nature medicine

TECHNICAL REPORT

https://doi.org/10.1038/s41591-019-0539-7

An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis

Po-Hsuan Cameron Chen[1,4], Krishna Gadepalli[1,4], Robert MacDonald[1,4], Yun Liu[1], Shiro Kadowaki[1], Kunal Nagpal[1], Timo Kohlberger[1], Jeffrey Dean[1], Greg S. Corrado[1], Jason D. Hipp[1,2], Craig H. Mermel[1*] and Martin C. Stumpe[1,3*]

The microscopic assessment of tissue samples is instrumental for the diagnosis and staging of cancer, and thus guides therapy. However, these assessments demonstrate considerable variability and many regions of the world lack access to trained pathologists. Though artificial intelligence (AI) promises to improve the access and quality of healthcare, the costs of image digitization in pathology and difficulties in deploying AI solutions remain as barriers to real-world use. Here we propose a cost-effective solution: the augmented reality microscope (ARM). The ARM overlays AI-based information onto the current view of the sample in real time, enabling seamless integration of AI into routine workflows. We demonstrate the utility of ARM in the detection of metastatic breast cancer and the identification of prostate cancer, with latency compatible with real-time use. We anticipate that the ARM will remove barriers towards the use of AI designed to improve the accuracy and efficiency of cancer diagnosis.

Microscopic examination of samples is the gold standard for the diagnosis of cancer, autoimmune diseases, infectious diseases and more. In cancer, the microscopic examination of stained tissue sections is critical for diagnosing and staging the patient's tumor, which informs treatment decisions and prognosis. In cancer, microscopy analysis faces three major challenges. As a form of image interpretation, these examinations are inherently subjective, exhibiting considerable inter- and intra-observer variability[1,2]. Moreover, clinical guidelines[3] and studies[4] have begun to require quantitative assessments as part of the effort towards better patient risk stratification[3]. For example, breast cancer staging requires the counting of mitotic cells, and quantification of the tumor burden in lymph nodes by measuring the largest tumor focus. However, despite being helpful in treatment planning, quantification is laborious and error-prone. Lastly, access to disease experts can be limited in both developed and developing countries[5], exacerbating the problem.

As a potential solution, recent advances in AI, specifically deep learning[6], have demonstrated automated medical image analysis with performance comparable to that by human experts[1,7–10]. Research has also shown the potential to improve diagnostic accuracy, quantitation and efficiency by applying deep learning algorithms to digitized whole-slide pathology images for cancer classification and detection[8,10,11]. However, the integration of these advances into cancer diagnosis is not straightforward because of two primary challenges: image digitization and the technical skills required to utilize deep learning algorithms. First, most microscopic examinations are performed using analog microscopes, and a digitized workflow requires significant infrastructure investments. Second, because of differences in hardware, firmware and software, the use of AI algorithms developed by others is challenging even for experts. As such, actual utilization of AI in microscopy frequently remains inaccessible.

Here, we propose a cost-effective solution to these barriers to entry of AI in microscopic analysis: an augmented optical light microscope that enables real-time integration of AI. We define 'real-time integration' as adding the capability of AI assistance without slowing down specimen review or modifying the standard workflow. We propose to superimpose the predictions of the AI algorithm on the view of the sample seen by the user through the eyepiece. Because augmenting additional information over the original view is termed augmented reality, we term this microscope the augmented reality microscope. Although we apply this technology to cancer diagnosis in this paper, the ARM is application-agnostic and can be utilized in other microscopy applications.

Aligned with ARM's function to serve as a viable platform for AI assistance in microscopy applications, the ARM system satisfies three major design requirements: spatial registration of the augmented information, system response time and robustness of the deep learning algorithms. First, AI predictions such as tumor or cell locations need to be precisely aligned with the specimen in the observer's field of view (FOV) to retain the correct spatial context. Importantly, this alignment must be insensitive to small changes in the user's eye position relative to the eyepiece (parallax-free) to account for user movements. Second, although the latest deep learning algorithms often require billions of mathematical operations[12], these algorithms have to be applied in real time to avoid unnatural latency in the workflow. This is especially critical in applications such as cancer diagnosis, where the pathologist is constantly and rapidly panning around the slide. Finally, many deep learning algorithms for microscope images were developed using other digitization methods, such as whole-slide scanners in histopathology[8,10,11]. We demonstrate that two deep learning algorithms for cancer detection and diagnosis, respectively, remain accurate when transferred to the ARM. These three core capabilities enable the seamless integration of AI into a traditional microscopy workflow.

[1]Google Health, Mountain View, CA, USA. [2]Present address: AstraZeneca, Gaithersburg, MD, USA. [3]Present address: Tempus Labs Inc., Chicago, IL, USA. [4]These authors contributed equally: Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald. *e-mail: cmermel@google.com

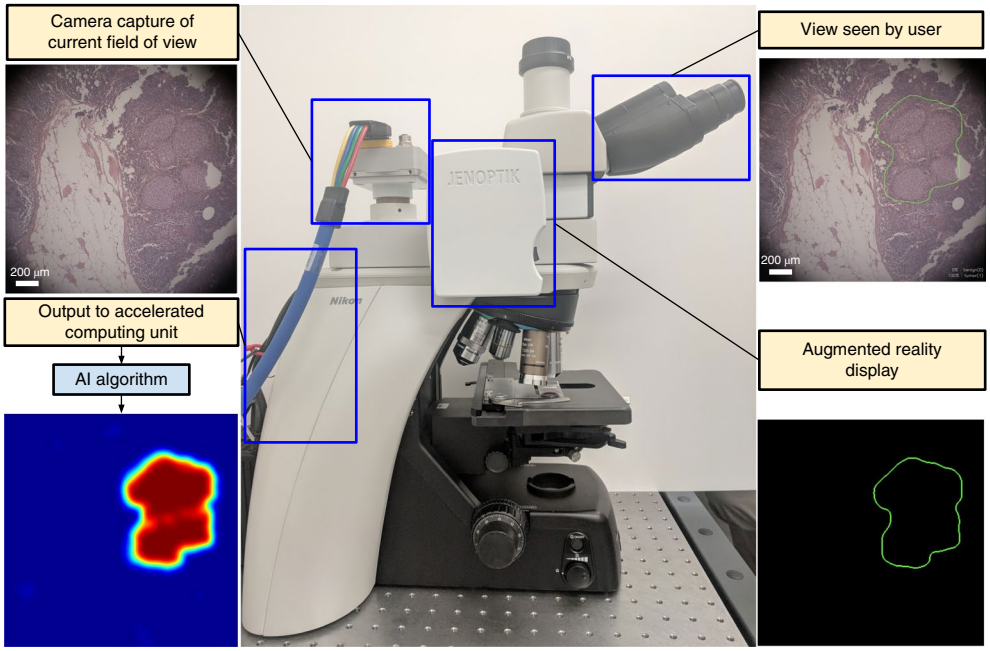


Fig. 1 | Hardware components of the ARM system enable real-time capture of FOV and display of information in the microscope eyepiece. The images of the sample are continuously captured. Next, a deep learning algorithm processes each image to produce an inference output (such as a heatmap) with an accelerated computing unit. Finally, the inference output is post-processed to display the most pertinent information without obscuring the original image. For example, outlines of various colors can be used to aid detection and diagnostic tasks, and text such as size measurements can also be displayed. Technical details can be found in Methods and Extended Data Fig. 4.

We designed and developed the ARM system with three major components: (1) an augmented microscope (Fig. 1); (2) a computer with a software pipeline for acquiring the microscope images, running the deep learning algorithms displaying microscopy results in real time; and (3) a set of trained deep learning algorithms. In this prototype, from an opto-mechanical perspective, the ARM includes a bright-field microscope (Nikon Eclipse Ni-U) augmented with two custom modules (Fig. 1). The first module is a camera that captures high-resolution images of the current FOV. Relay optics were selected and positioned to ensure that the sample was in focus at both the microscope eyepiece and the camera. The second module is a microdisplay that superimposes digital information into the original optical path. Parallax-free performance required alignment of the microdisplay to the virtual sample plane within 1 mm. From a computer hardware and software perspective, the ARM includes a computer with a high-speed image grabber (BitFlow CYT) and an accelerated computing unit (NVidia Titan Xp GPU). The ARM system leverages custom software pipelining to maximize utilization of different hardware components for different tasks and to improve responsiveness (Fig. 2a). Including the computer, the overall cost of the ARM system is one to two orders of magnitude lower than for conventional whole-slide scanners (for example, thousands of dollars versus hundreds of thousands), without incurring the workflow changes and delays associated with digitization. Furthermore, due to the modular design of the system, it can be easily retrofitted to most microscopes (for a second example using an Olympus microscope, see Extended Data Fig. 1).

Lastly, the application of the deep learning algorithm comprises two phases: training and inference (Extended Data Fig. 2). The training phase involves training an algorithm using a large dataset, while the inference phase involves processing an image with the trained deep learning algorithm. Because the microscope FOV (5,120 × 5,120 pixels) is larger than the typical image size used to train deep learning algorithms (<1,000 × 1,000 pixels), exhaustive sliding-window inference is generally required to process the entire FOV. To accelerate inference, we applied the concept of fully convolutional networks (FCN)[13] to the deep learning architecture of InceptionV3 (ref. [14]), which we call InceptionV3-FCN (see Methods). Relative to the original architecture, this modification eliminates 75% of the computation while remaining artifact-free (Fig. 2b and Extended Data Fig. 3), and can be applied to other architectures by following a few design principles in addition to the standard FCN conversion. The combination of pipelining and FCN improved the latency of the ARM system, from 2,126 to 296 ms, and the frame rate from 0.94 frames per second (fps) to 6.84 fps (Fig. 2c). Additional software optimizations (Methods) reduced the latency to 37 ms (27 fps). In our experience, this enables real-time updating of the augmented information to support a rapid workflow. Furthermore, improvements in computing accelerators will naturally lead to further reduction in latency and increase in frame rate over time.

To investigate the potential of ARM as a platform, we developed and tested deep learning algorithms for two clinical tasks: the detection of metastatic breast cancer in lymph nodes[15,16] and the identification of prostate cancer in prostate specimens[17]. These tasks affect breast cancer and prostate cancer staging, respectively, and thus inform therapy decisions. Figure 3a shows several sample FOVs through the ARM for these tasks.

