LETTERS
https://doi.org/10.1038/s41591-019-0715-9
nature medicine

Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks

Todd C. Hollon[1], Balaji Pandian[2], Arjun R. Adapa[2], Esteban Urias[2], Akshay V. Save[3], Siri Sahib S. Khalsa[1], Daniel G. Eichberg[4], Randy S. D'Amico[5], Zia U. Farooq[6], Spencer Lewis[2], Petros D. Petridis[3], Tamara Marie[7], Ashish H. Shah[4], Hugh J. L. Garton[1], Cormac O. Maher[1], Jason A. Heth[1], Erin L. McKean[1,8], Stephen E. Sullivan[1], Shawn L. Hervey-Jumper[1,15], Parag G. Patil[1], B. Gregory Thompson[1], Oren Sagher[1], Guy M. McKhann II[5], Ricardo J. Komotar[4], Michael E. Ivan[4], Matija Snuderl[9], Marc L. Otten[5], Timothy D. Johnson[10], Michael B. Sisti[5], Jeffrey N. Bruce[5], Karin M. Muraszko[1], Jay Trautman[6], Christian W. Freudiger[6], Peter Canoll[11], Honglak Lee[12], Sandra Camelo-Piragua[13] and Daniel A. Orringer[1,14]*

Intraoperative diagnosis is essential for providing safe and effective care during cancer surgery[1]. The existing workflow for intraoperative diagnosis based on hematoxylin and eosin staining of processed tissue is time, resource and labor intensive[2,3]. Moreover, interpretation of intraoperative histologic images is dependent on a contracting, unevenly distributed, pathology workforce[4]. In the present study, we report a parallel workflow that combines stimulated Raman histology (SRH)[5–7], a label-free optical imaging method and deep convolutional neural networks (CNNs) to predict diagnosis at the bedside in near real-time in an automated fashion. Specifically, our CNNs, trained on over 2.5 million SRH images, predict brain tumor diagnosis in the operating room in under 150 s, an order of magnitude faster than conventional techniques (for example, 20–30 min)[2]. In a multicenter, prospective clinical trial ($n = 278$), we demonstrated that CNN-based diagnosis of SRH images was noninferior to pathologist-based interpretation of conventional histologic images (overall accuracy, 94.6% versus 93.9%). Our CNNs learned a hierarchy of recognizable histologic feature representations to classify the major histopathologic classes of brain tumors. In addition, we implemented a semantic segmentation method to identify tumor-infiltrated diagnostic regions within SRH images. These results demonstrate how intraoperative cancer diagnosis can be streamlined, creating a complementary pathway for tissue diagnosis that is independent of a traditional pathology laboratory.

Approximately 15.2 million people are diagnosed with cancer across the world annually and more than 80% will undergo surgery[1]. In many cases, a portion of the excised tumor is analyzed during surgery to provide preliminary diagnosis, to ensure that the specimen is adequate for rendering final diagnosis and to guide surgical management. In the USA, there are over 1.1 million biopsy specimens annually[8], all of which must be interpreted by a contracting pathology workforce[9]. The conventional workflow for intraoperative histology, dating back over a century[3], necessitates tissue transport to a laboratory, specimen processing, slide preparation by highly trained technicians and interpretation by a pathologist, with each step representing a potential barrier to delivering timely and effective surgical care.

By harnessing advances in optics[5] and artificial intelligence (AI), we developed a streamlined workflow for microscopic imaging and diagnosis that ameliorates each of these barriers. SRH is an optical imaging method that provides rapid, label-free, submicrometer-resolution images of unprocessed biologic tissues[5]. SRH utilizes the intrinsic vibrational properties of lipids, proteins and nucleic acids to generate image contrast, revealing diagnostic microscopic features and histologic findings poorly visualized with hematoxylin and eosin (H&E)-stained images, such as axons and lipid droplets[7], while eliminating the artifacts inherent in frozen or smear tissue preparations[6].

Advances in fiber-laser technology have enabled the development of a Food and Drug Administration-registered system for generating SRH images that can be used in the operating room. We have demonstrated that SRH images reveal microscopic architectural features comparable to conventional H&E images[6]. Given this finding, we recently deployed clinical SRH imagers in our operating rooms, making histologic data readily available during surgery[6,10].

[1]Department of Neurosurgery, University of Michigan, Ann Arbor, MI, USA. [2]School of Medicine, University of Michigan, Ann Arbor, MI, USA. [3]College of Physicians and Surgeons, Columbia University, New York, NY, USA. [4]Department of Neurological Surgery, University of Miami, Miami, FL, USA. [5]Department of Neurological Surgery, Columbia University, New York, NY, USA. [6]Invenio Imaging, Inc., Santa Clara, CA, USA. [7]Department of Pediatrics Oncology, Columbia University, New York, NY, USA. [8]Department of Otolaryngology, University of Michigan, Ann Arbor, MI, USA. [9]Department of Pathology, New York University, New York, NY, USA. [10]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [11]Department of Pathology & Cell Biology, Columbia University, New York, NY, USA. [12]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [13]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [14]Department of Neurosurgery, New York University, New York, NY, USA. [15]Present address: Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. *e-mail: Daniel.Orringer@nyulangone.org

Whether histologic images are obtained via SRH or frozen sectioning, diagnostic interpretation has required the expertise of a trained pathologist. Both globally and within the USA, there is an uneven distribution of expert pathologists available to provide intraoperative diagnosis. For example, many centers performing brain tumor surgery do not employ a neuropathologist, and further shortages are expected given the 42% vacancy rate in neuropathology fellowships[4]. Moreover, although final pathologic diagnosis is increasingly driven by molecular rather than morphologic criteria[11], intraoperative diagnosis relies heavily on interpretation of cytologic and histoarchitectural features. We hypothesized that the application of AI could be used to expand access to expert-level intraoperative diagnosis in the ten most commonly encountered brain tumors and augment the ability of pathologists to interpret histologic images.

We have previously demonstrated that SRH images are particularly well suited for computer-aided diagnosis using hand-engineered feature extractors, with random forest and multilayer perceptron classifiers[6,10,12]. However, the manual feature engineering inherent in these methods is challenging, requires domain-specific knowledge and poses a major bottleneck toward achieving human-level accuracy and clinical implementation[13]. In contrast, deep neural networks utilize trainable feature extractors, which provide a learned and optimized hierarchy of image features for classification. Human-level accuracy for image classification tasks has been achieved through deep learning in the fields of ophthalmology[14], radiology[15], dermatology[16] and pathology[17–19].

Consequently, we designed a three-step intraoperative tissue-to-diagnosis pipeline (Fig. 1) consisting of: (1) image acquisition, (2) image processing and (3) diagnostic prediction via a CNN (see Supplementary Video 1). A fresh, unprocessed surgical specimen is passed off the surgical field and a small sample (for example, 3 mm^3) is compressed into a customized microscope slide. After inserting the slide into the SRH imager, images are acquired at two Raman shifts: 2,845 cm^{-1} and 2,930 cm^{-1}. SRH images are then processed via a dense sliding window algorithm to generate overlapping, single-scale, high-resolution and high-magnification patches used for CNN training and inference. In the prediction stage, individual patches are passed through the trained Inception-ResNet-v2 network, a benchmarked neural network that combines inception modules and residual connections in a deep CNN architecture for image classification[20].

Over 2.5 million labeled patches from 415 patients were used for CNN training (see Extended Data Fig. 1). The CNN was trained to classify tissue into 13 histologic categories, organized into a taxonomy that includes output and inference nodes focusing on commonly encountered brain tumors (see Extended Data Fig. 2). To provide a final patient-level diagnostic prediction, an inference algorithm was developed to map all patches from a specimen to a single probability distribution over the diagnostic classes (see Extended Data Fig. 3).

Noting the commentary on the importance of rigorous clinical evaluations of deep-learning-based algorithms[21], we executed a two-arm, prospective, multicenter, noninferiority clinical trial comparing the diagnostic accuracy of pathologists interpreting conventional histologic images (control arm) with the accuracy of SRH image classification by the CNN (experimental arm) (see Extended Data Fig. 4 and Supplementary Table 1). Fresh brain tumor specimens were collected, split intraoperatively into sister specimens, and randomly assigned to the control or experimental arm. Sister specimens in the control arm were processed via conventional frozen section and smear preparation techniques and interpreted by board-certified pathologists. Sister specimens in the experimental arm were imaged with SRH and diagnosis was predicted by the CNN. The number of patients included was 278 and the primary endpoint was overall multiclass diagnostic accuracy, using final clinical diagnosis as the ground truth. Overall diagnostic accuracy was 93.9% (261/278) for the conventional H&E histology arm and 94.6% (264/278) for the SRH plus CNN arm, exceeding our primary endpoint threshold for noninferiority (>91%) (Fig. 2).

Notably, the CNN was designed to predict diagnosis independent of clinical or radiographic findings, which were reviewed by study pathologists and are often of central importance in diagnosis. Of the 14 errors in the SRH plus CNN arm, 9 were glial tumors, which often have overlapping morphologic characteristics but highly divergent clinical presentations and radiographic appearances. Of the 17 errors in the conventional H&E arm, 10 were malignant gliomas incorrectly classified by pathologists as metastatic tumors, gliosis/treatment effect or pilocytic astrocytoma. In addition, the CNN correctly classified all 17 of the cases in which the pathologist's diagnosis was incorrect (see Extended Data Fig. 5). Moreover, pathologists correctly diagnosed all 14 cases misdiagnosed in the CNN/SRH arm. These results indicate that CNN-based classification of SRH images could aid pathologists in the classification of challenging specimens.

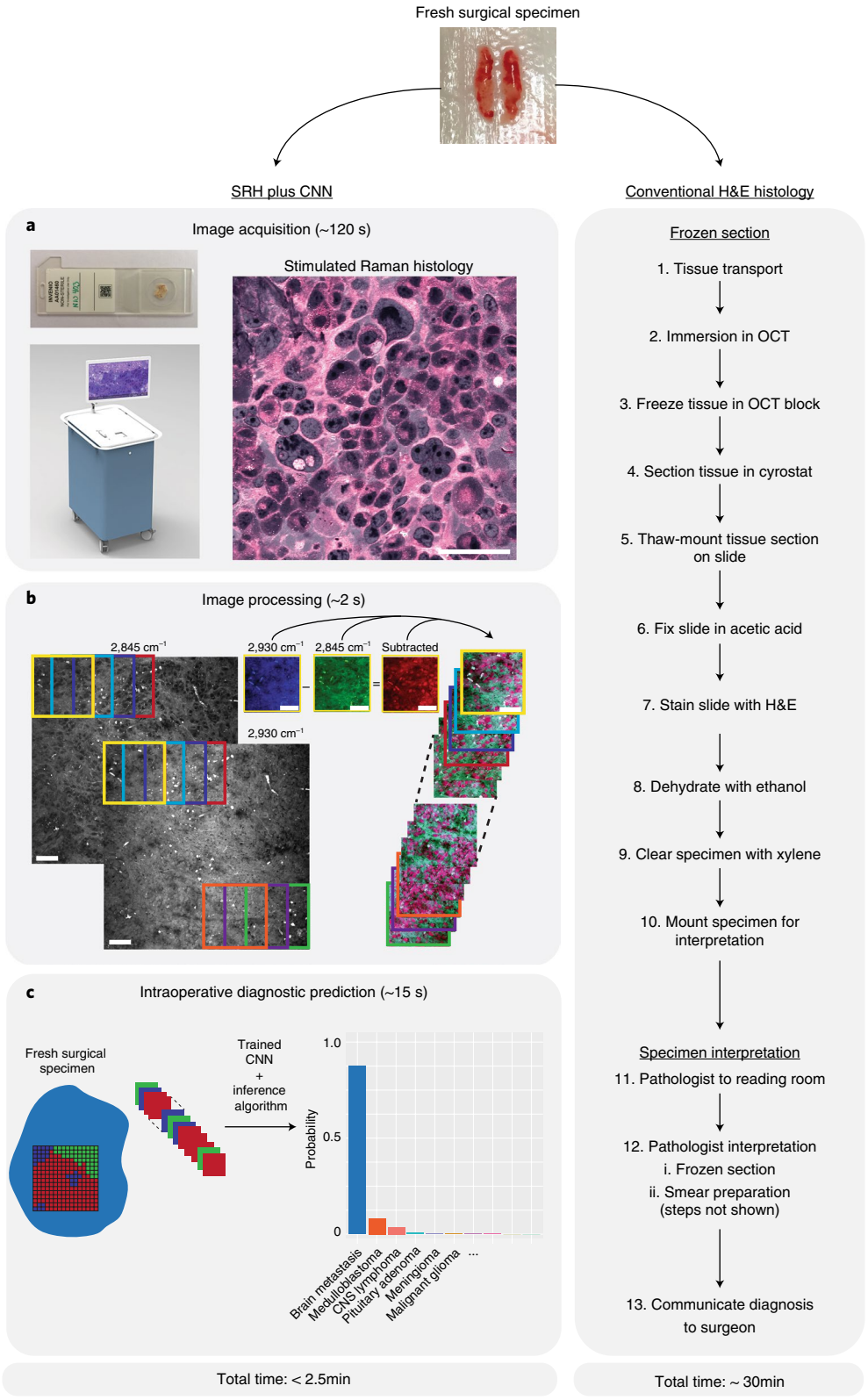
Although the CNN output classes in the present study would cover more than 90% of all brain tumors diagnosed in the USA[22], the diversity and scarcity of rare tumors preclude training of a fully universal CNN for brain tumor diagnosis. Understanding the limitations of our CNN, we developed and implemented a Mahalanobis distance-based confidence scoring system to detect rare tumors[23]. Of the patients enrolled in our trial, 13 were diagnosed with 9 rare tumor types. Our method for rare tumor detection identified all 13 tumors as entities distinct from the output diagnostic classes (see Extended Data Fig. 6).

