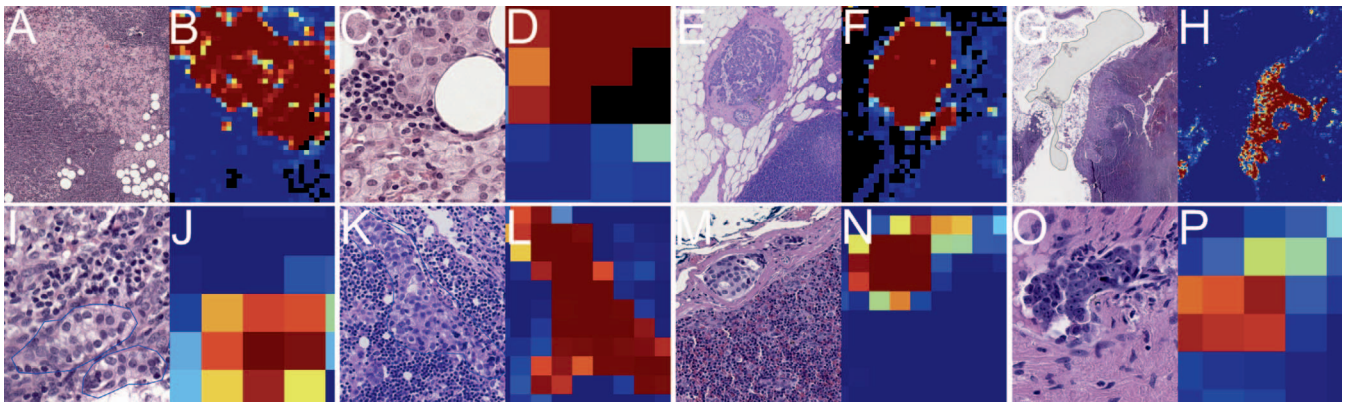

Figure 3. *Pairs of sample hematoxylin-eosin–stained images (left) and corresponding LYmph Node Assistant (LYNA) algorithm prediction heatmaps (right). On the heatmap, blue indicates lower predicted likelihood of tumor, red indicates higher likelihood, and colors in between indicate intermediate likelihood (color bar at the bottom of Figure 1). A through D, Slide Normal_086 (mislabeled in dataset as benign). A and B, At ×5 magnification, LYNA detects a micrometastasis in the top half, ignoring the histiocytes in the bottom right quadrant. C and D, View of the tumor-histiocyte boundary at ×40 magnification. E and F, Slide Normal_144 (mislabeled in dataset as benign). Extranodal extension is visible at ×5 magnification and is identified by LYNA as tumor. G and H, DS2 slide (labeled macrometastasis). The macrometastasis is visible at ×2.5 magnification and is identified by LYNA despite numerous issues in same field of view: air bubble, cutting artifacts, hemorrhagic, and necrotic and poorly processed tissue. I and J, Slide Test_011 (labeled as lobular micrometastasis), with the reference Camelyon16 annotation (verified by immunohistochemistry [IHC]) displayed as a thin blue outline around the tumor. Even at ×40 magnification, the focus is small and has poor contrast relative to histiocytes at the top left corner. LYNA identified this focus as tumor (albeit with moderate confidence). K and L, Slide Test_110 (labeled as lobular micrometastasis), with the reference Camelyon16 annotation (verified by IHC) displayed as a thin blue outline around the tumor. At ×20 magnification, the small size and unusual morphology of the tumor are apparent. LYNA identified this focus as tumor. M and N, DS2 slide (labeled ITC). The slide contained multiple ITC foci (as small as 50 μm), one of which is shown at ×20 magnification. LYNA detected all of the foci. O and P, DS2 slide (conservatively labeled ITC). This was a challenging case, because the foci (shown at ×40 magnification) did not appear on other levels. LYNA detected false positives in this slide with higher confidence, and thus we consider this a false negative.*

Table 2. Summary of Quantitative Evaluation at the Slide and Individual-Tumor Level

Method	Slide-Level Area Under Receiver Operating Characteristic Curve (AUC)	Macrometastasis Sensitivity (Tumor-Level; 0 FP)	Macrometastasis and Micrometastasis (Tumor-Level) Sensitivity (%) at x Average FP per Slide (95% Confidence Intervals)			
			0 FP	0.25 FP	1 FP	8 FP
LYNA (our algorithm)	**99.3** (98.1, 100)	**100** (100, 100)	**71.6** (65.9, 78.2)	**86.2** (80.3, 91.9)	**91.1** (87.3, 94.7)	**95.6** (92.8, 98.2)
Camelyon16 winning algorithm	**99.4** (98.3, 99.9)	[a]	[a]	77.3	81.3	82.7
Camelyon16 runner-up algorithm	97.6 (94.1, 99.9)	[a]	[a]	66.7	74.7	83.1
Single pathologist (without time constraint)	96.6 (92.7, 99.8)	[a]	**72.4** (64.3, 80.4)	[a]	[a]	[a]
Average of 11 pathologists (simulated clinical time constraint)	81.0 (73.8, 88.4)	[a]	[a]	[a]	[a]	[a]

Abbreviations: FP, false positives; LYNA, LYmph Node Assistant.

