



OPEN

DATA DESCRIPTOR

fastMRI+, Clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data

Ruiyang Zhao^{1,2,3}, Burhaneddin Yaman^{1,4}, Yuxin Zhang^{1,2,3}, Russell Stewart^{1,5}, Austin Dixon^{1,6}, Florian Knoll⁷, Zhengnan Huang⁷, Yvonne W. Lui⁷, Michael S. Hansen¹✉ & Matthew P. Lungren^{1,5}

Improving speed and image quality of Magnetic Resonance Imaging (MRI) using deep learning reconstruction is an active area of research. The fastMRI dataset contains large volumes of raw MRI data, which has enabled significant advances in this field. While the impact of the fastMRI dataset is unquestioned, the dataset currently lacks clinical expert pathology annotations, critical to addressing clinically relevant reconstruction frameworks and exploring important questions regarding rendering of specific pathology using such novel approaches. This work introduces fastMRI+, which consists of 16154 subspecialist expert bounding box annotations and 13 study-level labels for 22 different pathology categories on the fastMRI knee dataset, and 7570 subspecialist expert bounding box annotations and 643 study-level labels for 30 different pathology categories for the fastMRI brain dataset. The fastMRI+ dataset is open access and aims to support further research and advancement of medical imaging in MRI reconstruction and beyond.

Background & Summary

Magnetic resonance imaging (MRI) is a widely utilized medical imaging modality critically important for a broad range of clinical diagnostic tasks including stroke, cancer, surgical planning, acute injuries, and more. Machine learning (ML) techniques have demonstrated opportunities to improve the MRI diagnostic workflow particularly in the image reconstruction task by saving time, reducing contrast, and leading in cases to FDA-cleared solutions¹⁻⁴. Among the myriad applications of machine learning in medical imaging being explored, deep learning-based MRI reconstruction is showing considerable promise and is moving towards clinical impact.

ML-based MRI reconstruction approaches often require data from “raw” fully sampled k-space datasets in order to generate ground truth images. Public MRI datasets like Calgary-Campinas Public Dataset⁵, MRNet⁶, OAI⁷, SKM-TEA⁸, and mridata.org are available to empower ML-related research. Also, various datasets can be found in multiple medical image research challenges, including MC-MRREC and RealNoiseMRI. Most of these datasets only provided reconstructed MRI images (Note SKM-TEA dataset also provides knee tissue label and pathology detection information) or limited amount of raw data. Thus, large datasets of raw MRI measurements are generally not widely available. To address this need and facilitate cross-disciplinary research in accelerated MRI reconstruction using artificial intelligence, the fastMRI initiative was developed. fastMRI is a collaborative project between Facebook AI Research (FAIR), New York University (NYU) Grossman School of Medicine, and NYU Langone Health which includes the wide release of raw MRI data and image datasets⁹. While the fastMRI data has enabled exploration of ML-driven accelerated MRI reconstruction^{10,11}, there is a lack of clinical pathology information to accompany the imaging data which has limited the reconstruction assessment

¹Microsoft Research, Redmond, USA. ²University of Wisconsin-Madison, Department of Radiology, Madison, USA. ³University of Wisconsin-Madison, Department of Medical Physics, Madison, USA. ⁴University of Minnesota, Department of Electrical and Computer Engineering, Minneapolis, USA. ⁵Stanford University, School of Medicine, Stanford, USA. ⁶Duke University, School of Medicine, Durham, USA. ⁷New York University, School of Medicine, New York, USA. ✉e-mail: Michael.Hansen@microsoft.com