To gain insight into the learned representations utilized by the CNN for image classification, we used activation maximization, which generates an image that maximally activates a neuron in any neural network layer, using iterations of gradient ascent in the input space (Fig. 3 and see Extended Data Fig. 7)[24]. Deep hidden layers detected nuclear and chromatin morphology, axonal density and

Fig. 1 | Intraoperative diagnostic pipeline using SRH and deep learning. The intraoperative workflows for both conventional H&E staining histology and SRH plus CNNs are shown in parallel. **a**, Freshly excised specimens are loaded directly into an SRH imager for image acquisition. Operation of the SRH imager is performed by a single user, who loads tissue into a carrier and interacts with a simple touch-screen interface to initiate imaging. Images are sequentially acquired at two Raman shifts, 2,845 cm^{-1} and 2,930 cm^{-1}, as strips. After strip stitching, the two image channels are registered and virtual H&E provides SRH mosaics for an intraoperative review by surgeons and pathologists. Time to acquire a 1×1-mm^2 SRH image is approximately 2 min. **b**, Image processing starts by using a dense sliding window algorithm with valid padding over the 2,845 cm^{-1} and 2,930 cm^{-1} images concurrently. Registered 2,845 cm^{-1} and 2,930 cm^{-1} image patches are subtracted pixelwise to generate a third image channel (2,930 cm^{-1} to 2,845 cm^{-1}) that highlights nuclear contrast and cellular density. Each image channel is post-processed to enhance image contrast and concatenated to produce a single three-channel RGB image for CNN input. **c**, To provide an intraoperative prediction of brain tumor diagnosis, each patch undergoes a feedforward pass through the trained CNN and takes approximately 15 s using a single graphics processing unit (GPU) for the 1×1-mm^2 SRH image. Our inference algorithm (see Extended Data Fig. 3) for patient-level diagnosis acts by retaining the high probability tumor regions within the image based on patch-level predictions, and filtering the nondiagnostic and normal areas. Patch-level predictions from tumor regions are then summed and renormalized to generate a patient-level probability distribution over the diagnostic classes. Our pipeline can provide a diagnosis in <2.5 min using a 1×1-mm^2 image, decreasing time to diagnosis by a factor of 10 compared with conventional intraoperative histology[2]. Scale bar, 50 μm.

histoarchitecture, indicating that our network learned recognizable, domain-specific feature representations. We sampled 1,000 SRH patches from normal brain tissue and 2 tumor classes to investigate class-specific, hidden-layer neuron activation. Neurons from a deep hidden layer (convolutional layer 159) with maximal mean activation for each class were recorded and the distribution of mean rectified linear unit activations was plotted.

The images generated through activation maximization reveal recognizable features for each histologic class. For example, green linear structures (neuron 148) represent lipid-rich axons found in gray matter. Neuron 12 was maximally active for malignant glioma and responds to high nuclear density and lipid droplets, features associated with higher-grade gliomas[25,26]. Neuron 101 was maximally activated by patches containing large nuclei with prominent nucleoli and



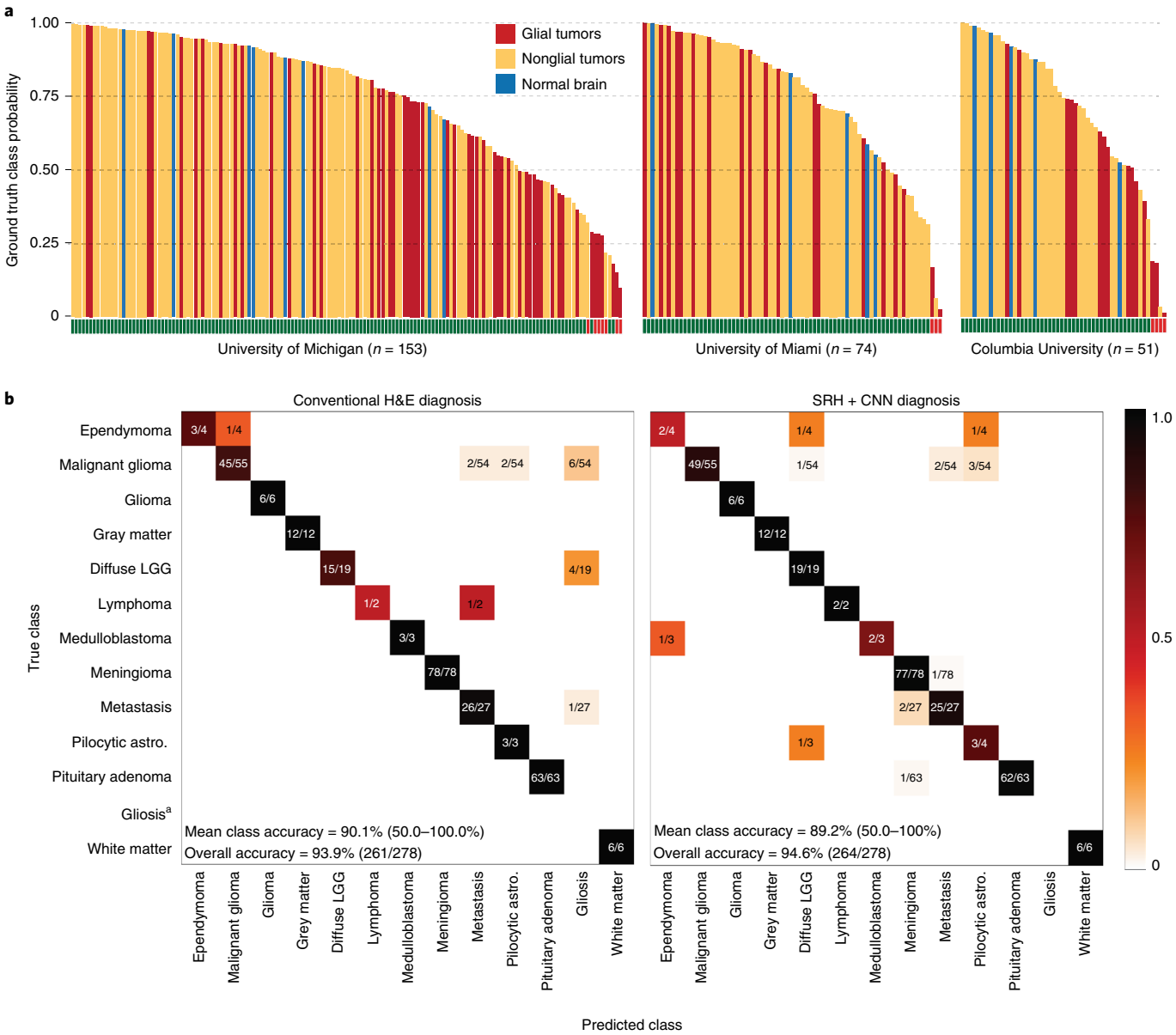


Fig. 2 | Prospective clinical trial of SRH plus CNN versus conventional H&E histology. a, The prediction probabilities for the ground truth classes are plotted in descending order by medical center with indication of correct (green) or incorrect (red) classification. **b**, Multiclass confusion matrices for both the control arm and the experimental arm. Mistakes in the control arm, traditional H&E histology with pathologist, were mostly misclassification of malignant gliomas (10/17). The glial tumors had the highest error rate in the SRH plus CNN arm (9/14). Less common tumors, including ependymoma, medulloblastoma and pilocytic astrocytoma, were also misclassified, probably due to a limited number of cases for model training, resulting in lower mean class accuracy compared with the control arm. These errors are likely to improve with additional SRH training data. Model performance on cases misclassified using conventional H&E histology can be found in Extended Data Fig. 6. The glioma inference class was used for the clinical trial in the setting where the control arm pathologist did not specify the glioma grade at the time of surgery, thereby allowing for one-to-one comparison between study arms. LGG, low-grade glioma. [a]No gliosis/treatment effect cases were enrolled during the clinical trial. This row is included because gliosis was a predicted label and to maintain the convention of square confusion matrices.

cytoplasmic vesicles commonly seen in metastatic tumor cells and pyramidal neurons. These results indicate that the CNN has learned the importance of specific histomorphologic, cytologic and nuclear features for image classification, including some features classically used by pathologists to diagnose cancer. In addition, we used *t*-distributed stochastic neighbor embedding to show that our histologic categories have similar internal CNN feature representations and form clusters based on diagnostic classes (see Extended Data Fig. 8).

We also implemented a semantic segmentation technique to provide pixel-level classification and demonstrate how CNN-based analysis could be used to highlight diagnostic regions within an SRH image (see Extended Data Fig. 9). By utilizing a dense sliding window algorithm, every pixel in an SRH image has an associated probability distribution over the diagnostic classes that is a function of the local overlapping patch-level predictions. Class probabilities can be mapped to a pixel intensity scale. A three-channel RGB overlay indicating tumor tissue, normal/non-neoplastic tissue and nondiagnostic regions allows for image overlay of pixel-level CNN predictions. Our segmentation method achieved a mean intersection-over-union (IOU) value of 61.6 ± 28.6 for the ground

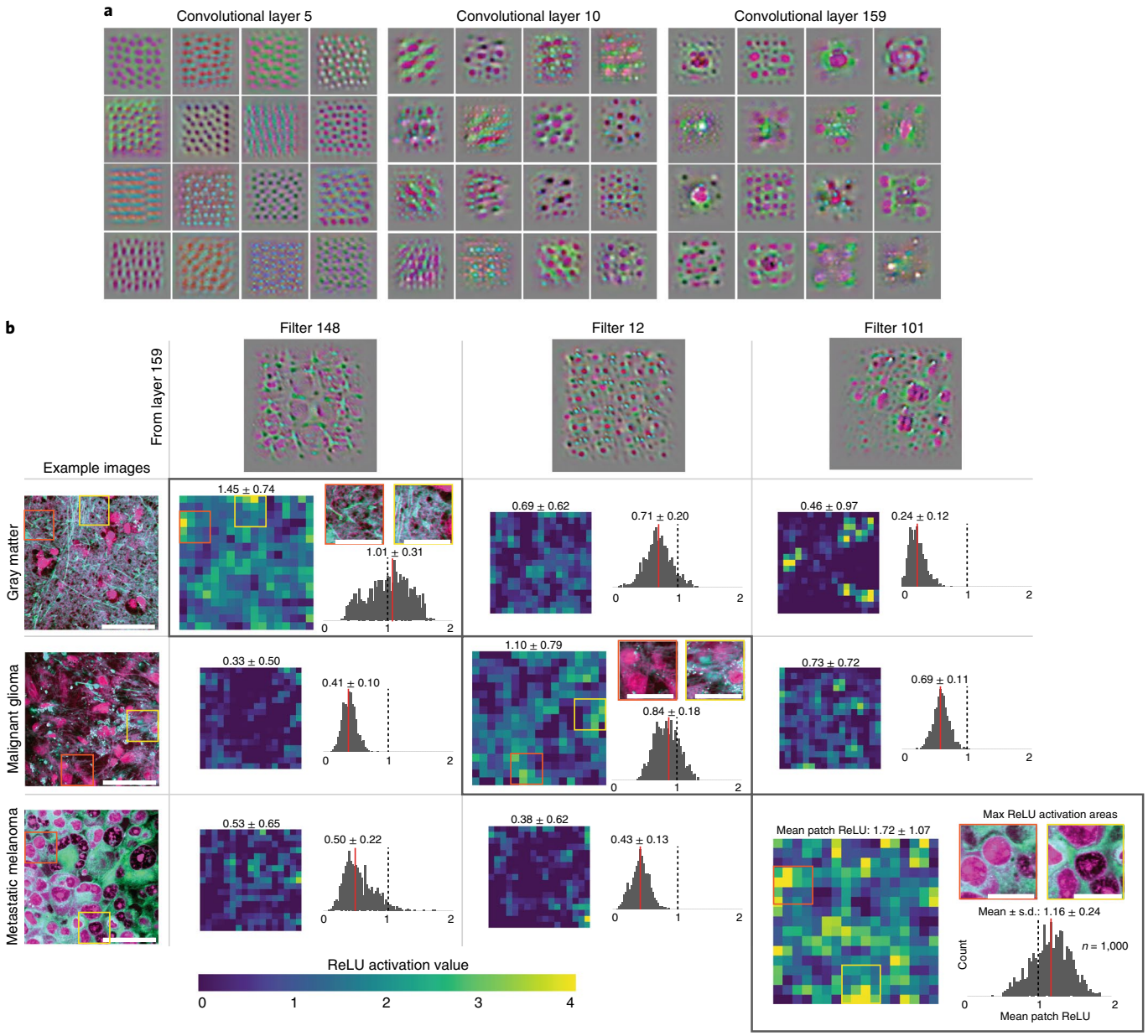


Fig. 3 | Activation maximization reveals a hierarchy of learned SRH feature representations. a, Images that maximize the activation of select filters from layers 5, 10 and 159 are shown. (Activation maximization images for each layer's filter bank can be found in Extended Data Fig. 7.) A hierarchy of increasingly complex and recognizable histologic feature representations can be observed. **b**, The activation maximization images for the filters 148, 12 and 101 in layer 159 are shown as column headings. These filters were selected because they are maximally active for the gray matter, malignant glioma and metastatic brain tumor class, respectively, with example images from each class shown as row labels. A spatial map of the rectified linear unit (ReLU) values for the class example images and corresponding mean ReLU value (±s.d.) is shown in each cell of the grid. Each cell also contains the distribution of mean activation values for 1,000 images randomly sampled from each diagnostic class. High-magnification crops from the example images that maximally activate each neuron are shown. Activation maximization images show interpretable image features for each diagnostic class, such as axons (neuron 148), hypercellularity with lipid droplets and high nuclear:cytoplasmic ratios (neuron 12), and large cells with prominent nucleoli and cytoplasmic vesicles (neuron 101). Example image scale bar, 50 μm; maximum ReLU activation area image scale bar, 20 μm.

truth diagnostic class and 86.0 ± 19.2 for the tumor inference class, for patients in our prospective cohort. Analysis of specimens collected at the tumor–brain interface in primary (Fig. 4) and metastatic brain tumors (see Extended Data Fig. 10) demonstrates how the CNN can differentiate tumor from noninfiltrated brain and nondiagnostic regions.