The LYNA algorithm predictions were compared with the pathologist-annotated ground truth masks provided by the Camelyon16 organizers. Values within the confidence intervals of LYNA are highlighted in bold.

[a] Not reported.

negative predictive value of 95% (52 of 55 slides). We did not collect pixel-level annotations for DS2 and thus were unable to compute the tumor-level sensitivity. However, we conducted an exhaustive analysis of the patch-level errors (detailed in a following section).

In general, we observed that LYNA was insensitive to a variety of artifacts such as cautery, air bubbles, cutting artifacts, hemorrhage, necrosis, and poor processing. A sample field of view containing many of these artifacts and LYNA's predictions is shown in Figure 3, G and H.

False-Negative and False-Positive Slides

Next, we exhaustively analyzed all of the false positives in Camelyon16 and DS2 by using a threshold that resulted in no slide-level false negatives. The threshold used for the 2 datasets was slightly different because the case mixes were different (eg, macrometastasis versus micrometastasis; Table 1). We then divided these errors by their causes into either the histology or scanning workflow component. These results are summarized in Table 3. The histology-related errors involved fixation and tissue processing quality, and floaters or contaminants; the errors originating from the scanning workflow included out-of-focus (OOF) germinal centers, histiocytes, and multinucleated giant cells. These were, in general, easy for pathologists to rule out during the review process.

With these no-false-negative thresholds (100% sensitivity in 49 tumor-containing cases), LYNA had 84% specificity (67 of 80 slides), and a positive predictive value of 79% (49 of 62 slides) in the Camelyon16 dataset. At the level of tumor detection, the macrometastasis sensitivity was 100% (22 of 22 slides) at zero patch-level false positives. In other words, all false negatives were small foci. Frequently, these small foci were technically above the 200-μm cutoff for micrometastasis, but contained far fewer than the other cutoff criterion for micrometastasis: 200 cells (usually <20 cells). In 1 slide (Test_051), many small extranodal tumor foci were surrounded by fat and in poor focus, resulting in missed detections. In cases such as this, the (larger) intranodal foci were detected correctly, leading to correct case- and slide-level diagnosis. In the DS2 dataset, LYNA had 85% specificity (44 of 52 slides), and a positive predictive value of 87% (54 of 62 slides).

Unpacking the Black Box

To understand how LYNA worked, we examined the most important pixels used to make the prediction for each input image. In the metastatic cancer, LYNA focused its attention primarily on nuclear features, such as pleomorphism and hyperchromasia (Figure 4, A through D). In fields of view containing giant cells, LYNA focused on the crowded and overlapping nuclei (Figure 4, E and F). In fields of view containing OOF nuclei, LYNA focused on nuclei from different focal depths that appeared stacked together (Figure 4, G and H). In a field of view containing a capsular nevus (discussed in more detail in the next section), LYNA focused on the stacked and crowded nuclei (Figure 4, I and J). In a field of view containing a floater, LYNA again focused on the pleomorphic, hyperchromatic nuclei (Figure 4, K and L).

False-Positive Regions of Interest

Next, we reviewed all of the top patch-level predictions (ROIs) for each slide (including those lower than the 100% slide-level sensitivity threshold above). We discovered that several predictions would have triggered additional work-

Table 3. Errors (False-Positive Regions Reported by the Algorithm) Categorized by Workflow			
Error Category	**Error**	**No. of Normal Slides (% Out of 80) Affected in Camelyon16 Test Dataset**	**No. of Normal Slides (% Out of 52) Affected in DS2 Dataset**
Preanalytic/pathology	Poor tissue processing (eg, fixation)	2 (3%)	8 (15%)
Pathology	Floater/contaminant	1 (1%)	0 (0%)
Scanning	OOF	10 (13%)	8 (15%)
Scanning/algorithm	OOF germinal centers	1 (1%)	1 (2%)
Scanning/algorithm	OOF histiocytes	5 (6%)	8 (15%)
Algorithm	Nevus	1 (1%)	0 (0%)
Algorithm	Multinucleated giant cells or clustered/overlapping histiocytes	5 (6%)	0 (0%)
Total false positives		13 (16%)	8 (15%)
Total false negatives (threshold was selected such that this was zero)		0 false negatives (out of 49 tumor slides)	0 false negatives (out of 54 tumor slides)

Abbreviation: OOF, out of focus.