Next, we verified that these algorithms were robust against differences in image quality and color balance. Specifically, the algorithms were developed using images from a different modality, whole-slide scanners, and applied to images captured in the microscope. We sampled FOVs from lymph node and prostate specimens, blinded to the output of the deep learning algorithms. In total, we selected 1,000 FOVs from 50 lymph node slides and 1,360 FOVs from 34 prostate slides, using the ×10 and ×20 objectives (Supplementary Table 1). These 'medium-power' objectives were selected because they are commonly used to search for regions of interest (ROIs) that can be examined in greater detail at higher magnification. For lymph node metastasis detection, the algorithm achieved an area under the receiver operating characteristic (ROC)

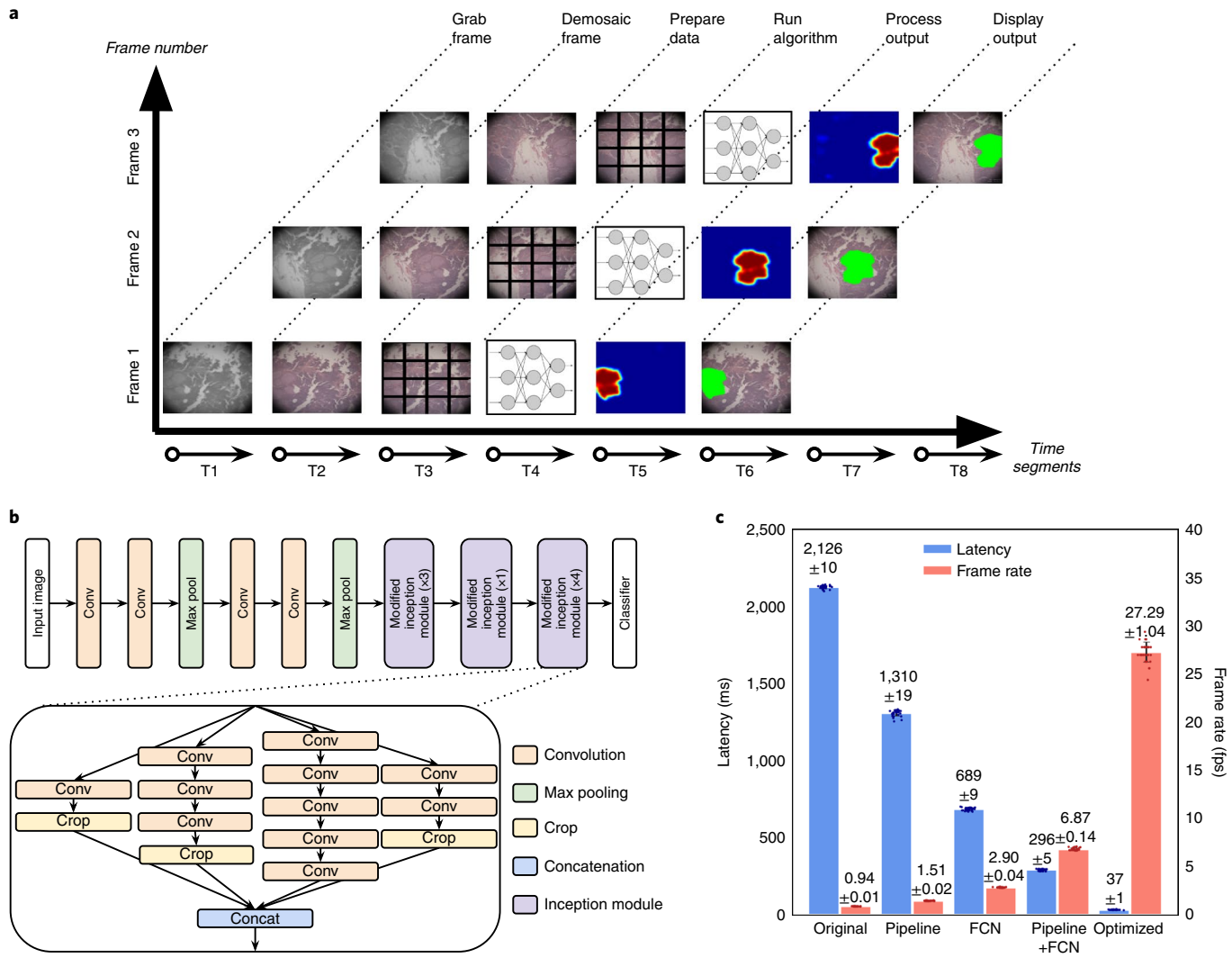


Fig. 2 | Software component of the ARM system improves responsiveness of computing-intensive deep learning algorithms. a, The process of generating predictions for a single FOV requires multiple stages: capturing the FOV, converting that image from the raw sensor data into three-color RGB pixel values, further image processing, running the deep learning algorithm on the images, processing the output and displaying the processed output. Because each stage has different computing requirements, software pipelining ensures that the appropriate hardware is utilized in each stage and reduces end-to-end latency. **b**, Modifying computing-intensive deep learning architectures to FCN reduces computing requirements. The key to doing this without introducing grid-like artifacts is careful cropping (see Methods and Extended Data Fig. 3). We chose to modify InceptionV3 as an example because it is the current state-of-the-art[8,10] in breast cancer metastasis detection and works well in other medical imaging such as dermatology[9] and ophthalmology[7]. **c**, Latency and frame-rate improvements from pipelining and FCN, the combination of both pipelining and FCN, and the combination of both plus further optimization (see section Additional software optimizations in Methods). The latency quantifies the absolute computational performance of the ARM system, while the frame rate more closely reflects the actual user-perceived responsiveness. For reference, eye-tracking studies have shown that for each FOV, pathologists look at several spots, each for 200–400 ms[26], indicating that this level of responsiveness is adequate for real-time usage. The error bars represent the standard deviation of 30 measurements of the full sequence of stages for a single FOV.

curve (AUC) of 0.92 (95% confidence interval, 0.86–0.96) at ×10 and an AUC of 0.97 (95% confidence interval, 0.95–0.99) at ×20. For prostate cancer detection, the algorithm achieved an AUC of 0.93 (95% confidence interval, 0.91–0.95) at ×10 and an AUC of 0.99 (95% confidence interval, 0.97–0.99) at ×20. The ROC curves are shown in Fig. 3b.

We have presented a novel augmented microscope with real-time AI capabilities to bridge the gap between AI algorithms and the traditional microscopy workflow. As a proof of concept, we have developed and evaluated deep learning algorithms for two applications: the detection of metastatic breast cancer in lymph nodes and the identification of prostate cancer. Further studies will be required to evaluate the impact of using the ARM in actual clinical workflows and with other microscope models (for example, different manufacturers). Because the ARM system is designed to align with the regular microscopy workflow, it will not be appropriate for all tasks. For example, whole-slide digitization may be more efficient for exhaustive AI-based interpretation of the whole specimen.

However, the ARM system can be used for a range of other applications, whether based on AI algorithms and/or solely utilizing the augmented reality capabilities. Other clinical applications that can benefit from the ARM include highly subjective tasks like stain quantification[18], disruptive tasks such as estimation of size measurements using a physical ruler[19], tasks that take place in low-resource environments and require (but lack) skilled personnel—such as infectious disease detection (for example, malaria[20] or tuberculosis[21])

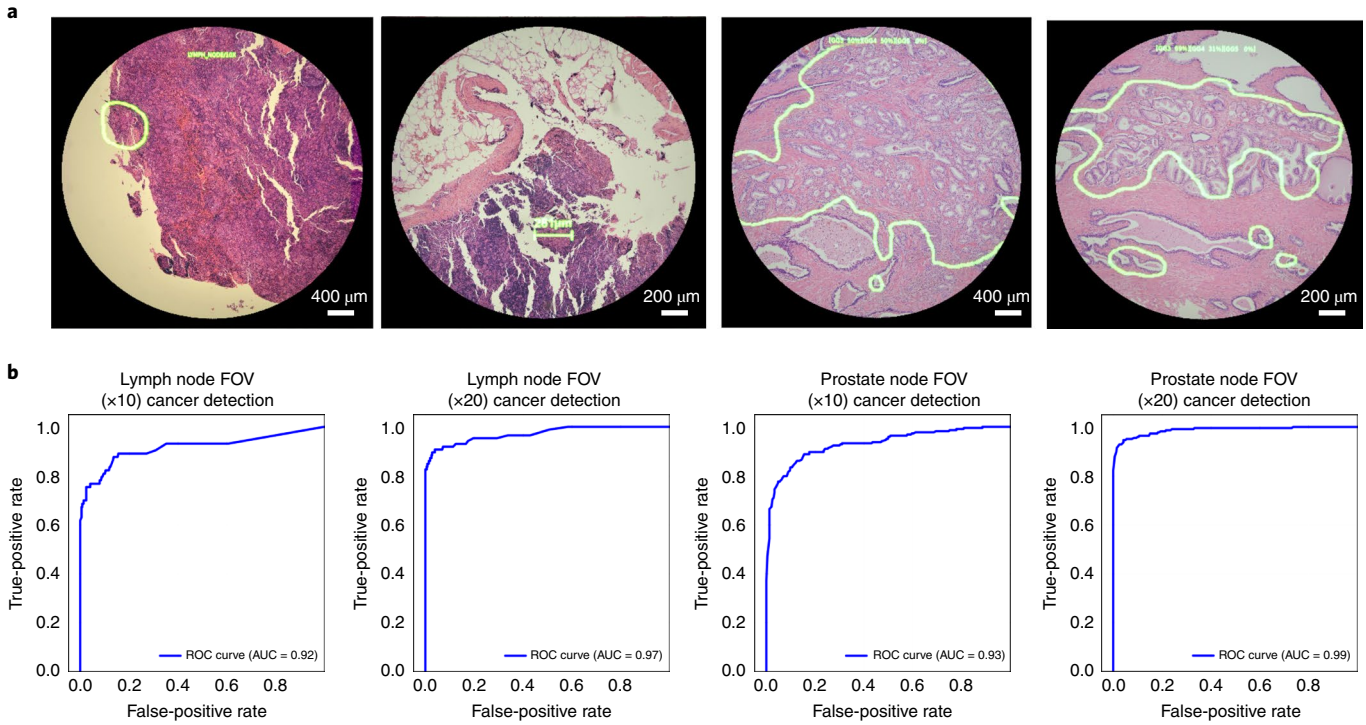


Fig. 3 | Qualitative and quantitative evaluation of lymph node metastasis and prostate cancer detection. a, Deep learning algorithm-derived information that can be displayed using the ARM. Left to right: outline of a small metastasis in lymph node, at medium power to aid detection; numerical measurements of metastasis size in lymph node to aid staging; outline of tumor regions in prostatectomies to aid detection; percentage breakdown of Gleason[27] 3,4,5-pattern involvement in the tumor area as a second opinion to aid Gleason grading[28] of prostate samples. Additional examples can be found in Extended Data Figs. 5,6. **b**, ROC curves evaluating the accuracy of detection of lymph node metastasis and prostate cancer. For each FOV, the output from the algorithm is a heatmap depicting the likelihood of cancer at each pixel location. The FOV likelihood is calculated by taking the maximum likelihood across all pixels. The FOV prediction is positive if the FOV likelihood is greater than or equal to a chosen threshold, and negative otherwise. By varying the threshold between 0 and 1, we generate the ROC curve for the true-positive rate against the false-positive rate. We report the corresponding performance metrics for high accuracy, precision and recall in Supplementary Table 2.

