Simulated Diagnostic Performance of Ultra-Low-Field MRI: Harnessing Open-Access Datasets to Evaluate Novel Devices

T. Campbell Arnold[1,2] *, Steven N. Baldassano[1,2] *, Brian Litt[1-3], Joel M. Stein[1,2,4]

1. Department of Bioengineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, PA, 19104, USA
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania, Philadelphia, PA, 19104, USA
3. Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
4. Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

***** These authors contributed equally to this work.

Corresponding author:
T. Campbell Arnold
Email: tcarnold@seas.upenn.edu
Address: 240 S. 33rd Street
Univ. of Pennsylvania
Philadelphia, PA 19104

Keywords: Ultra-Low-field MRI, Low-field MRI, point-of-care MRI, simulated clinical trial, virtual clinical trial, deep learning

Abbreviations: Ultra-low-field strength (ULF), high-field strength (HF), low-grade glioma (LGG), high-grade glioma (HGG), multiple sclerosis (MS), fluid-attenuated inversion recovery (FLAIR), signal to noise ratio (SNR), area under curve (AUC), receiver operating characteristic (ROC), class activation maps (CAMs), early feasibility studies (EFS), investigational device exemption (IDE)

Article length
Abstract: 241, Introduction: 500, Methods: 1110, Results: 849, Discussion: 1005, Conclusion: 99
Tables: 2, Figures: 6, References: 43
Article total: 3563

Abstract

The purpose of this study is to demonstrate a method for virtually evaluating novel imaging devices using machine learning and open-access datasets, here applied to a new, ultra-low-field strength (ULF), 64mT, portable MRI device. Paired 3T and 64mT brain images were used to develop and validate a transformation converting standard clinical images to ULF-quality images. Separately, 3T images were aggregated from open-source databases spanning four neuropathologies: low-grade glioma (LGG, N=76), high-grade glioma (HGG, N=259), stroke (N=28), and multiple sclerosis (MS, N=20). The transformation method was then applied to the open-source data to generate simulated ULF images for each pathology. Convolutional neural networks (DenseNet-121) were trained to detect pathology in axial slices from either 3T or simulated 64 mT images, and their relative performance was compared to characterize the potential diagnostic capabilities of ULF imaging. Algorithm performance was measured using area under the receiver operating characteristic curve. Across all cohorts, pathology detection was similar between 3T and simulated 64mT images (LGG: 0.97 vs. 0.98; HGG: 0.96 vs. 0.95; stroke: 0.94 vs. 0.94; MS: 0.90 vs 0.87). Pathology detection was further characterized as a function of lesion size, intensity, and contrast. Simulated images showed decreasing sensitivity for lesions smaller than 4 cm^2 (~2.25 cm in diameter). While simulations cannot replace prospective trials during the evaluation of medical devices, they can provide guidance and justification for prospective studies. Simulated data derived from open-source imaging databases may facilitate testing and validation of new imaging devices.

Highlights

- Ultra-low-field, point-of-care MRI has potential to detect a range of pathologies including brain tumors, strokes, and multiple sclerosis. However, determining the diagnostic capabilities and appropriate use case for such devices requires further prospective studies.
- Open-source image datasets provide a powerful tool for accelerating imaging research and enable simulated trials that can guide prospective clinical trials or device development.

1. Introduction

Modern medical imaging has become a mainstay of patient care, particularly in the diagnosis and management of patients with neurologic disease. While the availability of imaging technology has dramatically increased worldwide in recent decades, the expense and operational complexity of standard imaging systems limits access in underserved areas and developing countries [1]. This so-called “radiology divide” leaves about ninety percent of the world’s population without access to magnetic resonance imaging (MRI) [2] and almost two-thirds of the population without even basic imaging technology such as ultrasound and X-ray radiography [3–5].

Ultra-low-field strength (ULF, <0.1T) MRI systems aim to make MRI more accessible, promising lower cost, portability, fewer magnetic field-related safety concerns, and ease of use [6]. Such devices could decrease healthcare expenditures, improve availability in underserved areas, and provide a convenient and ionizing-radiation-free modality for routine or monitoring

studies. ULF MRI systems may be suitable for hospitalized patients, such as those in intensive care units or isolation wards, for whom transport to a standard clinical scanner carries unacceptable risk [7–9]. More broadly, ULF MRI units could potentially be used in ambulances, emergency departments, physician's offices or rural clinics [10,11].

While ULF MRI presents clear practical advantages, these systems produce images with lower signal-to-noise than their high-field strength (HF) counterparts and are largely designed to complement, and not replace, standard MRI. Prior to deployment for clinical use, the diagnostic capabilities of novel imaging technologies such as ULF MRI should be evaluated across a wide range of patients and pathology. The standard approach for device evaluation, improvement, and optimization requires recruiting large numbers of patients and manual image review by radiologists. This process is costly and time-consuming, which can limit the device development cycle. Moreover, selecting target use cases is difficult without basic information about device sensitivity. A complementary approach is to simulate ULF images from existing HF data, leveraging large publicly accessible HF imaging databases. Such databases, typically compiled for machine learning competitions [12–15] or collaborative research programs [16,17], span broad ranges of pathology and offer a wealth of information for retrospective analysis [18].

Here, we propose a generalizable approach to use open-access imaging databases to evaluate novel imaging devices. We employ an empiric method to transform existing open-access images to a custom domain. We use this simple method to convert high-resolution 3T MR images aggregated from several open-access databases to lower-quality images designed to mimic images acquired on an ULF MRI scanner. Separate convolutional neural networks are then trained to detect pathology in axial slices from the HF or simulated ULF images, and detection performance is compared between image pairs to characterize the potential diagnostic

capabilities of ULF MRI. While automated lesion detection in simulated images does not guarantee detection on the actual device, simulated performance may help indicate whether pursuing a prospective study for a given application is promising. Applied here to ULF MRI, this simulated trial approach offers a broadly applicable means for evaluating and optimizing novel medical imaging technologies to complement traditional clinical trials.

2. Materials & Methods

2.1. Data collection

We collected data from six adult patients with known or suspected hydrocephalus imaged same-day on a 64mT ULF MRI scanner (Hyperfine) and a standard 3T HF MRI system (Siemens). Data was collected as part of an ongoing research study approved by the University of Pennsylvania Institutional Review Board and patients provided informed consent. Axial fluid-attenuated inversion recovery (FLAIR) images covering the whole brain were collected on each scanner. Image resolution was 1.5 x 1.5 x 5 mm for ULF images and ranged from isotropic to 0.78 x 0.78 x 3 mm for HF images.