Label	Annotation Count	Subject Count
Meniscus		
Meniscus Tear	5658	663
Displaced Meniscal Tissue	232	56
Bones and Cartilage		
Bone-Subchondral Edema	986	196
Bone Lesion	183	29
Bone-Fracture/Contusion/Dislocation	1060	119
Cartilage Full Thickness Loss/Defect	615	122
Cartilage Partial Thickness Loss/Defect	2985	588
Ligaments		
ACL High Grade Sprain	678	101
ACL Low-Mod Grade Sprain	765	153
MCL High Grade Sprain	11	4
MCL Low-Mod Grade Sprain	285	121
PCL High Grade Sprain	18	3
PCL Low-Mod Grade Sprain	142	40
LCL Complex High Grade Sprain	14	3
LCL Complex Low-Mod Grade Sprain	130	48
Other		
Joint Effusion	1311	142
Joint Bodies	38	11
Periarticular Cysts	864	161
Muscle Strain	65	11
Soft Tissue Lesion	90	10
Patellar Retinaculum High Grade Sprain	24	4
Artifact	/	13

Table 1. Knee label summary. *Artifact is study-level label.

Image Level Label	Annotation Count	Subject Count
Absent Septum Pellucidum	3	1
Craniectomy	32	4
Craniotomy	1025	99
Craniotomy with Cranioplasty	43	3
Dural Thickening	351	30
Edema	369	44
Encephalomalacia	161	18
Enlarged Ventricles	300	38
Extra-Axial Mass	104	11
Intraventricular Substance	8	1
Likely Cysts	17	5
Lacunar Infarct	113	32
Mass	380	46
Nonspecific Lesion	757	124
Nonspecific White Matter Lesion	1826	173
Normal Variant	73	21
Paranasal Sinus Opacification	40	8
Pineal Cyst	2	1
Possible Artifact	505	52
Posttreatment Change	1262	99
Resection Cavity	199	27

Table 2. Brain image-level label summary. *Likely Cysts is applied to small lesions (approximately 1 cm or less in diameter) which are difficult to distinguish from parenchymal, simple parenchymal neuronal cyst, and prominent perivascular space.

Study Level Label	Subject Count
Global Ischemia	1
Small Vessel Chronic White Matter Ischemic Change	221
Motion Artifact	33
Possible Demyelinating Disease	2
Colpocephaly	2
White Matter Disease	2
Innumerable Bilateral Focal Brain Lesions	2
Extra-Axial Collection	9
Normal for Age	371

Table 3. Brain study-level label summary.

approaches to validate quantitative metrics such as peak signal-to-noise ratio (pSNR)/structural similarity index measure (SSIM), leaving important questions regarding how various pathologies are represented in ML-based reconstruction unanswered¹². For instance, low sensitivity and stability to clinically relevant features stall their clinical-aware applications¹²⁻¹⁴.

In this paper, we present wide availability of a complementary dataset of annotations, fastMRI+, consisting of human subspecialist expert clinical bounding box labelled pathology annotations for knee and brain MRI scans from the fastMRI multi-coil dataset: specifically encompassing 16154 bounding box annotations and 13 study-level labels for 22 different pathology categories on knee MRIs, as well as 7570 bounding box annotations and 643 study-level labels for 30 different pathology categories on brain MRIs. This new dataset is open and accessible to all for educational and research purposes with the intent to catalyse new avenues of clinically relevant, ML-based reconstruction approaches and evaluation.

Methods

MRI image dataset. The fastMRI dataset is an open-source dataset, which contains raw and DICOM data from MRI acquisitions of knees and brains, described in detail elsewhere⁹. The images used in this study were directly obtained from the fastMRI dataset, reconstructed from fully sampled, multi-coil k-space data (both knee and brain). The fastMRI dataset was managed and anonymized as part of a study approved by the NYU School of Medicine Institutional Review Board. Image reconstruction was performed by inverse Fast Fourier Transform of each individual coil and coil combination with root sum square (RSS) for the purpose of creating pre-annotation images in fastMRI+. The reconstructed images were subsequently converted to DICOM format for human expert reader (radiologist) annotation.

Annotations. Annotation was performed using a commercial browser-based annotation platform (MD.ai, New York, NY) which allowed adjustment of brightness, contrast, and magnification of the images. Readers used personal computers to view and annotate the images using the mentioned annotation platform.

A subspecialist board certified musculoskeletal radiologist with 6 years in practice experience performed annotation for the knee dataset and a subspecialist board certified neuroradiologist with 2 years in practice experience performed annotation for the brain dataset. Annotation was performed with bounding box annotation to include the relevant label for a given pathology on a slice-by-slice level. When more than one pathology was identified in a single image slice, multiple bounding boxes were used.