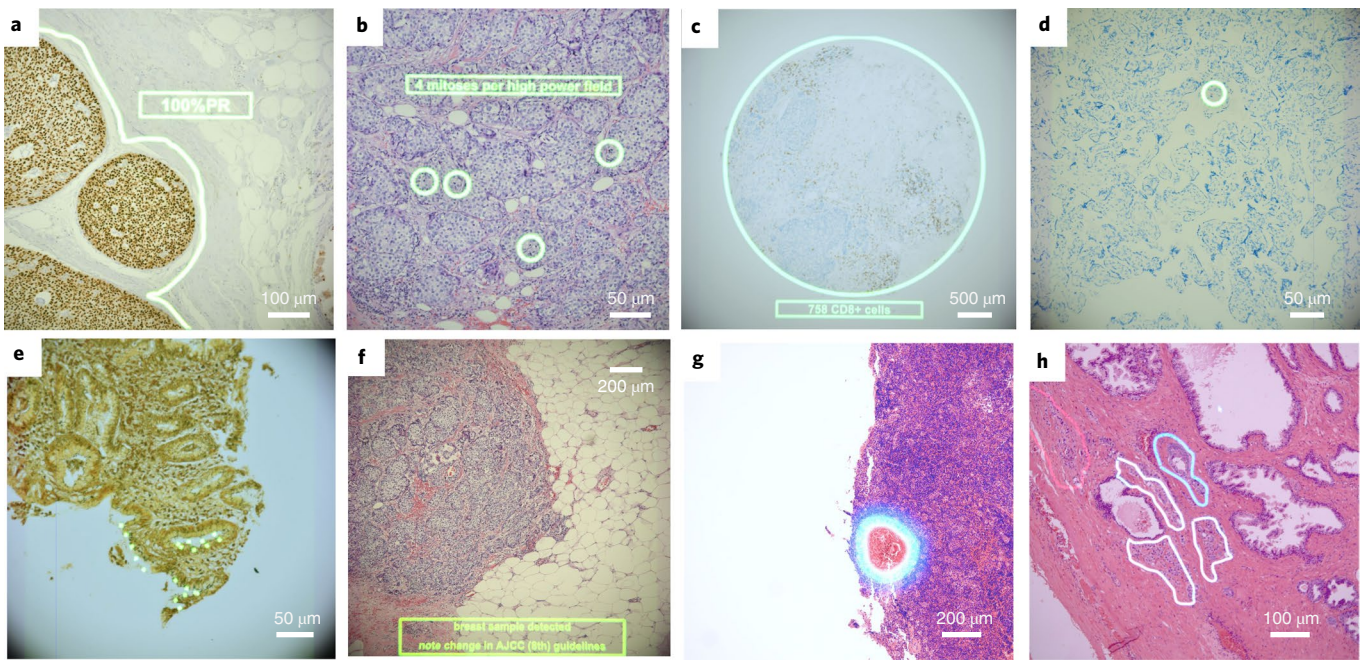


Fig. 4 | Sample future applications that leverage the ARM system's capabilities. a–**c**, AI-based stain quantification of progesterone receptor (PR) (**a**), mitosis counting (**b**) and cell counting (**c**). **d**,**e**, AI-assisted detection of microorganisms (*Mycobacterium tuberculosis* in a sputum smear (**d**) and *Helicobacter pylori* in a tissue section (**e**)). **f**, The ability to display notifications. **g**, The ability to display multi-colored images, such as colored contours to convey different levels of uncertainty. **h**, The ability to outline predictions of multiple categories simultaneously, such as Gleason patterns 3, 4 and 5.

and tedious tasks such as cell or mitosis counting[22] (Fig. 4). Beyond the clinic, the ARM could potentially be useful as a teaching tool by leveraging reverse image search tools[23] that can help trainees quickly search reference resources and answer the question 'what is this histologic feature that I am looking at?' More experienced doctors could also leverage the ARM for clinical research to prospectively validate AI algorithms not previously approved for patient care, such as mutational status[24] or microsatellite instability predictions[25]. To conclude, we anticipate that the ARM will enable the seamless integration of AI into the microscopy workflow and improve the efficiency and consistency of the microscopic examination of biological specimens for the diagnosis of cancer and other diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0539-7.

Received: 19 June 2019; Accepted: 2 July 2019;
Published online: 12 August 2019

References

1. Elmore, J. G. et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015).
2. Brimo, F., Schultz, L. & Epstein, J. I. The value of mandatory second opinion pathology review of prostate needle biopsy interpretation before radical prostatectomy. *J. Urol.* **184**, 126–130 (2010).
3. Amin, M. B. 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.* **67**, 93–99 (2017).
4. Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).
5. Wilson, M. L. et al. Access to pathology and laboratory medicine services: a crucial gap. *Lancet* **391**, P1927–P1938 (2018).
6. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
7. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
8. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
9. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
10. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. *arXiv,* https://arxiv.org/abs/1703.02442 (2017).
11. Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193.e7 (2018).
12. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
13. Shelhamer, E., Long, J. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 640–651 (2017).
14. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
15. Liu, Y. et al. Artificial intelligence-based breast cancer nodal metastasis detection. *Arch. Pathol. Lab. Med.* **143**, 859–868 (2018).
16. Steiner, D. F. et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
17. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit. Med.* **2**, 48 (2019).
18. Vandenberghe, M. E. et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci. Rep.* **7**, 45938 (2017).
19. Russ, J. C. *Computer-Assisted Microscopy: The Measurement and Analysis of Images* (Springer Science & Business Media, 2012).
20. Pirnstill, C. W. & Coté, G. L. Malaria diagnosis using a mobile phone polarized microscope. *Sci. Rep.* **5**, 13368 (2015).
21. Quinn, J. A. et al. Deep convolutional neural networks for microscopy-based point of care diagnostics. In *Proceedings of the International Conference on Machine Learning for HealthCare* **56,** 271–281 (2016).
22. Xie, W., Noble, J. A. & Zisserman, A. Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **6**, 283–292 (2018).
23. Hegde, N. et al. Similar image search for histopathology: SMILY. *NPJ Digit. Med.* **2**, 56 (2019).
24. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
25. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
26. Krupinski, E. A. et al. Eye-movement study and human performance using telepathology virtual slides: implications for medical education and differences with experience. *Hum. Pathol.* **37**, 1543–1556 (2006).
27. Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urol.* **111**, 58–64 (1974).
28. Epstein, J. I., Allsbrook, W. C. Jr, Amin, M. B. & Egevad, L. L., ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma. *Am. J. Surg. Pathol.* **29**, 1228–1242 (2005).

Acknowledgements

We would like to thank the following pathologists who provided initial user feedback: M. Amin, S. Binder, T. Brown, M. Emmert-Buck, I. Flament, N. Olson, A. Sangoi and J. Smith, as well as colleagues who provided assistance with engineering components and paper writing: T. Boyd, A. Chai, L. Dong, W. Ito, J. Kumler, T.-Y. Lin, M. Moran, R. Nagle, D. Stephenson, S. Sudhir, D. Sykora and M. Weakly.

Author contributions

P.-H.C.C. led the deep learning algorithm development and evaluation. K.G. and S.K. led the software integration. R.M. led the optics development. Y.L. prepared data for the lymph node metastasis algorithm. K.N. prepared data for the prostate cancer algorithm. T.K. prepared data for the optical focus assessment algorithm. J.D., G.S.C. and C.H.M. provided strategic guidance. J.D.H. provided clinical guidance. M.C.S. conceived the idea and led the overall development. All authors contributed equally to writing the manuscript.

Competing interests

P.-H.C.C., R.M., K.G., Y.L., S.K., K.N., T.K., J.D., G.S.C., and C.H.M. are employees of Google and own Alphabet stock. J.D.H. is an employee of AstraZeneca. M.C.S. is an employee of Tempus Labs.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0539-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0539-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.H.M.

Peer review information: Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Opto-mechanical design. The schematic of the optic design is shown in Extended Data Fig. 4. Component design and selection were driven by final performance requirements. The camera and display devices were chosen for effective cell- and gland-level feature representation. The camera (Adimec S25A80) included a 5,120×5,120-pixel color sensor with high sensitivity and global shutter capable of capturing images at up to 80 frames s^{-1}. Camera images were captured by an industrial frame-grabber board (Cyton CXP-4) with a Peripheral Component Interconnect Express interface to the computer. The microdisplay (eMagin SXGA096, 1,292×1,036 pixels) was mounted on the side of the microscope and imaged with an achromatic condenser (Nikon MBL71305) at a location tuned to minimize parallax and ensure that the specimen and display image were simultaneously in focus. The microdisplay includes a high-definition multimedia interface for receiving images from the computer. Due to the limited brightness of this display, the second beam splitter (BS2) was chosen to transmit 90% of the light from the display and 10% from the sample, which resulted in good contrast between the projected image and the specimen image when operating the microscope light source at approximately half of its maximum intensity. The opto-mechanical design used here can be readily retrofitted into most standard bright-field microscopes.

Software and hardware system. The application driving the entire system runs on a standard off-the-shelf computer, with a BitFlow frame grabber connected to a camera for live image capture and an NVidia Titan Xp GPU for running deep learning algorithms. The process from frame grabbing to the final display is shown in Fig. 2. To improve responsiveness, the system is implemented as a highly optimized, pipelined, multi-threaded process, resulting in low overall latency. The software is written in C++ and TensorFlow[29].

The primary pipeline consists of a set of threads that continuously grab an image frame from the camera, debayer it (that is, convert the raw sensor output into a red–green–blue (RGB) color image), prepare the data, run the deep learning algorithm, process the results and finally display the output. Other preprocessing steps, such as flat-field correction and white balancing, can be done in this thread as well for cameras that cannot perform them directly on-chip. To reduce the overall latency, these steps run in parallel for a sequence of successive frames—that is, the display of frame 'N', generation of heatmap of frame 'N+1' and running an algorithm on frame 'N+2' all happen in parallel (Fig. 2).

In addition to this primary pipeline, the system also runs a background control thread. One purpose of this thread is to determine whether the camera image is sufficiently in focus to yield accurate deep learning algorithm results. The system uses a convolutional neural network-based out-of-focus detection algorithm to assess focus quality. A second purpose of this thread is to determine the currently used microscope objective, so that the deep learning algorithms tuned for the respective magnification are used. An automated detection of the current microscope objective was implemented, without the need for manual specification of the magnification. Additionally, settings for white balance and exposure time on the camera can be set to optimal profiles for the respective lens.

