Neuroinformatics
https://doi.org/10.1007/s12021-024-09659-5

RESEARCH

An Automated Tool to Classify and Transform Unstructured MRI Data into BIDS Datasets

Alexander Bartnik[1] · Sujal Singh[1] · Conan Sum[1] · Mackenzie Smith[1] · Niels Bergsland[1] · Robert Zivadinov[1] · Michael G. Dwyer[1]

Accepted: 7 March 2024

Abstract
The increasing use of neuroimaging in clinical research has driven the creation of many large imaging datasets. However, these datasets often rely on inconsistent naming conventions in image file headers to describe acquisition, and time-consuming manual curation is necessary. Therefore, we sought to automate the process of classifying and organizing magnetic resonance imaging (MRI) data according to acquisition types common to the clinical routine, as well as automate the transformation of raw, unstructured images into Brain Imaging Data Structure (BIDS) datasets. To do this, we trained an XGBoost model to classify MRI acquisition types using relatively few acquisition parameters that are automatically stored by the MRI scanner in image file metadata, which are then mapped to the naming conventions prescribed by BIDS to transform the input images to the BIDS structure. The model recognizes MRI types with 99.475% accuracy, as well as a micro/macro-averaged precision of 0.9995/0.994, a micro/macro-averaged recall of 0.9995/0.989, and a micro/macro-averaged F1 of 0.9995/0.991. Our approach accurately and quickly classifies MRI types and transforms unstructured data into standardized structures with little-to-no user intervention, reducing the barrier of entry for clinical scientists and increasing the accessibility of existing neuroimaging data.

Keywords Magnetic Resonance Imaging · BIDS · Data Curation · Reproducibility · Machine Learning · Automation

Introduction

Neuroimaging is an important tool in both clinical and research settings for expanding our understanding of the human brain and has seen ever-increasing usage in both settings, contributing to the very large amounts of imaging data available for study. Due to its ease of use and noninvasive nature, magnetic resonance imaging (MRI) is particularly well suited for large-scale data collection and is one of the most commonly collected imaging modalities (Smith-Bindman et al., 2019). This has led to the creation of large-scale publicly available MRI repositories, as well as an untapped pool of MRI collected in the clinical routine, which primarily remains in data silos as of now. Clinical imaging in particular is an intriguing option for imaging studies, as it bypasses the cost and time spent collecting images for new studies.

However, there are several logistical barriers to organizing large amounts of MRI data, including the lack of standard naming conventions describing acquisition types across sites and studies. Annotation of MRI data is further complicated by the lack of standards used to describe the acquisition plane or quality of a scan in addition to its acquisition type. While there have been a number of collaboratory initiatives in recent years to bring together clinical images with other clinical data to study various diseases, such as MSBase (Butzkueven et al., 2006) or the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008), the manual annotation of MRI types throughout these datasets requires a tremendous amount of work in order to organize these images into usable formats (van Ooijen, 2019).

The organization of these large-scale imaging datasets is critical for analysis and sharing of both data as well as results and typically stems from a combination of examining the

✉ Michael G. Dwyer
mgdwyer@buffalo.edu

[1] Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, 77 Goodell St, Buffalo, NY 14203, USA

Published online: 26 March 2024

semantic strings used to describe the data and the reported acquisition parameters used to acquire it. There are several commonly acquired MRI "types" that are used for various purposes both clinically and in the research setting, with these types typically having a continuum of subtypes depending on the scanner manufacturer and parameters used to acquire the scan. Different kinds of analyses require different MRI acquisition types as input. For example, brain tissue segmentation via the SIENAX software (Smith et al., 2002) requires a T1 weighted (T1w) image, while many methods of focal lesion analysis in multiple sclerosis (MS) require images collected via a T2 weighted (T2w)-fluid-attenuated inversion recovery (FLAIR) sequence. Because the type of an MRI scan depends on the parameters used to acquire it, it is often possible to infer the acquisition type of a scan from the acquisition parameters reported by the scanner, though these parameters may vary greatly from one MRI scanner manufacturer to another. Additionally, organization of MRI scans to be used for further analyses is often a manual process and requires considerable time and effort, which is cumbersome at the scale of many large public datasets. However, while MRI is often used in clinical research, many clinical researchers are not trained to recognize these different acquisition types or interpret their acquisition parameters. This renders the process prone to error and may result in inaccuracies in analysis and hamper the ability to share data.

Biomedical imaging has long used the standard established by the Digital Imaging and Communications in Medicine (DICOM) (Mildenberger et al., 2002), which provides a comprehensive means of storing, transmitting, and reporting biomedical imaging data, including MRI. A scan's acquisition parameters are stored in the DICOM file headers, with pertinent parameter values linked to a given tag – each tag is in the format of (group, element). DICOM tags are either standard, which is when the group is an even number or private, which is when the group is an odd number. Standard tags are defined by the DICOM standard (e.g., all MRI DICOMs will report the TE in tag (0018,0081) whereas private tags can be used by the manufacturer for storing additional information as needed. However, inferring the acquisition type from acquisition parameters stored in DICOM headers by hand is cumbersome, so the Protocol Name (0018,1030) or Series Description (0008,103E) tags are often used by MRI technicians to leave free text annotations of the acquisition type in the DICOM header. Since these tags allow for arbitrary strings, their values are entirely at the discretion of the technician or imaging center, with no standardization within or among sites. This makes it increasingly difficult to classify MRI acquisition types using these tags, as there are an innumerous number of permutations available to annotate the series description or protocol name. This lack of consistency in the naming and reporting of acquisition types poses a problem for neuroimaging researchers when working with new images from studies or from public datasets. Moreover, it is often the case that tags containing free-form text (e.g., Protocol Name, Series Description) are stripped before sharing real-world clinical data with other sites since they may potentially have patient-specific information in them. As such, there is a non-negligible risk of inadvertently exposing protected health information (PHI) when sharing clinical DICOMs.