All 1172 fastMRI knee MRI raw dataset studies were reconstructed and clinically annotated for fastMRI+. Each knee examination consisted of a single series (either proton density (PD) or T2-weighted) of coronal images where bounding box labels were placed on each slice where representative pathology was identified^{15,16}. Effort was made to try to include all the pathology within the bounding box while limiting the normal surrounding anatomy. If the examination contained significant clinically limiting artifacts, then the annotation for “Artifact” was added as a study-level label. In these instances, an interpolation tool was used in which the first and last slice were each labelled and the user interface interpolated the labels on intervening slices. If no relevant pathology was identified on an examination, no labels were provided.

A sub selection of 1001 out of 5847 fastMRI brain MRI raw dataset studies were selected randomly for annotation. Each brain examination included a single axial series (either T2-weighted FLAIR, T1-weighted without contrast, or T1-weighted with contrast) where bounding box labels were placed on each image in which representative pathology or normal anatomical variant was identified^{17,18}. As in knee examinations, effort was made to try to include all the pathology within the bounding box while limiting the normal surrounding anatomy. In some cases, the pathology or normal anatomic variant displayed within a given examination was so extensive or diffuse that a study-level label was used to characterize the relevant images or the entire exam inclusive of the finding (i.e., diffuse white matter disease). The study-level label, in these instances, replaced the use of a bounding box. If no relevant pathology was identified on a given examination, no labels were provided.

Note there are several limitations to this dataset that bear acknowledgement. First, while the annotators are subspecialist radiologists in practice at leading academic medical centers, the lack of multiple annotators/repeated annotations to determine inter-rater/intra-rater reliability metrics or ensure consensus agreement is a limitation and should be considered in the use of these labels. Further work may include multiple annotations

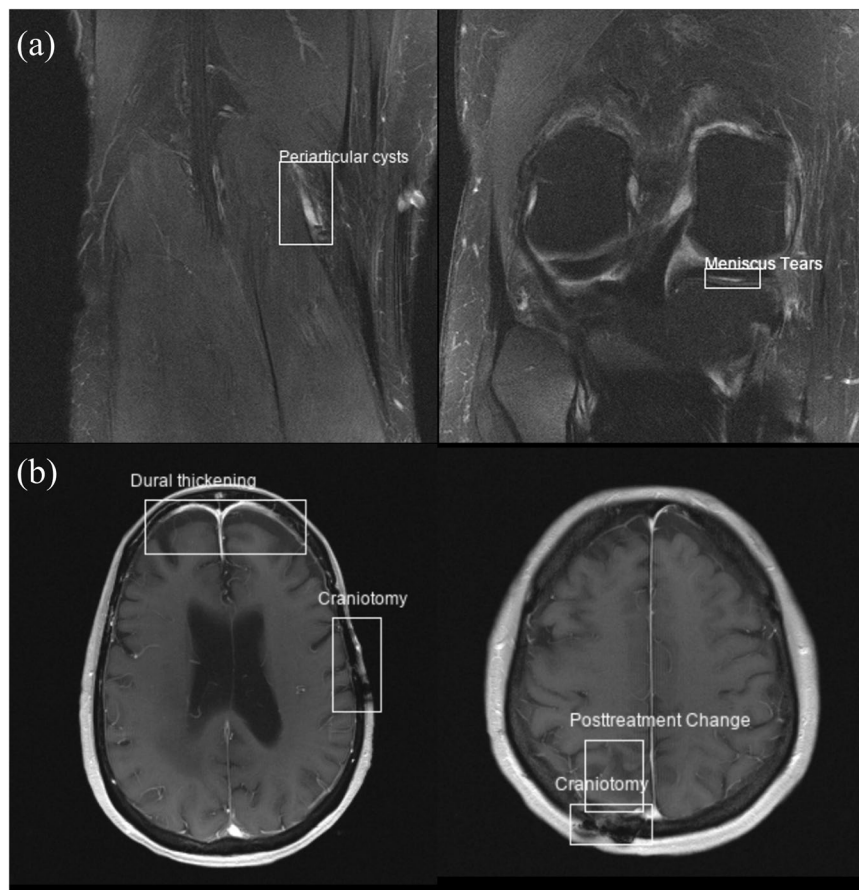


Fig. 1 Example annotations (labels and bounding boxes) from the fastMRI+ dataset shown superimposed on both knee (a) and brain (b) reconstructed images from the fastMRI dataset.