Additional software optimizations. To further improve the responsiveness, we modified the demosaic process (for example, 2×2×1 array of 'RG–GB') to output a three-channel RGB image of half the height and width (for example, 1×1×3 rather than 2×2×3), to avoid artificial upsampling of the information content. This procedure halved the resolution in each dimension (from 5,120×5,120×one-pixel bayer images to 2,560×2,560×three-pixel RGB images), improving the inference time. We also leveraged a lower numerical precision (floating point 16) to further improve speed.

Out-of-focus detection. We developed an out-of-focus detection algorithm using the proposed InceptionV3-FCN. The network was trained on 216,000 image patches randomly chosen from tissue regions of 27,000 whole-slide images digitized with an Aperio AT2 (pixel size 0.252×0.252 µm). Each patch was labeled by three independent non-pathologist human raters to be either in or out of focus[30].

Data sets for deep learning algorithm development. For breast cancer metastasis detection, we obtained training data from the Cancer Metastases in Lymph Nodes (Camelyon) 2016 challenge dataset[8]. This dataset comprises 215 whole-slide images from slides digitized by one of two whole-slide scanners—a 3DHISTECH Pannoramic 250 Flash II (pixel size 0.243×0.243 μm^2) or a Hamamatsu XR C12000 (pixel size 0.226×0.226 μm^2). The dataset contains pixel-level ground truth diagnoses of both tumor and benign. For prostate cancer identification, we obtained 75 radical prostatectomy slides from The Cancer Genome Atlas (TCGA)[31] and a further 376 radical prostatectomy slides from another source. The slides were digitized with an Aperio AT2 scanner (pixel size 0.252×0.252 μm^2). These whole-slide images were annotated by pathologists by outlining regions as benign, Gleason pattern 3, Gleason pattern 4 or Gleason pattern 5.

Data sets for deep learning algorithm evaluation. To evaluate the deep learning algorithm performance, we obtained FOVs from 50 slides for lymph node and 34 slides for prostate, from two independent sources, as testing data (Supplementary Table 1). Each slide came from a different patient case. In total, we collected 1,000 FOVs for lymph node slides and 1,360 FOVs for prostate slides.

For the former, we selected FOVs to represent several categories of benign tissue—capsule and subcapsular sinus, medullary sinuses, adipose tissue, lymphocytes, follicles, broken edges of tissue and regions with artifacts. The reference standard labels for these slides were established by reviews from two pathologists and adjudication by a third, using pancytokeratin (AE1/AE3) immunohistochemistry staining for reference. We collected FOVs from ten locations, including a maximum of five tumor-containing locations where available. For each location we collected two FOVs, using the ×20 and ×10 objective, respectively.

For the prostate slides, we selected FOVs to represent a diversity of histopathological processes ranging from benign, to inflammation (including prostatitis), to premalignant, to various tumor histological Gleason patterns (3, 4 and 5). The reference standard labels for these slides were established by reviews from three pathologists, using PIN4 immunohistochemistry staining where available. Similar to the lymph node dataset, we collected 20 locations (40 FOVs) per slide, including a maximum of ten tumor-containing locations (20 FOVs) where available.

Color variability in the data sets. The color distributions of stained tissue slides can vary widely because of variability in factors such as tissue processing, staining protocol and image capturing device. To quantify the degree of variability in the test set, we plotted quantitative measures of the color (hue, saturation and brightness) for all training slides from whole-slide scanners and test images from the ARM microscope. The color distribution exhibited marginal overlap in the lymph node training and test data sets, and minimal overlap in the prostate training and test data sets (Extended Data Figs. 7,8).

Deep learning algorithm design. *Constraints of large image size.* The deep learning component of the ARM system is designed to convey the algorithm interpretation of the current FOV, which is 5,120×5,120 pixels in this prototype (2,560×2,560 pixels in the version with additional software optimizations). This size is dependent on the camera's sensor, and high resolution yields crisper images. However, this large image size presents a computational challenge for the latest deep learning algorithms, which contain millions of parameters and require billions of floating point operations even for images of size 300 pixels. Because the number of mathematical operations typically scale (approximately) proportionately with the number of input pixels, increasing the input image width and height by a factor of ten increases the computing requirements by a factor of 100. Thus, developing deep learning algorithms to directly deal with images of such size is currently intractable. Together with our goal of presenting the ARM system as a platform that can utilize other deep learning algorithms, we decided on an alternative, the patch-based approach.

Limitations of standard patch-based approach. The patch-based approach crops the input image into smaller patches for training and applies the algorithm in a sliding window across these patches for inference. In histopathology, the interpretation of a ROI generally involves looking at areas of the image near the ROI for additional context, and this holds true for both human experts and deep learning algorithms. Because of this requirement, application of the deep learning algorithm for a center ROI of size x generally involves feeding as input a patch of size greater than x, and patch-based inference involves re-processing the overlapping 'context' regions for nearby patches (Extended Data Fig. 9). Thus, the standard patch-based approach, although feasible at training (smaller patch size), results in poor computational efficiency at inference (from repeated computations).

Limitations of patch-based approach in fully convolutional mode. A solution to improving computational efficiency during inference is the concept of FCN, which modifies a deep learning algorithm (more generally, an artificial neural network) to utilize only operations that are invariant to the input image size, such as convolutions and pooling. In this manner, a network designed to train with input of size 300 produces valid output even with larger input size, such as 5,000 at inference. However, valid output does not imply consistent output (Extended Data Fig. 9), defined as achieving the same overall output grid of predictions whether the network was used in FCN mode (5,000 at inference) or not (300 at inference, applied with a sliding window to form the grid of outputs). In our example figure, naively applying the FCN results in grid-like artifacts that are not present when not leveraging the FCN. This artifact is caused by the popular 'same' padding option for convolutions, which preserves the input and output sizes to that operation by padding the input with an appropriate number of zeros at the border. These zeros do not cause issues when the training and inference patch sizes are the same, but application of the trained network in an FCN mode replaces the additional zeros with additional 'context' from the image (Extended Data Fig. 9). This mismatch in training and inference causes the grid-like artifacts.

Proposed solution. To solve this padding issue with FCNs applied to networks with 'same' padding convolutions, our proposed modification changes these

paddings to 'valid' (no zero-padding) at training time, allowing an artifact-free FCN modification. As an additional contribution, we show that this is possible even with networks that branch into multiple pathways and merge using a channel-wise concatenation. Where each branch originally used valid padding to maintain the same spatial size for the concatenation, we modified this to crop the branches appropriately to the output size of the smallest branch (Extended Data Fig. 9). This detail is important for the latest networks, many of which contain these branches[14]. The patch-level AUC and predicted heatmaps among InceptionV3, naive FCN and proposed FCN are shown Extended Data Fig. 10.

Other applications of the proposed approach. Finally, we note that although we have presented this modified FCN approach as motivated by the image size of a few thousand pixels across, the same concept applies (at a larger scale) for whole-slide scanner images, which are approximately 100,000 × 100,000 pixels. Usage of our approach can significantly accelerate the application of deep learning algorithms to such large images.

Deep learning algorithm patch-based training. The deep learning image analysis workflow includes two phases—algorithm development and algorithm application, as illustrated in Extended Data Fig. 2. For algorithm development, we trained the neural networks on digitized pathology slides with patches of size 911 × 911 pixels at magnifications ×4, ×10, ×20 and ×40 for lymph node and ×4, ×10 and ×20 for prostate, and the corresponding labels indicating the diagnosis—for example, tumor/benign or Gleason grades. By changing the weights of the neural network to reduce the difference between the predicted results and the corresponding labels, the neural network learned to recognize patterns and distinguish between images with different labels. During neural network training, we also scaled the input (scanner) images by the appropriate factor to match the pixel resolution from the scanner images (~0.25 µm per pixel) to the pixel resolution from the microscope camera (~0.11 µm per pixel). The alternative of scaling the ARM image pixels in real time was rejected, to minimize computation at inference in the ARM system. For algorithm application, the neural network was fed images of size 5,120 × 5,120 pixels captured from the microscope camera. The output from the network is a heatmap depicting the likelihood of cancer at each pixel location. The heatmap can be displayed directly using a colormap, or thresholded to obtain an outline that is then displayed as an overlay on the sample. ARM users favored these outlines over heatmaps, to avoid occluding the view of the underlying sample. The ARM system is capable of displaying either visualization mode, and has the ability to quickly switch the augmented display off to examine the sample without deep learning assistance.

Deep learning algorithm evaluation. We evaluate the algorithm performance for tumor detection within the FOVwith the following metrics: ROC curves (true- versus false-positive rate), AUC, accuracy, precision and recall (TP: true-positive; TN: true-negative; FP: false-positive; FN: false-negative):

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$
$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$
$$\text{Recall or true-positive rate} = \text{TP}/(\text{TP} + \text{FN})$$
$$\text{False-positive rate} = \text{FP}/(\text{FP} + \text{TN})$$

For each FOV, the output from the network is a heatmap depicting the likelihood of cancer at each pixel location. FOV likelihood was calculated by taking the maximum likelihood across all pixels. The FOV prediction was considered positive if FOV likelihood was larger than a chosen threshold, and negative if otherwise. By sweeping the threshold from 0 to 1, we generated the ROC curve of the true- against the false-positive rate.

The FOVs used for evaluation were collected to cover a diverse range of histological types. For lymph node evaluation, the images included lymphoid cells, connective tissue, blood or lymphatic vessels, fat and metastatic breast cancer. For prostate evaluation, the images included benign prostatic glands, blood or lymphatic vessels, fat, inflammation and prostate cancer of varying Gleason patterns. Note that the image quality acquired from the microscope and whole-slide scanner differ significantly with respect to the level of focus, exposure time and so on. In this study, we collected images with correct focus and exposure level for evaluation, as in regular microscopic tissue review workflow.

Application-agnostic platform. The ARM is an application-agnostic platform that operates based on the input pixel size of the pre-trained neural network. For networks with an input pixel size of at least 2,560 pixels (configurable; this is the ARM's image sensor resolution after optimization), the network will be used directly to make predictions. For pre-trained neural networks with an input pixel size smaller than 2,560 pixels, the system will stride the network across the whole FOV and assemble all the network inference results back into a single FOV. A network that is fully convolutional can be run in either mode by specifying the input size as either 2,560 (for a single-pass inference) or a smaller size for strided, 'sliding-window' inference. Therefore, a single ARM add-on module is compatible with with any new convolutional neural network that follows this common format of image-in and prediction-out.