Preanalytic: sampling collection, fixation; Pathology: cutting, staining, coverslipping; Scanning: focus, color balance, stitching artifacts; Algorithm: amount and diversity of data.

flows such as deeper levels of H&E or IHC in actual clinical practice. Therefore, LYNA flagging these areas for review would benefit the clinical workflow by enabling pathologists to determine the next steps.

The first category of actionable false positives consisted of floaters or contaminants (henceforth termed *floaters* for brevity), which were detected by LYNA in 4 slides: Test_018, Test_044, Test_054, and Test_101 (Figure 5, A through D). The first 3 are suspicious and likely cancer, while the last is consistent with normal colonic tissue. While determining whether to penalize the algorithm for identifying this and to better understand the clinical significance of these results, we conducted the following study. We asked 5 board-certified pathologists to review these 4 lymph node images as per their usual clinical workflow. To verify that any missed floaters were true errors instead of not being reported, we asked the pathologists if they had seen the foci on their initial review. All of the pathologists arrived at the correct slide-level diagnosis (negative) for all 4 slides. Two of the 5 pathologists detected the floater for the first case, none detected the floater for the second case, all detected the floater for the third case, and 3 of the 5 detected the floater for the fourth case. Of the 4 cases, the 5 pathologists detected floaters in 1 of 4, 1 of 4, 2 of 4, 3 of 4, and 3 of 4 cases, respectively, on initial review, indicating that the use of LYNA might help detect a significant number of floaters or contaminants that might otherwise have been missed in routine clinical settings.

In addition, LYNA detected epithelioid-appearing cells in the capsule in 2 images (Test_037 and Test_063; Figure 5, E and F). The challenge organizers (with access to the original slides, IHC, and the original case, eg, melanoma or breast cancer) confirmed that these were capsular nevi. Previous studies have shown that these cells, when present in small