Separately, retrospective axial FLAIR images obtained at 3T for a range of pathologies were aggregated from several open-access sources [12–15,19]. Pathologies consisted of high-grade glioma (HGG, N=259), low-grade glioma (LGG, N=76), stroke (N=28), and multiple sclerosis (MS, N=20) [12–15]. Each dataset contained manual segmentations of lesions, which generally manifest as hyperintense areas on FLAIR imaging. These datasets incorporate a range of lesion sizes and signal intensities that can be quantified using the provided lesion segmentations. Note that HGG and LGG lesion segmentations on FLAIR imaging include areas that may represent vasogenic edema or non-enhancing infiltrative neoplasm as well as enhancing

components when present. Additionally, non-lesional control subjects (N=5) were drawn from the OASIS3 dataset [19]. **Table 1** contains information about pathology datasets used in this retrospective study and related web addresses as of publication are listed in section *2.8 Code and data availability*.

2.2. Simulated image generation: High-field to ultra-low-field MRI

To generate simulated ULF images from HF data, we employed a simple image transformation using 3T and 64mT image pairs from three subjects. Transformation steps are listed in **Fig. 1A** and include registration, brain extraction, re-slicing, Gaussian smoothing, and noise filtering. To match ULF image quality, smoothing and noise filters were parameterized and differences in histogram features between real and simulated images were minimized. The objective function was comprised of the first three statistical moments (mean, standard deviation, and skewness). An example HF/ULF pair and the simulated ULF image can be seen in **Fig. 1B** with intensity histograms shown in **Fig. 1C**. The transformation was validated using data from three additional subjects. The image transformation method was quantitatively assessed using the gradient entropy, *F*, as a measure of perceived diagnostic image quality. This metric was derived by McGee et al. [20]:

$$F = -\sum_{ij} h_{i,j} \log_2[h_{i,j}], \tag{1}$$

$$h_{i,j} = \frac{|[1 \quad -1] * g_{i,j}|}{\sum_{ij} |[1 \quad -1] * g_{i,j}|}. \tag{2}$$

where $g_{i,j}$ is the pixel value at coordinate i,j and * represents the convolution operation. Of 24 metrics evaluated, McGee et al. found gradient entropy (Equation 1) to be the most strongly correlated with perceived image quality when applied to structural MRI. The

transformation was applied to the HF images collected from open-access datasets to produce simulated ULF images for glioma, stroke, and multiple sclerosis.

2.3. Modulating lesion contrast

To assess the relationship between lesion contrast and detection accuracy, we prepared additional simulated ULF images from the HGG dataset by decreasing signal intensity within the segmented lesions. Lesion intensity values were scaled independently from surrounding brain tissue in 20% increments over a range from 100% (original intensity) to 0% (isointense with background tissue). Isointensity was defined as mean lesion intensity equal to mean intensity of non-lesional tissue in the same slice. As described subsequently, separate classifiers were trained to identify lesions at each intensity level, allowing for the decoupling of intensity contrast and structural abnormalities in classifier performance. Note that because of structural abnormalities caused by large tumors, such as mass effect, midline shift, and ventricular effacement, as well as intensity heterogeneity within lesions, even isointense lesions may retain some structural and signal alterations after contrast modulation. To minimize the effect of intensity heterogeneity, we restricted our analysis to subjects within the top 50% of lesion homogeneity (N=130), as defined by within tumor signal to noise ratio (SNR). Accurate detection of isointense lesions should therefore be primarily driven by structural abnormalities.

2.4. Model Architecture

A convolutional neural network model was used to identify pathology in each high-field and simulated ultra-low-field dataset. Model construction and training was carried out using the Keras API [21] with TensorFlow [22,23] backend. Model architecture consisted of the

DenseNet-121 network [24,25] with initial weights pre-trained on the ImageNet database [26] and four additional densely-connected layers using Xavier initialization. This architecture was consistent across datasets (Supplemental Fig. 1).

2.5. Model Training

For each dataset, a unique model was trained to perform binary classification of axial slices (lesion present vs. lesion absent). Separate models were trained on the HF and simulated ULF images. Slices were labeled as "lesion present" if at least one pixel from the ground-truth lesion segmentations was present. Each dataset was divided (~9:1 split) into "training" and "test" datasets (Supplemental Table 1). Each subject was confined to either the training or test dataset. Training image order was randomly shuffled. All reported performance metrics are derived from held-out test data. Models were trained for 100 epochs using the Nadam optimizer, a learning rate of 0.002 with decay [27], and a batch size of 32. Batch size was chosen to accommodate VRAM of Titan X GPU. Training data were augmented using random horizontal flipping. Training parameters were consistent across datasets.

2.6. Model Evaluation

Classification performance was evaluated using two metrics: (1) area under curve (AUC) of the receiver operating characteristic (ROC) and (2) F1 score (harmonic mean of precision and recall). A random chance null model (performance averaged over 1000 trials) was included for comparison.

Class activation maps (CAMs) were generated from shallow and deep convolutional layers to identify discriminative image regions [28]. CAMs were constructed using the output

feature map of a given convolutional layer with each feature map channel weighted by the lesion class gradient. Conceptually, CAMs help to interpret model function by visualizing image areas that are driving the model's classification decision.

In addition to slice-by-slice classification performance, models were evaluated on a per-subject basis. For each test subject, the models assigned a classification score to all axial slices. A sliding convolutional filter was used to determine the mean classification score over several adjacent slices (approximately 1.5 cm in the z-axis) (Supplemental Fig. 2).

2.7. Statistical Analysis

ROC curves were compared using DeLong's test, implemented using the pROC R package [29]. Significance of logistic regression parameters was determined by Wald test.

2.8. Code and data availability

All code related to simulated image generation, classifier design, and statistical analysis can be found at: https://github.com/penn-cnt/Arnold_simulated_clinical_trial. We are grateful to the researchers that published the well-curated, publicly available datasets used in this study. As of publication, these data can be found at MS-SEG 2008 [12]: http://www.ia.unc.edu/MSseg/, BraTS 2019 [13]: https://www.med.upenn.edu/cbica/brats2019.html, MS-SEG 2016 [14]: https://portal.fli-iam.irisa.fr/english-msseg/, ISLES 2015 [15]: http://www.isles-challenge.org/ISLES2015/, OASIS3 [19]: https://www.oasis-brains.org/.