Statistical analysis. The error bars for performance metrics (for example, AUC, accuracy, precision and recall) represent confidence intervals that were computed based on 5,000 bootstrap replications. The error bars for model inference frame rate and latency represent the standard deviation of the measurements of 30 independent inferences. Please see Life Sciences Reporting Summary for more details.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

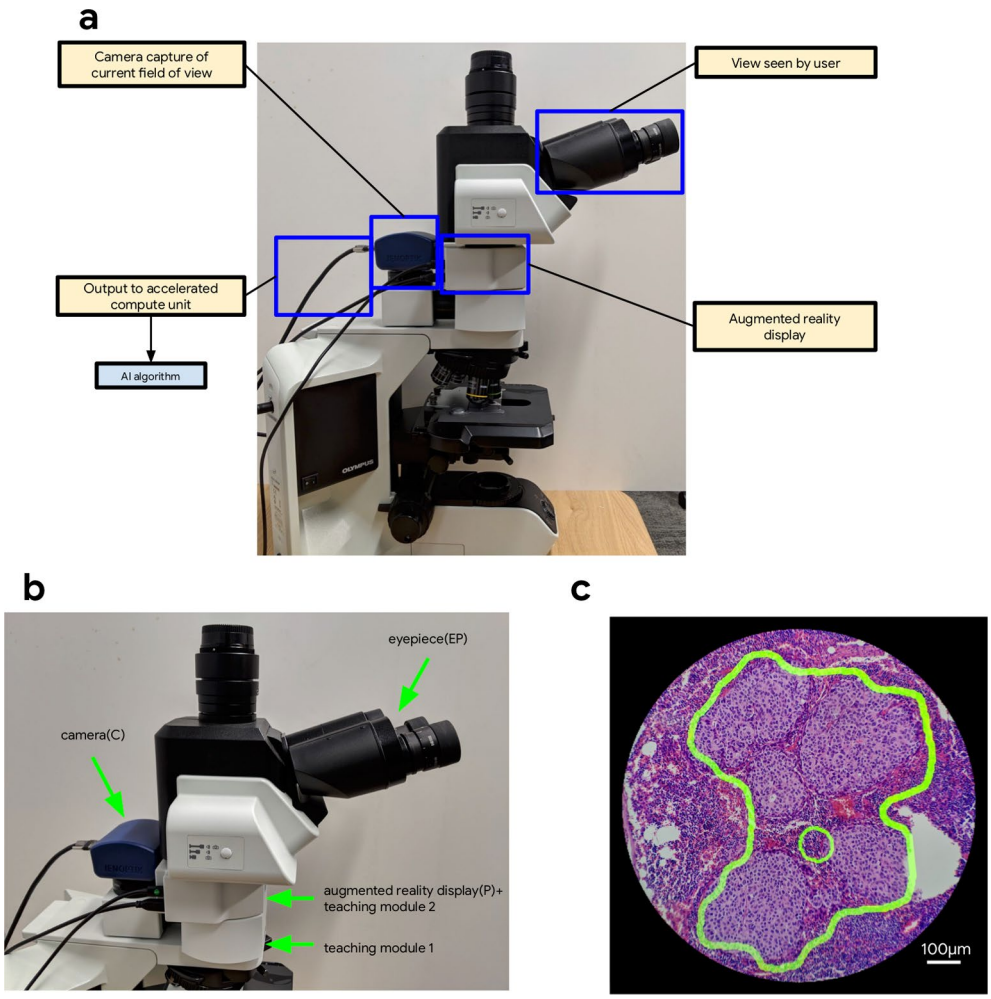
The Camelyon16 dataset utilized to develop the deep learning algorithms used in this study is available from the Camelyon challenge[8] (https://camelyon16.grand-challenge.org/). The prostate dataset from TCGA that was used to develop the deep learning algorithms used in this study is available from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/), which is based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/). The remainder of the prostate and the lymph node data sets is not publicly available due to restrictions in data-sharing agreements with the data sources. The use of de-identified tissue for this study was approved by the Institutional Review Board.

Code availability

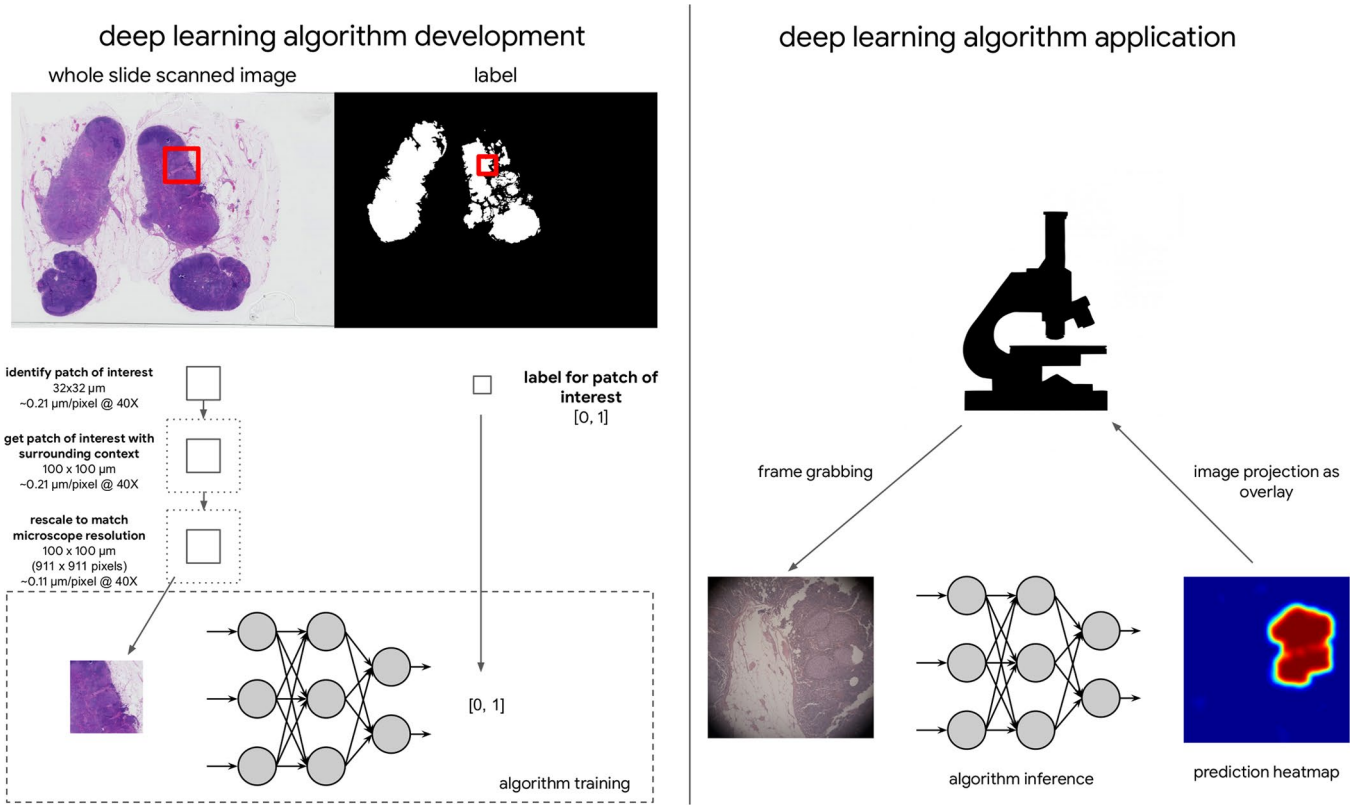
The deep learning architecture will be made available at https://github.com/google-research/google-research/tree/master/nopad_inception_v3_fcn. The deep learning framework used here (TensorFlow) is available at https://www.tensorflow.org/. The camera grabber drivers (BitFlow) are available at http://www.bitflow.com/. The software used for basic image processing (OpenCV) is available at https://opencv.org/. The Python library used for computation and plotting of the performance metrics (SciPy, NumPy and MatPlotLib) is available at https://www.scipy.org/, http://www.numpy.org/ and https://matplotlib.org/, respectively.

References

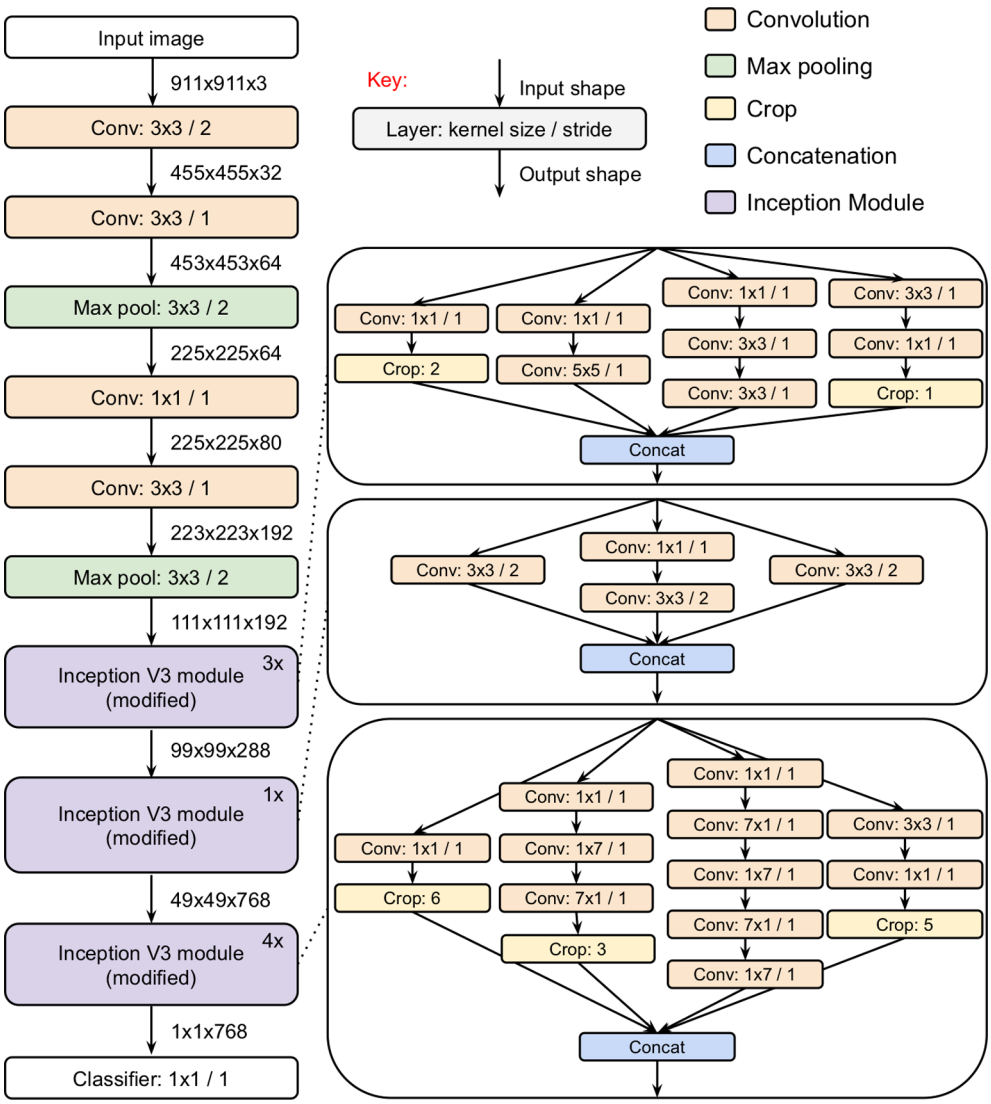
29. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation* (2016).
30. Kohlberger, T. et al. Whole-slide image focus quality: automatic assessment and impact on AI cancerdetection. *arXiv,* https://arxiv.org/abs/1901.04619 (2019).
31. Gutman, D. A. et al. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J. Am. Med. Inform. Assoc.* **20**, 1091–1098 (2013).
32. van Der Laak, J. A., Pahlplatz, M. M., Hanselaar, A. G. & de Wilde, P. C. Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy. *Cytometry* **39**, 275–284 (2000).