by multiple readers to further refine the clinical labels applied in fastMRI+. Additionally, the fastMRI knee MRI raw dataset contained only coronally acquired series while the brain MRI dataset contained only axially acquired series, each in a variety of pulse sequences and coils. Most knee/brain pathologies that are visible in the non-coronal/non-axial planes are also visible in coronal/axial planes, though not as well seen or as well characterized. For instance, patellofemoral cartilage in the knee and optic neuritis in the brain. While sufficient for annotation, it is important to note that true diagnostic interpretation in MRI for the included pathologies typically demands multi-sequence and multi-planar images for clinically accurate interpretation. What is more, only bounding boxes indicating knee and brain diseases were exported and reported in this work which may limit the research applications of this dataset. Full segmentation of structures would be more laborious and would be a potential subject of future work. Thus, the annotations provided by fastMRI+ may be incomplete. In the future, raw MRI datasets containing fully sampled multi-planar and multi-sequence data would enable optimal clinical annotation.

Statistical analysis. Label distribution analysis was conducted for both knee and brain datasets showing detailed label descriptions at the same time. Table 1 shows annotation count and subject count for corresponding image-level knee labels. Note ‘Artifact’ is a study-level label for the entire study rather than a label of individual images. Table 2 shows annotation count and subject count for corresponding image-level brain labels. Table 3 shows subject count for corresponding subject-level brain labels. Note subject count was provided to show the prevalence of given pathology.

Data Records

We created separate annotation files for the 1172 validation knee datasets and 1001 brain datasets, all based on the fastMRI source data⁹. The annotation files (knee.csv and brain.csv) can be accessed from both fastmri-plus Synapse repository¹⁹ and fastMRI-plus GitHub repository (<https://github.com/microsoft/fastmri-plus>) in CSV formats. Four CSV files are included in the ‘Annotations’ folder. File names of all radiologist-interpreted dataset are stored in knee_file_list.csv and brain_file_list.csv, respectively. Annotations are contained in knee.csv and brain.csv. In each annotation CSV file, the file names (i.e., column ‘Filename’) are aligned with the naming in the fastMRI dataset. For each annotation, file name, slice number, bounding box information, and disease label are provided. The bounding box information includes four parameters, x, y, width (pixel), and height (pixel),

representing the x and y coordinates of the upper-left corner, the width and height of the bounding box. Unit of the bounding box parameters is ‘pixel’. Study-level labels are marked as ‘Yes’ in column ‘Study Level’ for slice 0 of the corresponding subjects with no specified bounding box information.

Technical Validation

A board-certified radiologist with 10 years of experience reviewed the overall quality of the MRI image dataset prior to annotation and clinical evaluation was performed by two additional board-certified subspecialist radiologists. We cleaned and validated raw annotation files following instructions from MD.ai Documentation (<https://docs.md.ai/>). Creation and publication of fastMRI+ code repository followed standard practices with release of open-source software. Specifically, files with annotations and associated tools and scripts were managed source code control, continuous integration tests, and code/data reviews.

Usage Notes

The bounding box information can be used to plot overlaid bounding boxes on images, as shown in Fig. 1. The clinical labels, together with the bounding box coordinates, can also be converted to other formats (e.g., YOLO format²⁰) in order to configure a classification or object detection problem. The open-source repository also contains an example Jupyter Notebook (‘ExampleScripts/example.ipynb’) of how to read the annotations and plot images with bounding boxes in Python.

Code availability

Scripts used to generate the DICOM images for radiologists can be accessed from (‘ExampleScripts/fastmri-to-dicom.py’) in the open-source GitHub repository. The detailed method used has been specified in the Methods section. More open-source tools for reconstructing the original fastMRI dataset, including standardized evaluation criteria, standardized code, and PyTorch data loaders can be found in the fastMRI GitHub repository (<https://github.com/facebookresearch/fastMRI>).

Received: 10 November 2021; Accepted: 11 March 2022;

Published online: 05 April 2022

References

1. Koonjoo, N. *et al.* Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. *Scientific reports* **11**(1), 1–16 (2021).
2. Knoll, F. *et al.* Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues. *IEEE signal processing magazine* **37**(1), 128–140 (2020).
3. Fujita, S. *et al.* Deep learning approach for generating MRA images from 3D quantitative synthetic MRI without additional scans. *Investigative radiology* **55**(4), 249–256 (2020).
4. Wang, T. *et al.* Contrast-enhanced MRI synthesis from non-contrast MRI using attention CycleGAN. *International Society for Optics and Photonics* **11600**, 116001L (2021).
5. Souza, R. *et al.* An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage* **170**, 482–494 (2018).
6. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* **15**, e1002699 (2018).
7. Peterfy, C. G., Schneider, E. & Nevitt, M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage* **16**, 1433–1441 (2008).
8. Desai, A. D. *et al.* Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021).
9. Knoll, F. *et al.* fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning. *Radiology: Artificial Intelligence* **2**(1), e190007 (2020).
10. Hammernik, K. *et al.* Learning a variational network for reconstruction of accelerated MRI data. *Magnetic resonance in medicine* **79**(6), 3055–3071 (2018).
11. Recht, M. P. *et al.* Using Deep Learning to Accelerate Knee MRI at 3 T: Results of an Interchangeability Study. *American Journal of Roentgenology* **215**(6), 1421–1429 (2020).
12. Knoll, F. *et al.* Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. *Magnetic resonance in medicine* **84**(6), 3054–3070 (2020).
13. Antun, V. *et al.* On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences* **117**, 30088–30095 (2020).
14. Darestani, M. Z., Chaudhari, A. S. & Heckel, R. Measuring robustness in deep learning based compressive sensing. *International Conference on Machine Learning*. PMLR. (2021).
15. Quatman, C. E., Hettrich, C. M., Schmitt, L. C. & Spindler, K. P. The clinical utility and diagnostic performance of magnetic resonance imaging for identification of early and advanced knee osteoarthritis: a systematic review. *The American journal of sports medicine* **39**(7), 1557–1568 (2011).
16. Oei, E. H., Nikken, J. J., Verstijnen, A. C., Ginai, A. Z. & Myriam Hunink, M. G. MR imaging of the menisci and cruciate ligaments: a systematic review. *Radiology* **226**(3), 837–848 (2003).
17. Mehan, W. A. Jr *et al.* Optimal brain MRI protocol for new neurological complaint. *PLoS one* **9**(10), e110803 (2014).
18. Dangouloff-Ros, V. *et al.* Incidental brain MRI findings in children: a systematic review and meta-analysis. *American Journal of Neuroradiology* **40**(11), 1818–1823 (2019).
19. Zhao, R. *et al.* fastMRI+, Clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Synapse* <https://doi.org/10.7303/syn26345776> (2021).
20. Jocher, G. *et al.* ultralytics/yolov5: v5.0 – YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. *Zenodo*. <https://doi.org/10.5281/zenodo.4679653> (2021).

Acknowledgements

The authors want to thank Desney Tan at Microsoft Research and Michael Recht at New York University for project support. Sincerest thanks to George Shih and Quan Zhou at MD.ai for providing annotation infrastructure.

Author contributions

R.Z., B.Y. and Y.Z. contributed equally to data processing, data analysis, and manuscript preparation. R.S. and A.D. contributed to data annotation work. F.K., Z.H., Y.L. authorized and facilitated access and usage of raw fastMRI dataset and contributed to manuscript editing. M.S.H. and M.P.L. coordinated and led all details of this project, manuscript composition, and editing.

Competing interests

The authors declare no competing interests.

Additional information

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

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022