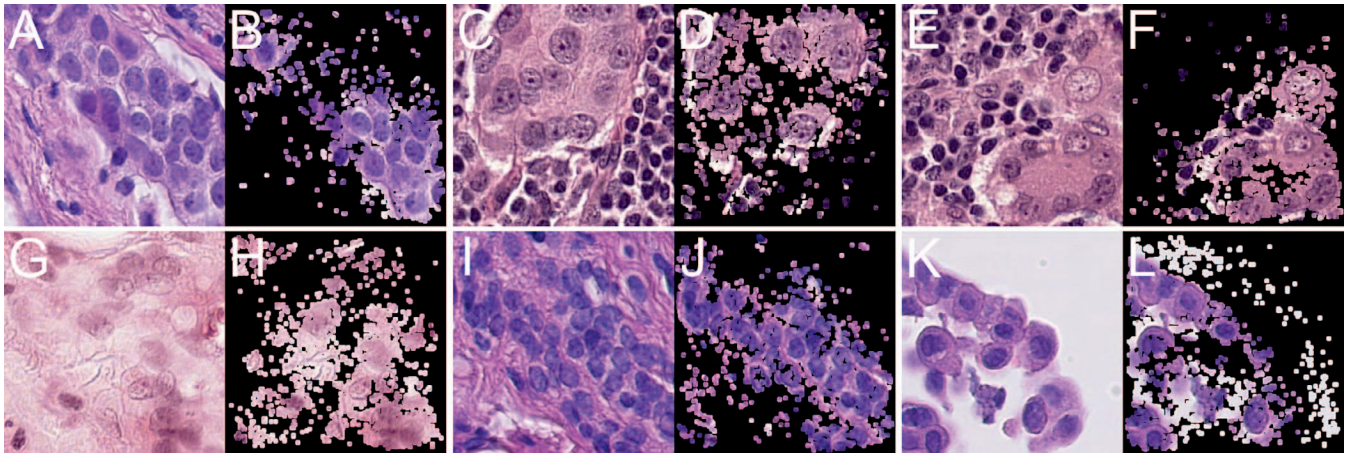

Figure 4. *Visualization of the pixels "focused on" by LYmph Node Assistant (LYNA) in hematoxylin-eosin–stained images (Methods). In each image pair, the original input image at ×40 magnification is shown on the left, and the same image cropped to retain only the most important pixels (according to LYNA) is shown. A and B, Slide Test_004 at high power (true positive). LYNA focuses on the several large tumor cells on the bottom right quadrant with hyperchromatic nuclei and prominent nucleoli. C and D, Slide Test_061 at high power (true positive). LYNA focuses on the cells with large nuclei while ignoring the lymphocytes. E and F, Slide Test_103 at high power (false positive). LYNA focuses on multinucleated, somewhat ductlike giant cell on the bottom right quadrant, and the large cell (likely benign) in the top right quadrant. G and H, Slide Test_123 at high power (false positive). LYNA focuses on large out-of-focus, overlapping histiocytes. I and J, Slide Test_063 at high power (false-positive nevus). LYNA focuses on the lobular-like arrangement of the nevus cells. K and L, Slide Test_044 at high power (false-positive floater, suggestive of adenocarcinoma). LYNA focuses on the malignant cells.*

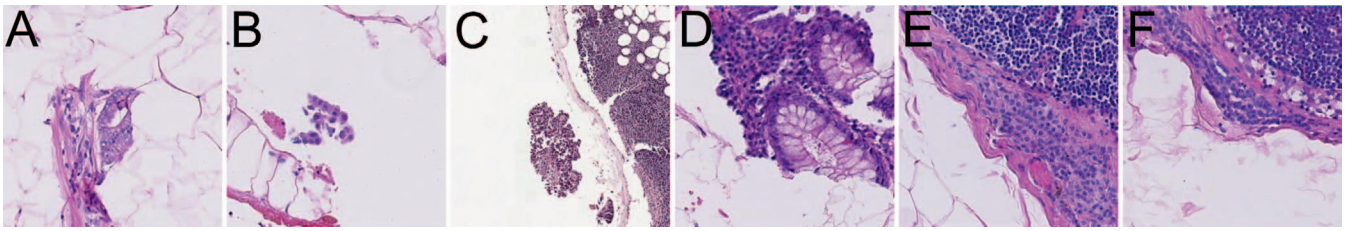

Figure 5. *Examples of actionable false positives in hematoxylin-eosin–stained images: raising the alarm for nonnative tissue (floaters and contaminants), both benign and suggestive of malignancy. Format: slide name, magnification, comment. A, Slide Test_018 at ×20 magnification, floater ≈ 315 μm. B, Slide Test_044 at ×20 magnification, floater ≈ 75 μm, possibly adenocarcinoma. C, Slide Test_054 at ×5 magnification, 3 floaters ≈ 510, 170, and 130 μm. D, Slide Test_101 at ×20 magnification, floater ≈ 410 μm, possibly normal colon. E, Slide Test_037 at ×20 magnification, nevus ≈ 380 μm. F, Slide Test_063 at ×20 magnification, nevus ≈ 275 μm.*

clusters in the capsule, can be diagnostically challenging on H&E alone when using glass slides.[22,23] They can be even more difficult when reviewed in a digital format where color contrast and focus can be slightly different than when reviewing the corresponding glass slides using a microscope.

Thus, while determining whether to penalize the algorithm for finding these capsular nevi and to better understand the significance of these results, we conducted the following study. We asked 5 board-certified pathologists for a review of each slide. After this unbiased review, we asked if their diagnosis or decisions changed after being directed to the specific region. For Test_037, 5 of 5 pathologists detected the ROI on first review, but requested IHC stains (5 of 5 for cytokeratin, and 3 of 5 for a melanocytic marker such as S100, Sox10, HMB-45, or MART1). Their decisions did not change when we directed them to the specific region. For Test_063, 4 of 5 pathologists detected the ROI on first review. Three of 5 requested cytokeratin IHC and 1 requested melanocytic markers in addition. Interestingly, one of the pathologists who had seen the region on the initial review (but considered it benign) requested an IHC stain when we directed them to the region. The remaining pathologist who had not identified the region on initial review considered the region a nevus when asked. We expect that with the assistance of LYNA, nevi or melanoma (if present) will be more consistently detected and be included in the differential for lymph nodes examined for metastatic disease.

DISCUSSION

Our analysis shows that LYNA generated both slide-level and patch-level predictions accurately while ignoring many types of artifacts and benign mimics of cancer (Figure 3, A through N; Table 2). LYNA detected all the macrometastases at a threshold corresponding to zero false positives, and a combined macrometastasis or micrometastasis sensitivity just under that of a human pathologist. Although detecting all of the individual tumor foci does not directly reflect clinical workflow (eg, after the detection of a macrometastasis, detecting additional smaller foci does not affect clinical staging), it is a useful proxy for tumor detection ability under the assumption that any given tumor could have appeared as the only focus in that case or slide. In this scenario, missed detections by either a pathologist or an algorithm will result in a false-negative case-level diagnosis. In addition, the tumor-level sensitivity assesses the ability to both detect metastasis-containing slides and also correctly locate each focus. This correctly penalizes an algorithm that produced a correct slide-level prediction by falsely identifying a benign region as tumor, but missing the true tumor focus.