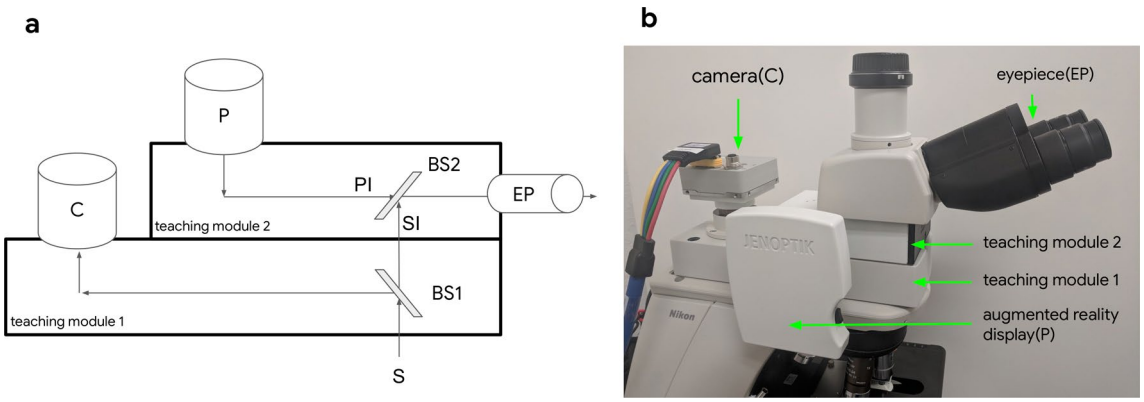
Extended Data Fig. 1 | Integration of the system on a different microscope (Olympus BX43). a, Hardware components of the integrated ARM system on the Olympus BX43. **b**, Photograph of the ARM system implementation labeled with the corresponding modules. **c**, Sample view of lymph node metastasis detection model through the lens.



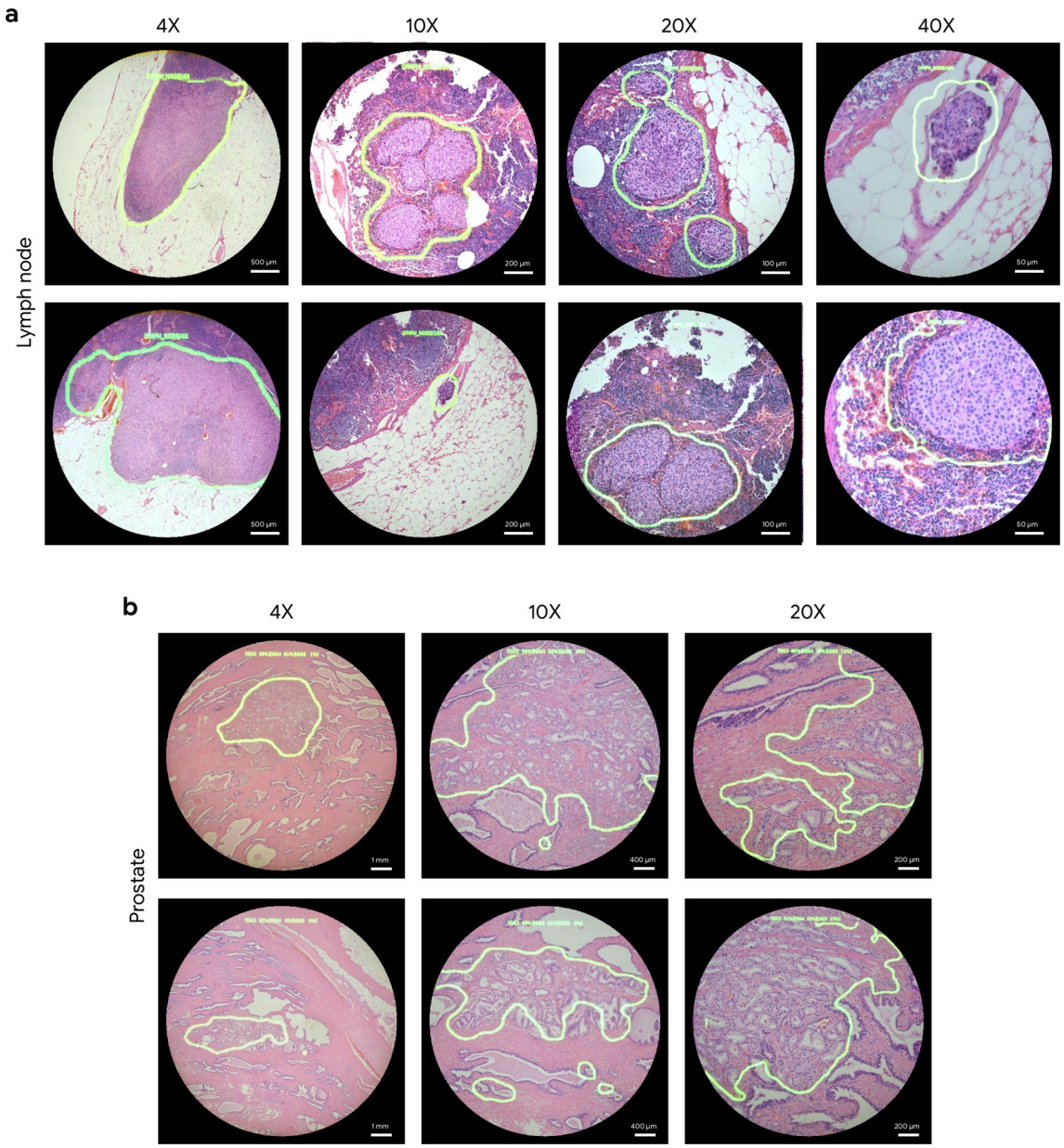
Extended Data Fig. 2 | Deep learning algorithm development and application. In the development phase, we first sample patches of size 911 × 911 pixels from digitized whole-slide imagery. The patches are then preprocessed to match the data distribution of microscope images. In the application phase, an image of size 2,560 × 2,560 pixels is provided to the network. The output of the network is a heatmap depicting the likelihood of cancer at each pixel location.



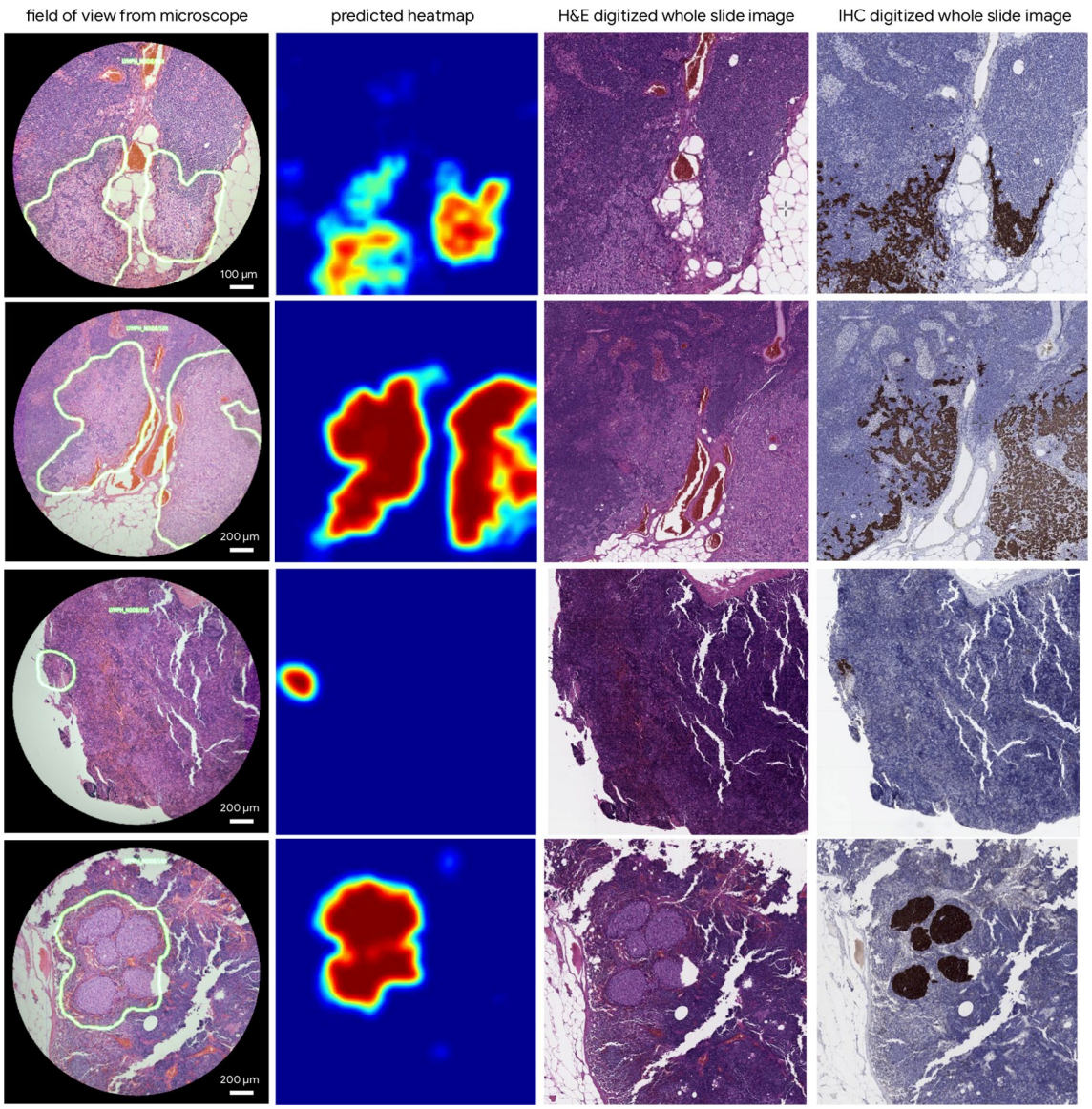
Extended Data Fig. 3 | Modified InceptionV3 network. Modified InceptionV3 network that avoids introduction of artifacts when run in 'fully convolutional mode' at inference. 'Crop' layers with parameter *k* crop a border of width *k* from its input. The principles we followed in the modifications were: (1) use of 'valid' rather than 'same' padding for all convolutions, to avoid introduction of artificial zeroes when the input size is increased at inference time; (2) differential cropping of the output of the branches in each Inception block as appropriate, to maintain the same spatial size (height and width of each feature block) for the channel-wise concatenation operation; and (3) increasing the receptive field to increase tissue context available for interpretation of the neural network.