A potential solution to this problem is the adoption of the Brain Imaging Data Structure specification (BIDS), which prescribes standardized methods for the storage and annotation of neuroimaging data, including strict naming conventions for MRI acquisition types (Gorgolewski et al., 2016). However, MRI data are typically derived from scanners in the DICOM file format, while BIDS requires images to be in the Neuroimaging Informatics Technology Initiative (NIfTI) file format that is more common in research settings. Conversion from DICOM to NIfTI is most often done via the dcm2niix software (Li et al., 2016), which also stores the metadata containing acquisition parameters in sidecar JavaScript Object Notation (JSON) files that are then used by BIDS, though the transformation from DICOM to BIDS is often tedious and subject to the same challenges imposed by the lack of standard naming conventions used to describe scans. Despite these challenges, BIDS has seen increasing adoption in the neuroimaging community in recent years, and several large neuroimaging repositories have made use of it, most notably OpenNeuro (Markiewicz et al., 2021). Still, many public datasets either predate BIDS or use their own organization schemes, and images produced by MRI scanners are often entirely unsorted, which severely limits the adoption of the BIDS standard because large amounts of data still need to be transformed into valid BIDS datasets.

Solving this problem would make it easier for clinical researchers and advanced neuroimaging scientists alike to work with new and large MRI datasets by harmonizing them according to an established standard. This would facilitate the sharing and reuse of existing data, as well has ease the use of existing large-scale public datasets for future research. Additionally, this would reduce the barrier of entry for clinical researchers seeking to use neuroimaging data in their studies but who may not have the expertise necessary to classify and sort MRI scans manually.

To solve this problem, we combined thousands of MRI scans from public datasets and in-house studies to train an XGBoost (Chen et al., 2015) model to automatically classify MRI scans according to nine acquisition types common to the clinical routine, using only acquisition parameters found in either DICOM file headers or BIDS sidecars. The model is then used to automate the transformation of unsorted

Table 1 Public datasets (train and test set)

Source	n (unique rows)	MPRAGE (unique rows)	SPGR (unique rows)	DWI (unique rows)	FLAIR (unique rows)	T2* (unique rows)	T2 FSE (unique rows)	PD (unique rows)	Fieldmap (unique rows)	fMRI (unique rows)
ADNI	4,435 (543)	1,194 (100)	476 (42)	939 (223)	1,029 (117)	0 (0)	319 (25)	319 (24)	0 (0)	159 (12)
OASIS3	5,945 (31)	1,104 (9)	0 (0)	519 (6)	298 (4)	695 (2)	687 (3)	0 (0)	1,206 (4)	1,436 (3)
PPMI	941 (9)	941 (9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	*11,321 (583)*	*3,239 (118)*	*476 (42)*	*1,458 (229)*	*1,327 (121)*	*695 (2)*	*1,006 (28)*	*319 (24)*	*1,206 (4)*	*1,595 (15)*

Table 2 Validation set

Source	n	MPRAGE	SPGR	DWI	FLAIR	T2*	T2 FSE	PD	Fieldmap	fMRI
In house	14,711 (1,339)	329 (4)	2,005 (356)	980 (368)	5,942 (377)	885 (88)	708 (305)	804 (12)	3,058 (197)	0 (0)
ADHD-200	840 (6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	(0)	0 (0)	840 (6)
Cam-CAN	643 (1)	0 (0)	0 (0)	643 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
PNC	501 (1)	0 (0)	0 (0)	501 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Open-Neuro	1,022 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	(0)	1,022 (8)
Total	*17,717* (1,355)	*329* (4)	*2,005* (356)	*2,184* (369)	*5,942* (377)	*0(0)*	*708* (305)	*804* (12)	*3,058 (197)*	*1,862* (14)

MRI data into valid BIDS datasets. Because the acquisition parameters that the model relies on are also found in BIDS sidecars, the transformation works for NIfTIs converted by dcm2niix as well as DICOMs. The software used for this transformation is publicly available and open source, as well as the model itself, allowing for continuous training on new data by users.

Methods

Data

Training Data

The model was trained and tested on 11,321 MRI scans obtained from ADNI ($n=4,435$), Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2011) ($n=941$), and Open Access Series of Imaging Studies (OASIS3) (LaMontagne et al., 2019) ($n=5,945$), with a train/test split of 9,056/2,265 scans. From these datasets, we identified the following scan types: T1-weighted magnetization-prepared rapid gradient echo (MPRAGE), T1 inversion recovery/fast-spoiled gradient echo (IR/F-SPGR), diffusion weighted imaging (DWI), T2-FLAIR, T2 turbo/fast spin echo (T/FSE), T2*, proton density (PD), field mapping, and functional MRI (fMRI) sequences. Table 1 describes the distribution of scan types across these datasets, as well as the total number of unique protocols.

Validation Data

The model was validated on a combination of in-house scans collected from various studies and publicly available datasets for a total of 17,717 MRI scans. These datasets came from different sources than the training and test datasets, allowing us to test how well the model generalizes and verify that it has not been trained simply to recognize scans from known datasets. We supplemented the in-house dataset with scans from the Attention-Deficit Hyperactivity Disorder (ADHD)-200 Consortium (Milham et al., 2012), Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository (Taylor et al., 2017), the Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2016), and several datasets publicly available on openneuro.org. The datasets included from OpenNeuro were ds003823, ds003791, ds003835, ds003851, and ds003871. Table 2 describes the distribution of scan types across these datasets, as well as the total number of unique protocols.

Model

XGBoost Model Parameters

To automate MRI scan classification, we used the gradient boosted decision tree algorithm XGBoost, using the Python library xgboost version 1.5.0. We used an XGBClassifier model with a gbtree booster, learning rate = 0.3, max depth = 6, number of estimators = 100, and a multiclass soft probability objective. Model parameters were tuned via grid search cross validation using the *GridSearchCV* function from scikit-learn (Pedregosa et al., 2011) version 0.23.1 to optimize the negative log loss function. The result is a model that can accept DICOM header tags as input and accurately predict the acquisition type of an MRI scan using only metadata.

To inform our classification model, we chose ten features that are required but potentially null (i.e., Type 2 as

per the DICOM standard) DICOM tags or may be found in BIDS sidecar JSON files for cases where DICOMs are not available. These features are TE, repetition time (TR), inversion time (TI), flip angle (FA), diffusion b value, and echo train length (ETL). Additionally, each of the following were treated as binary variables drawn from the Scanning Sequence tag: presence of spin echo (SE) sequence, presence of inversion recovery (IR) sequence, presence of gradient echo (GR) sequence, and presence of echo-planar (EP) sequence.

Model Evaluation

To assess the validity of the model, we compared predicted scan types to ground truth and calculated the accuracy, precision, recall, and F1. Because acquisition types were not distributed evenly across the validation set (e.g., many more FLAIR and fMRI than PD), we considered the micro-average precision, recall, and F1s in addition to macro-averaged scores. Micro-averaged metrics take the average score across all predictions regardless of acquisition type, while macro-averaged metrics first take the average for each acquisition type, then take the average of those averages across all acquisition types.

Model Maintenance and Continuous Training

To help ensure that the model remains generalizable into the future over a variety of real-world MRI data, we provide a framework to easily further train the model on new data. We accomplish this by making use of XGBoost's ability to continuously train on new data, even without the original training data. This framework requires the new training data to be in valid BIDS format, as the acquisition classes are inferred from BIDS paths.

DICOM-to-BIDS Transformation

Because our XGBoost model only produces an integer encoding for one of nine MRI scan types, its general usability is somewhat limited. Ideally, a classifier system would also automatically generate a valid BIDS dataset from an unsorted collection of DICOMs.

To do this, we use the dcm2niix software to convert DICOM scans to the NIfTI format required by BIDS, keeping the sidecar JSON files that dcm2niix produces. The DICOM-to-BIDS transformer program accepts a path to a directory as user input and parses that directory for all valid DICOM files, using pydicom (Mason et al., 2022) to check the validity of DICOM files. The program uses the Patient ID DICOM tag (0010,0020) to track individual MRI sessions and maps them to their respective subjects using the Patient's Name DICOM tag (0010,0010), storing them in a Python dictionary (Fig. 1). Scans with the same Patient ID are collected, and the path to an individual scan's NIfTI and sidecar is mapped to their respective sessions. The result is a mapping from each subject to a list of sessions, with each session mapped to a list of scans, mirroring the hierarchy of a BIDS dataset. To identify scans likely to be use for analyses we heuristically check each session for the presence of localizer and calibration scans using nibabel (https://github.com/nipy/nibabel), filtering out any scans with slice thickness greater than 6 mm. The remaining scans are fed into the XGBoost model to predict the acquisition type, which are mapped to the naming convention prescribed by BIDS (e.g., MPRAGE is mapped to sub-id/session-id/anat/T1w_acq-MPRAGE, FLAIR is mapped to sub-id/session-id/anat/FLAIR, fMRI is mapped to sub-id/session-id/func/bold, etc.). The appropriate BIDS dataset directories (i.e., subject and session directories) are created according to the hierarchy defined in the dictionary map, and the NIfTI and sidecar files are renamed according to the mapping derived from the XGBoost model.

To safeguard against misclassification, the tool outputs the model's certainty in its predictions, providing users with some measure of the confidence level associated with each classification. Specifically, when performing the BIDS transformation, the tool writes the certainty of prediction for all MRI types to a file for each scan, following BIDS naming conventions (e.g., 'sub-01_ses-01_acq-MPRAGE_run-9_T1w_desc-predcertainty.csv' for a T1w MPRAGE scan). Likewise, the tool outputs the Shapley Additive Explanation (SHAP) values (Lundberg & Lee 2017) denoting each feature's – e.g., TE, TR, FA – contribution to the model's prediction, which are written to a file following BIDS naming conventions (e.g. 'sub-01_ses-01_acq-MPRAGE_run-4_T1w_desc-featurecontributions.csv').

While some BIDS-enabled applications (BIDS app) are capable of processing multiple scans of the same type, differentiated by BIDS as separate "runs," many expect only one candidate scan as an input. The DICOM-to-BIDS transformer allows the user to specify if all runs should be kept in the final BIDS dataset, or if only a candidate scan should remain. The program determines which scan is the candidate by comparing the acquisition times of all scans within a session of the same type, keeping the scan with the latest acquisition time via DICOM tag (0008,0032). If multiple scans have the same acquisition time, the scan with the lowest Series Number via DICOM tag (0020,0011) is kept. This is done to select the scan that has the best chance of having the highest quality, offsetting the possibility that a scan was interrupted or corrupted by excessive subject motion and repeated later in the scanning session.

```
{
    "sub-patient_name01": [
        {
            "session_id": "session-patient_id",
            "scans": [
                {
                    "type": "FLAIR",
                    "dcm_path": "./path/to/dcmdir",
                    "nifti_path": "./path/to/nifti.nii.gz",
                    "bids_nifti_path": "./bids/sub-patient_name01/session-patient_id/anat/sub-patient_name01_FLAIR.nii.gz",
                    "bids_sidecar_path": "./bids/sub-patient_name01/session-patient_id/anat/sub-patient_name01_FLAIR.json"
                },
                {
                    "type": "MPRAGE",
                    "dcm_path": "./path/to/dcmdir",
                    "nifti_path": "./path/to/nifti.nii.gz",
                    "bids_nifti_path": "./bids/sub-patient_name01/session-patient_id/anat/sub-patient_name01_T1w.nii.gz",
                    "bids_sidecar_path": "./bids/sub-patient_name01/session-patient_id/anat/sub-patient_name01_T1w.json"
                }
            ]
        }
    ],
    "sub-patient_name02": [
        {
            "session_id": "session-patient_id",
            "scans": [
                {
                    "type": "SPGR",
                    "dcm_path": "./path/to/dcmdir",
                    "nifti_path": "./path/to/nifti.nii.gz",
                    "bids_nifti_path": "./bids/sub-patient_name02/session-patient_id/anat/sub-patient_name02_SPGR.nii.gz",
                    "bids_sidecar_path": "./bids/sub-patient_name02/session-patient_id/anat/sub-patient_name02_SPGR.json"
                },
                {
                    "type": "rsfMRI",
                    "dcm_path": "./path/to/dcmdir",
                    "nifti_path": "./path/to/nifti.nii.gz",
                    "bids_nifti_path": "./bids/sub-patient_name02/session-patient_id/func/sub-patient_name02_task-rest_bold.nii.gz",
                    "bids_sidecar_path": "./bids/sub-patient_name02/session-patient_id/anat/sub-patient_name02_task-rest_bold.json"
                }
            ]
        }
    ]
}
```


Fig. 1 Hierarchy of DICOM to BIDS mapping. Following conversion from DICOM to NIfTI, the Patient Name and Patient ID DICOM tags are mapped to subjects and sessions in BIDS, respectively. Metadata is used to automatically classify the scan type, filling out the rest of the BIDS filename

All code and the pretrained model are publicly available at https://gitlab.com/abartnik/scan_classifier. Training, test, and validation data derived from external sources may be obtained from ADNI (https://adni.loni.usc.edu/), OASIS3 (https://www.oasis-brains.org/), PPMI (https://www.ppmi-info.org/), ADHD-200 (http://preprocessed-connectomes-project.org/adhd200/), Cam-CAN (https://www.cam-can.org/), PNC (https://www.med.upenn.edu/bbl/philadelphianeurodevelopmentalcohort.html), and OpenNeuro.org (https://openneuro.org/). Validation data derived from in-house studies is available at https://gitlab.com/abartnik/scan_classifier.

Results

Model Performance

The model is able to predict the acquisition type of an MRI scan using only ten features derived from DICOM headers with 99.475% accuracy. The micro-averaged precision – the average precision across all acquisition types at once – was 0.995 and the micro-average recall was 0.995. The macro-averaged precision – the average of each acquisition type's individual precision – was 0.994 and the macro-averaged recall was 0.989. The micro-averaged F1 for the model was 0.995 and the macro-averaged F1 was 0.991. Table 3 provides an overview of the precision, recall, and F1s for each acquisition type, while Table 4 shows the confusion matrix

Table 3 Performance per acquisition type in validation set

Acquisition type	Precision	Recall	F1
T2*	0.998	0.915	0.955
DWI	0.999	0.996	0.997
T2 FLAIR	1.00	0.999	0.999
Field Map	1.00	0.999	0.999
fMRI	0.999	1.00	0.999
T2 FSE	0.987	0.999	0.993
T1 MPRAGE	1.00	1.00	1.00
PD	0.999	0.995	0.997
T1 SPGR	0.963	1.00	0.981

for actual vs. predicted acquisition types for the validation set.

Model Summary

The most important features for the model to predict acquisition type are the relaxation times, followed by TI and FA. In contrast, the presence of different scanning sequences (SE, GR, etc.) played a less important role. Since XGBoost is an ensemble machine learning technique that takes advantage of multiple decision trees, it is possible to examine each tree and see which features are commonly used in the first decision. The features that influenced the first decisions were overwhelmingly TR and TE, accounting for the first decision in 62.3% of all trees in the ensemble. Figure 2 displays the features ranked by importance to the model, along with the scores XGBoost calculates to assign importance of a feature.

Table 4 Model predictions confusion matrix

	T2*	DWI	T2 FLAIR	Field Map	fMRI	T2 FSE	T1 MPRAGE	PD	T1 SPGR
T2*	810	0	0	0	0	0	0	0	75
DWI	0	2114	0	0	1	9	0	0	0
T2 FLAIR	0	0	5941	0	0	0	0	1	0
Field Map	0	0	0	3056	0	0	0	0	2
fMRI	0	0	0	0	1862	0	0	0	0
T2 FSE	0	1	0	0	0	707	0	0	0
T1 MPRAGE	0	0	0	0	0	0	329	0	0
PD	2	2	0	0	0	0	0	800	0
T1 SPGR	0	0	0	0	0	0	0	0	2005

X axis – predicted; Y axis – actual

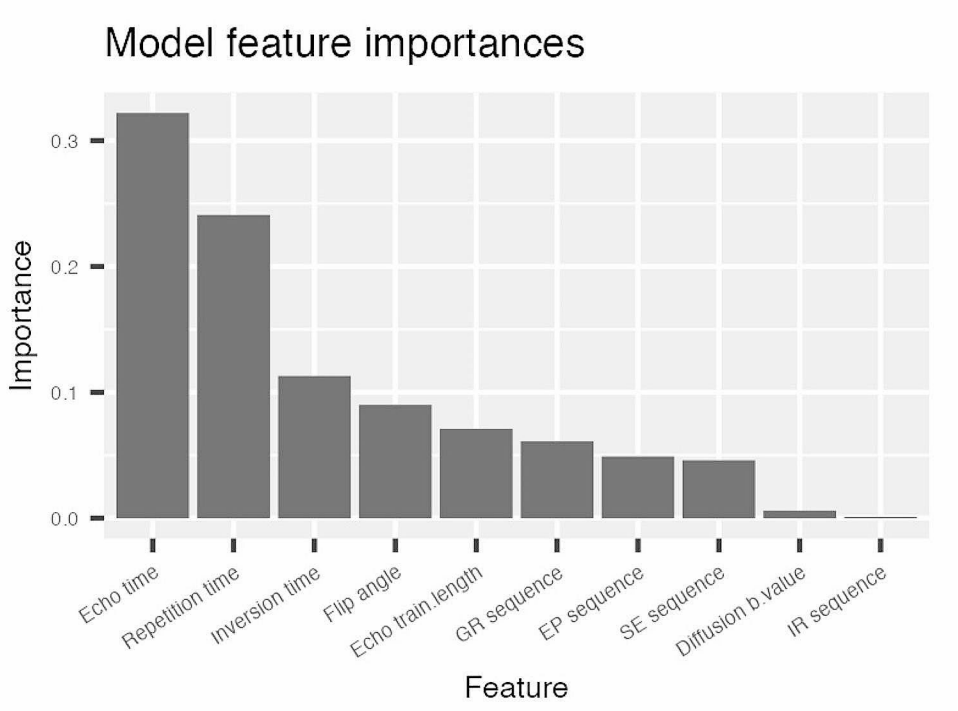

Fig. 2 Features ranked by importance. The most important features for the model to classify the acquisition type of an MRI scan relate to relaxation time (TE, TR, and TI). This has face validity, given our understanding of how different MRI types are acquired largely by tuning the scanner to the relaxation times of tissue

Model Deployment

The scan classification model is available in two main forms: a standalone app that returns the name of the acquisition type based on a single input image, or an app that transforms a raw DICOM directory into a valid BIDS dataset. Additionally, the scan classifier is available as a Python library, which is used by both apps. Because the model works only on metadata available from file headers or sidecars rather than entire images, classification is very fast. This is due in large part to the nature of XGBoost – which predicted the acquisition types of the entire 17,717 image validation set in 139ms – but also to the small amount of file opening operations (classifying a single scan via the standalone app took on average 2.93s, calculated via GNU time https://www.gnu.org/software/time/).

Discussion

Here we present a fully automated tool capable of identifying MRI acquisition types commonly seen in the clinical routine with high accuracy and transforming them into valid BIDS datasets. This level in accuracy is achieved by using XGBoost to classify common MRI types from just ten features typically found in most DICOM headers and BIDS sidecars, the most important being relaxation time. Varying the relaxation time is a fundamental part of acquiring different types of MRI – namely T1w vs. T2w vs. PD – and its importance in the model's ability to differentiate between MRI types emphasizes how well the model mirrors this reality. The model is able to further delineate other common subtypes of T1w (e.g., MPRAGE, SPGR) and T2w (e.g., FLAIR) images using only a few more features, such as TI, FA, and scanning sequence.

XGBoost is an ideal candidate for solving this problem because XGBoost models excel in datasets containing many observations with relatively few features. XGBoost achieves this via "boosting" the model, a process involving a sequence of decision trees whose predictions inform the rest, and the accuracy of these predictions is used to weight subsets of training data according to their rate of misclassification. Because the model relies on only ten features, it is broadly applicable even in cases where metadata in DICOM headers or BIDS sidecars is incomplete. This also helps mitigate the problem of some metadata only being contained in private DICOM tags that vary among MRI manufacturers by breaking the acquisition down to the most basic and commonly available parameters. By providing users with the means to assess classification uncertainty and understand the underlying factors contributing to each prediction (i.e., SHAP values), we aim to enhance the usability of our tool and facilitate more informed decision-making when dealing with misclassifications. Understanding the level of certainty associated with each classification is crucial, while SHAP values help users identify potential sources of error. Additionally, both certainty estimates and SHAP values provides valuable insights that can inform future training iterations, allowing for targeted improvements to the model's performance and generalizability.

The tool is able to generate BIDS datasets from unsorted DICOM directories or NIfTIs, provided the NIfTIs have the BIDS sidecar generated by dcm2niix, greatly alleviating the time a researcher would need to spend manually curating a BIDS dataset. This approach has several key advantages that should make it a useful tool for clinical researchers by removing as many steps involved with manual curation of BIDS datasets as possible. Primarily, the use of a pretrained model eliminates the need to open DICOM headers or BIDS sidecars to determine the acquisition parameters or sequences *a priori*. This is particularly valuable for clinical researchers or inexperienced neuroimaging scientists who may not be familiar with MRI acquisition parameters or what they mean. DICOM headers are also particularly difficult to open on their own for an inexperienced researcher, outside of image viewer software that may be cumbersome to use to quickly ascertain the acquisition parameters of many images. Additionally, a pretrained model does not require a configuration file containing possible ranges of acquisition parameters that a user must fill out for every dataset, which has the advantage of limiting how much a clinical researcher who may not have much computational experience would need to work with unfamiliar filetypes (e.g., JSON). Taken together, these create a largely hands-free approach to generating valid BIDS datasets that clinical researchers with little-to-no experience working with MRI data can then use in their studies via BIDS-enabled applications.

The proposed approach is important for clinical and translational research involving neuroimaging by facilitating analysis and sharing of imaging data for researchers without extensive experience and expert neuroimaging scientists alike. Non-experts would more easily be able to include new and existing imaging data in their studies, since the software requires very little manual tuning. This greatly decreases the barrier to entry for researchers who wish to include imaging data in their studies. By reducing the steps involved with manually curating imaging data into BIDS datasets, neuroimaging researchers will be better able to share data and analyses in collaborative studies. Because the model has been trained on MRI types common to the clinical routine, this has the additional effect of helping neuroimaging researchers work with clinical imaging data that may otherwise be left in silos, as well as making it easier

to work with large datasets. Furthermore, being able to automatically generate a valid BIDS dataset facilitates the process of uploading study data to publicly hosted neuroinformatics databases such as OpenNeuro (which requires datasets to be in valid BIDS format) or the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) (Luo et al., 2009), where the dataset would get a valid DOI and promote re-analysis and open science practices. Finally, since the model is fully available to the public, it is possible for any researcher to continually train it on new datasets, further increasing its accuracy and generalizability across new data.

One of the greatest strengths for identifying scans on hard-coded metadata is the ability to bypass the limitation of relying on the Series Description DICOM tag (0008,103E) encoded in arbitrary strings, as is often the case when not using an automated tool to organize DICOMs. This problem applies to NIfTIs as well, as the Series Description is often used by dcm2niix in the NifTI filename and even carries over to the BIDS sidecar that is generated. Because the series description is encoded as a string, it is subject to variance within and between datasets. This introduces inconsistencies among images in the same dataset, potentially leading to DICOMs being sorted incorrectly. Furthermore, the free-form nature of the Series Description tag (if available) means that its exact value is largely dependent on the scanning center. In addition to hampering the organization of scans within a study, this also greatly limits the ability to share scans between sites and studies. Use of regular expressions helps alleviate this, but also introduces an unnecessary level of complexity for most clinical. Additionally, relying only on acquisition parameters stored in headers and sidecars allows the model to run agnostic of scan quality. This bypasses any issues caused by motion correction or pathology that would limit the performance of a classifier trained to recognize scans based on the image data itself. This has the potential problem of still requiring manual quality control of scans after classifying and sorting them, but the automated transformation to BIDS allows one to then leverage the MRIQC (Esteban et al., 2017) BIDS-app to aid in streamlined quality control.

This is not the first tool to automatically generate BIDS datasets from unsorted DICOMs – there are several mature tools available that have seen widespread use throughout the neuroimaging community, some of the most used software being HeuDiConv (Halchenko and others 2018), dcm2bids (Bedetti et al. 2022), BIDScoin (Zwiers et al., 2022), and DeepDicomSort (van der Voort et al., 2021). Additionally, a similar approach has been used by Gauriau et al., wherein a random forest model was trained to classify common MRI acquisition types with similar accuracy, further supporting the feasibility of our approach (Gauriau et al., 2020). With the exception of DeepDicomSort, these approaches rely on user input – typically in the form of configuration files – to specify heuristics of values to match in DICOM tags in order to map them to a specific acquisition type (e.g., a TI of 1900ms – 2100ms in the header is mapped to T2 FLAIR). The goal of these software tools is largely to alleviate the complexity and manual effort required in sorting a known DICOM dataset into a valid BIDS form. Of these, HeuDiConv is perhaps the most mature and widely used, playing a key role in such large-scale neuroimaging initiatives as the Flywheel Platform (Tapera et al., 2021) and ReproNim (Kennedy et al., 2019). Like many other DICOM-to-BIDS solutions, HeuDiConv (https://github.com/nipy/heudiconv) relies on heuristics in user-supplied configuration files to classify acquisition types, and was likely the first to use this approach to great success. This approach works very well for experienced neuroimaging scientists and allows HeuDiConv to adapt to nearly any imaging dataset, but relies on the user to understand the process of MRI acquisition and how to find these acquisition parameters in DICOM headers, limiting its use for clinical researchers. The heuristic files may be in the form of ReproNim heuristics or Python scripts, maintaining this barrier of entry for clinical researchers. Finally, many of the heuristics rely on the values encoded as arbitrary strings in the Series Description and Protocol Name (0018,1030) DICOM tags, which may be prone to errors and inconstancies within and between datasets, as described previously.

Similarly, dcm2bids (https://github.com/UNFmontreal/Dcm2Bids) is a mature and actively developed project that uses heuristics to automatically transform raw DICOM directories into valid BIDS datasets. While the heuristics files are stored in a relatively simpler JSON format, they still require the user to be able to access DICOM headers and know ahead of time what the range of acquisition parameters are.

The BIDScoin (https://github.com/Donders-Institute/bidscoin) software is unique in that it offers a graphical user interface (GUI), making the tool much easier for clinical researchers to use. In addition to a GUI, BIDScoin also provides a suite of modular command line tools to help the user convert their data in BIDS datasets. Since it relies generally on a heuristics approach, BIDScoin is limited by its ability to generalize across Series Descriptions or Protocol Names, which are some of the main attributes it uses for sorting DICOMs. Finally, BIDScoin expects DICOMs to be sorted ahead of time into BIDS-compliant subject or session directories, which still requires manual effort form the user.

DeepDicomSort (https://github.com/Svdvoort/DeepDicomSort) is unique in that it uses a convolutional neural network to classify MRI types by the appearance of the images themselves, rather than acquisition parameters from

metadata. Like our model, DeepDicomSort was trained largely on data from ADNI with high accuracy. However, a limitation of DeepDicomSort is that it does not recognize fMRI, which is an important part of neuroimaging research and is commonly included in many large public datasets. Since DeepDicomSort is a nearly handsfree approach that uses only image appearance and was trained on ADNI data, it may be possible to work in conjunction with our metadata-based model to more robustly classify MRI types with higher accuracy.

While the presented software has the potential to quickly classify and organize raw MRI DICOMs and greatly simplify the process of transforming data in valid BIDS datasets, there are several key limitations that should be mentioned. While the training and test datasets are large and comprehensively cover many common acquisition types, the model was trained on only three public datasets with fairly consistent acquisition protocols. This limits the heterogeneity of training data and hampers the generalizability of the model to new data, particularly across new scanning sites and manufacturers. However, the images in these public datasets have been collected from multiple scanners and sites, and the validation set comprises data from five separate sources.

Because the model is trained only to recognize acquisition parameters found in DICOM headers, there is no way to determine the quality of images or select the scan most suitable for analysis when multiple of the same type exist in the same exam. Furthermore, many MRI scanners provide reconstructed and reoriented series, creating entries that the model considers duplicates. Current heuristics allow for the selection of primary scans in favor of reconstructions, though these may of suitable quality for analysis. Future work should improve the heuristics used and expand on the model's current capabilities to allow for the sorting and selecting of candidate scans based on image quality. This may potentially be done by using something like MRIQC to derive values for signal-to-noise ratio or contrast-to-noise ratio once the dataset has been converted to BIDS. Likewise, the model is unable to recognize post-contrast enhancement in scans, as is sometimes administered in the clinical routine for MS. Future work should also be done to take the presence of post-contrast into account when selected candidate scans.

While the included scan types are commonly acquired in research, our validation focused on acquisition types that are most commonly used clinically and fit in the pre-existing ecosystem of BIDS-apps, targeting one-size-fits all apps like the nipreps framework (Esteban et al., 2019). By providing a framework to facilitate further training of the model on new data, it is only possible to train the model on acquisition types already present in the model. In order to add new acquisition types to the model, it would be necessary to train the entire model from scratch. This tool also has limited applicability to working with several public datasets, which typically provide only images in the form of NIfTIs rather than DICOMs and may not always provide the sidecars that the model requires.

Conclusion

We have developed a model that accurately identifies MRI acquisition types commonly seen in clinical routine from using just a DICOM header or a BIDS sidecar. The model is fully available to the public, making it possible for any researcher to continually train it on new datasets, further increasing its accuracy and generalizability across new data. By automating the transformation of DICOM to BIDS, the tool has the potential to greatly facilitate the process of uploading study data to publicly hosted neuroinformatics databases, as well as promote re-analysis and open science practices.

Author Contributions A.B. wrote the main manuscript text, including tables and figures, and conceptualized the approach.A.B., M.S., N.B., and M.D. identified model features.A.B. and C.S. trained and validated the model.A.B. and S.S. implemented the model in a tool to transform MRI data to BIDS.All authors reviewed the manuscript.

Declarations

Competing Interests The authors declare no competing interests.

References

Bedetti, C., arnaudbore, Guay, S., Carlin, J., Nick, Dastous, A. (2022, May). UNFmontreal/Dcm2Bids: 2.1.7. Zenodo. https://doi.org/10.5281/zenodo.6596007.

Butzkueven, H., Chapman, J., Cristiano, E., Grand'Maison, F., Hoffmann, M., Izquierdo, G., et al. (2006). MSBase: An international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Multiple Sclerosis Journal*, *12*(6), 769–774.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0 4-2*, *1*(4), 1–4.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9), e0184661.

Esteban, O., Wright, J., Markiewicz, C. J., Thompson, W. H., Goncalves, M., Ciric, R. (2019). NiPreps: enabling the division of labor in neuroimaging beyond fMRIPrep, 7–9.

Gauriau, R., Bridge, C., Chen, L., Kitamura, F., Tenenholtz, N. A., Kirsch, J. E., et al. (2020). Using DICOM Metadata for Radiological Image Series categorization: A feasibility study on large clinical brain MRI datasets. *Journal of Digital Imaging*, *33*(3), 747–762. https://doi.org/10.1007/s10278-019-00308-x.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging

experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44.

Halchenko, Y. O. (2018). & others. Open Source Software: Heudiconv. *Zenodo. doi*, *10*.

JackJr., C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049.

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., et al. (2019). *Everything matters: The ReproNim Perspective on reproducible neuroimaging*. Frontiers in Neuroinformatics.

LaMontagne, P. J., Benzinger, T. L. S., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C. (2019). OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *MedRxiv*, 2012–2019.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, *264*, 47–56. https://doi.org/10.1016/j.jneumeth.2016.03.001.

Lundberg, S. M., & Lee, S. I. *A Unified Approach to Interpreting Model Predictions*. Advances in neural information processing systems 30 (2017).

Luo, X. J., Kennedy, D. N., & Cohen, Z. (2009). Neuroimaging Informatics Tools and resources Clearinghouse (NITRC) Resource announcement. *Neuroinformatics*, *7*(1), 55–56. https://doi.org/10.1007/s12021-008-9036-8.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The Parkinson progression marker Initiative (PPMI). *Progress in Neurobiology*, *95*(4), 629–635. https://doi.org/10.1016/j.pneurobio.2011.09.005.

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *eLife*, *10*, e71774. https://doi.org/10.7554/eLife.71774.

Mason, D., scaramallion, Suever, Vanessasaurus, J. (2022). pydicom/pydicom: pydicom 2.3.0. https://doi.org/10.5281/ZENODO.6394735.

Mildenberger, P., Eichelberg, M., & Martin, E. (2002). Introduction to the DICOM standard. *European Radiology*, *12*(4), 920–927. https://doi.org/10.1007/s003300101100.

Milham, M., Fair, D., Mennes, M., & Mostofsky, S. (2012). The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Satterthwaite, T. D., Connolly, J. J., Ruparel, K., Calkins, M. E., Jackson, C., Elliott, M. A., et al. (2016). The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage*, *124*, 1115–1119. https://doi.org/10.1016/j.neuroimage.2015.03.056.

Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P. M., Federico, A., & De Stefano, N. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage*, *17*(1), 479–489. https://doi.org/10.1006/nimg.2002.1040.

Smith-Bindman, R., Kwan, M. L., Marlow, E. C., Theis, M. K., Bolch, W., Cheng, S. Y., et al. (2019). Trends in Use of Medical Imaging in US Health Care systems and in Ontario, Canada, 2000–2016. *JAMA - Journal of the American Medical Association*, *322*(9), 843–856. https://doi.org/10.1001/jama.2019.11456.

Tapera, T. M., Cieslak, M., Bertolero, M., Adebimpe, A., Aguirre, G. K., Butler, E. R., et al. (2021). *FlywheelTools: Data Curation and Manipulation on the Flywheel platform*. Frontiers in Neuroinformatics.

Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., et al. (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, *144*, 262–269. https://doi.org/10.1016/j.neuroimage.2015.09.018.

van der Voort, S. R., Smits, M., & Klein, S. (2021). DeepDicomSort: An automatic sorting algorithm for Brain magnetic resonance Imaging Data. *Neuroinformatics*, *19*(1), 159–184. https://doi.org/10.1007/s12021-020-09475-7.

van Ooijen, P. M. A. (2019). In E. R. Ranschaert, S. Morozov, & P. R. Algra (Eds.), *Quality and Curation of Medical images and data BT - Artificial Intelligence in Medical Imaging: Opportunities, applications and risks* (pp. 247–255). Springer International Publishing. https://doi.org/10.1007/978-3-319-94878-2_17.

Zwiers, M. P., Moia, S., & Oostenveld, R. (2022). BIDScoin: A User-Friendly Application to Convert Source Data to Brain Imaging Data Structure. *Frontiers in Neuroinformatics*, 15(January). https://doi.org/10.3389/fninf.2021.770608.