One of the evaluation metrics, the "zero false positive" tumor sensitivity reveals additional insight into the algorithm's mode of operation and how it and other similar algorithms can be best evaluated and used. First, note in Table 2 that the confidence intervals are wide, but decrease with increasing false positives. Because LYNA (and other similar algorithms) operate via exhaustive search of every slide at high-power magnification (about 100,000 fields of view per slide), there are many opportunities for error for each slide. One false positive is equivalent to a specificity of 99.999%. Therefore, if this performance metric is used for primary evaluation, it can be highly variable across studies. When the threshold is loosened, LYNA achieves a significantly higher tumor sensitivity with narrower confidence intervals. This also suggests that pathologists can leverage LYNA's exhaustive search and resultant high tumor-level sensitivity by first reviewing the top predicted tumor regions for each slide, then ignoring false positives and interpreting only the true positive regions, such as for size measurements and vascular invasion. In this manner, algorithms such as LYNA can raise "alerts" for ROIs (such as tumor) and leave the interpretation of the tissue to pathologists.

Actionable False-Positive ROI

In our study, we find that in addition to metastases, LYNA detected 2 types of actionable false positives: floaters/contaminants and capsular nevi. The detection of floaters and contaminants, while often of marginal clinical interest, may occasionally prompt institution- or pathologist-specific protocols to investigate the origins of these findings.[23–25] Our study of 4 slides indicates that consistent detection is challenging, likely because of the small size of these fragments and often random location on the slide. Sensibly, the detection rate of floaters and contaminants directly correlated with the size of these fragments: in order from smallest to largest floater, the detection rate was 0 of 5, 2 of 5, 3 of 5, and 5 of 5. In particular, the case with the smallest floater (≈75 μm; Figure 5, B) was not detected by any of the 5 pathologists on initial review. While LYNA was not developed to detect floaters, the fact that it does identify several (that were missed by pathologists simulating a routine workflow) presents another benefit of leveraging algorithms during sign-out.

Based on our study, capsular nevi were similarly challenging to diagnose by H&E alone: the 2 cases prompted cytokeratin IHC analysis from 5 of 5 pathologists for both images, and melanocytic IHC analysis from 3 of 5 pathologists in the first