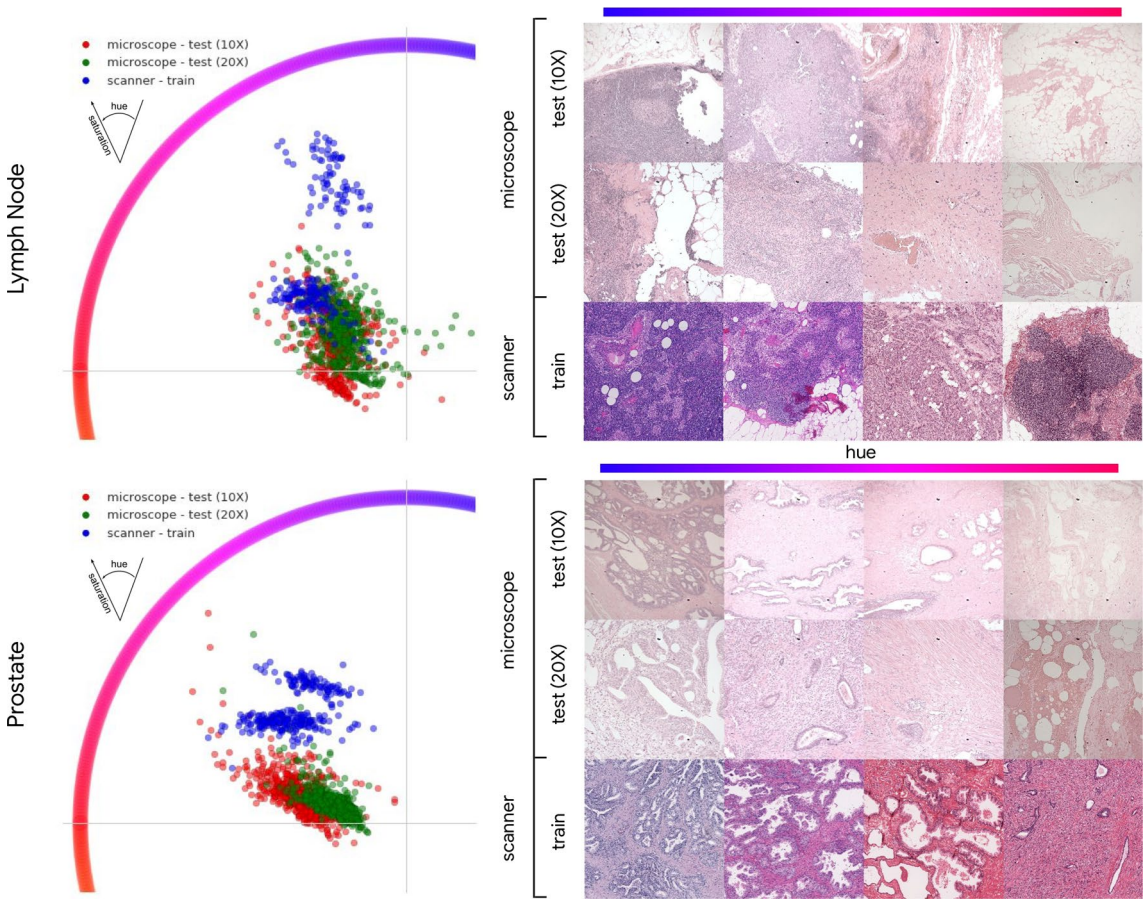
Extended Data Fig. 4 | System overview. a, Schematic of the optic pathway. The standard upright microscope illuminates the specimen (S) from behind and captures the image rays with a conventional objective. These rays propagate upward, in a collimated state, towards the oculars. A teaching module (Nikon Y-IDP) with a beam splitter (BS1) was inserted into the optical pathway in the collimated light space. This module was modified to accept a microscope camera (C), so that the specimen image relayed from BS1 was in focus at the camera sensor when the specimen was also in focus for the microscope user. A second customized teaching module (Nikon T-THM) was inserted between the oculars and the first teaching module. The beam splitter in this module (BS2) was rotated 90° to combine light from the specimen image (SI) with that from the projected image (PI) from the microdisplay (P). The augmented reality display includes a microdisplay and collimating optics, which were chosen to match the display size with the ocular size (22 mm). In this prototype, we tested two microdisplays—one that supports arbitrary colors (RGB) and another, brighter, display that supports only the green channel. The position of the collimator was adjusted to position the microdisplay in the virtual focal plane of the specimen. This collocation of SI and PI in the same plane minimizes relative motion when the observer moves, a phenomenon known as parallax. Note that BS1 needs to precede BS2 in the optical pathway from objective to ocular, so that camera C sees a view of the specimen without the projection PI. The observer looking through the eyepiece (EP) sees PI superimposed onto SI. **b**, Photograph of the actual implementation labeled with the corresponding modules.



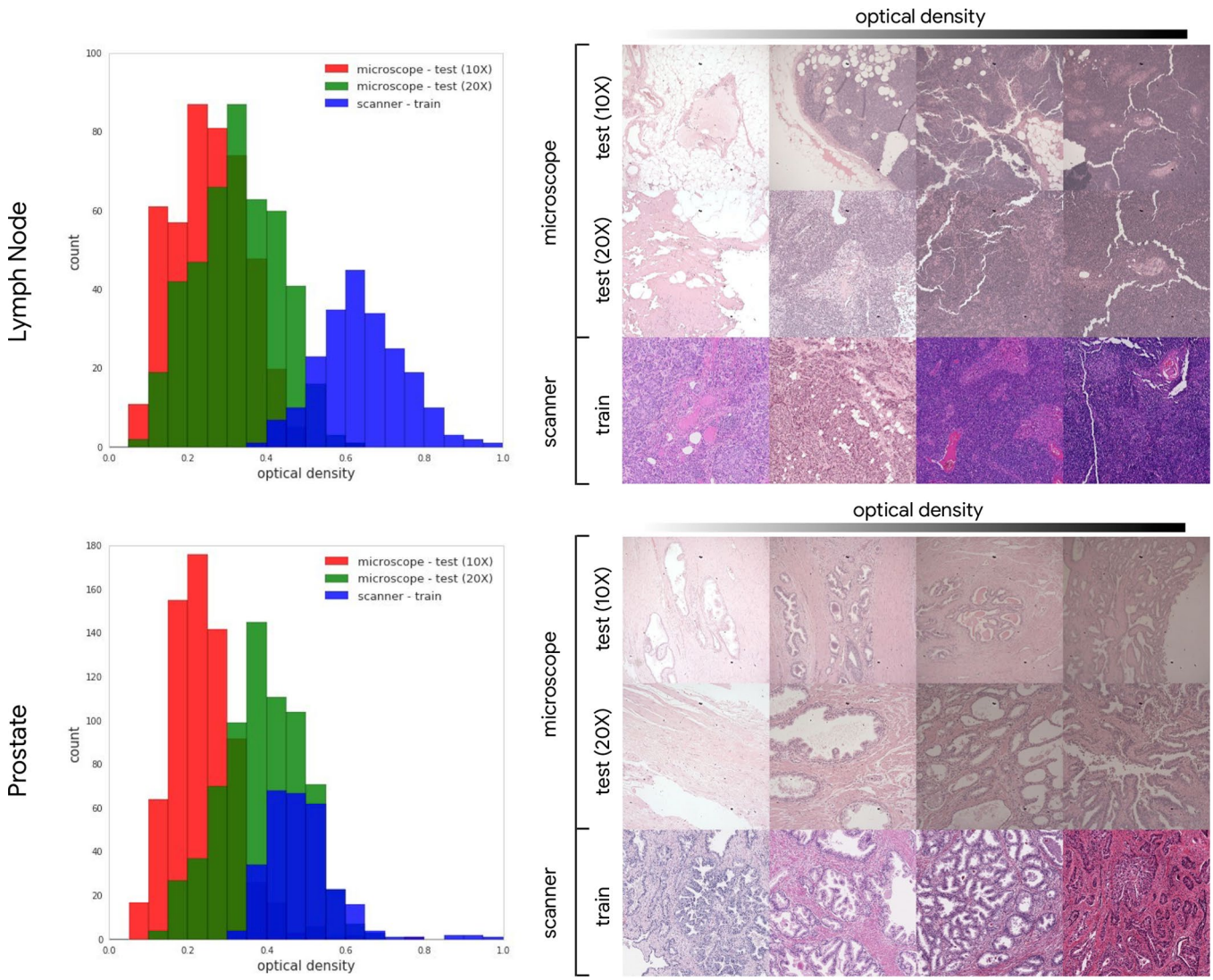
Extended Data Fig. 5 | Sample views through the lens. The images show actual views through the lens of the ARM, with green outlines highlighting the predicted tumor region. **a**, Left to right: lymph node metastasis detection at ×4, ×10, ×20 and ×40. **b**, Left to right: prostate cancer detection at ×4, ×10 and ×20.