3. Results

3.1. Image Transformation Validation

The image transformation method was validated on data from three additional subjects not included during the transformation fitting step. For each subject, a standard 3T FLAIR image was transformed into a simulated ULF image for comparison against the authentic ULF ground truth. Representative images from each subject are shown in **Fig. 2**.

Perceived diagnostic image quality was quantitatively assessed using the entropy of the MR image gradient. Lower gradient entropy indicates sharper edges and correlates strongly with higher perceived image quality on MRI [20]. Gradient entropy of HF images (mean ± standard deviation: 5.91±0.54) was significantly lower than both the real ULF images (7.89± 0.90) and simulated ULF (7.42±0.73) images ($p<0.0001$, t-test). While there was a statistical difference between the gradient entropy of real and simulated ULF images ($p<0.05$), the effect size was dramatically reduced compared to the original HF images (0.47 vs 1.98). While perceived quality was modestly higher in simulated ULF images, there was substantial overlap of gradient entropy with real ULF images, indicating similar image quality between simulated and real images.

3.2. Comparing pathology detection in standard and simulated ultra-low-field images

Deep learning models were trained to perform binary classification of images in each disease cohort using either HF or simulated ULF MRI. ROC curves for each cohort are shown in **Fig. 3**. Despite significant image degradation, per-slice classifier performance was similar for HF and simulated ULF MRI across all pathologies (**Table 2**). As expected, accuracy on subtle pathology (MS lesions) was lower than more prominent pathology for both HF and simulated ULF MRI datasets. The performance achieved is comparable to previous benchmarks using deep learning for detection of brain masses [30,31], and MS lesions [32], and significantly exceeded null models in all cohorts tested.

3.3. Characterizing ultra-low-field pathology detection

To broadly characterize pathology detection capabilities in simulated ULF images, detection sensitivity was aggregated across all pathologies and modeled using a logistic regression as a function of lesion size and intensity as shown in **Fig. 4A & 4B**. For both HF and simulated ULF images, sensitivity was more strongly associated with lesion size, though both parameters reached statistical significance (standard: z_{size}=13.5, p_{size}<2e-16, $z_{intensity}$=9.2, $p_{intensity}$<2e-16; simulated ultra-low-field: z_{size}=8.5, p_{size}<2e-16, $z_{intensity}$=3.9, $p_{intensity}$<9e-5). HF imaging outperformed simulated ULF imaging for detection of smaller or less intense lesions as shown in **Fig. 4C**. While performance did not vary significantly between HF and simulated ULF images across the cohorts as a whole, sensitivity differences in this subgroup analysis suggests a performance drop-off when using ULF imaging for 1-4 cm^2 lesions (**Fig. 4A & 4B**).) These findings are agnostic to pathology type and may serve as generalizable performance guidelines for FLAIR at 64mT in yet-untested patient populations.

3.4. Patient-level classification

In addition to per-slice performance, we assessed pathology detection on a per-subject basis. Algorithms were evaluated over 17 held-out subjects (three MS, four stroke, five HGG, five LGG) and achieved 100% sensitivity in both HF and simulated ULF images. The LGG classifier was also evaluated using five control subjects, and correctly identified all subjects as non-lesional (100% specificity in both HF and simulated ULF images) as shown in **Fig. 5A & 5B**.

3.5. Class activation mapping

We used class activation mapping to probe which image regions were driving algorithm decisions as shown in **Fig. 5C**. As expected, areas containing pathology are the primary drivers of classification at both shallow and deep layers. The deep CAM for the MS model also reveals that this model attends to periventricular areas known to be clinically important for lesion identification. These findings are reassuring that the tested models are detecting pathology of interest as expected and may have empirically captured features of the typical disease distribution. It is important to note that CAMs serve as an approximate representation of model attention, and each convolutional model layer has a unique CAM. While interpretation of CAMs alone is difficult due to the nonlinear nature of neural networks, the CAMs and the subject level visualizations provide convergent evidence that our models are attending to pathological features.

3.6. Determining the effect of lesion intensity on detection

A potential limitation of the simulated image generation method is that it does not account for possible changes in ULF lesion to background tissue contrast relationships due to differences in relaxation times or pulse sequences. Here, we quantify detection robustness by measuring performance over a range of lesion contrasts (outlined in *2.3. Modulating lesion contrast*) for HGG images. This approach also allows us to assess the relative importance of lesion to background tissue contrast in comparison to structural distortion from large brain tumors.

ROC curves for detection of HGGs are shown in **Fig. 6**. Compared to detection of full-intensity tumors (AUC=0.972), there was a statistically significant decrease in performance for

tumors with relative contrast of 60% or less (AUC_{80}=0.972, p=NS; AUC_{60}=0.943, p=0.01; AUC_{40}=0.905, p=2.0e-5; AUC_{20}=0.901, p=2.2e-6; AUC_{0}=0.874, p=1.7e-8). While contrast has a substantial impact on lesion detection, an AUC of 0.874 and F1 of 0.71 is achieved even for isointense lesions, indicating that in this dataset even with reduced lesion contrast large pathology could be identified due to structural deformation.

4. Discussion

In this study we propose a generalizable methodology for evaluating novel imaging modalities, applied here to ULF MRI. By leveraging open-access datasets, this virtual trial paradigm permits rapid, low-cost assessment of a device's potential diagnostic capabilities across a range of pathologies. We assert that this approach can help to address challenges in medical device development, regulatory approval, and clinical trial design.

ULF MRI offers an exciting opportunity for improving imaging accessibility in low-resource environments and enables point-of-care MRI. These scanners have relatively low manufacturing and operating costs and may help stem the increasing contribution of medical imaging to healthcare expenditures [33]. To accelerate device development cycles and reduce the cost of bringing devices to market, it is pivotal that tools are developed to allow rapid prototyping, efficient regulatory approval, and expedited deployment. Virtual clinical trials can efficiently evaluate medical imaging technology by simulating patients, imaging systems, and image interpreters [34]. For example, in breast tomography, Barufaldi et al. developed a virtual breast phantom and analytical pipeline that can simulate a clinical trial for several hundred patients per day [35]. While not a replacement for traditional prospective studies, simulated clinical trials may offer significant value as the FDA considers devices prior to extensive

prospective data collection. Simulated trials could contribute to early feasibility studies (EFS) or provide supporting evidence for investigational device exemption (IDE) approval. Additionally, simulated trials could identify key patient populations or indications to prioritize for standard clinical trials. When considering a particular disease process, a simulated trial approach could help to establish benchmarks that a proposed device (such as a scanner or pulse sequence) must meet to provide sufficient diagnostic performance.

Based on simulated ULF images, our study suggests ULF MRI scanners should detect many brain lesions with comparable performance to standard MR imaging. However, simulated ULF imaging was sensitive to lesion size. Accuracy was lower for 1-4 cm^2 lesions (approximately 2.25 cm diameter and below) as shown in **Fig. 4A**. These findings indicate that ULF MRI may perform adequately for identifying macro-scale pathology (most gliomas, medium-large vessel stroke, etc.) or measuring major brain structures, but may be less reliable for more subtle pathologies (MS lesions, embolic infarcts).

Importantly, we expect ULF MRI to complement and not replace standard clinical MRI. For this reason and as a proof of concept, our analysis is limited to basic diagnostic capabilities (presence vs absence of pathology), defining a range of expected size and signal intensity thresholds, rather than more complex image interpretation such as precise lesion segmentation or tracking lesion evolution over time. While we compare performance of ULF MRI to 3T MR devices, the practical alternative in certain use cases (ICU patients, underserved communities, in-office disease tracking) would be portable CT scans or no imaging at all. In these settings, ULF MRI may have advantages over CT such as increased tissue contrast and lack of ionizing radiation exposure.

This work underscores the power of open-access clinical databases to facilitate translational research. Platforms for data sharing, such as XNAT Central [17], iEEG Portal [36], and crowdsourced competitions [37] have led to rapid advances in many fields. While public databases provide diverse repositories of patient data with sufficient sample size to train deep learning algorithms, most medical imaging data remains federated across institutions [38]. Further data-sharing efforts designed explicitly for evaluating devices and software for regulatory approval could reduce the cost and time necessary to bring innovative imaging technology to the clinic. Recently radiology has shifted toward centralizing algorithms while maintaining individual data ownership [39]. While this approach may facilitate algorithm validation across research groups, it precludes creative use of multi-institutional data for applications beyond algorithm testing.

The simulated trial paradigm presented here is meant to serve as a framework for applying pre-existing datasets and deep learning to explore the expected performance of novel diagnostic devices. However, the approach brings with it several important limitations and methodological considerations. Most importantly, the utility of simulated data is directly linked to the transformation method quality. Here, we implemented a relatively simple histogram matching based algorithm. While the present method approximates tissue intensity and resolution in ULF images, it does not account for other potential field strength, pulse sequence and device-specific artifacts that may affect lesion conspicuity and image quality. More advanced transformation methods, including generative adversarial networks (GANs), synthetic MRI, and other quantitative methods [40–43] could certainly improve simulation quality. Future investigations should incorporate more advanced transformation methods and image quality validation by expert radiologist review.

However, transformation algorithms that learn by example, such as GANs, require large amounts of data. Methods that simulate images with a low N can be advantageous in certain situations [42]. Specifically, low-data requirement methods can be useful for evaluating new devices or during the prototyping process, where available data are scarce. While data-driven methods may produce closer matched simulations in the long run, these methods may not be applicable for all applications. Importantly, the simulated trial approach is not meant to serve as a replacement for prospective clinical trials. While more advanced methods such as GANs may improve image-to-image simulation quality, there will still likely be a gap between real and simulated images, especially in patients with pathology [43]. Simulations can provide useful guidance, including expected outcomes for prospective trials, however these are retrospective analyses and cannot provide the same level of scientific evidence as clinical trials.

Another consideration is the specific deep learning algorithm implemented. The purpose of the algorithm in this study was to provide a standardized baseline for analysis and comparison, and not necessarily to optimize detection for any particular use case. Approaches such as data augmentation or incorporation of data from multiple adjacent slices could likely provide incremental improvements in performance. Additionally, this work is limited by the sequences and pathologies that are publicly available. Results are likely sequence dependent and patients included in these datasets may not be reflective of the disease more broadly. Publicly available patient data is often collected at tertiary care centers and may be more likely to contain advanced pathology or larger lesions.

5. Conclusion

In this study, we have proposed a method for evaluating imaging devices via simulated trials, incorporating domain transfer and automated pathology detection, and demonstrated its application to a new ULF MRI device. This method allows for rapid evaluation of actual or proposed diagnostic devices. Importantly, simulated trials can provide guidance and justification for prospective studies. In our simulations, we found that gliomas, strokes and multiple sclerosis could be detected in ultra-low-field quality images. These findings justify prospective studies evaluating these pathologies on the device. This work additionally highlights the importance of centralized data sharing for device design and validation.

Acknowledgment

We thank Jonathan Rothberg, PhD, and the team at Hyperfine Research (Guilford, GT), particularly Samantha By, PhD and Brian Welch, PhD, for the use of the ULF MRI scanner. We also thank the Penn Neuroradiology Research Core for assisting with patient recruitment and scanning. This work was supported by a sponsored research agreement with Hyperfine. Additional funding was provided by the NIH (T32NS091006-01), the HHMI-NIBIB Interfaces Initiative (5T32EB009384-10), Jonathan and Bonnie Rothberg, The Mirowski Family Fund, and Neil and Barbara Smit.

Disclosures of Conflicts of Interest

Joel Stein is the Principal Investigator on a Sponsored Research Agreement to test the Hyperfine ULF MRI device in an ambulatory setting for imaging hydrocephalus. Dr. Stein does

not personally receive any compensation from this company. Brian Litt is an unpaid Scientific Advisor to 4Catalyzer, a Jonathan Rothberg founded incubator, which initiated Hyperfine as one of its companies. Dr. Litt is a Co-Founder of EpilepsyCo, a separate Jonathan Rothberg founded company. His time consulting for EpilepsyCo is compensated under a consulting agreement. Both Dr. Stein's and Dr. Litt's interactions with these companies, and those of their trainees, are performed in strict accordance with the policies and conflict of interest management rules of the University of Pennsylvania and are reviewed annually.

CRediT Author Contributions

T. Campbell Arnold: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Writing – Review & Editing**. Steven N. Baldassano**: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Writing – Review & Editing**. Brian Litt**: Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition. **Joel M. Stein**: Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition.

References

[1] Ogbole GI, Adeyomoye AO, Badu-Peprah A, Mensah Y, Nzeh DA. Survey of magnetic resonance imaging availability in West Africa. Pan Afr Med J 2018;30. https://doi.org/10.11604/pamj.2018.30.240.14000.

[2] Marques JP, Simonis FFJ, Webb AG. Low□field MRI: An MR physics perspective. J Magn Reson Imaging 2019;49:1528–42. https://doi.org/10.1002/jmri.26637.

[3] Mollura DJ, Shah N, Mazal J. White paper report of the 2013 RAD-AID conference: I. J.

Am. Coll. Radiol., vol. 11, 2014, p. 913–9. https://doi.org/10.1016/j.jacr.2014.03.026.

[4] Mollura D, Lungren MP. Radiology in global health. New York, NY: Springer; 2014.

[5] Maru DS-R, Schwarz R, Andrews J, Basu S, Sharma A, Moore C. Turning a blind eye: the mobilization of radiology services in resource-poor regions. Global Health 2010;6:18. https://doi.org/10.1186/1744-8603-6-18.

[6] Campbell-Washburn AE, Ramasawmy R, Restivo MC, Bhattacharya I, Basar B, Herzka DA, et al. Opportunities in Interventional and Diagnostic Imaging by Using High-Performance Low-Field-Strength MRI. Radiology 2019;293:384–93. https://doi.org/10.1148/radiol.2019190452.

[7] Sheth KN, Mazurek MH, Yuen MM, Cahn BA, Shah JT, Ward A, et al. Assessment of brain injury using portable, low-field magnetic resonance imaging at the bedside of critically ill patients. JAMA Neurol 2020:E1–7. https://doi.org/10.1001/jamaneurol.2020.3263.

[8] Turpin J, Unadkat P, Thomas J, Kleiner N, Khazanehdari S, Wanchoo S, et al. Portable Magnetic Resonance Imaging for ICU Patients. Crit Care Explor 2020;2:e0306. https://doi.org/10.1097/cce.0000000000000306.

[9] Heiss R, Grodzki DM, Horger W, Uder M, Nagel AM, Bickelhaupt S. High-performance low field MRI enables visualization of persistent pulmonary damage after COVID-19. Magn Reson Imaging 2021;76:49–51. https://doi.org/10.1016/j.mri.2020.11.004.

[10] Deoni SCL, Bruchhage MMK, Beauchemin J, Volpe A, D'Sa V, Huentelman M, et al. Accessible pediatric neuroimaging using a low field strength MRI scanner. Neuroimage 2021;238:118273. https://doi.org/10.1016/j.neuroimage.2021.118273.

[11] Shen FX, Wolf SM, Bhavnani S, Deoni S, Elison JT, Fair D, et al. Emerging Ethical

Issues Raised by Highly Portable MRI Research in Remote and Resource-Limited International Settings. Neuroimage 2021:118210. https://doi.org/10.1016/j.neuroimage.2021.118210.

[12] Styner M., Lee J., Chin B., Chin M.S., Commowick O., Tran H., Markovic-Plese S., Jewells V. WS. MIDAS Journal - 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. Insight J 2008.

[13] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge 2018.

[14] Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-31911-7.

[15] Maier O, Menze BH, von der Gablentz J, Häni L, Heinrich MP, Liebrand M, et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med Image Anal 2017;35:250–69. https://doi.org/10.1016/j.media.2016.07.009.

[16] Allen B, Seltzer SE, Langlotz CP, Dreyer KP, Summers RM, Petrick N, et al. A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. J Am Coll Radiol 2019;16:1179–89. https://doi.org/10.1016/j.jacr.2019.04.014.

[17] Herrick R, Horton W, Olsen T, McKay M, Archie KA, Marcus DS. XNAT Central: Open sourcing imaging research data. Neuroimage 2016;124:1093–6. https://doi.org/10.1016/j.neuroimage.2015.06.076.

[18] Prevedello LM, Halabi SS, Shih G, Wu CC, Kohli MD, Chokshi FH, et al. Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. Radiol Artif Intell 2019;1:e180031. https://doi.org/10.1148/ryai.2019180031.

[19] LaMontagne PJ, Benzinger TLS, Morris JC, Keefe S, Hornbeck R, Xiong C, et al. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. MedRxiv 2019:2019.12.13.19014902. https://doi.org/10.1101/2019.12.13.19014902.

[20] McGee KP, Manduca A, Felmlee JP, Riederer SJ, Ehman RL. Image metric-based correction (autocorrection) of motion effects: analysis of image metrics. J Magn Reson Imaging 2000;11:174–81. https://doi.org/10.1002/(sici)1522-2586(200002)11:2<174::aid-jmri15>3.0.co;2-3.

[21] Chollet F. Keras. 2015.

[22] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. USENIX Symp Oper Syst Des Implement 2016:265–83.

[23] GoogleResearch. TensorFlow: Large-scale machine learning on heterogeneous systems. Google Res 2015.

[24] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks 2016.

[25] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning 2017.

[26] Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database, Institute of Electrical and Electronics Engineers (IEEE);

2010, p. 248–55. https://doi.org/10.1109/cvpr.2009.5206848.

[27] Dozat T. Incorporating Nesterov Momentum into Adam. ICLR Work 2016;1:2013–6.

[28] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016- Decem, IEEE Computer Society; 2016, p. 2921–9. https://doi.org/10.1109/CVPR.2016.319.

[29] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.

[30] Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging 2020;51:175–82. https://doi.org/10.1002/jmri.26766.

[31] Tandel GS, Biswas M, Kakde OG, Tiwari A, Suri HS, Turk M, et al. A review on a deep learning perspective in brain cancer classification. Cancers (Basel) 2019;11. https://doi.org/10.3390/cancers11010111.

[32] Yoo Y, Tang LYW, Brosch T, Li DKB, Kolind S, Vavasour I, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. NeuroImage Clin 2018;17:169–78. https://doi.org/10.1016/j.nicl.2017.10.015.

[33] Smith-Bindman R, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. Health Aff 2008;27:1491–502. https://doi.org/10.1377/hlthaff.27.6.1491.

[34] Abadi E, Segars WP, Tsui BMW, Kinahan PE, Bottenus N, Frangi AF, et al. Virtual

clinical trials in medical imaging: a review. J Med Imaging 2020;7:1. https://doi.org/10.1117/1.jmi.7.4.042805.

[35] Barufaldi B, Bakic PR, Higginbotham D, Maidment ADA. OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis. In: Chen G-H, Lo JY, Gilat Schmidt T, editors. Med. Imaging 2018 Phys. Med. Imaging, vol. 10573, SPIE; 2018, p. 194. https://doi.org/10.1117/12.2294935.

[36] Wagenaar JB, Brinkmann BH, Ives Z, Worrell GA, Litt B. A multimodal platform for cloud-based collaborative research. Int. IEEE/EMBS Conf. Neural Eng. NER, 2013, p. 1386–9. https://doi.org/10.1109/NER.2013.6696201.

[37] Baldassano SN, Brinkmann BH, Ung H, Blevins T, Conrad EC, Leyde K, et al. Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings. Brain 2017;140:1680–91. https://doi.org/10.1093/brain/awx098.

[38] Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging 2017;30:392–9. https://doi.org/10.1007/s10278-017-9976-3.

[39] Greenspan H, Van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Trans Med Imaging 2016;35:1153–9. https://doi.org/10.1109/TMI.2016.2553401.

[40] Taylor AJ, Salerno M, Dharmakumar R, Jerosch-Herold M. T1 Mapping Basic Techniques and Clinical Applications. JACC Cardiovasc Imaging 2016;9:67–81. https://doi.org/10.1016/j.jcmg.2015.11.005.

[41] Blystad I, Warntjes JBM, Smedby O, Landtblom A-M, Lundberg P, Larsson E-M.

Synthetic MRI of the brain in a clinical setting. Acta Radiol 2012;53:1158–63. https://doi.org/10.1258/ar.2012.120195.

[42] Wu Z, Chen W, Nayak KS. Minimum Field Strength Simulator for Proton Density Weighted MRI. PLoS One 2016;11:e0154711. https://doi.org/10.1371/journal.pone.0154711.

[43] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019;58:101552. https://doi.org/10.1016/J.MEDIA.2019.101552.

Tables

Dataset	BraTS 2019	ISLES 2015	MICCAI 2008	MS-SEG-2016
Pathology	Glioma (HGG & LGG)	Stroke	Multiple Sclerosis (MS)	Multiple Sclerosis (MS)
Subjects	335 (259 HGG, 76 LGG)	28	10	10
Center/Scanners	19	2	2	3
Classes	3 (enhanced, non-enhanced, edema)	1 (infarct)	1 (MS lesion)	1 (MS lesion)

Table 1. List of open-access neuropathology datasets used in this study. Abbreviations: High Grade Glioma (HGG), Low Grade Glioma (LGG), Multiple Sclerosis (MS).

Pathology	Standard AUC	Low-Field AUC	p_{AUC}	Standard F1	Low-Field F1	Null Model F1
HGG	0.972	0.978	0.16	0.920	0.896	0.481±0.01
LGG	0.957	0.949	0.61	0.885	0.880	0.468±0.02
Stroke	0.936	0.940	0.83	0.772	0.761	0.485±0.02
MS	0.896	0.873	0.49	0.766	0.745	0.442±0.02

Table 2. Performance metrics for each pathology type using standard or simulated low-field images. Abbreviations: High Grade Glioma (HGG), Low Grade Glioma (LGG), Multiple Sclerosis (MS), Area Under the Curve (AUC).

Figures

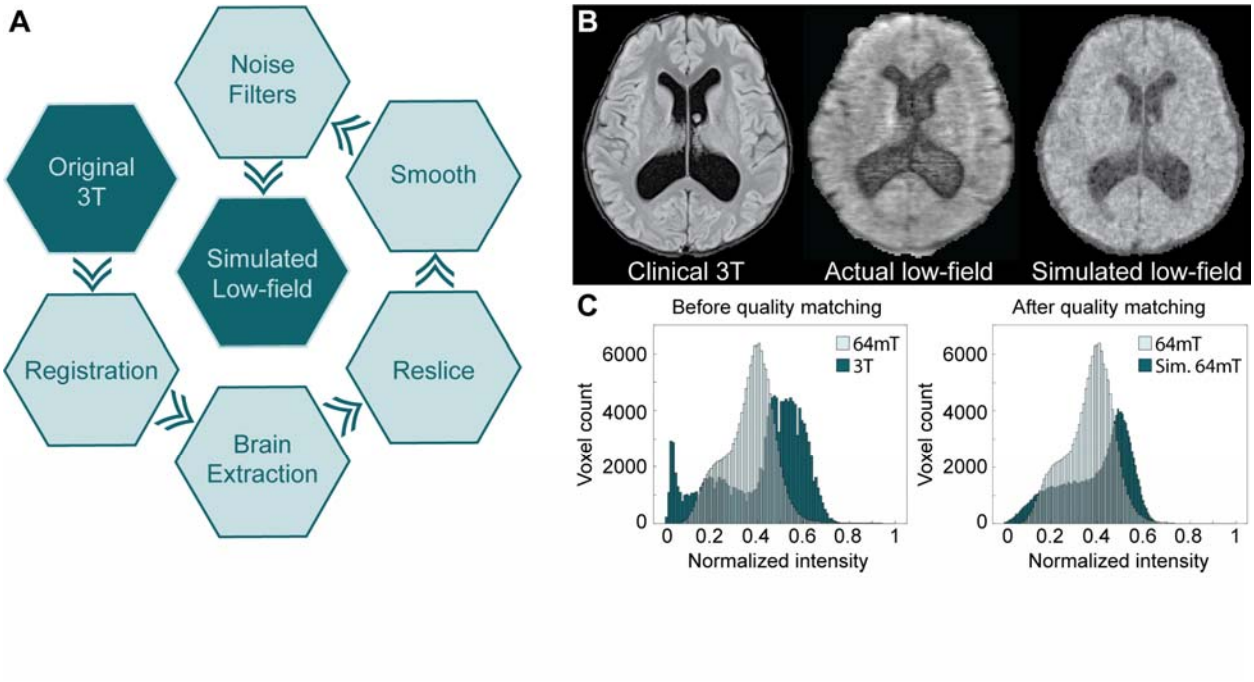

Fig. 1. Generating simulated ultra-low-field (ULF) images. **(A)** Steps in the image processing pipeline from the original 3T image (upper-left) to the simulated ULF version (center). **(B)** Example skull-stripped axial FLAR images from a clinical 3T (left) and a 64mT ULF MRI scanner (center). The 3T image was passed through the image transformation pipeline to produce the simulated ULF image (right). **(C)** To generate simulated ULF images resembling actual ULF images, histogram features (mean, standard deviation, and skewness) were used to guide image transformation. The intensity histogram distributions relative to actual 64 mT images are shown before (left) and after (right) transformation.

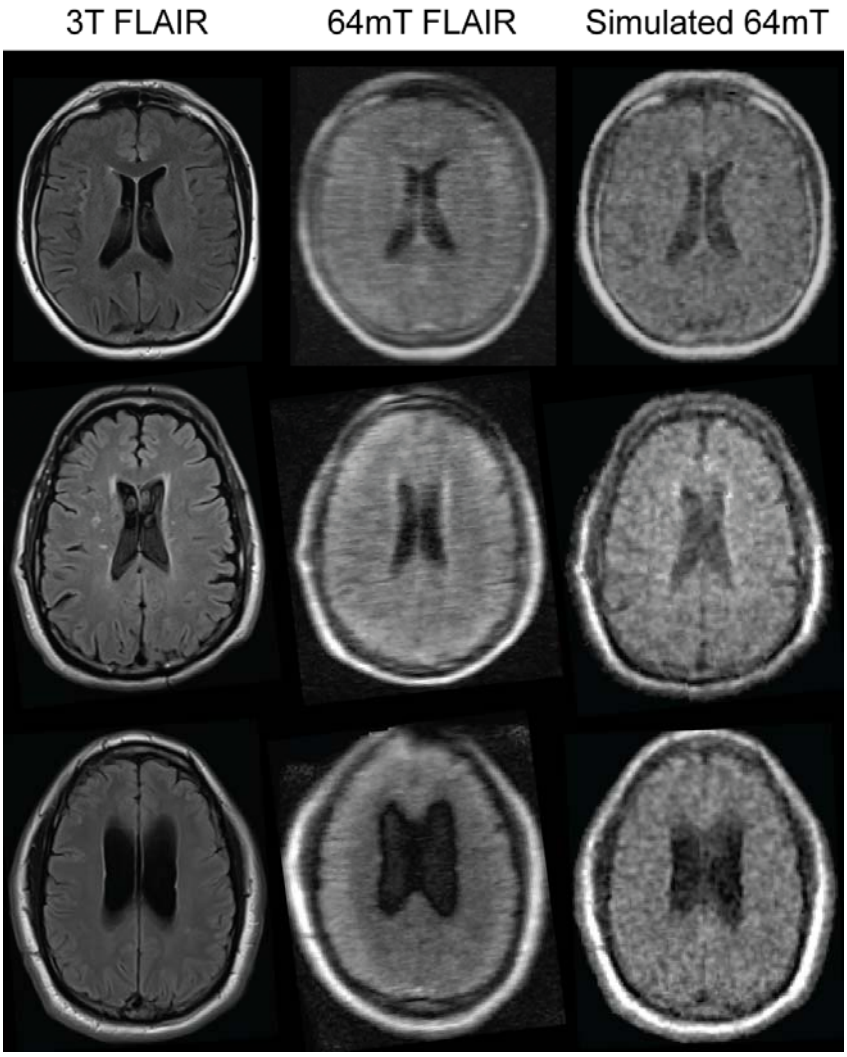

Fig. 2. Validation of image transformation method. 3T images (column 1), simulated 64mT images (column 2), and actual 64mT images (column 3) are shown for three additional subjects. These subjects were not included in the training set used to develop the automated transformation.

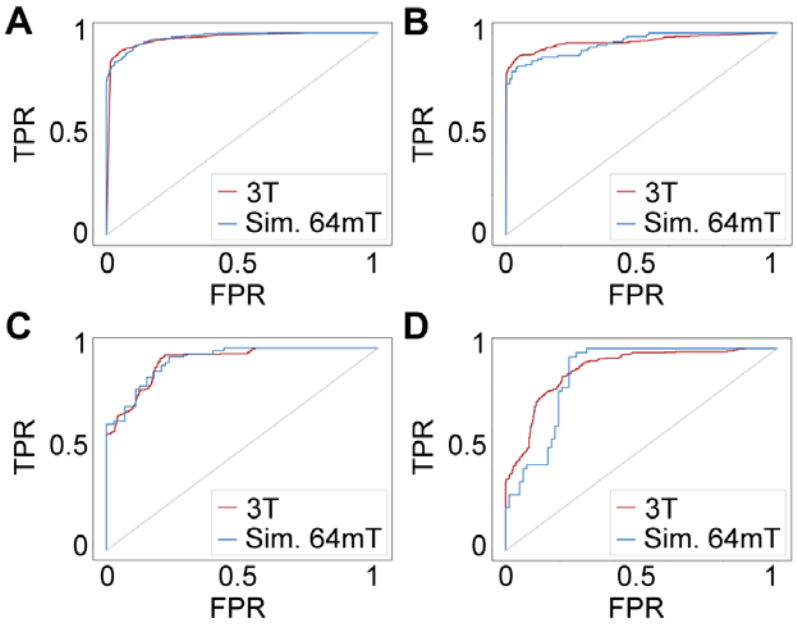

Fig. 3. Pathology detection performance. Receiver operating characteristic (ROC) curves shown for binary classification of images with and without pathology for high-grade glioma **(A)**, low-grade glioma **(B)**, ischemic stroke **(C)**, and multiple sclerosis **(D)**. Curves are shown for standard (3T) and simulated low-field images. Abbreviations: True positive rate (TPR), False positive rate (FPR), Simulated (Sim.).

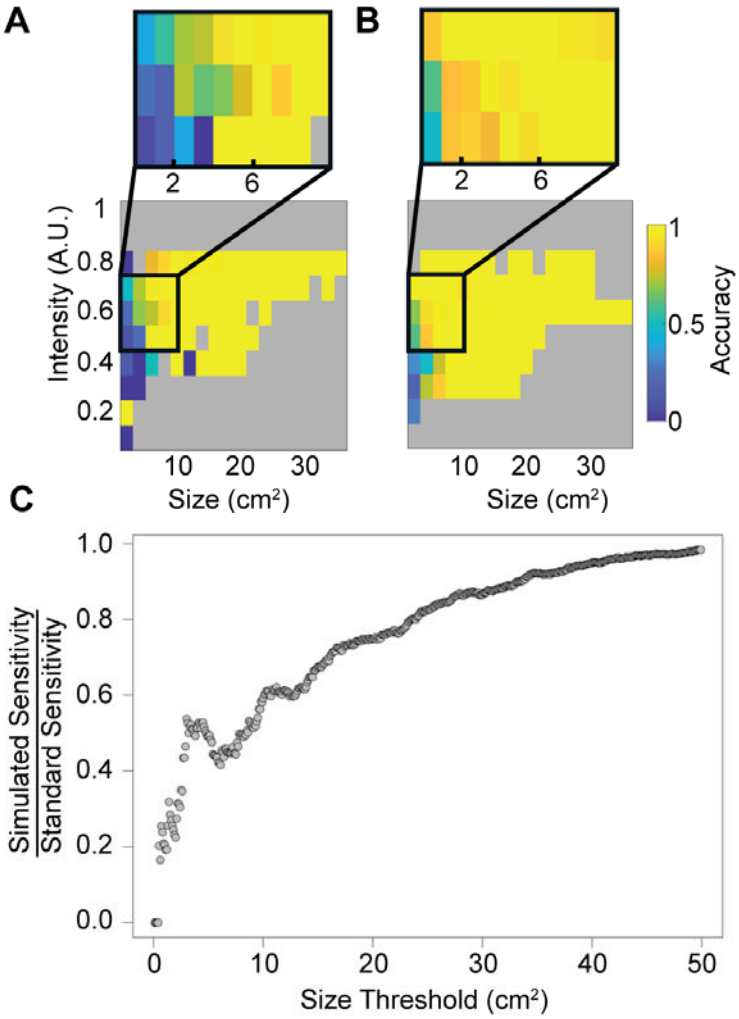

Fig. 4. Detection sensitivity as a function of lesion size and scaled intensity. Sensitivity of the deep learning classifiers for detecting lesions in the validation set is shown in the **(A)** simulated ULF and **(B)** standard HF images. Areas highlighting discrepancies between the datasets are highlighted in image insets. **(C)** Sensitivity of lesion detection in simulated ULF images relative to HF images. Each point represents the sensitivity ratio measured on all lesions smaller than the given size threshold. Note that sensitivity is similar between image types when averaged over all lesions but differs significantly when restricted to smaller lesions.

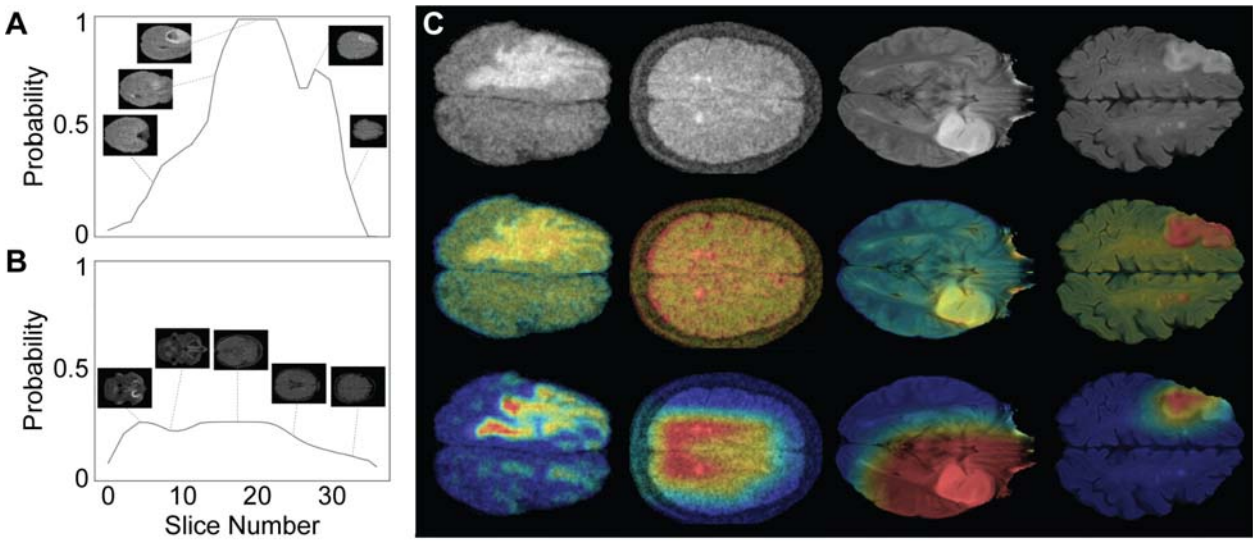

Fig. 5. Model validation and interpretability. Panel A and B provide examples of per-subject pathology detection. Convolutional filters were used to generate average lesion probability values across several adjacent axial slices. A threshold value for subject-level classification was determined empirically by maximizing per-subject classification accuracy in the training set. Sample plots are shown for a subject with HGG **(A)** and a control patient **(B)** using simulated low-field imaging. **(C)** Class activation mapping. Row 1: Sample images of high-grade glioma (low-field), multiple sclerosis (low-field), low-grade glioma (standard), and ischemic stroke (standard). Row 2: CAMs generated from shallow network layer for each pathology. Row 3: CAMs generated from deepest convolutional network layer for each pathology.

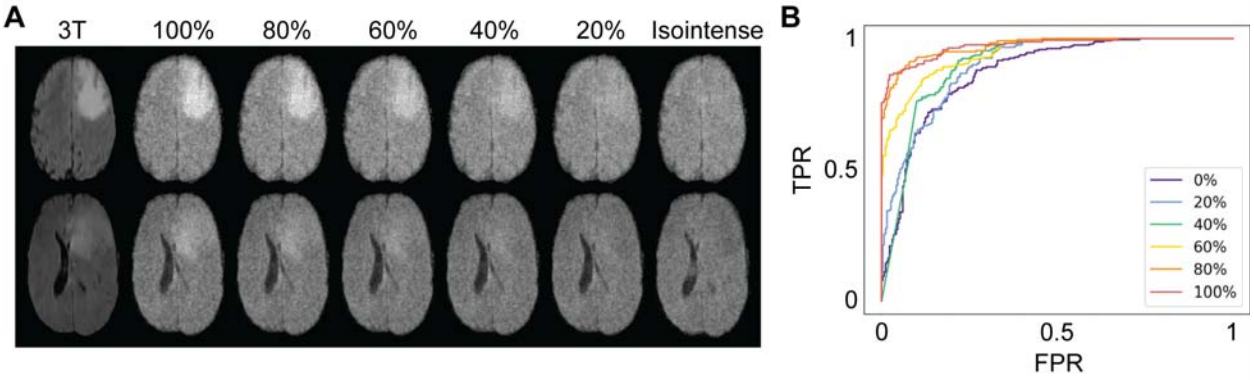

Fig. 6. Intensity modulation to explore effects of intensity contrast. **(A)** Intensity values of the pathology segmentation were modulated over a range from normal intensity (100%) to isointense (tumor = background). **(B)** For each image subset, a classifier was trained to distinguish pathological and normal slices. AUC varied directly with lesion contrast but remained significantly better than chance even in the isointense cohort, which likely reflects structural deformations (such as ventricular effacement and midline shift as in the bottom row of panel A) or residual signal intensity heterogeneity.