Our semantic segmentation technique parallels that of Chen and colleagues who reported the development of an augmented reality microscope with real-time AI-based prostate and breast cancer diagnosis using conventional light microscopy[27]. Both methods superimpose diagnostic predictions of an AI algorithm on a microscopic image, calling the clinician's attention to areas containing diagnostic information and providing insight into how AI could ultimately streamline tissue diagnosis.

In conclusion, we have demonstrated how combining SRH with deep learning can be employed to rapidly predict intraoperative brain tumor diagnosis. Our workflow provides a transparent means of delivering expert-level intraoperative diagnosis where

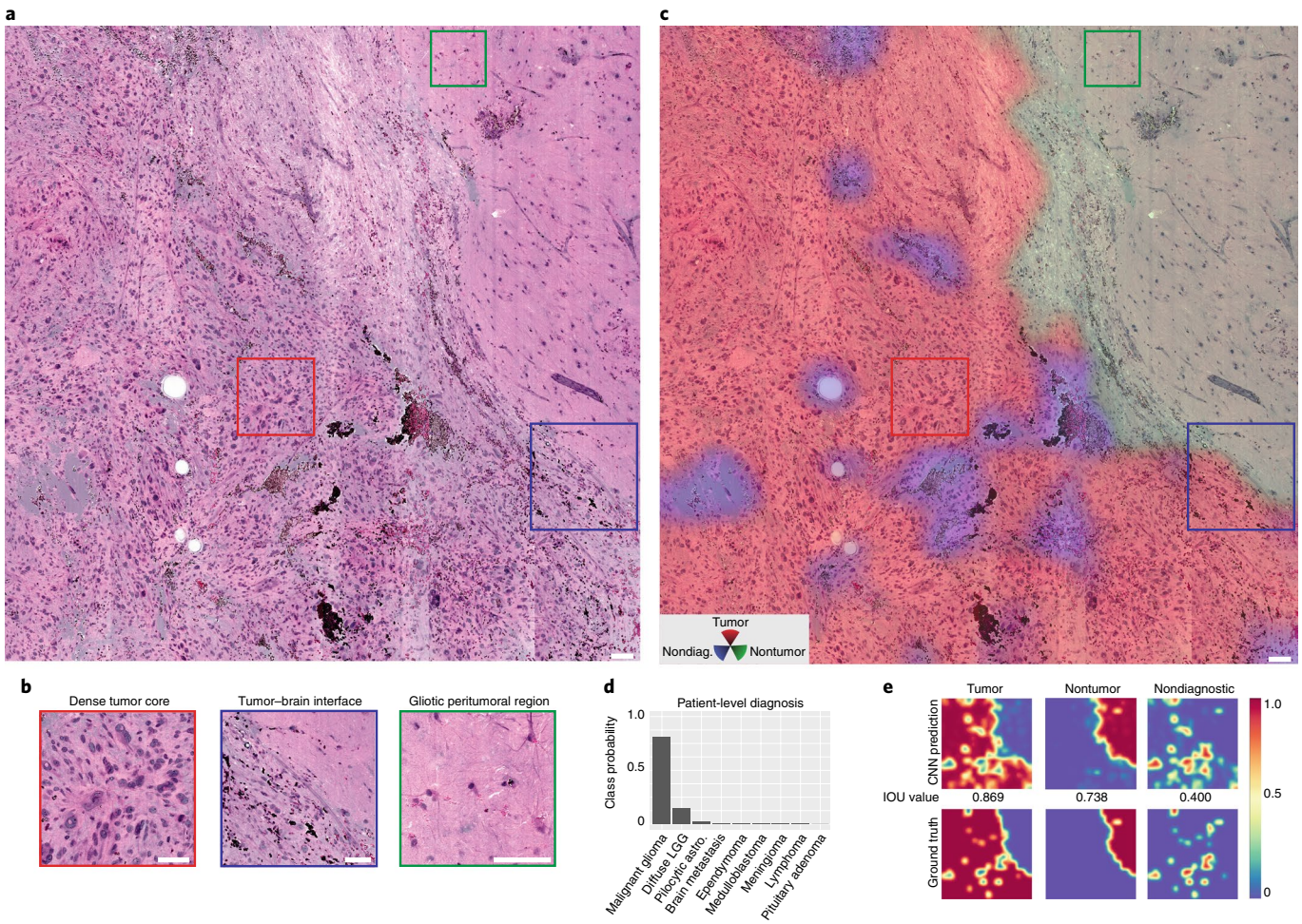


Fig. 4 | Semantic segmentation of SRH images identifies tumor-infiltrated and diagnostic regions. a, Full SRH mosaic of a specimen collected at the brain–tumor interface of a patient diagnosed with glioblastoma, WHO IV. **b**, Dense hypercellular glial tumor with nuclear atypia is seen diffusely on the left and peritumoral gliotic brain with reactive astrocytes on the right of the specimen. SRH imaging of fresh specimens without tissue processing preserves both the cytologic and the histoarchitectural features, allowing for visualization of the brain–tumor margin. **c**,**d**, Three-channel RGB CNN-prediction transparency is overlaid on the SRH image for surgeon and pathologist review intraoperatively (**c**), with associated patient-level diagnostic class probabilities (**d**). astro., astrocytoma; LGG, low-grade glioma. **e**, Inference class probability heatmaps for tumor (IOU = 0.869), nontumor (IOU = 0.738) and nondiagnostic (IOU = 0.400) regions within the SRH image are shown with ground truth segmentation. The brain–tumor interface is well delineated using CNN semantic segmentation and can be used in the operating room to identify diagnostic regions, residual tumor burden and tumor margins. Scale bar, 50 μm.

neuropathology resources are scarce, and improving diagnostic accuracy in resource-rich centers. The workflow also allows surgeons to access histologic data in near real-time, enabling more seamless use of histology to inform surgical decision-making based on microscopic tissue features.

In the future we anticipate that AI algorithms can be developed to predict key molecular alterations in brain tumors such as *O*-methylguanine DNA methyltransferase (MGMT) methylation, isocitrate dehydrogenase (IDH) and α-thalassemia/mental-retardation-syndrome-X-linked gene (*ATRX*) status. In addition, it is possible that SRH will ultimately incorporate spectroscopic detection of the metabolic effects of diagnostic genetic mutations, such as accumulation of 2-hydroxyglutarate in IDH-mutated gliomas. In the interim, however, we note that SRH preserves the integrity of imaged tissue for downstream analytic testing and integrates well within the modern practice of molecular diagnosis.

Although our workflow was developed and validated in the context of neurosurgical oncology, many histologic features used to diagnose brain tumors are found in the tumors of other organs. Consequently, we predict that a similar workflow incorporating optical histology and deep learning could apply to dermatology[28], head and neck surgery[29], breast surgery[30] and gynecology[31], where intraoperative histology is equally central to clinical care. Importantly, our AI-based workflow provides unparalleled access to microscopic tissue diagnosis at the bedside during surgery, facilitating detection of residual tumor, reducing the risk of removing histologically normal tissue adjacent to a lesion, enabling the study of regional histologic and molecular heterogeneity, and minimizing the chance of nondiagnostic biopsy or misdiagnosis due to sampling error[32,33].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0715-9.

Received: 8 August 2019; Accepted: 24 November 2019;
Published online: 6 January 2020

References

1. Sullivan, R. et al. Global cancer surgery: delivering safe, affordable, and timely cancer surgery. *Lancet Oncol.* **16**, 1193–1224 (2015).
2. Novis, D. A. & Zarbo, R. J. Interinstitutional comparison of frozen section turnaround time. A College of American Pathologists Q-Probes study of 32868 frozen sections in 700 hospitals. *Arch. Pathol. Lab. Med.* **121**, 559–567 (1997).
3. Gal, A. A. & Cagle, P. T. The 100-year anniversary of the description of the frozen section procedure. *JAMA* **294**, 3135–3137 (2005).
4. Robboy, S. J. et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch. Pathol. Lab. Med.* **137**, 1723–1732 (2013).
5. Freudiger, C. W. et al. Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. *Science* **322**, 1857–1861 (2008).
6. Orringer, D. A. et al. Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy. *Nat. Biomed. Eng.* **1**, ii (2017).
7. Ji, M. et al. Rapid, label-free detection of brain tumors with stimulated Raman scattering microscopy. *Sci. Transl. Med.* **5**, 201ra119 (2013).
8. *Top 100 Lab Procedures Ranked by Service* (2017); https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/Downloads/LabCHARG17.pdf?agree=yes&next=Accept
9. Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F. & Park, J. Y. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA Netw. Open* **2**, e194337 (2019).
10. Hollon, T. C. et al. Rapid intraoperative diagnosis of pediatric brain tumors using stimulated Raman histology. *Cancer Res.* **78**, 278–289 (2018).
11. Louis, D. N. et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
12. Ji, M. et al. Detection of human brain tumor infiltration with quantitative stimulated Raman scattering microscopy. *Sci. Transl. Med.* **7**, 309ra163 (2015).
13. Krizhevsky, A. et al. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems **25**, 1097–1105 (Curran Associates, Inc., 2012).
14. 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).
15. Titano, J. J. et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).
16. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
17. Litjens, G. et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).
18. 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).
19. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *Proc. 2015 IEEE International Conf. Computer Vision (ICCV)* 1026–1034 (IEEE Computer Society, 2015).
20. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *AAAI* **4**, 12 (2017).
21. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
22. Ostrom, Q. T. et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro-Oncology* **19**, v1–v88 (2017).
23. Lee, K., Lee, K., Lee, H. & Shin, J. A Simple Unified Framework for Detecting Out-of-distribution Samples and Adversarial Attacks. *Proc. 32nd International Conference on Neural Information Processing Systems* 7167–7177 (2018).
24. Erhan, D, Bengio, Y, Courville, A. & Vincent, P. Visualizing Higher-Layer Features of a Deep Network. Technical Report, Univeristé de Montréal (2009).
25. Lu, F.-K. et al. Label-free neurosurgical pathology with stimulated Raman imaging. *Cancer Res.* **76**, 3451–3462 (2016).
26. Kohe, S., Colmenero, I., McConville, C. & Peet, A. Immunohistochemical staining of lipid droplets with adipophilin in paraffin-embedded glioma tissue identifies an association between lipid droplets and tumour grade. *J. Histol. Histopathol.* **4**, 4 (2017).
27. Chen, P.-H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
28. Viola, K. V. et al. Mohs micrographic surgery and surgical excision for nonmelanoma skin cancer treatment in the Medicare population. *Arch. Dermatol.* **148**, 473–477 (2012).
29. Hoesli, R. C., Orringer, D. A., McHugh, J. B. & Spector, M. E. Coherent Raman scattering microscopy for evaluation of head and neck carcinoma. *Otolaryngol. Head Neck Surg.* **157**, 448–453 (2017).
30. Carter, C. L., Allen, C. & Henson, D. E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63**, 181–187 (1989).
31. Ratnavelu, N. D. G. et al. Intraoperative frozen section analysis for the diagnosis of early stage ovarian cancer in suspicious pelvic masses. *Cochrane Database Syst. Rev.* **3**, CD010360 (2016).
32. Sottoriva, A. et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl Acad. Sci. USA* **110**, 4009–4014 (2013).
33. Dammers, R. et al. Towards improving the safety and diagnostic yield of stereotactic biopsy in a single centre. *Acta Neurochir.* **152**, 1915–1921 (2010).

Methods

Study design. The main objectives of the study were to: (1) develop an intraoperative diagnostic computer vision system that combines clinical SRH and a deep-learning-based method to augment the interpretation of fresh surgical specimens in near real-time, and (2) perform a multicenter, prospective clinical trial to test the diagnostic accuracy of our clinical SRH system combined with trained CNNs. 'Near real-time' diagnosis was defined as a nonclinically substantial delay from the time of tissue removal from the resection cavity to tissue diagnosis (that is, 2–3 min). Patient enrollment for intraoperative SRH imaging began 1 June 2015. Inclusion criteria for intraoperative imaging included: (1) male or female; (2) subjects undergoing central nervous system (CNS) tumor resection or epilepsy surgery at Michigan Medicine, New York–Presbyterian/Columbia University Medical Center or the University of Miami Health System; (3) subject or durable power of attorney able to give informed consent; and (4) subjects in whom there was additional specimen beyond what was needed for routine clinical diagnosis. We then trained and validated a benchmarked CNN architecture on an image classification task to provide rapid and automated evaluation of fresh surgical specimens imaged with SRH. CNN performance was then tested using a two-arm, prospective, noninferiority trial conducted at three tertiary medical centers with dedicated brain tumor programs. A semantic segmentation method was developed to allow for surgeon and pathologist review of SRH images with integrated CNN predictions.

Stimulated Raman histology. All images used in the present study were obtained using a clinical stimulated Raman scattering (SRS) microscope[5]. Biomedical tissue is excited with a dual-wavelength fiber laser with a fixed wavelength pump beam at 790 nm and a Stokes beam tunable from 1,015 nm to 1,050 nm. This configuration allows for spectral access to Raman shifts in the range from 2,800 cm^{-1} to 3,130 cm^{-1} (ref. [35]). Images are acquired via beam scanning with a spatial sampling of 450 nm $pixel^{-1}$, 1,000 pixels per strip and an imaging speed for 0.4 Mpixel(s) per Raman shift. The NIO Laser Imaging System (Invenio Imaging, Inc.), a clinical fiber-laser-based SRS microscope, was used to acquire all images in the prospective clinical trial. For SRH, samples were imaged sequentially at the two Raman shifts: 2,850 cm^{-1} and 2,950 cm^{-1}. Lipid-rich brain regions (for example, myelinated white matter) demonstrate high SRS signal at 2,845 cm^{-1} due to CH_2 symmetric stretching in fatty acids. Cellular regions produce high 2,930 cm^{-1} intensity and large $signal_{2,930}$:$signal_{2,845}$ ratios to high protein and DNA content. A virtual H&E look-up table is applied to transform the raw SRS images into SRH images for intraoperative use and pathologic review. A video of intraoperative SRH imaging with automated CNN-based prediction can be found in Supplementary Video 1. The NIO Imaging System (Invenio Imaging, Inc.) is delivered ready to use for image acquisition. SRH images can be reviewed locally using the integrated high-definition monitor, remotely via the health system's picture archiving and communication system or a cloud-based image viewer that allows images to be reviewed anywhere with a high-speed internet connection of less than 30 s.

Image preprocessing and data augmentation. The 2,845 cm^{-1} image was subtracted from the 2,930 cm^{-1} image, and the resultant image was concatenated to generate a three-channel image (2,930 cm^{-1} minus 2,845 cm^{-1}, red; 2,845 cm^{-1}, green; 2,930 cm^{-1}, blue). A 300 × 300-$pixel^2$ sliding window algorithm with 100-pixel step size (both horizontal and vertical directions) and valid padding was used to generate image patches. This single-scale sliding window method over high-resolution, high-magnification images has the following advantages: (1) accommodates the image input size of most CNN architectures without downsampling; (2) allows for efficient, graphical processing, unit-based, model implementation; (3) boosts the number of training and inference images by approximately an order of magnitude; (4) allows for better learning of high-frequency image features; and is (5) faster and (6) easier to implement compared with multi-scale networks. Previous multi-scale CNN implementations have not yielded better performance for image classification tasks involving histologic images[36]. In addition, the use of larger, lower-magnification images complicates image label assignment in the setting of multiple class labels applying to separate regions within a single image (that is, white matter, tumor tissue, nondiagnostic gliotic tissue, and so on), which introduces an additional tunable hyperparameter to identify an optimal class-labeling strategy. This problem is effectively avoided using high-magnification patches, where multiple class labels for a single image are rare. To optimize image contrast, the bottom and top 3% of pixels by intensity from each channel were clipped and images rescaled. All image patches in the training, validation and testing datasets were reviewed and labeled by study authors (T.C.H, S.S.K., S.L, A.R.A. and E.U.). To accommodate class imbalance due to variable incidence rates between the CNS tumors included in the present study, oversampling was used for the underrepresented classes. We used multiple, label-preserving, affine transformations for data augmentation, including any uniformly distributed, random combination of rotation, shift and reflection. All images were mean zero centered by subtracting the channel mean of the training set.

Image datasets. The present study included four image datasets obtained from four SRH imagers: (1) University of Michigan (UM) images from a prototype clinical SRH microscope[6]; (2) UM images from one NIO Imaging System; (3) Columbia University images from a second NIO Imaging System; and (4) University of Miami images from a third NIO Imaging System. Distribution of tumor classes by both number of patches and patients used for CNN training and validation can be found in Extended Data Fig. 1. A total of 296 patients were imaged using the prototype SRH microscope and 339 using the NIO Imaging System. Final tissue diagnosis was provided by each institution's board-certified neuropathologists. Only UM images were used for model training and validation. Images acquired at Columbia University and the University of Miami were used only in the prospective clinical trial to test model performance on SRH images acquired at other medical centers and optimize assessment of CNN generalizability within the present study.

CNN training. A total of 13 diagnostic classes were selected that (1) represent the most common CNS tumors[11,22] and (2) optimally inform intraoperative decisions that bring about surgical goals. Classes included malignant glioma (glioblastoma and diffuse midline glioma, World Health Organization (WHO) grade IV), diffuse lower-grade gliomas (oligodendrogliomas and diffuse astrocytomas, WHO grades II and III), pilocytic astrocytoma, ependymoma, lymphoma, metastatic tumors, medulloblastoma, meningioma, pituitary adenoma, gliosis/reactive astrocytosis/treatment effect, white matter, gray matter and nondiagnostic tissue. We implemented the Google (Google LLC) Inception-ResNet-v2 architecture with 55.8 million trainable parameters randomly initialized. Similar to previous studies, our preliminary experiments using pretrained weights from the ImageNet challenge did not improve model performance, probably due to the large domain difference and limited feature transferability between histologic images and natural scenes (see Extended Data Fig. 6)[36,37]. The network was trained on approximately 2.5 million unique patches from 415 patients using a categorical cross-entropy loss function weighted using inverse class frequency. A randomly selected 16-patient validation set, imaged using the NIO Imaging System at UM, was used for hyperparameter tuning and model selection based on patch-level classification accuracy. We used the Adam optimizer with an initial learn rate of 0.001, β_1 of 0.9 and β_2 of 0.999 (exponential decay rates), ε of 10^{-8} (constant for numerical stability) and a 32-image batch size. An early stopping callback was used with a minimum validation accuracy increase of 0.05 and 5 epoch patients (see Extended Data Fig. 1). Training, validation and testing were done using the high-level Python-based neural network API, Keras (v.2.2.0), with a TensorFlow (v.1.8.0)[38] backend running on two NVIDIA GeForce 1080 Ti graphical processing units.

Patient-level diagnosis and inference algorithm. Patch-level predictions from each patient need to be mapped to a mosaic-, specimen- or patient-level diagnosis to provide a final intraoperative classification (see Extended Data Fig. 3). The set of diagnostic patch softmax output vectors from a specimen or patient is summed elementwise and renormalized to produce specimen-level or patient-level probability distribution. To account for normal brain and pathologic tissue contained within the same specimen, a thresholding procedure was used, such that, if the probability of a normal specimen was >90%, a normal label was assigned. Otherwise, the normal class probabilities were set to zero, the probability distribution was renormalized and the final diagnosis was the expected value of the renormalized distribution. Our inference algorithm leverages the fact that normal brain tissue and nondiagnostic regions have similar histologic features among all patients, resulting in high patch-level classification accuracy for normal brain, and eliminating the need to train an additional classifier based on the patch-level probability histograms[39]. Similar to previous publications using deep learning for medical diagnosis[15,16], a taxonomy of inference classes was used to allow for classification at various clinically relevant levels of granularity (see Extended Data Fig. 2). The probability of any parent/inference class is the sum of its child node probabilities.

Mahalanobis distance-based confidence score. The most common brain tumor types were used for model training and includes >90% of all CNS tumors diagnosed in the USA[22]; however, rare tumor types will be encountered in the clinical setting. Therefore, in addition to a posterior probability distribution over the CNN output classes, we aimed to provide a confidence score to detect tumor samples that are far away from the training distribution, to detect rare tumor types not included during training. We induce class-conditional gaussian distributions with respect to mid- and upper-level features (that is, layer outputs) of our CNN under gaussian discriminant analysis, which results in a confidence score based on the Mahalanobis distance[23]. Without any modification to our pertained network, we obtain a generative model by converting the penultimate layer, for example, to a class-conditional distribution that follows a multivariate gaussian distribution. Specifically, we compute 13 class-conditional gaussian distributions, one for each histologic class, with a tied covariance matrix using our training set. Using these induced class-conditional gaussian distributions, we calculate a confidence score, $M(\mathbf{x})$, using the Mahalanobis distance between the test specimen $\mathbf{x}$ and the closest class-conditional gaussian distribution:

$$M(x) = \min_c (f(x) - \bar{\mu}_c)^{\intercal} \hat{\Sigma}^{-1} (f(x) - \bar{\mu}_c)$$

where $\bar{\mu}_c$ is the class mean, $f(\mathbf{x})$ is the output from the penultimate layer and $\hat{\Sigma}$ is the tied covariance matrix. The specimen-level confidence is the mean patch-level

confidence score. To improve performance and increase separability of common and rare tumor classes as previously described[23], we implemented an ensemble method that included output from seven layers: convolutional layers 159, 195, 199 and 203, final average pooling layer, final dense layer and softmax output layer. Mahalanobis distance-based confidence scores for each layer were then used as features to train a linear discriminant classifier on our training set and rare tumor specimens imaged before starting prospective trial enrollment.

Prospective clinical trial design. A noninferiority trial was designed to rigorously validate our proposed intraoperative diagnostic pipeline. An expected accuracy of 96%, a δ (inferiority limit) of 5%, α (alpha level) of 0.05 and statistical power of 0.9 were used to calculate a minimum patient sample size of 264, with the primary endpoint of overall multiclass diagnostic accuracy (see Extended Data Fig. 4). Prospective enrollment began on 6 April 2018 and closed on 26 February 2019, with a total of 302 patients enrolled. Clinical trial inclusion criteria were the same for intraoperative SRH imaging. Exclusion criteria were: (1) poor quality of specimen on visual gross examination due to excessive blood, coagulation artifact, necrosis or ultrasonic damage or (2) specimen classified as out of distribution by the linear discriminant analysis classifier using the Mahalanobis distance-based confidence score. A total of 278 patients were included in the trial. The conventional intraoperative H&E diagnosis was used in the control arm and the SRH imaging plus CNN was used for the experimental arm. The final histopathologic diagnosis was used to label patients into the appropriate patient-level ground truth class. For example, a patient with final WHO classification of glioblastoma, WHO IV, is classified into the malignant glioma class, or a final diagnosis of diffuse astrocytoma, WHO II, is classified into the diffuse lower-grade glioma class. The strategy does not bias either study arm and allows for a multiclass accuracy value to be calculated for each study arm. Three instances arose where the control arm diagnosis was limited to 'glioma' without further specification. To allow for a one-to-one comparison between the two study arms, the 'glial tumor' inference class was used in the experimental arm for these cases. To eliminate the possibility of sampling error in the control arm, all incorrectly classified specimens underwent secondary review by two board-certified neuropathologists (S.C.P. and P.D.C.) to ensure that the specimen was of sufficient quality to make a diagnosis and that tumor tissue was present in the specimen. After completion of the trial, we prospectively imaged eight stereotactic needle brain biopsies to validate that our workflow in the operations we were sampling was based on stereotactic navigation rather than gross inspection of the tissue (see Supplementary Fig. 1). SRH with automated CNN diagnosis can play an essential role in these cases to confirm diagnostic tissue sampling, provide intraoperative histologic data and cut total surgical time in half (that is, from 60–90 min to 20–30 min).

Activation maximization. Activation maximization allows for qualitative evaluation of learned representations in deep neural network architectures[24]. The objective is to generate an image that maximally activates a neuron or filter in a CNN hidden layer given a set of fixed, trained weights, such that:

$$x^{*} = \underset{x\ s.t.\ R(x)}{\operatorname{argmax}}\ h_j^l(x,\theta)$$

where x is the input image, θ denotes the neural network weights, $h_j^l(x,\theta)$ is the activation of a jth neuron in hidden layer l and $R(x)$ a regularization term. An image, x^{*}, can be generated by computing the gradient of $h_j^l(x,\theta)$ and updating the pixel values of x using iterations of gradient ascent. Our regularization term included weight decay, gaussian blur and dark pixel clipping to improve image clarity and interpretability[40]. We used 500 iterations of gradient ascent for each of the images shown in Fig. 3 and Extended Data Fig. 7. We chose convolutional layer 159, a deep hidden layer with sufficient spatial information to identify regions of low and high activation within a single image, to evaluate class-specific activation.

Probability heatmaps and semantic segmentation of SRH. Class probability heatmaps can localize diagnostic tissue and spatially identify areas with different predicted class labels (for example, normal versus tumor-infiltrated tissue). Our single-scale sliding window approach allows for an intuitive image patch-to-heatmap pixel mapping that: (1) is computationally efficient; (2) yields a ninefold increase in heatmap pixel spatial resolution relative to patch size; and (3) integrates a local neighborhood of overlapping patch predictions for semantic segmentation. For example, a 1,000×1,000-pixel^2 SRS image is divided into a 10×10-pixel^2 grid. The image area contained within each heatmap pixel will overlap with one (grid corners) to nine (inner 6×6 grid) neighboring patches due to valid padding and 100-pixel step size (see Extended Data Fig. 9). The softmax output vector from each overlapping patch is summed and renormalized to give a probability distribution for each heatmap pixel. This procedure yields a prediction heatmap for each output class to produce a 10×10-xk array, where k is the number of output classes. This method can be repeated to produce heatmaps for arbitrarily large SRH images. The IOU metric was used to evaluate segmentation performance.

To produce effective prediction overlays for pathologist review, probabilities were uniformly mapped to a 0–255 scale for three diagnostic classes (that is, nondiagnostic, nontumor inference class and tumor inference class) to generate a three-channel RGB transparency overlay (α=40%).

Statistics and reproducibility. All measures on central tendency were reported as mean ± s.d. CNN training was replicated 10 times and the model with the highest validation accuracy was selected for use in the prospective clinical trial. Pearson's correlation coefficient was used to measure linear correlations. A full R code for calculated trial sample size can be found in our code repository (see below). Please see the 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

A University of Michigan Institutional Review Boards protocol (no. HUM00083059) was approved for the use of human brain tumor specimens in the present study. To obtain these samples or SRH images, contact D.A.O. A code repository for network training, evaluation and visualizations is publicly available at https://github.com/toddhollon/srh_cnn.

References

34. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014* 818–833 (2014).
35. Freudiger, C. W. et al. Stimulated Raman scattering microscopy with a robust fibre laser source. *Nat. Photonics* **8**, 153–159 (2014).
36. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. *arXiv* [cs.CV] (2017). https://arxiv.org/abs/1703.02442
37. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Proc. 27th International Conference on Neural Information Processing Systems* **2**, 3320–3328 (2014).
38. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. *OSDI* **16**, 265–283 (2016).
39. Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2424–2433 (2016).
40. Qin, Z. et al. How convolutional neural networks see the world: a survey of convolutional neural network visualization methods. *Math. Found. Comput.* **1**, 149–180 (2018).

Acknowledgements

We thank T. Cichonski for manuscript editing. This work was supported by the National Institutes of Health National Cancer Institute (grant no. R01CA226527-02), Neurosurgery Research Education Fund, University of Michigan MTRAC and The Cook Family Foundation.

Author contributions

T.C.H., S.C.-P. and D.A.O. conceived of the study, designed the experiments and wrote the article. B.P., H.L., A.R.A., E.U., Z.U.F., S.L., P.D.P., T.M., M.S., P.C. and S.S.S.K. assisted in writing the article. C.W.F. and J.T. built the SRH microscope. T.C.H., A.R.A., E.U., A.V.S., T.D.J., P.C. and A.H.S. analyzed the data. T.D.J. and T.C.H. performed statistical analyses. D.A.O., S.L.H.-J., H.J.L.G., J.A.H., C.O.M., E.L.M., S.E.S., P.G.P., M.B.S., J.N.B., M.L.O., B.G.T., K.M.M., R.S.D., O.S., D.G.E., R.J.K., M.E.I. and G.M.M. provided surgical specimens for imaging. All authors reviewed and edited the manuscript.

Competing interests

D.A.O. is an advisor and shareholder of Invenio Imaging, Inc., a company developing SRH microscopy systems. C.W.F., Z.U.F. and J.T. are employees and shareholders of Invenio Imaging, Inc.

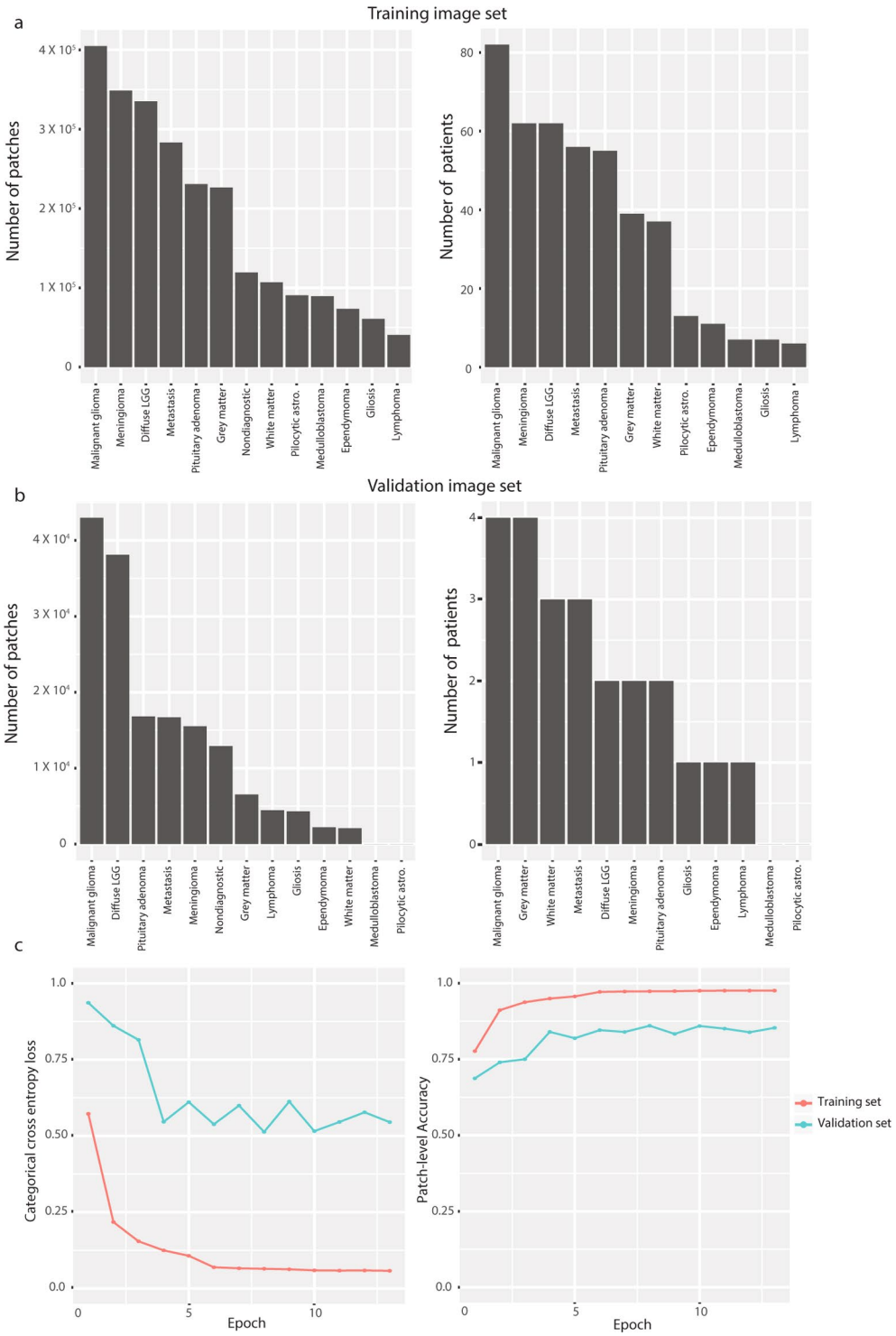
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0715-9.

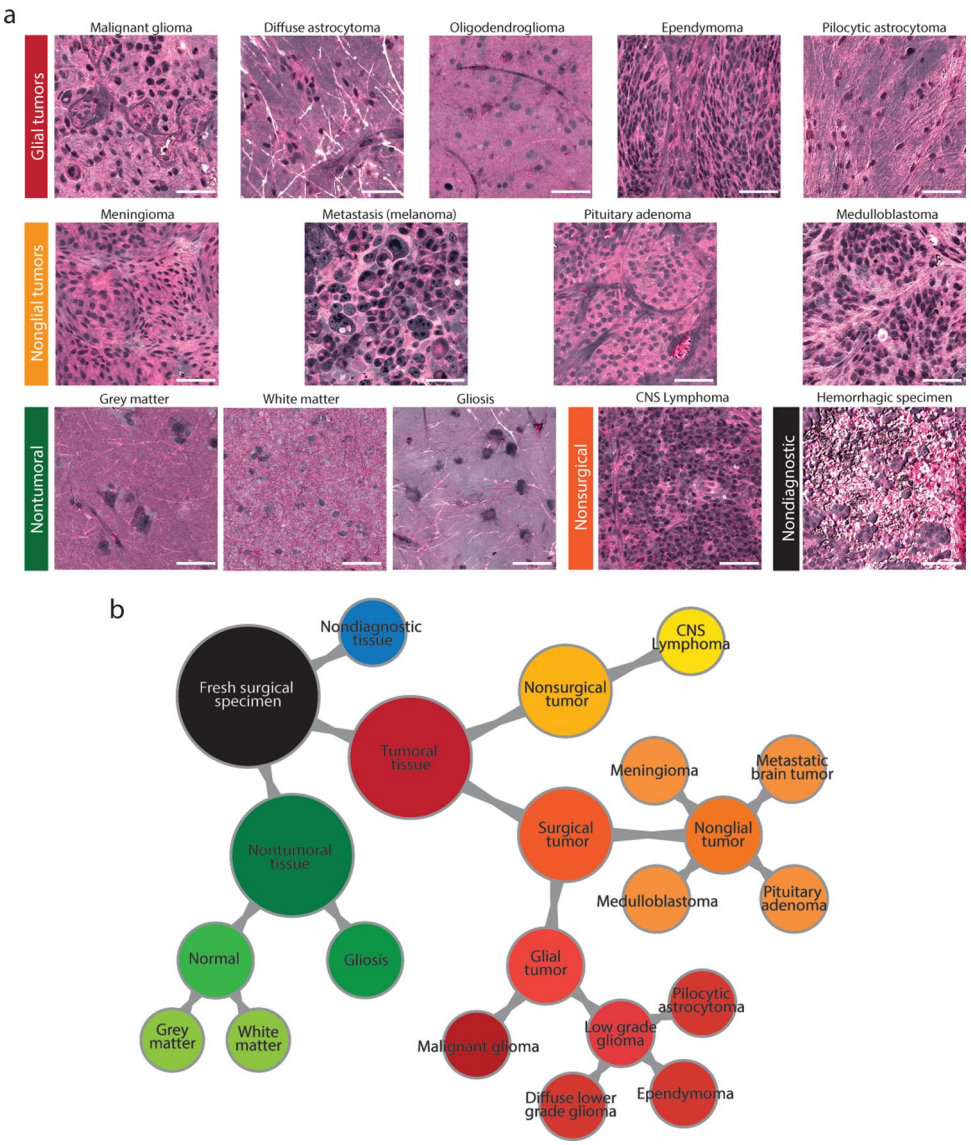
Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0715-9.

Correspondence and requests for materials should be addressed to D.A.O.

Peer review information B. Benedetti and J. Carmona were the primary editors on this article, and managed its editorial process and peer review in collaboration with the rest of the editorial team.



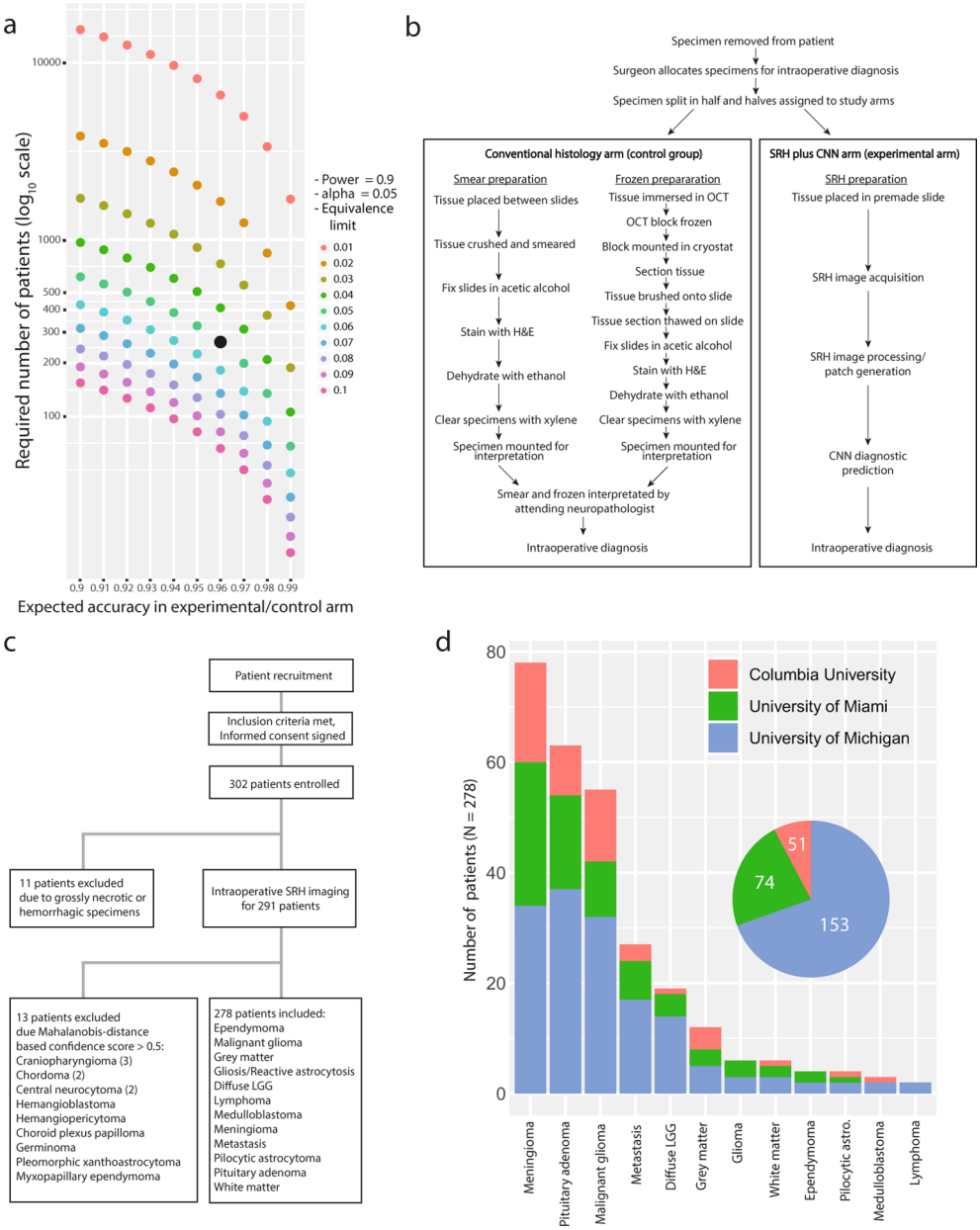
Extended Data Fig. 1 | SRH image dataset and CNN training. The class distribution of (**a**) training and (**b**) validation set images are shown as number of patches and patients. Class imbalance results from different incidence rates among human central nervous system tumors. The training set contains over 50 patients for each of the five most common tumor types (malignant gliomas, meningioma, metastasis, pituitary adenoma, and diffuse lower grade gliomas). In order to maximize the number of training images, no cases from medulloblastoma or pilocytic astrocytoma were included in the validation set and oversampling was used to augment the underrepresented class during CNN training. **c**, Training and validation categorical cross entropy loss and patch-level accuracy is plotted for the training session that yielded the model used for our prospective clinical trial. Training accuracy converges to near-perfect with a peak validation accuracy of 86.4% following epoch 8. Training procedure was repeated 10 times with similar accuracy and cross entropy convergence. Additional training did not result in better validation accuracy and early stopping criteria were reached.



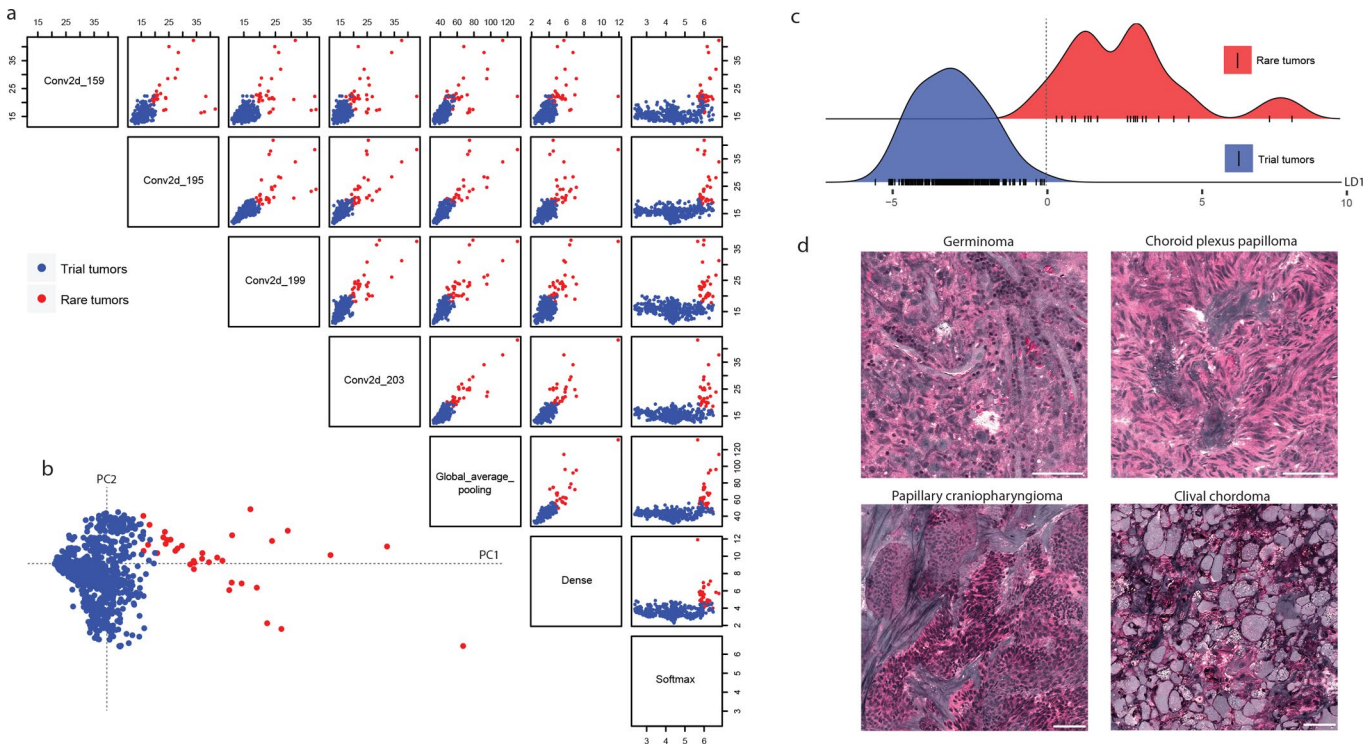
Extended Data Fig. 2 | A taxonomy of intraoperative SRH diagnostic classes to inform intraoperative decision-making. a, Representative example SRH images from each of the 13 diagnostic class are shown. Both diffuse astrocytoma and oligodendroglioma are shown as examples of diffuse lower grade gliomas. Classic histologic features (i.e., piloid process in pilocytic astrocytomas, whorls in meningioma, and microvascular proliferation in glioblastoma) can be appreciated, in addition to features unique to SRH images (e.g., axons in gliomas and normal brain tissue). Scale bar, 50 μm. **b**, A taxonomy of diagnostic classes was selected specifically to inform intraoperative decision-making, rather than to match WHO classification. Essential intraoperative distinctions, such as tumoral versus nontumoral tissue or surgical versus nonsurgical tumors, allow for safer and more effective surgical treatment. Inference node probabilities inform intraoperative distinctions by providing coarse classification with potentially higher accuracy due to summation of daughter node probabilities[16]. The probability of any inference node is the sum of all of its daughter node probabilities.

```
1: Inputs
2:    patches (set of arrays): a set of images from a patient
3:    model (computational graph): trained CNN
4:
5: Outputs
6:    distribution (dictionary): a mapping of diagnostic classes to probabilities
7:
8: procedure PREDICTION(patches, model)
9:    predictions ← []
10:   for patch in patches do
11:      softmax_output ← model(patch)
12:      if argmax(softmax_output) == "nondiagnostic" then
13:         continue
14:      else
15:         append softmax_output to predictions
16:   end for
17:   return predictions
18:
19: procedure RENORMALIZE(predictions)
20:   summed_dist ← sum(predictions)
21:   for class in predictions do
22:      predictions.class ← sum(predictions.class) / summed_dist
23:   end for
24:   return predictions
25:
26: procedure DIAGNOSIS(patches, model)
27:   renorm_prediction ← RENORMALIZE(PREDICTION(patches, model))
28:   if sum(renorm_prediction.normal) > 0.9 then
29:      return renorm_prediction
30:   else
31:      renorm_prediction.normal ← 0
32:      return RENORMALIZE(renorm_prediction)
33:
34: return DIAGNOSIS(patches, model)
```

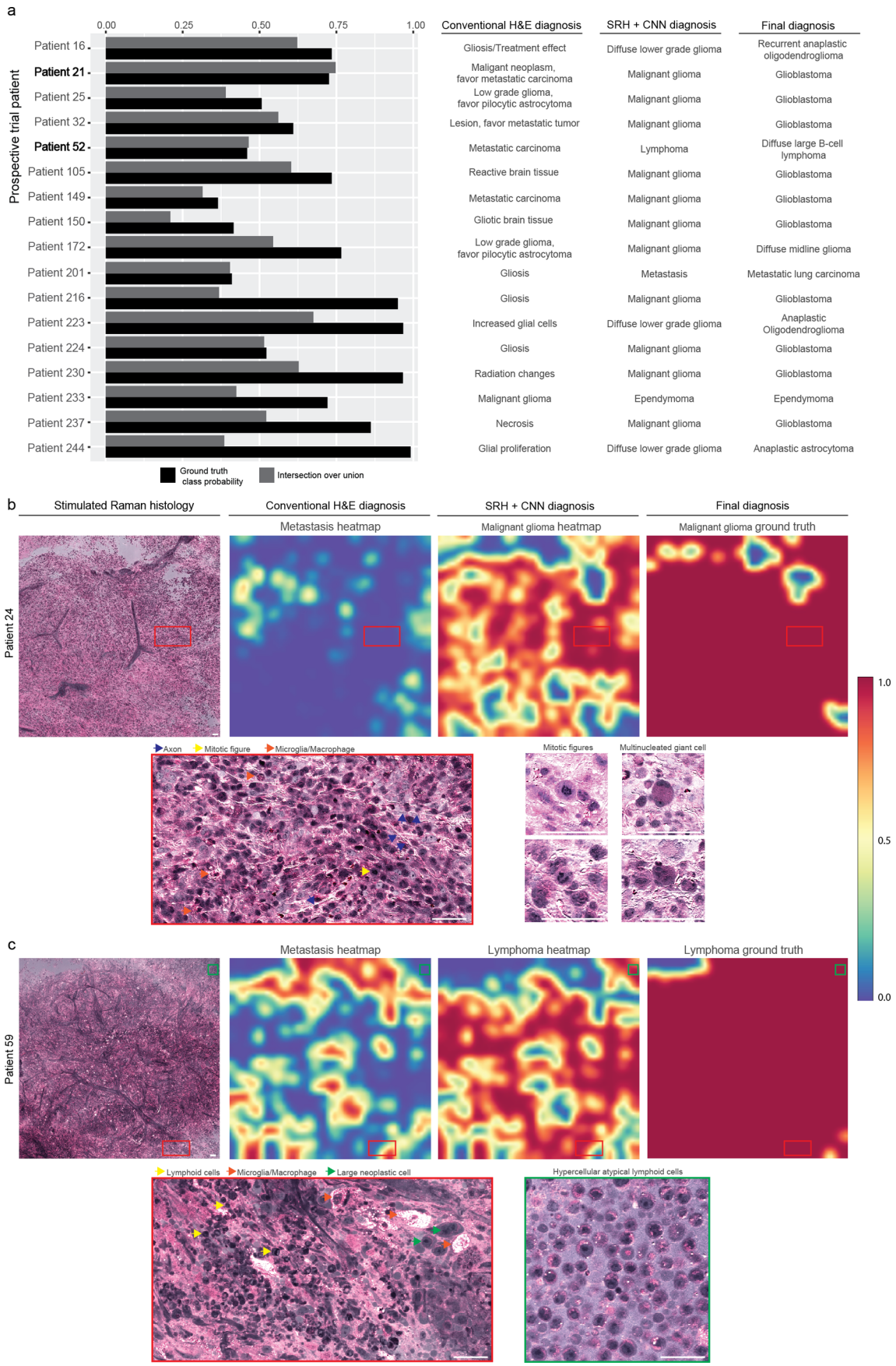
Extended Data Fig. 3 | Inference algorithm for patient-level brain tumor diagnosis. A patch-based classifier that uses high-magnification, high-resolution images for diagnosis requires a method to aggregate patch-level predictions into a single intraoperative diagnosis. Our inference algorithm performs a feedforward pass on each patch from a patient, filters the nondiagnostic patches (line 12), and stores the output softmax vectors in an $R^{N \times 13}$ array. Each column of the array, corresponding to each class, is summed and renormalized (line 22) to produce a probability distribution. We then used a thresholding procedure such that if greater than 90% of the probability density is nontumor/normal, that probability distribution is returned. Otherwise, the normal/nontumor class (gray matter, white matter, gliosis) probabilities are set to zero (line 31), the distribution renormalized, and returned. This algorithm leverages the observation that normal brain and nondiagnostic tissue imaged using SRH have similar features across patients resulting in high patch-level classification accuracy. Using the expected value of the renormalized patient-level probability distribution for the intraoperative diagnosis eliminates the need to train an additional classifier based on patch predictions.



Extended Data Fig. 4 | Prospective clinical trial design and recruitment. a, Minimum sample size was calculated under the assumption that pathologists' multiclass diagnostic accuracy ranges from 93% to 97% based on our previous experiments[6] and that a clinically significant lower accuracy bound was less than 91%. We, therefore, selected an expected accuracy of 96% and equivalence/noninferiority limit, or delta, of 5%, yielding a noninferiority threshold accuracy of 91% or greater. Minimum sample size was 264 (black point) patients using an *alpha* of 0.05 and a power of 0.9 (*beta* = 0.1). **b**, Flowchart of specimen processing in both the control and experimental arms is shown. **c**, A total of 302 patients met inclusion criteria and were enrolled for intraoperative SRH imaging. Eleven patients were excluded at the time of surgery due to specimens that were below the necessary quality for SRH imaging. A total of 291 patients were imaged intraoperatively and 13 patients were subsequently excluded due to a Mahalanobis distance-based confidence score (See Extended Data Figure 5), resulting in a total of 278 patients included. **d**, Meningioma, pituitary adenomas, and malignant gliomas were the most common diagnoses in our prospective cohort. University of Michigan, University of Miami, and Columbia University recruited 55.0%, 26.6%, 18.4% of the total patients, respectively.

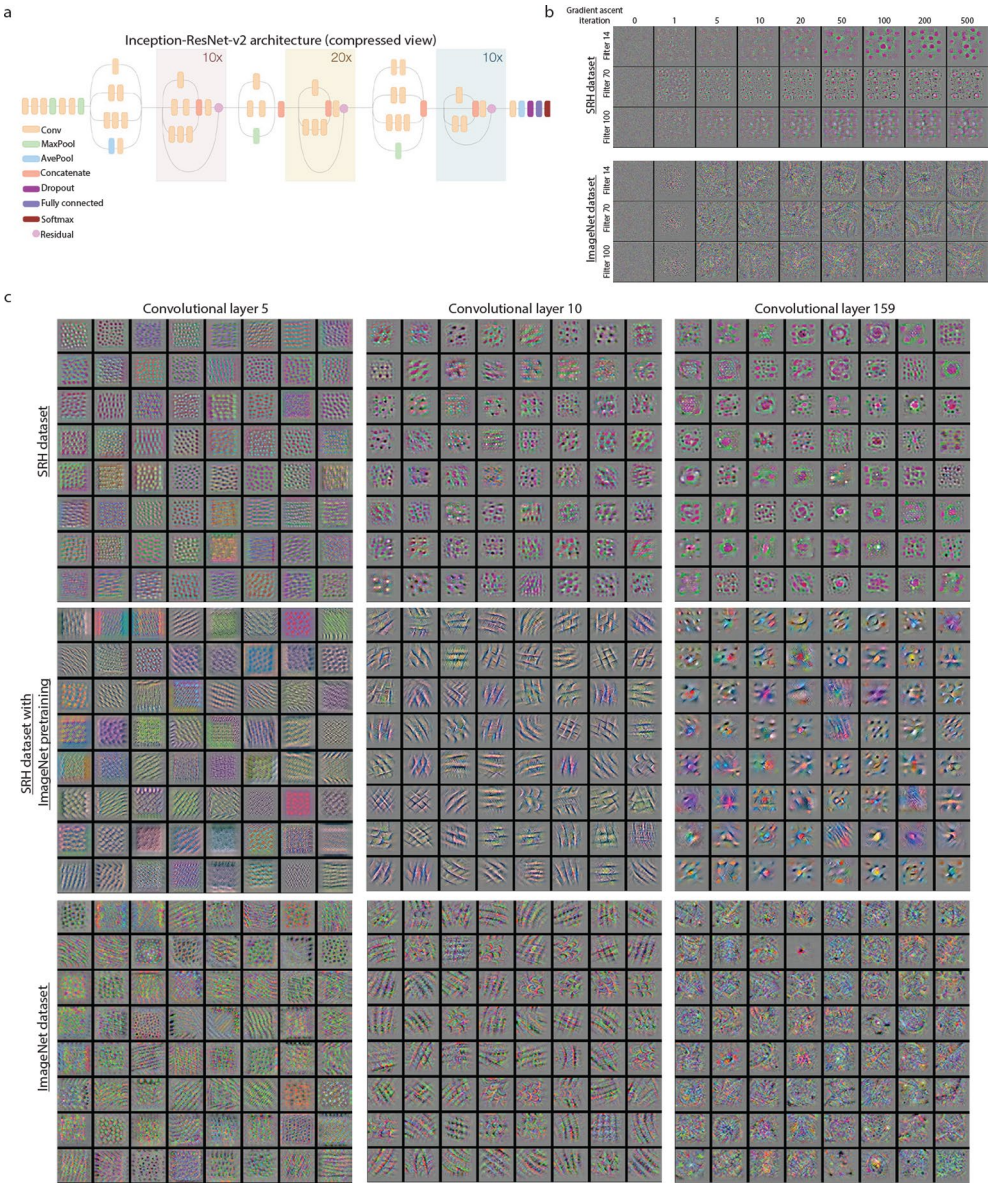


Extended Data Fig. 5 | Mahalanobis distance-based confidence score. a, Pairwise comparison and **b**, principal component analysis of class-conditional, Mahalanobis distance-based, confidence score for each layer output included in the ensemble. The confidence score from the mid- and high-level hidden features are correlated, which demonstrate that out-of-distribution samples result in greater Mahalanobis distances throughout the network. As previously described and observed in our results, out-of-distribution (i.e. rare tumors) are better detected in the representation space of deep neural networks, rather than the "label-overfitted" output space of the softmax layer[23]. **c**, Specimen-level predictions (black hashes, n = 478) and kernel density estimate from the trained LDA classifier for all specimens imaged during the trial period projected onto the linear discriminant axis. Trial and rare tumor cases were linearly separable resulting in all 13 rare tumor cases imaged during the trial period correctly identified. **d**, SRH mosaics of rare tumors imaged during the trial period are shown. Germinomas show classic large round neoplastic cells with abundant cytoplasm and fibrovascular septae with mature lymphocytic infiltrate. Choroid plexus papilloma shows fibrovascular cores lined with columnar cuboidal epithelium. Papillary craniopharyngioma have fibrovascular cores with well-differentiated monotonous squamous epithelium. Clival chordoma has unique bubbly cytoplasm (i.e. physaliferous cells). Scale bar, 50 μm.

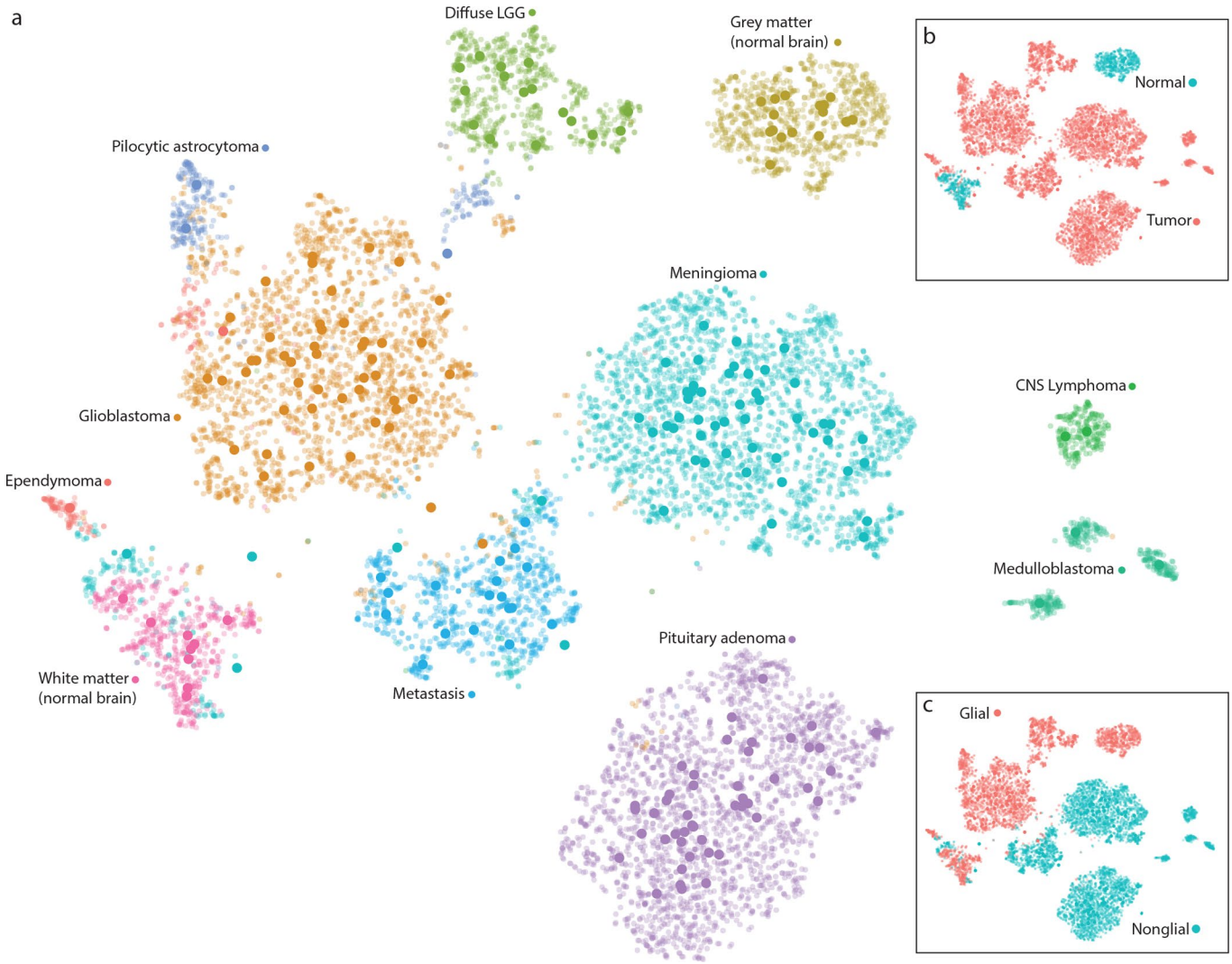


Extended Data Fig. 6 | see figure caption on next page.

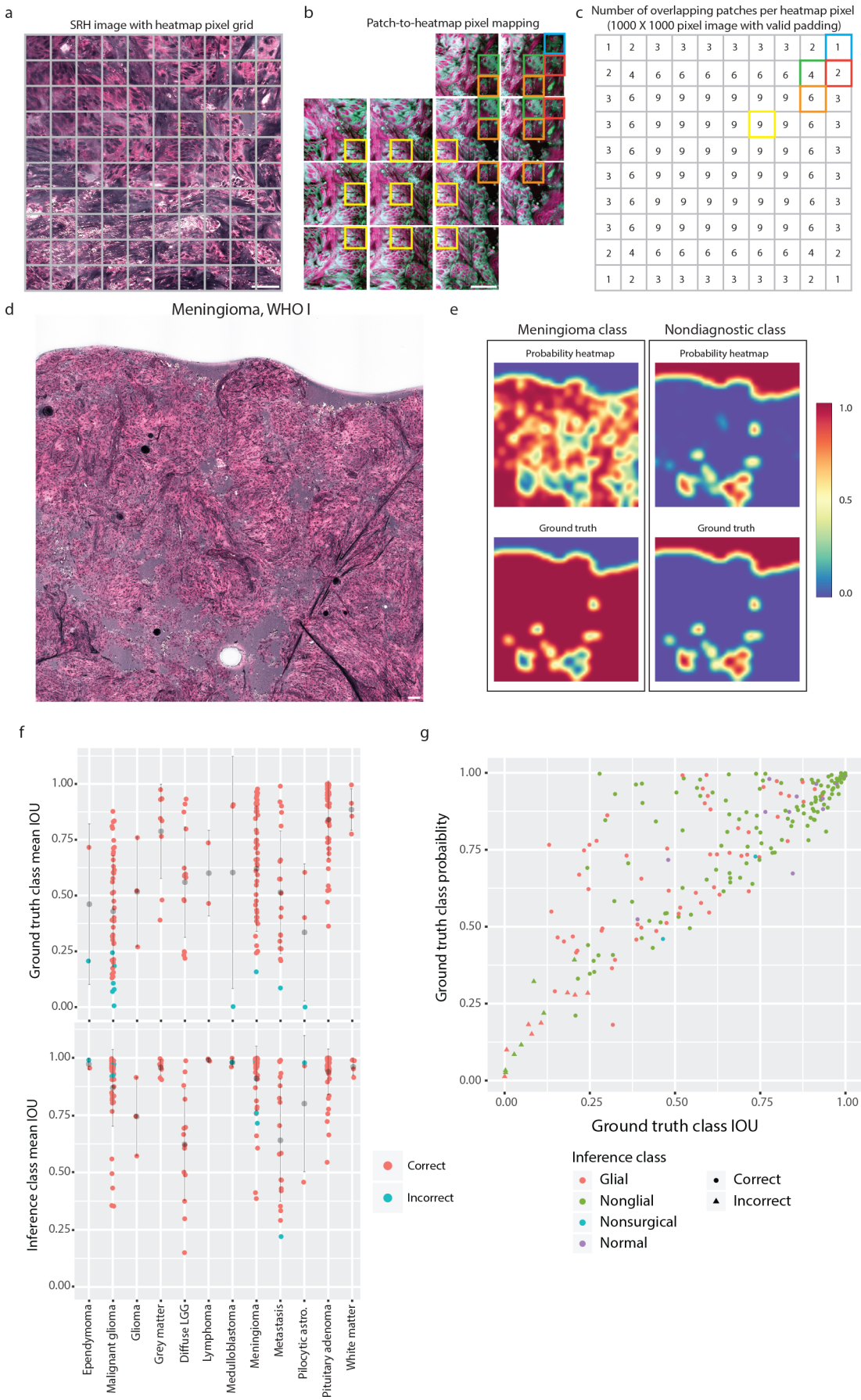
Extended Data Fig. 6 | Error analysis of pathologist-based classification of brain tumors. a, The true class probability and intersection over union values for each of the prospective clinical trial patients incorrectly classified by the pathologists. All 17 were correctly classified using SRH plus CNN. All incorrect cases underwent secondary review by two board-certified neuropathologists (S.C.P., P.C.) to ensure the specimens were 1) of sufficient quality to make a diagnosis and 2) contained tumor tissue. **b**, SRH mosaic from patient 21 (glioblastoma, WHO IV) is shown. Pathologist classification was metastatic carcinoma; however, CNN metastasis heatmap does not show high probability. Malignant glioma probability heatmap shows high probability over the majority of the SRH mosaic, with a 73.4% probability of patient-level malignant glioma diagnosis. High-magnification views show regions of hypercellularity due to tumor infiltration of brain parenchyma with damaged axons, activated lipid-laden microglia, mitotic figures, and multinucleated cells. **c**, SRH mosaic from patient 52 diagnosed with diffuse large B-cell lymphoma predicted to be metastatic carcinoma by pathologist. While CNN identified patchy areas of metastatic features within the specimen, the majority of the image was correctly classified as lymphoma. High-magnification views show atypical lymphoid cells with macrophage infiltration. Regions with large neoplastic cells share cytologic features with metastatic brain tumors, as shown in Fig. 3. Scale bar, 50 μm.



Extended Data Fig. 7 | Activation maximization to elucidate SRH feature extraction using Inception-ResNet-v2. a, Schematic diagram of Inception-ResNet-v2 shown with repeated residual blocks compressed. Residual connections and increased depth resulted in better overall performance compared to previous Inception architectures. **b**, To elucidate the learned feature representations produced by training the CNN using SRH images, we used activation maximization[24]. Images that maximally activate the specified filters from the 159th convolutional layer are shown as a time series of iterations of gradient ascent. A stable and qualitatively interpretable image results after 500 iterations, both for the CNN trained on SRH images and for ImageNet images. The same set of filters from the CNN trained on ImageNet are shown in order to provide direct comparison of the trained feature extractor for SRH versus natural image classification. **c**, Activation maximization images are shown for filters from the 5th, 10th, and 159th convolutional layers for CNN trained using SRH images only, SRH images after pretraining on ImageNet images, and ImageNet images only. The resulting activation maximization images for the ImageNet dataset are qualitatively similar to those found in previous publications using similar methods[34]. CNN trained using only SRH images produced similar classification accuracy compared to pretraining and activation maximization images that are more interpretable compared to those generated using a network pretrained on ImageNet weights.

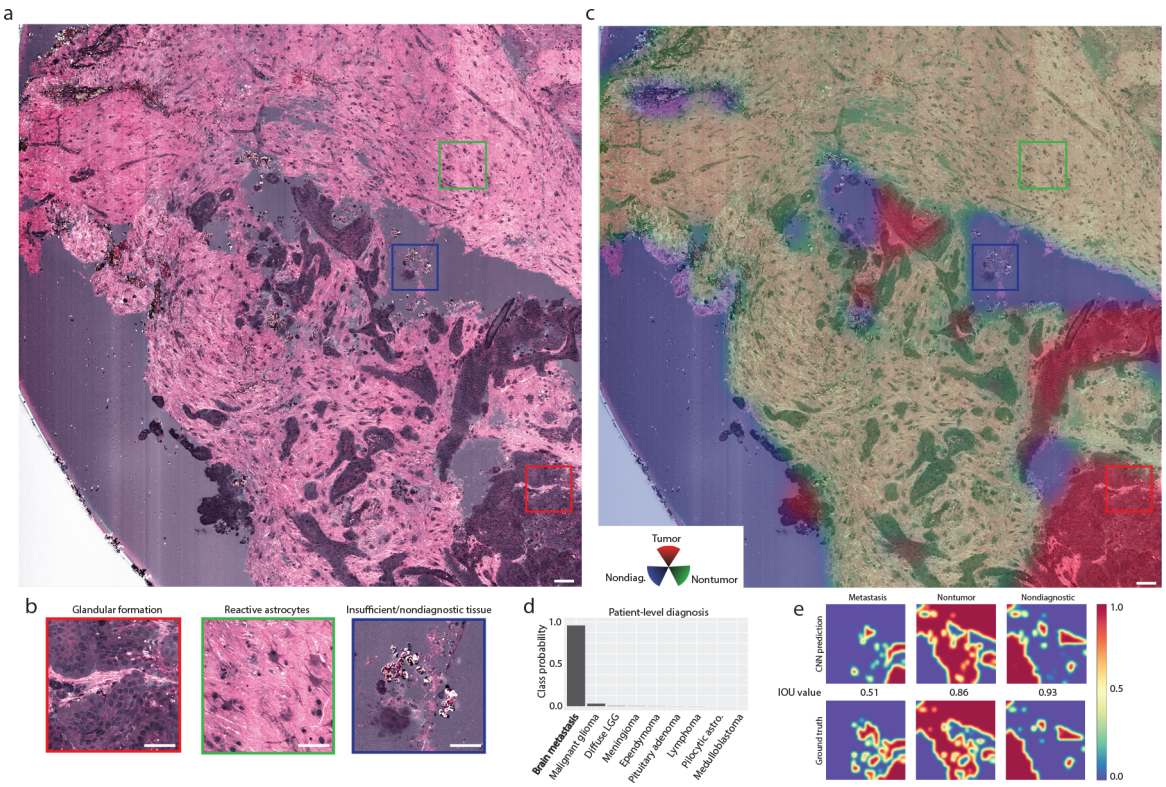


Extended Data Fig. 8 | t-SNE plot of internal CNN feature representations for clinical trial patients. We used the 1536-dimensional feature vector from the final hidden layer of the Inception-ResNet-v2 network to determine how individual patches and patients are represented by the CNN using t-distributed stochastic neighbor embedding (t-SNE), an unsupervised clustering method to visualize high-dimensional data. **a**, One hundred representative patches from each trial patient (n = 278) were sampled for t-SNE and are shown in the above plot as small, semi-transparent points. Each trial patient is plotted as a large point located at their respective mean patch position. Recognizable clusters form that correspond to individual diagnostic classes, indicating that tumor types have similar internal CNN representations. **b**, Gray and white matter form separable clusters from tumoral tissue, but also from each other. lipid-laden myelin in white matter has significantly different SRH features compared to gray matter with axons and glial cells in a neuropil background. **c**, Diagnostic classes that share cytologic and histoarchitectural features form neighboring clusters, such as malignant glioma, pilocytic astrocytoma, and diffuse lower grade glioma (i.e., glial tumors). Lymphoma and medulloblastoma are adjacent and share similar features of hypercellularity, high nuclear:cytoplasmic ratios, and little to no glial background in dense tumor.



Extended Data Fig. 9 | see figure caption on next page.

Extended Data Fig. 9 | Methods and results of SRH segmentation. a, A 1000 × 1000-pixel SRH image is shown with the corresponding grid of probability heatmap pixels that results from using a 300 × 300-pixel sliding window with 100-pixel step size in both horizontal and vertical directions. Scale bar, 50 μm. **b**, An advantage of this method is that the majority of the heatmap pixels are contained within multiple image patches and the probability distribution assigned to each heatmap pixel results from a renormalized sum of overlapping patch predictions. This has the effect of pooling the local prediction probabilities and generates a smoother prediction heatmap. **c**, For our example, each pixel of the inner 6 × 6 grid has 9 overlapping patches from which the probability distribution is determined. **d**, An SRH image of a meningioma, WHO grade I, from our prospective trial is shown as an example. Scale bar, 50 μm. **e**, The meningioma probability heatmap is shown after bicubic interpolation to scale image to the original size. Nondiagnostic prediction and ground truth is for the same SRH mosaic and is shown. **f**, The SRH semantic segmentation results of the full prospective cohort (n = 278) are plotted. The upper plot shows the mean IOU and standard deviation (i.e., averaged over SRH mosaics from each patient) for ground truth class (i.e., output classes). Note that the more homogenous or monotonous histologic classes (e.g., pituitary adenoma, white matter, diffuse lower grade gliomas) had higher IOU values compared to heterogeneous classes (e.g., malignant glioma, pilocytic astrocytoma). The lower plot shows the mean inference class IOU and standard deviation (i.e., either tumor or normal inference class) for each trial patient. Mean normal inference class IOU for the full prospective cohort was 91.1 ± 10.8 and mean tumor inference class IOU was 86.4 ± 19.0. **g**, As expected, mean ground truth class IOU values for the prospective patient cohort (n = 278) were correlated with patient-level true class probability (Pearson correlation coefficient, 0.811).



Extended Data Fig. 10 | Localization of metastatic brain tumor infiltration in SRH images. a, Full SRH mosaic of a specimen collected at the brain–tumor margin of a patient with a metastatic brain tumor (non-small cell lung adenocarcinoma). **b**, Metastatic rests with glandular formation are dispersed among gliotic brain with normal neuropil. **c**, Three-channel RGB CNN-prediction transparency is overlaid on the SRH image for pathologist review intraoperatively with associated (**d**) patient-level diagnostic class probabilities. **e**, Class probability heatmap for metastatic brain tumor (IOU 0.51), nontumor (IOU 0.86), and nondiagnostic (IOU 0.93) regions within the SRH image are shown with ground truth segmentation. Scale bar, 50 μm.

nature research

Corresponding author(s): Daniel Orringer

Last updated by author(s): Oct 10, 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.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

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	Proprietary software used in the NIO Imaging System (Invenio Imaging, Inc) Python 3.6.5, tensorflow-gpu 1.8.0, keras 2.2.4, scipy 1.1.0, numpy 1.14.3, open-cv 3.4.1, imageio 2.4.1, skimage 0.13.1, pydicom 1.1.0, pandas 0.23.0
Data analysis	Python 3.6.5, tensorflow-gpu 1.8.0, keras 2.2.4, sklearn 0.19.1, scipy 1.1.0, numpy 1.14.3, open-cv 3.4.1, imageio 2.4.1, skimage 0.13.1, matplotlib 2.2.2, pydicom 1.1.0, pandas 0.23.0 R 3.4.4, ggplot2 3.0.0, dplyr 0.7.6, tidyr 0.8.1, purrr 0.2.5, caret 6.0

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

All code is available from the corresponding author upon request. A University of Michigan IRB protocol (HUM00083059) was approved for the use of human brain tumor specimens in this study. To obtain these samples or SRH images, contact D.A.O.

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

Life sciences study design

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

Sample size	A noninferiority trial was designed to rigorously validate our proposed intraoperative diagnostic pipeline. An expected accuracy of 96%, a noninferiority threshold/delta of 5%, alpha 0.05, and power of 0.9 were used to calculate a minimum patient sample size of 264 with the primary endpoint of overall multiclass diagnostic accuracy
Data exclusions	No data was excluded other than those patients that meet exclusion criteria: Exclusion criteria: 1) Poor quality of specimen on visual gross examination due to excessive blood, coagulation artifact, necrosis, or ultrasonic damage. 2) Surgeon declares that all collected specimens must be allocated for clinical purposes 3) Final histopathologic diagnosis not included in CNN output classes. Output classes include: a. Malignant glioma b. Diffuse lower grade glioma c. Ependymoma (Non-myxopapillary) d. Pilocytic astrocytoma e. Medulloblastoma f. Metastasis g. Pituitary adenoma h. Lymphoma i. Meningioma j. Normal brain (grey and white matter) k. Gliosis/Astrocytosis 4) Any unanticipated malfunction of SRH imager at the time of surgery
Replication	Results were replicated using multiple health centers with multiple stimulated Raman histology imagers.
Randomization	N/A
Blinding	N/A

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Inclusion criteria for intraoperative SRH imaging includes: 1) Male or female of any age 2) Subjects undergoing central nervous system tumor resection or epilepsy surgery at Michigan Medicine, New York-Presbyterian/Columbia University Medical Center, or the University of Miami Health System 3) Subject/parent/durable power of attorney able to give informed consent

	4) Subjects in which there is additional specimen beyond what is needed for routine clinical intraoperative diagnosis and final histopathologic diagnosis
Recruitment	All patients (or their guardians) with newly diagnosed or recurrent brain lesions/suspected brain tumors were approached for recruitment. Recruitment occured both in the outpatient and inpatient setting. Physicians and advanced practice providers were trained to recruit patients and obtained study consent. Risks and benefits of the trial were detailed. Patients were reassured that the trial is to assess a diagnostic modality only and is noninterventional. We explained that no additional tissue/brain will be injured as a result of study enrollment and any tissue collected would have either been in excess of what is needed for diagnosis or discarded. All collected PHI was password protected and encrypted to minimize any possible harm to the patient.
Ethics oversight	Complete Institutional Review Board oversight.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A. The clinical trial was diagnostic/non-interventional and conducted at Michigan Medicine, New York-Presbyterian/Columbia University Medical Center, and the University of Miami Health System
Study protocol	Full study protocol can be obtained from the corresponding author upon request.
Data collection	Study conducted at Michigan Medicine, New York-Presbyterian/Columbia University Medical Center, and the University of Miami Health System. Prospective enrollment began on April 6, 2018 and closed on February 26, 2019.
Outcomes	Primary outcome measure was overall diagnostic accuracy. Secondary outcome was mean class diagnostic accuracy.