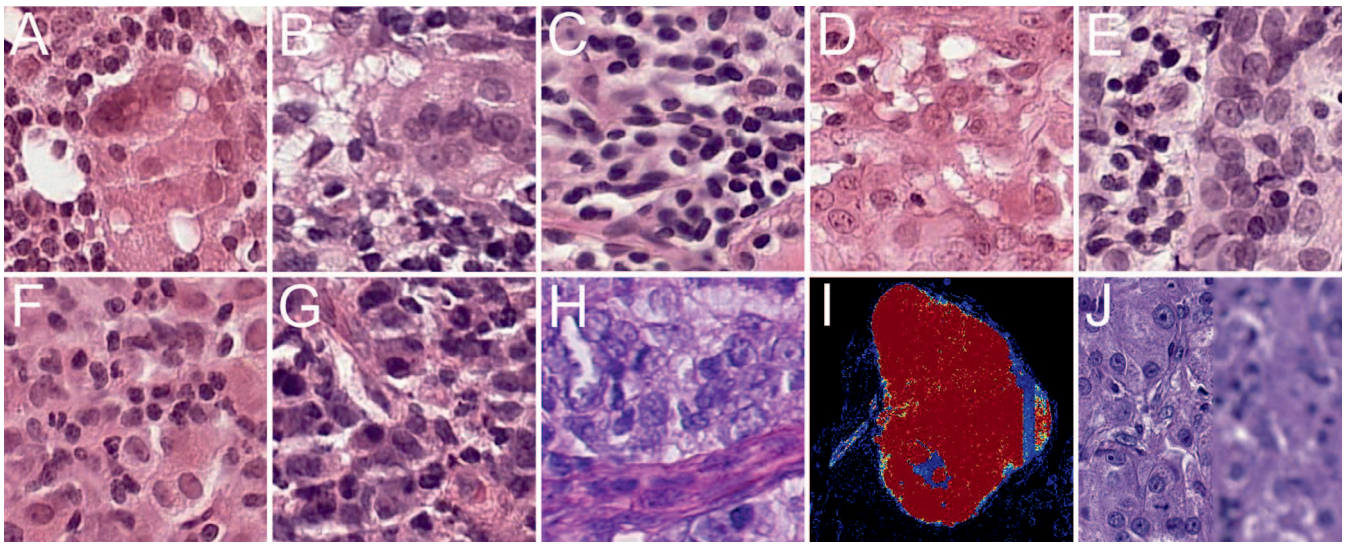

Figure 6. *Other examples of errors in hematoxylin-eosin–stained images, many involving local tissue areas that are out of focus (OOF) to some degree. On the heatmap, blue indicates lower predicted likelihood of tumor, red indicates higher likelihood, and colors in between indicate intermediate likelihood (color bar at the bottom of Figure 1). All panels are displayed at ×40 magnification, except (I), which is a heatmap. A, Slide Test_112 at high power, multinucleated giant cell with artifacts. B, Slide Test_060 at high power, multinucleated giant cell. C, Slide Test_130 at high power, poor staining, OOF. D, Slide Test_007 at high power, poor staining, OOF histiocytes in sinus, retraction artifact. E, Slide Test_006 at high power, OOF histiocytes in sinus. F, Slide Test_096 at high power, OOF histiocytes in sinus. G, Slide Test_053 at high power, OOF germinal center with vessel. H, Slide Test_019 at high power, OOF germinal center. I and J, Slide in DS2. I, LYNA identifies a macrometastasis, except for a negative strip near the right border. J, On closer inspection, the entire strip of tissue is OOF—to a degree at which a review at high magnification is not feasible.*

image and 1 of 5 pathologists in the second image. These data indicate that the capsular nevi prompted suspicion of breast cancer metastasis as well as a second malignancy of melanoma, or potentially a mislabeled slide. While our algorithm was not trained to detect these suspicious-looking cells from H&E alone, the fact that it does identify them could also be of benefit to practicing pathologists.

Other False-Positive ROI: Out of Focus

In our study, other classes of false positives (such as OOF giant cells, histiocytes, and germinal centers; see Figure 6, A through H) were easy for pathologists to rule out. However, in a select subset of instances, some cells were concerning enough to warrant either a glass-slide review or IHC staining to rule out tumor cells. A common theme among these false positives was "local" OOF that affects either individual cells or cellular compartments, and "regional" OOF that affects larger patches of tissue, such as entire scan lanes (Figure 6, I and J) or entire slides. We have also observed entire slides that were OOF ("global" OOF) resulting from dust or dirt on the glass slide or cover slip. These were resolved by cleaning the slide and rescanning. Local OOF might cause otherwise benign tissue to have morphologic characteristics of tumor (such as indistinct cell to cell boundaries and packed nuclei). This level of OOF stems from the fundamental fact that each tissue section's thickness can exceed the optics' depth of field. This issue is exacerbated at higher magnifications, since the depth of field inversely decreases with the magnification—this is why we are constantly refocusing when reviewing glass slides. Such "local" OOF affects the algorithm more so than the pathologist in most, but not all, scenarios (difficult cases are still challenging to evaluate on a digitized slide with no ability to refocus). Correspondingly, solving this OOF issue may involve the use of scanners that enable refocusing. Regional OOF, on the other hand, can obscure high-power review of tissue entirely for both a pathologist and LYNA (Figure 6, J). In contrast to local OOF, regional and global OOF are problems that must be solved by improved scanner capability to correctly detect the focal plane of the tissue.

Unpacking the Black Box and Extension to Nonbreast Cancer

Next, based on the "unpacking the black box" study above, LYNA appeared to be learning sensible morphologic features of malignancy in lymph nodes. For example, LYNA seemed to be sensitive to large and pleomorphic nuclei and ductlike structures. We reasoned that LYNA might also generalize to other cancers present in lymph nodes. Although an in-depth analysis of multiple cancer types is beyond the scope of the current study, we had unintentionally digitized a few slides from nonbreast cancer cases in the course of validating our results. Nodes from 3 of these cases were positive for metastatic cancer, and LYNA correctly detected all 3. The 3 metastatic cases were respectively adenocarcinoma of the colon (Figure 7, A and B), signet ring cell carcinoma of the colon (Figure 7, C and D), and papillary thyroid carcinoma (Figure 7, E and F). Although anecdotal, these results suggest that despite not having been developed with nonbreast specimens, LYNA may generalize to other metastatic cancers in the lymph node, possibly by identifying common tumor morphologic characteristics. Also interestingly, signet ring cell carcinoma of the breast might similarly be detected by LYNA. We hypothesize that further development using other cancer types as training data will enable development of a general cancer detection algorithm for lymph nodes.

Utility of Image Analysis Algorithms in Clinical Practice

Despite some debate, people agree that computer-assisted diagnosis using technologies such as deep learning could be used to augment pathologists' workflows.[24–26] One

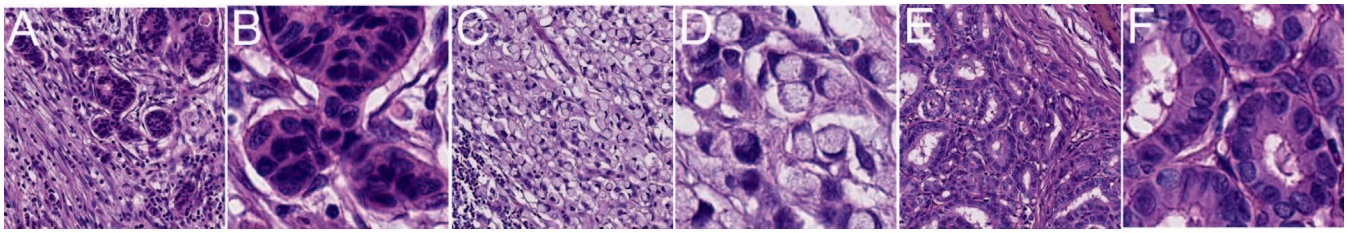

Figure 7. *LYNA detects several nonbreast tumors within hematoxylin-eosin–stained images of lymph nodes. A and B, Metastatic adenocarcinoma of the colon at ×10 magnification (A) and ×40 magnification (B). C and D, Metastatic signet ring cell carcinoma of the colon at ×10 magnification (C) and ×40 magnification (D). E and F, Metastatic papillary thyroid carcinoma at ×10 magnification (E) and ×40 magnification (F).*

slide-level use case is to automatically flag negative or challenging slides for IHC staining before pathologist review as a way to streamline the pathologist's review process.[11] When used in this manner, all of the negative cases could be verified by up-front IHC, at a cost excess of 3 slides in both datasets studied (staining 83 slides to verify the 80 negative slides in Camelyon16 and 55 slides to verify 52 negatives in DS2). Conversely, the algorithm could be used to prioritize the review of positive cases to speed up sign-out for cases with positive nodes. Our data suggest that reviewing the 62 slides predicted to have the highest likelihood of tumor in both datasets would capture all of the positives, corresponding to reviewing 48% (62 of 129) of the slides in Camelyon16 and 58% (62 of 106) of the slides in DS2. Skipping review of the remaining slides (all are negative in our study) is in principle possible, but would require additional validation of these results in other datasets. Another use case could be a "second read" that flags missed metastasis for review, particularly as part of an institutional or individual Quality Assurance protocol. Finally, a patch-level–assisted read mode could guide pathologists to highly suspicious regions, similar to the role of a junior resident indicating ROIs with marking ink on a physical glass slide. Concretely, at the patch level, LYNA predictions could be filtered to show only the highest tumor-likelihood regions based on a (potentially adjustable) threshold depending on the use case. For example, a "looser" threshold that indicates more regions might be desirable for an in-depth review, while a "tighter" threshold that highlights fewer regions might be desirable for a second read of negative cases. These filtered predictions can then be displayed in a small number of colors (such as 1 or 2) to help prioritize review.

On another note, the implementation of digital pathology and computer algorithms such as this could enable more accurate data collection for future American Joint Committee on Cancer guidelines.[21] In breast cancer, for example, measuring metastatic foci using their largest dimension is error prone because unless the focus is a perfect sphere, the true largest dimension can be missed, based on the sectioning protocol and specimen orientation. Automated analysis using computer algorithms would enable more exhaustive sectioning and more accurate measurements of tumor size for both clinical workflow and research purposes.

We have conducted a thorough analysis of the error modes of the LYNA algorithm to allow pathologists to evaluate its strengths and weaknesses, much as we need to understand potential false positives and negatives of an IHC stain before clinical use.[27,28] In addition, we have evaluated LYNA on its generalizability to specimens from a different institution, prepared by using a different protocol, and digitized by using a different scanner. Finally, we have identified the image features that LYNA triggers on in order to "open the black box." We hope that this work provides a template for the evaluation of future histopathology artificial intelligence algorithms for clinical use.

Limitations

Despite promising results, our study contains some limitations. First, LYNA operates on a 75-μm field of view. This means that LYNA lacks context about the anatomic position of the current field of view and will be unable to automatically make position-dependent determinations such as extranodal extension and lymph-vascular invasion. More generally, LYNA is currently unable to compare the current field of view with similar cells in less ambiguous regions of the same slide or case as a pathologist would. Moreover, despite valuable insight, our study would be improved by more cases and slides to detect the rarer error modes. Finally, this work does not directly evaluate the effects on work efficiency or accuracy when using LYNA for diagnosis of lymph node slides. This is the subject of future work.[29]

We thank Greg Corrado, PhD, and Philip Nelson, PhD, for their advice and guidance in enabling this work, Craig Mermel, MD, PhD, for helpful comments on the manuscript, James Wren, MPH, for administrative support, and Josh Pomorski, BS, for data collection. We thank members of the Google AI Pathology team for software infrastructure and logistical support, and slide digitization services. Gratitude also goes to pathologists Kathy Brady, MD, Imok Cha, MD, Steve Cordero, MD, Chris Kim, MD, and one other pathologist for assistance in interpreting images as part of the floater or nevi studies, or the DS2 dataset. Thanks also go to Hossein Talebi, PhD, for helpful discussions about color normalization. Last but not least, we are grateful to the Camelyon16 organizers for creating the challenge, data access, and helpful discussions in clarifying image findings and performance evaluation.

References

1. Apple SK. Sentinel lymph node in breast cancer: review article from a pathologist's point of view. *J Pathol Transl Med*. 2016;50(2):83–95.

2. Mukhopadhyay S, Feldman MD, Abels E, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am J Surg Pathol*. 2018;42(1):39–52.

3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.

4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Paper presented at: 25th International Conference on Neural Information Processing Systems; December 2012; Lake Tahoe, NV.

5. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015;115(3):211–252.

6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.

7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.

8. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22): 2211–2223.

9. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016; 7(1):29.

10. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: whole-slide imaging and beyond. *Annu Rev Pathol*. 2013;8:331–359.

11. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. 2016; 6:26286.

12. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. arXiv [csCV]. March 2017. http://arxiv.org/abs/1703.02442. Accessed March 11, 2018.

13. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. arXiv [q-bioQM]. June 2016. http://arxiv.org/abs/1606.05718. Accessed March 11, 2018.

14. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210.

15. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA*. 2017;318(22):2184–2186.

16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv [csCV]. December 2015. http://arxiv.org/abs/1512.00567. Accessed March 11, 2018.

17. van Der Laak JA, Pahlplatz MM, Hanselaar AG, de Wilde PC. Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy. *Cytometry*. 2000;39(4):275–284.

18. Pitie F, Kokaram A. The linear Monge-Kantorovitch linear colour mapping for example-based colour transfer. Paper presented at: 4th European Conference on Visual Media Production; November 2007; London, United Kingdom.

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

20. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv [csLG]. June 2017. http://arxiv.org/abs/1706.03825. Accessed March 11, 2018.

21. Amin MB, Greene FL, Edge SB, et al. The eighth edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93–99.

22. Davis J, Patil J, Aydin N, Mishra A, Misra S. Capsular nevus versus metastatic malignant melanoma: a diagnostic dilemma. *Int J Surg Case Rep*. 2016;29:20–24.

23. Bautista NC, Cohen S, Anders KH. Benign melanocytic nevus cells in axillary lymph nodes: a prospective incidence and immunohistochemical study with literature review. *Am J Clin Pathol*. 1994;102(1):102–108.

24. Granter SR, Beck AH, Papke DJ Jr. AlphaGo, deep learning, and the future of the human microscopist. *Arch Pathol Lab Med*. 2017;141(5):619–621.

25. Granter SR, Beck AH, Papke DJ Jr. Straw men, deep learning, and the future of the human microscopist: response to "Artificial Intelligence and the Pathologist: Future Frenemies?" *Arch Pathol Lab Med*. 2017;141(5):624.

26. Sharma G, Carter A. Artificial intelligence and the pathologist: future frenemies? *Arch Pathol Lab Med*. 2017;141(5):622–623.

27. Listrom MB, Dalton LW. Comparison of keratin monoclonal antibodies MAK-6, AE1:AE3, and CAM-5.2. *Am J Clin Pathol*. 1987;88(3):297–301.

28. Christensen WN, Boitnott JK, Kuhajda FP. Immunoperoxidase staining as a diagnostic aid for hepatocellular carcinoma. *Mod Pathol*. 1989;2(1):8–12.

29. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018. In press.