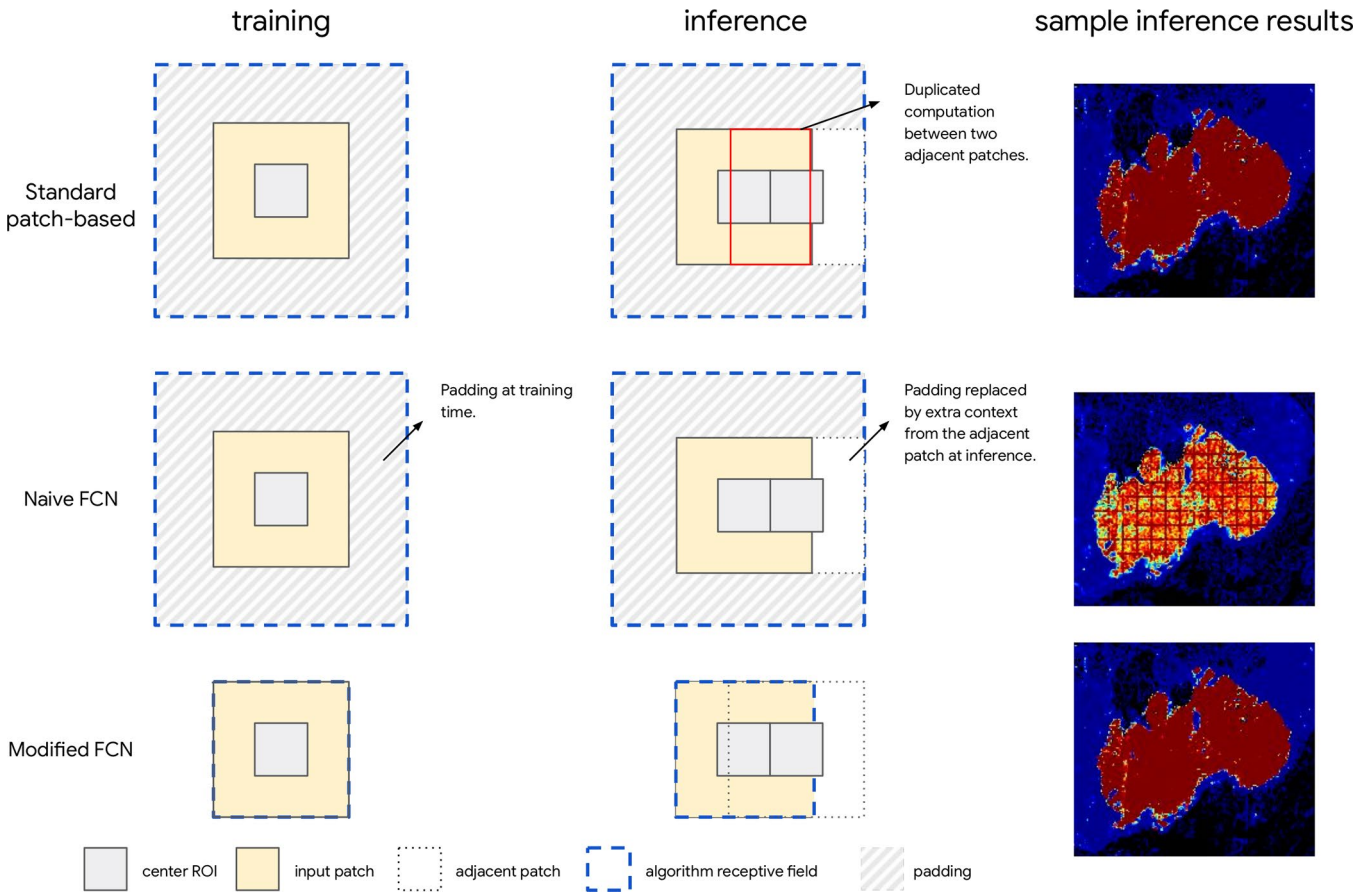
Extended Data Fig. 6 | Lymph node cancer detection at ×10 compared to the corresponding immunohistochemistry as the reference standard. Left to right: FOV as seen from the ARM, predicted heatmap, corresponding FOV from a digital scanner and corresponding FOV of an immunohistochemistry (pancytokeratin AE1/AE3 antibody)-stained slide from a digital scanner. This immunohistochemistry stain highlights the tumor cells in brown.



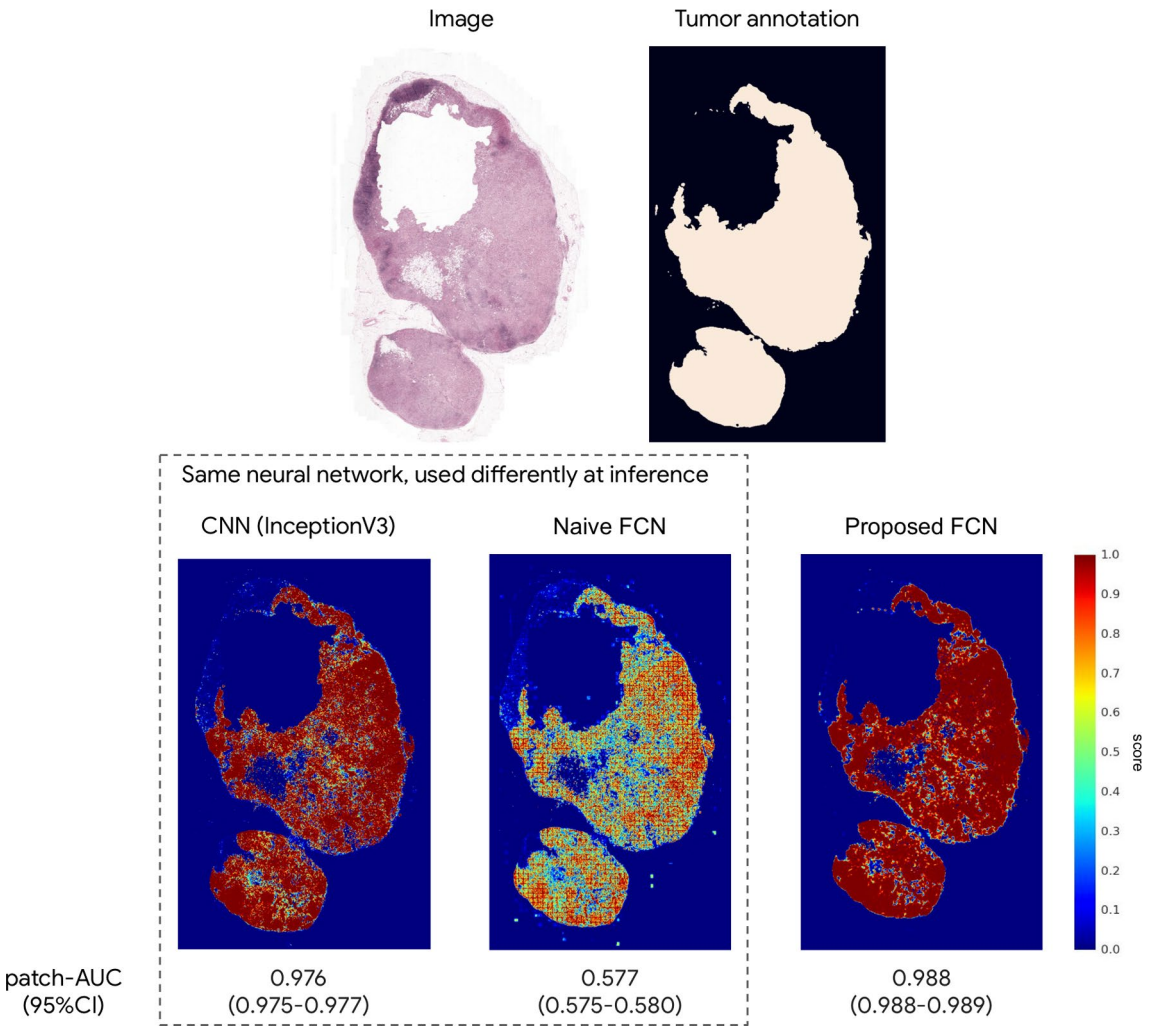
Extended Data Fig. 7 | Visualization of color distribution of slides used in the training and test sets. In the polar scatter plots, the angle represents the hue (color) while the distance from origin represents the saturation. Each point represents the average hue and saturation of an image after mapping RGB values to optical densities followed by a hue–saturation–density (HSD) color transform. The HSD transform is similar to hue–saturation–value, but corrects for the logarithmic relationship between light intensity and stain amount and has been shown to better represent stained slides[10,32]. The training set from a digitized scanner is shown in blue, and test sets from microscopy are shown in red and green.



Extended Data Fig. 8 | Visualization of optical density distribution of slides used in the training and test sets. In the histogram plots, the *x* axis represents the luma (brightness) of the image. Each point represents the average optical density of an image after mapping RGB values to optical densities followed by HSD color transform[32]. The training set from a digitized scanner is shown in blue, and test sets from microscopy are shown in red and green.



Extended Data Fig. 9 | Training and inference comparison across three design choices for the deep learning algorithm. The standard patch-based approach crops the input image into smaller patches for training, and applies the algorithm in a sliding window across these patches for inference. This results in poor computational efficiency at inference from repeated computations across adjacent patches. Naive FCN eliminates the adjacent computations but causes the grid-like artifacts due to the mismatched context between training and inference. Modified FCN removes paddings in the network, ensuring consistent context between training and inference. This proves artifact-free inference results with no repeated computations between adjacent patches.



Extended Data Fig. 10 | Qualitative and quantitative comparison of patch-level AUC between convolutional neural network, naive FCN and proposed FCN. Confidence intervals (CIs) were calculated with 5,000 bootstrap replications.

nature research

Corresponding author(s): Craig Mermel, Martin Stumpe

Last updated by author(s): Jun 26, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The deep learning architecture is available upon request and will be made available at https://github.com/google-research/google-research/nopad_inception_v3_fcn. The deep learning framework used here (TensorFlow) is available at https://www.tensorflow.org/. The camera grabber drivers (BitFlow) are available at http://www.bitflow.com/. The software used for basic image processing (OpenCV) is available at https://opencv.org/.
Data analysis	The Python library used for computation and plotting of the performance metrics (SciPy, NumPy, and MatPlotLib) are available under https://www.scipy.org/, http://www.numpy.org/, https://matplotlib.org/, respectively.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Camelyon16 dataset that were used to develop the deep learning algorithms used in this study is available from the Camelyon challenge6 (https://camelyon16.grand-challenge.org/). The prostate dataset from TCGA that were used to develop the deep learning algorithms used in this study is available from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/), which is based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/). The rest of the prostate dataset and the lymph node dataset are not publicly available due to restrictions in the data sharing agreements with the data sources. The use of de-identified tissue for this study was approved by the Institutional Review Board (IRB).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A power calculation is not performed because this study doesn't aim to compare performances between different approaches but to evaluate the absolute performance of the proposed method.
Data exclusions	No data were excluded.
Replication	The algorithm performances are evaluated on 4 datasets with 2360 fields of view in total.
Randomization	Not applicable to this study.
Blinding	We sampled fields of view from lymph node and prostate specimens, blinded to the output of the deep learning algorithms.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging