Journal of Digital Imaging (2023) 36:289–305
https://doi.org/10.1007/s10278-022-00690-z

ORIGINAL PAPER

Improving the Automatic Classification of Brain MRI Acquisition Contrast with Machine Learning

Julia Cluceru[1] · Janine M. Lupo[1] · Yannet Interian[2] · Riley Bove[3,4] · Jason C. Crane[1]

Received: 1 December 2021 / Revised: 22 June 2022 / Accepted: 22 July 2022 / Published online: 8 August 2022

Abstract

Automated quantification of data acquired as part of an MRI exam requires identification of the specific acquisition of relevance to a particular analysis. This motivates the development of methods capable of reliably classifying MRI acquisitions according to their nominal contrast type, e.g., T1 weighted, T1 post-contrast, T2 weighted, T2-weighted FLAIR, proton-density weighted. Prior studies have investigated using imaging-based methods and DICOM metadata-based methods with success on cohorts of patients acquired as part of a clinical trial. This study compares the performance of these methods on heterogeneous clinical datasets acquired with many different scanners from many institutions. RF and CNN models were trained on metadata and pixel data, respectively. A combined RF model incorporated CNN logits from the pixel-based model together with metadata. Four cohorts were used for model development and evaluation: MS research ($n=11{,}106$ series), MS clinical ($n=3244$ series), glioma research ($n=612$ series, test/validation only), and ADNI PTSD ($n=477$ series, training only). Together, these cohorts represent a broad range of acquisition contexts (scanners, sequences, institutions) and subject pathologies. Pixel-based CNN and combined models achieved accuracies between 97 and 98% on the clinical MS cohort. Validation/test accuracies with the glioma cohort were 99.7% (metadata only) and 98.4 (CNN). Accurate and generalizable classification of MRI acquisition contrast types was demonstrated. Such methods are important for enabling automated data selection in high-throughput and big-data image analysis applications.

Keywords Image processing · Image retrieval · Image classification · Machine learning · Deep learning · Magnetic resonance imaging

Introduction

Automated quantification of data from a magnetic resonance imaging (MRI) exam typically requires identification of the specific type of acquisition relevant to a particular analysis. MRI enables images to be collected with a wide range of physical properties and with varying contrast [1, 2]. A typical MRI exam comprises multiple types of MR sequences or series, each with its own set of acquisition parameters that determines tissue contrast, highlighting different properties of the imaged anatomy. Although the most common and clinically relevant neuroimaging MRI contrasts are T1 weighted (T1), T1 weighted after administration of a gadolinium-based contrast agent (T1C), T2 weighted (T2), T2-weighted fluid attenuated inversion recovery (T2-FLAIR), and proton density-weighted (PD) images (Fig. 1A), there are many additional kinds of MRI contrasts typically acquired, including diffusion-weighted and T2*-weighted images. The numerous parameters that define each acquisition type are therefore encoded as text in the DICOM [3] image header alongside the image pixels. Unfortunately, the heterogeneity in the way in which many of these parameters are entered and encoded in proprietary vendor formats make it challenging to unambiguously identify a specific

✉ Jason C. Crane
jason.crane@ucsf.edu

1 Center for Intelligent Imaging, Department of Radiology & Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA

2 MS in Analytics Program, University of San Francisco, San Francisco, CA, USA

3 Department of Neurology, MS and Neuroinflammation Clinic, University of California San Francisco, San Francisco, CA, USA

4 Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, USA

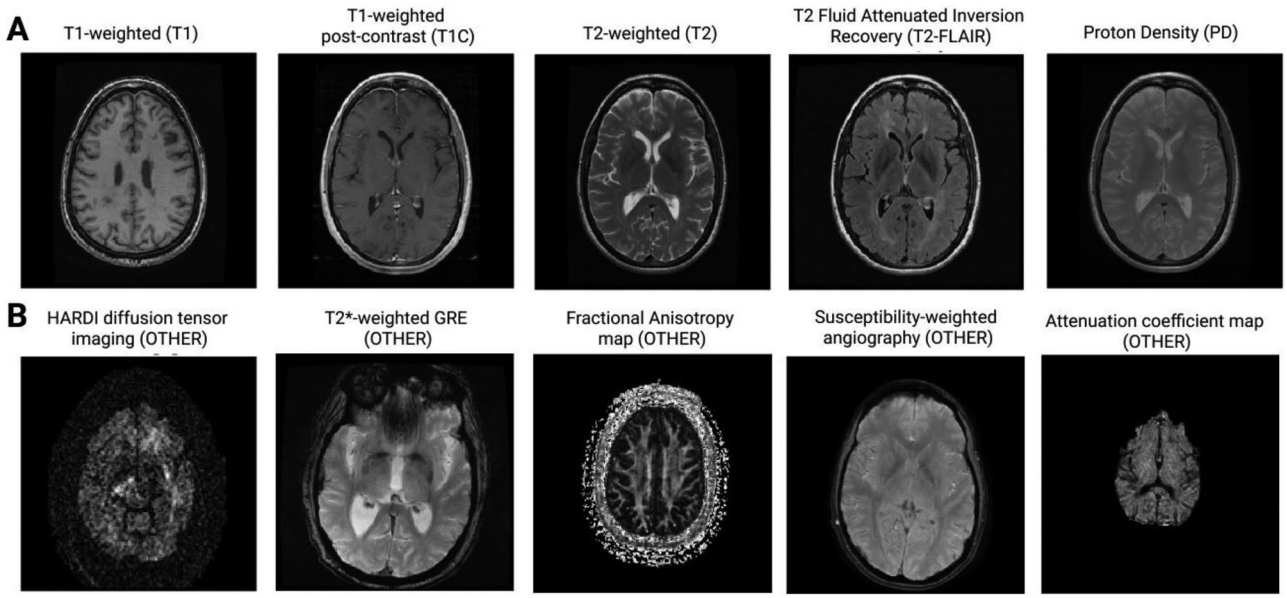

Fig. 1 Examples of images in each category. **A** T1, T1C, T2, T2-FLAIR, and PD are the most common MRI contrasts acquired during neuroimaging exams. **B** Category "OTHER" comprised many kinds of MRI sequences, including T2*-weighted images that can look similar to T2 or PD. Note: though this figure includes all axial images, the model was trained, validated, and tested on the raw DICOM files that could be acquired in coronal, sagittal, or oblique planes as well

MRI acquisition type for analysis based on the header information alone.

The development of automated methods capable of prospectively analyzing real-world, heterogeneously acquired MRI data thus motivates the need for strategies to reliably and programmatically classify MRI acquisitions according to their nominal acquisition contrast types, e.g., T1, T1C, T2, T2-FLAIR, PD [1, 4, 5]. For example, monitoring of neurological disease progression and of response to therapy using MRI involves quantification of serial changes that occur on images acquired with similar tissue contrast (e.g., lesion volume based on T1-weighted or T2-weighted contrast) [6]. Development of artificial intelligence (AI) and machine learning (ML) models often depends on the availability of large quantities of specific classes of data [7, 8] for training, validation, and testing. MRI contrast classification models are therefore an important element of content-based image retrieval (CBIR) systems [9–11], enabling large medical centers to leverage vast amounts of prospectively acquired, but retrospectively analyzed, data for population-level computational health research. Similarly, using such ML for prospective inference on real-world data will require data classification models to identify the relevant input data for the models in order to generate meaningful results. For example, an ML model trained to segment lesions on T2-FLAIR images must be able to identify just the T2-FLAIR acquisition from an MRI exam comprising potentially 10–20 different acquisitions.

Although relevant features that identify the image contrast of an MRI series may consist of both metadata from the DICOM header and intrinsic properties of the imaging pixel data, the header data is not always consistent. Further, it may conflict with the intrinsic imaging contrast, especially when longitudinal examinations include data acquired on different scanners (manufacturer, field strength, software) with a variety of scan protocols. There are also a vast range of physical scanning parameters associated with each acquisition type, and descriptive fields such as the DICOM "series description" (0008,103E) are free-text attributes subject to local conventions, technologist choice, and even human error [12]. As a result, the header features associated with MRI scans are extremely heterogeneous with DICOM attributes that often do not explicitly identify the type of acquisition, limiting the ability to automatically retrieve images acquired with the same contrast weighting using text-based approaches and presenting challenges for development of high-throughput methods for automated analysis of MRI data.

Background Literature

Only a few prior studies focus on automatically classifying MRI series acquisition contrast, all underscoring the challenging nature of the problem, stemming from the high degree of variability in real-world imaging data. Together, they reinforce the increasing need for such methods in content-based image retrieval (CBIR) [9, 13, 14] and automated image analysis pipelines, particularly as the field embraces high-throughput, big-data, and machine

learning–based methods. However, each study takes a different approach to modeling and cohort selection with concomitant implications for application to real-world use cases with outstanding considerations being applicability to different input data (orientation, number of slices, etc.), computational cost, and model generalizability. Among these, three studies used convolutional neural network (CNN) architectures to classify MR images based on the imaging pixel data alone. Remedios et al. [13] used a multi-step approach to identify T1, T2, and T2-FLAIR followed by classification of pre- vs. post-contrast T1 and T2-FLAIR images. The authors used 3418 series representing multiple pathologies, 4 sites, and 5 scanners to achieve an overall accuracy of 97.5%. Ranjbar et al. [15] classified brain tumor MR images from 20 institutions into four contrasts (T1, T1C, T2, T2-FLAIR) with excellent accuracy (99.2%). The data in the analysis were acquired using clinical research protocols and included only data acquired with those 4 acquisition contrast classes. The third study, by Pizarro et al. [16], achieved similar results (> 99% accuracy) also adding proton density (PD) and magnetic transfer ON and OFF and even deploying the algorithm in an image-processing pipeline. The description of the dataset includes over 100 institutions with over 45,000 MRI series, though they mention only that they were acquired through clinical trials and do not describe the pathology or heterogeneity in scanner acquisitions. Clinical trials are often more uniform in their scanning acquisition parameters across institutions, and they do not necessarily reflect real-world clinical performance. In addition, the methodology requires 30 axial slices as input and a second neural network in order to classify a single MRI series into its contrast, limiting the applicability of this algorithm to volumes containing over 30 slices. Finally, though these studies achieve highly accurate results, they do not address classification of brain MRI series that are acquired with other contrast mechanisms, which is necessary for real-world clinical applications.

Another recent study by Gauriau et al. achieved accuracies between 97.5 and 99.96% from MRI-associated DICOM metadata alone [12]. This approach is appealing because inference is less computationally expensive and is faster compared with methods based on pixel-level imaging data. Their model was trained and validated on two very large datasets from different institutions and achieved very high accuracy. The study used ground truth contrast labels defined from the DICOM "Series Description" attribute, which is known to be unreliable for real-world exams. The authors acknowledge the limitation of this approach, noting that for up to 10% of series, the contrast mechanism could not be accurately identified by the series description alone.

Preliminary efforts in the present work to classify MRI data from the UCSF clinical radiology PACS using the "Series Description" alone were even less successful than those reported above. There were 18 times as many unique series descriptions observed in clinically acquired multiple sclerosis (MS) exams compared with the number of unique descriptions observed in MRI exams from a longitudinal MS research study, despite the selected clinical cohort comprising 60% fewer exams. The average number of series per "Series Description" decreased from 179 for the research cohort to only 3 for the clinical cohort. This is not surprising as clinical trial protocols result in relatively consistent acquisition parameters including consistent use of "Series Descriptions."

In addition to the primary challenge of accurately classifying the MRI acquisition contrast in brain exams, the problem is further complicated by missing or mislabeled anatomical region. The DICOM "Body Part Examined" (0018,0015) tag is frequently missing or contains incorrect labels which confounds CBIR from PACS.

These challenges motivated the present work aimed at developing and validating methods to automatically classify brain MR images according to their specific types of acquisition contrast using a combination of metadata and pixel-based machine learning approaches. We included images acquired for patients with MS and with gliomas. Specifically, the main objective was to obtain high accuracy for arbitrary real-world MRI exams of MS patients sampled from the UCSF clinical PACS. These data are not subject to the strict protocols of clinical trials and are therefore much more challenging to automatically classify. This distinction is crucial, as this scenario describes the real clinical setting in which models and analysis pipelines are deployed. A secondary objective of this study was to test the performance of the present models on datasets acquired from vastly different pathological profiles by testing the classification algorithm on other disease cohorts including exams from patients with high-grade gliomas. The prediction accuracy among four approaches was compared in order to determine the optimal contrast classification model: 1) a metadata only, rule-based approach; 2) a metadata-only machine learning model; 3) an imaging-only convolutional neural network; and 4) a combined ensemble model that uses both metadata and imaging data. The hypothesis of this study is that the combined model using metadata and imaging data together will obtain the highest classification accuracy.

Methods

Datasets

Four MRI datasets were included in this analysis: 1) an MS research (MSR) dataset consisting of 1731 exams [17]; 2) a glioma research (GR) dataset consisting of

179 newly diagnosed and recurrent glioma exams [18, 19]; 3) a post-traumatic stress disorder research dataset (ADNIR) consisting of 116 exams from the publicly available ADNIDOD dataset [20]; and 4) a clinical MS (MSC) dataset consisting of 311 exams that are representative of typical real-world institutional PACS data acquired on multiple scanners and external sites (Table 1). The ADNIR data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) designed to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. MSC exams retrieved from the UCSF PACS were acquired with varying scan protocols at UCSF and external sites, resulting in a total of 104 unique combinations of "Manufacturer" (0008,0070), "Manufacturer Model Name" (0008,1090), "Software Versions" (0018,1020), and "Magnetic Field Strength" (0018,0087); 814 values of "Protocol Name" (0018,1030); and 32% of the studies being acquired at external institutions. Datasets MSR, GR, and ADNIR were obtained with well-defined research acquisition protocols. GR has much more extensive pathology per exam compared with MSR, MSC, or ADNIR. All UCSF data used in this work were deidentified prior to analysis and used according to IRB approval. The deidentification process did not remove or modify any of the DICOM attributes used for the metadata analysis in this work.

Labels

Each MRI series in the MSR and MSC datasets was assigned one of the following weak labels using a rule-based model: T1 weighted (T1), T1 post-contrast weighted (T1C), T2 weighted (T2), proton density (PD), and T2 FLAIR (T2_FLAIR). A catch-all category of other contrast types called "OTHER" was used to describe additional series such as diffusion-weighted and T2*-weighted images often acquired based on the application (Fig. 1B). No distinction was made between labeling spin-echo and spoiled gradient recalled echo images as they both provide the same information for interpretation by users in the context of the target UCSF *Bridge* application for MS. Images with substantial artifacts, e.g., motion or ringing, were also labeled as OTHER even if originally acquired with one of the 5 relevant contrasts. Image volumes and preliminary labels from both MSC and MSR datasets were subsequently visually verified by two brain imaging scientists (J.C.C. and J.G.C.). Visual review was performed using an in-house tool developed in Python that presented the images in a medical image viewer together with a window containing the DICOM metadata to aid in the classification. The reader was able to window-level, change slice and FOV, read DICOM attributes, and then select the classification from a drop-down menu which stored the information in a flat-file format for use in training. Although the GR and ADNIR datasets were already labeled, approximately 25% of MRI series from these datasets were randomly chosen for manual visual review to ensure correctness (J.G.C.). MRI localizers [21] were excluded based on their DICOM Series Description attribute together with the number of imaged slices in the series because they are typically acquired with the PD MRI contrast class but rarely used clinically or in research due to their low resolution or limited field of view.

Training and Testing Splits

Because the primary objective was to assess the ability of a model to accurately classify the MSC dataset which represents the real-world heterogenous data located in a clinical PACS, MSC was randomly split by exam into 33.3% training, 33.3% validation, and 33.3% test sets. This allowed for representation of heterogeneous, "messy" data in training while also creating validation and test sets with similar distributions. The datasets MSR and ADNIR were acquired with research protocol and are much more uniform; thus, testing the algorithm on these datasets would not have provided insight into the performance of the algorithm on real-world data, and therefore 100% of both datasets were used for training. GR was a relatively small and homogeneous research cohort and was included in this study primarily for testing how well the model would perform on data with

Table 1 Data distribution. Training, validation, and test splits for each are defined in the text

Dataset	Dataset type	Number of exams	Total series	T1	T1C	T2	T2 FLAIR	PD	OTHER
MS research (MSR)	Research protocol	1731	11,106	3562	1679	1332	887	1392	2254
MS clinical (MSC)	Heterogeneous	311	3244	655	722	384	593	75	815
Glioma (GR)	Research protocol	180	612	145	170	126	167	0	4
PTSD (ADNIR)	Research protocol	116	477	101	0	117	125	0	134

more extensive pathology compared with the MS cohort. GR was therefore split into 50% validation and 50% test sets to evaluate whether the developed models were robust enough to accurately predict image contrast, even when presented with MR images containing more extensive pathology than they had previously been trained on.

Rule-Based Classification

MRI image contrast is determined by multiple scanning parameters such as the "Echo Time" (0018,0081) and "Repetition Time" (0018,0080) that are stored as attributes in the DICOM header. In addition, administered contrast agents are often indicated in the DICOM header, e.g., "Contrast Bolus Agent" (0018,0010). Although the "Series Description" (0008,103E) tag may explicitly define the acquisition contrast in some data sets, it is highly variable and subject to operator entry error. A rule-based model using metadata from DICOM attributes was developed to derive weak contrast classification labels from the MSR and MSC cohorts. This model was based on a priori knowledge [1, 22, 23] of DICOM acquisition parameter values ("Echo Time," "Repetition Time," "Flip Angle" (0018,1314), "Inversion Time" (0018,0082), "Scanning Sequence" (0018,0020), and "Contrast Bolus Agent") used to scan with specific image contrast weighting [24]. The rule-based approach was developed in-house in Python using Pydicom [25] to extract DICOM header attributes.

Metadata-Only Model Development

A set of DICOM metadata attributes, was extracted from the datasets using the Pydicom python package [25] and is listed in Table 2. The majority of the attributes were taken from the DICOM "MR Image Module" [26] and relate to the physics of the MR acquisition that could impact tissue contrast. Other DICOM tags relating to the image dimensionality and resolution were included to help differentiate some lower resolution functional imaging series from higher-resolution structural images. The "Contrast Bolus Agent" tag was included to assist with identification of T1C images, though the field is not used consistently. "SOP Class UID" was included as a convenience to filter DICOM "Secondary Capture" [27] objects during preprocessing with these image types automatically labeled as "NA." Several engineered features representing properties of each series ("Number Of Files," "Number of Images," "Number of Volumes") were added to assist with differentiation of functional and multi-volume acquisitions (e.g., perfusion imaging [28, 29]). Missing string-type attributes were replaced with "None" and empty numeric attributes were replaced with the mean of the feature in the training data. The majority of these attributes were numeric; those that were string-type features were hashed to numeric values using the SHA256 algorithm in the Python hashlib library [30]. This was done to automate the mapping of string-based features representing enumerated values (e.g., "Scanning Sequence") to numeric feature values amenable to modeling. This permitted new enumerated attributes to be added without requiring an explicit mapping and was considered important for creating a model training strategy that could be extended to other modalities or anatomical domains. Support vector machine (SVM) and random forest (RF) models using these features were developed in the scikit-learn python package (svm.LinearSVC, svm.SVC, ensemble.RandomForestClassifier) [31, 32]. Fivefold cross-validation (model_selection.cross_val_score [33]) on the training data was used to evaluate which algorithm was best suited for predicting MRI contrast from DICOM metadata.

For RF models, randomized cross validation (RandomizedSearchCV in sklearn) [33] was used to search for the optimal set of hyperparameters (n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, bootstrap) using the training set only. An RF using all training data (instead of 4/5) was retrained using the optimal hyperparameters returned from the search. Impurity-based feature importance scores derived from the trained RF are biased toward high-cardinality features, and because of this the "permutation importance" function (inspection.permutation_importance) [34] in sklearn was used to calculate the relative feature importance for the training, validation, and test sets. Briefly, permutation importance is defined as the decrease in a model score function when a single-input feature is shuffled [35]. Features were permuted five times. Finally, the algorithm was tested on the validation and test MSC and GR datasets. This algorithm is referred to as the "metadata only" model.

Imaging-Based Model Development

Image-based methods were trained using only the two-dimensional center slice of each volume as input. The original unprocessed DICOM slice passing through the center DICOM LPS location was used as the center slice of the volume. During training, the center slice was transformed with random horizontal and vertical flips ($p = 0.5$), random affine rotations and translations, and slight alterations to brightness, contrast, saturation, and hue and resized to 224×224. A ResNet-50 convolutional neural network (CNN) architecture [36, 37], pre-trained on ImageNet [38], was chosen to initialize the model weights. Data were normalized using the ImageNet normalization means and standard deviations [39]. The final layer of the ResNet-50 was replaced with a fully connected layer with 6 outputs representing the 6 contrast categories. A cosine differential learning rate with a maximum value of 0.0003 was used together with a weight decay coefficient of 0.0001. The model was trained for 40

Table 2 Metadata-only model feature importances for training, validation, and test sets. Derived features such as NumberOfVolumes has DICOM tag = None

Metadata only					
Rank	***DICOM tag***	***Name***	***Feature importance—train***	***Feature importance—valid***	***Feature importance—test***
0	(0018,0010)	ContrastBolusAgentBinarized	0.01199	0.17513	0.16201
1	(0018,0081)	EchoTime	0.10565	0.04165	0.03331
2	(0018,0082)	InversionTime	0.00128	0.02531	0.02608
3	(0018,0080)	RepetitionTime	0.00074	0.01325	0.01304
4	(0018,0095)	PixelBandwidth	0.00060	0.00618	0.01049
5	(0020,0011)	SeriesNumber	0.00022	0.00088	0.00567
6	(0018,1314)	FlipAngle	0.00047	0.00530	0.00567
7	(0018,0020)	ScanningSequence	0.00043	0.00780	0.00454
8	(0018,0091)	EchoTrainLength	0.00014	0.00338	0.00241
9	(0018,0093)	PercentSampling	0.00005	−0.00044	0.00184
10	(0018,0022)	ScanOptions	0.00046	0.00162	0.00113
11	(0018,0088)	SpacingBetweenSlices	−0.00019	0.00074	0.00099
12	(0018,0089)	NumberOfPhaseEncodingSteps	0.00000	0.00000	0.00085
13	(None)	NumberOfFiles	−0.00005	0.00000	0.00071
14	(0018,0094)	PercentPhaseFieldOfView	0.00011	0.00044	0.00057
15	(0018,0025)	AngioFlag	0.00000	0.00000	0.00057
16	(0018,0083)	NumberOfAverages	0.00019	0.00000	0.00028
17	(0018,0050)	SliceThickness	−0.00002	−0.00029	0.00028
18	(0020,1002)	ImagesInAcquisition	0.00006	0.00088	0.00028
19	(0018,0086)	EchoNumbers	0.00002	0.00015	0.00014
20	(0018,0024)	SequenceName	0.00000	0.00074	0.00000
21	(0018,0087)	MagneticFieldStrength	0.00000	0.00044	0.00000
22	(0018,1310)	AcquisitionMatrix	0.00002	0.00000	0.00000
23	(0008,0016)	SOPClassUID	0.00002	0.00000	0.00000
24	(0018,1251)	TransmitCoilName	0.00000	0.00000	0.00000
25	(0018,0085)	ImagedNucleusQuantized	0.00000	0.00000	0.00000
26	(None)	NumberOfVolumes	−0.00002	0.00000	0.00000
27	(0018,0021)	SequenceVariant	0.00006	0.00059	0.00000
28	(0018,0023)	MRAcquisitionType	0.00003	−0.00029	0.00000
29	(0018,0015)	BodyPartExamined	0.00000	0.00000	-0.00014
30	(0028,0030)	PixelSpacing	0.00014	0.00000	-0.00043
31	(0018,1312)	InPlanePhaseEncodingDirection	−0.00003	0.00029	-0.00043
32	(None)	NumberOfImagePositions	0.00014	−0.00059	-0.00043
33	(0018,0084)	ImagingFrequency	0.00017	0.00074	-0.00128
34	(0028,0011)	Columns	0.00002	0.00074	-0.00213
35	(0028,0010)	Rows	0.00014	0.00088	-0.00227

epochs, but early stopping was employed such that the highest accuracy model on the validation dataset was saved. All deep learning experiments were implemented using PyTorch 1.9. This model is referred to as the "imaging only" model.

MRI exams may comprise different anatomies and even nominally neurological exams may include spinal cord images. The DICOM "Body Part Examined" (0018,0015) tag is unreliable with only 21% of MSC series containing this tag. Of the images that were not brains, based on visual inspection, 23% contained this DICOM tag. Of those non-brain images with this DICOM tag, 64% were incorrectly labeled as "BRAIN" or "HEAD" though primarily comprising spinal anatomy. A binary classifier was therefore also developed to assist with preliminary selection of brain vs. not-brain images, which was of primary relevance to the motivating use cases. Briefly, 2362 exams (21,114 series) from the above cohorts were manually labeled as "BRAIN" or "OTHER" by visual inspection as described above. The center slices from each series were transformed with random horizontal and vertical flips ($p = 0.5$), random affine

Table 3 Final model comparison results (MS clinical dataset, MSC)

	Cohort: MSC data only							
	Overall accuracy		*Testing class accuracy*					
Classifier	**Validation**	**Test**	**T1**	**T1C**	**T2**	**T2 FLAIR**	**PD**	**Other**
Rule based	73.9%	74.5%	55.7%	47.5%	33.6%	99.0%	62.1%	93.7%
Metadata RF	94.0% (94.5%)	95.4% (95.6%)*	97.3%	95.5%	98.4%	100.0%	72.0%	91.4%
Imaging CNN	97.4% (98.0%)	96.9% (97.1%)	99.1%	95.5%	100.0%	95.6%	88.0%	96.9%
Imaging CNN + metadata RF	97.7% (97.9%)	97.5% (97.7%)	99.1%	96.8%	99.2%	99.5%	88.0%	95.6%

*Accuracy after correcting mislabeled cases

rotations, resized to 224 × 224, and normalized by mean centering and standard deviation scaling. A ResNet-50 convolutional neural network (CNN) architecture [36, 37], pretrained on ImageNet [38], was chosen to initialize the model weights. The model achieved 99% accuracy on the validation set. This model was used for inference during preprocessing to limit analysis to brain images.

Model Analysis and Evaluation

Each center MR image was sent through the trained contrast classifier CNN network and the final 6 logit outputs were saved before application of the softmax function. First, t-distributed stochastic neighbor embedding (t-SNE) [40] was performed on these final 6 features using scikit-learn [41]. Regions of the t-SNE clusters were visually investigated in order to assess 1) whether there were obvious visual differences between images among regions within the same cluster and 2) whether the misclassified images had visual similarities to their neighbors. The final 6 logit values were also combined together with the metadata features using a RF classifier as described above. The new RF model was refitted on the training data, resulting in a combined imaging plus metadata-based machine learning model. This model is referred to as the "combined" model. A command line Python utility was developed for using the trained models to infer contrast classifications prospectively as part of image processing pipelines. The application takes the path to a DICOM exam as input and returns the predicted contrast classification for each series, or "NA" for non-imaging series, for example, those containing "Secondary Capture" DICOM images. The utility is able to run on either GPU or exclusively CPU-enabled hosts.

Results

Rule-Based Classification Results

The rule-based approach utilizing a priori knowledge [1, 22, 23] of acquisition parameters obtained from DICOM attributes achieved 73.9%, 74.5%, 81.8%, and 78.2% accuracies on MSC validation, MSC test, GR validation, and GR test sets, respectively (Tables 3 and 4). Overall and per-class accuracy for the rule-based model are listed in Tables 3 and 4; notably, there is no "training," but training, validation, and test sets are separated to serve as a comparison point for the following models that require training.

Modeling Results

The main objective of the present work was to obtain high-accuracy classification on the MSCdataset as it represents a real-world multi-site clinical data set. Table 3 presents validation and test set accuracies as well as per-class accuracies for the test set for the metadata-only, imaging-only, and combined models. The models that included imaging features

Table 4 Final model comparison results (Glioma Research Dataset, GR)

Cohort: glioma (GR)								
	Overall accuracy		*Testing class accuracy*					
Classifier	**Validation**	**Test**	**T1**	**T1C**	**T2**	**T2 FLAIR**	**PD**	**Other**
Rule based	81.8%	78.2%	88.2%	97.6%	13.8%	95.0%	NA	100.0%
Metadata RF	99.7%	99.7%	100.0%	100.0%	100.0%	100.0%	NA	50.0%
Imaging CNN	98.4%	98.4%	100.0%	100.0%	92.2%	100.0%	NA	100.0%
Imaging CNN + metadata RF	94.7%	94.1%	100.0%	100.0%	71.0%	100.0%	NA	100.0%

outperformed the metadata-only models in validation and test set accuracies (> 97% compared to ~ 95%).

Metadata Modeling Results

The metadata-only modeling strategy utilized an RF based on the accuracy from both fivefold cross validation compared with SVM and subsequent RF hyperparameter search experiments. The final RF model was trained with the following input parameters [31]: number of estimators = 450, minimum number of samples per split = 2, minimum number of samples in each leaf = 4, maximum number of features = sqrt, maximum depth of trees = 66, and bootstrapping on. Permutation importance calculations for each feature are listed in Table 2 and depicted in Fig. 2. The feature importance calculations suggest that removing the "ContrastBolusAgentBinarized" field decreased accuracy by 1.2%, 17.6%, and 16.2% (training, validation, test), while removing "EchoTime" decreased the accuracy by approximately 10.6%, 4.1%, and 3.3%, respectively. "InversionTime" and "RepetitionTime" were increasingly important in the validation and test sets compared with the training set (Table 2). Note that permutation analysis may be limited in cases of correlated features, where the permutation of only one of the correlated features may reflect a misleadingly low importance

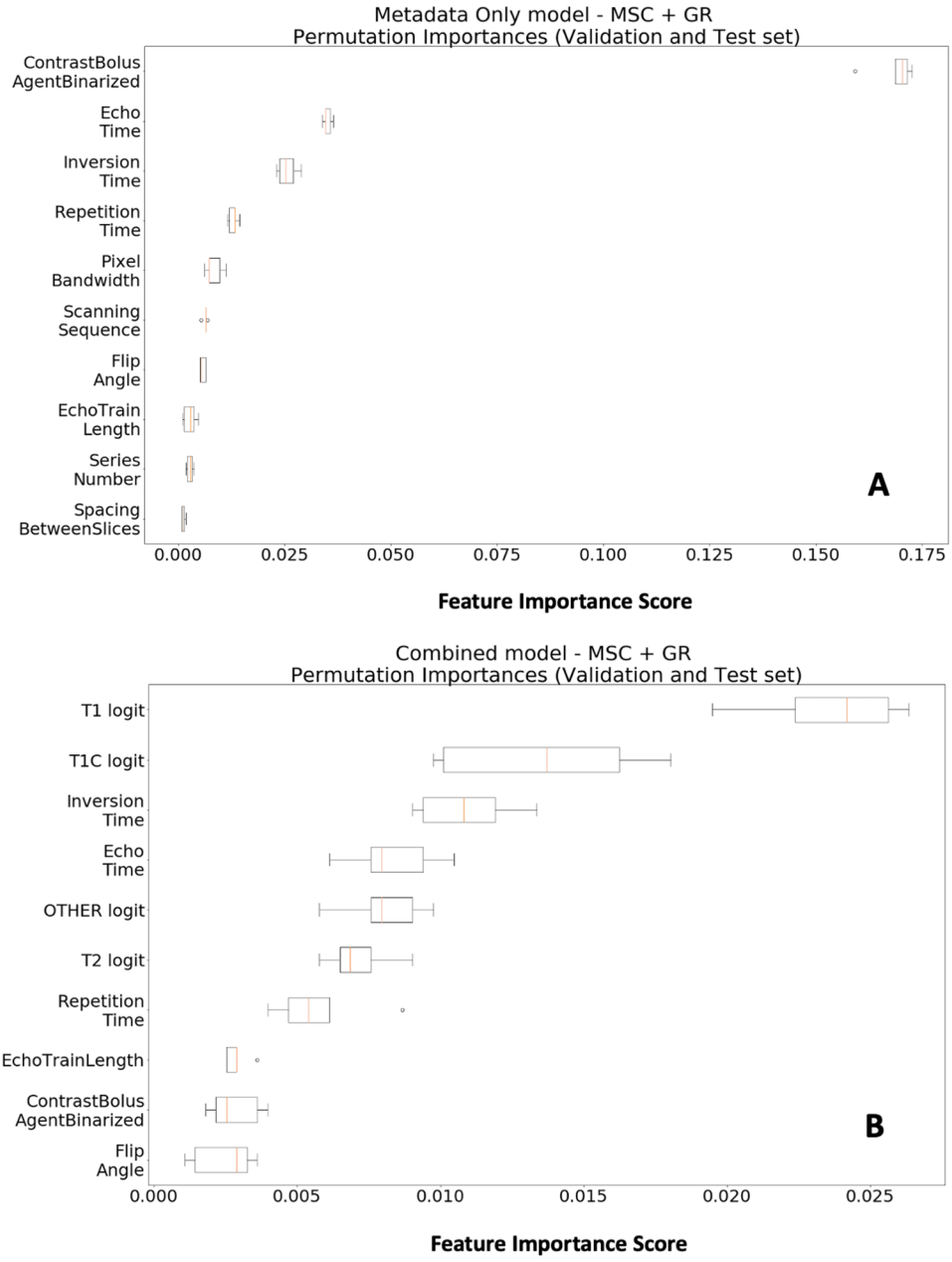

Fig. 2 RF feature importance graphs on validation and test sets. The permutation feature importance was calculated for the random forest experiments. **A** The metadata-only model feature importance calculations depict that removal of the contrast bolus agent feature decreased the accuracy between 15 and 17.5%, while removal of the next most important feature—echo time—decreased the accuracy by approximately 3%. **B** The combined model feature importance calculations depict that when imaging features are added, they decrease the importance of the contrast bolus agent feature. It can be inferred that the T1 and T1C logits, now the two most important features, contain the information that the contrast bolus agent feature contained metadata-only model in (A)

[35]. The final result on the MSC dataset achieved 99.7%, 94.0%, and 95.4% on the training, validation, and test set, respectively (Table 3). The most common mistake made on the MSC dataset was the incorrect classification of images containing artifacts (due to motion, Gibbs ringing, or aliasing) that were labeled as OTHER instead of their original acquisition contrast. This was not surprising given that the algorithm had no access to pixel-level imaging data. The next most common mistake was classifying T1 as T1C and vice versa. Upon inspection of all misclassified MRI series, a few series were found to have the incorrect ground truth label but were classified into their correct contrast by the classifier. After adjusting for the incorrect ground truth labels, the MSC validation and test set accuracies increased to 94.5% and 95.6%, respectively. On the GR dataset, the metadata-only model performed better (99.7% validation, 99.7% test) than models that included imaging data, with only 2 misclassified images in total: one from each of the GR validation and test sets (Table 5).

Image-Based Modeling Results

The imaging-only CNN achieved 2.5% higher average accuracy over the MSC validation and test sets compared to the metadata-only model, with a final MSC validation and test set accuracy of 97.4% and 96.9%. Similarly to the metadata model, the most common mistake was misclassifying T1 for T1C and vice versa; this misclassification is visualized in the t-SNE analysis (Fig. 3E). Misclassification analysis revealed that the CNN properly identified 7, 6, and 1 (train, validation, and test) images that had been assigned an incorrect ground truth label, a much higher number than the metadata-only or combined models detected (Table 6). When adjusting for the proper identification of these images, the imaging-only CNN achieved 98.0% and 97.1% on the validation and test MSC set. Although the imaging-only CNN performed worse than the metadata-only GR dataset, classifying 10 MRI series incorrectly (5 each from the GR validation and test set), all of these mistakes were on sagittal T2-weighted images obtained with higher contrast than what was seen by the network in training (Fig. 3I).

Combined Modeling Results

The combined imaging and metadata model consisted of a trained RF using the features of the metadata-only model together with the 6 logit values outputted from the image-based model CNN. Compared with metadata-only and imaging-only models, the combined model performed the best on the MSC dataset achieving 97.7% and 97.5% on validation and test data, respectively. When adjusting

Table 5 Misclassification analysis of GR where: (1) Correct indicates cases where the model predicted the correct label for a series that had its ground truth mislabeled; (2) Artifact indicates cases that were misclassified due to severe imaging artifacts; (3) OTHER as original contrast indicates cases that were intentionally labeled OTHER due to severe imaging artifact that we would not want the classifier to deliver downstream to clinicians; (4) "High-res, high-contrast, 3D T2 as other" indicates specific acquisition types labelled as T2 that were classified as OTHER

		Misclassification analysis: GR				
		Correct	*Artifact*	*Other as original contrast*	*High-res, high-contrast, 3D T2 as other*	*Unknown reason*	*Total*
Metadata RF	*Valid*	0	0	0	0	1	1
	Test	0	0	0	0	1	1
Imaging CNN	*Valid*	0	0	0	5	0	5
	Test	0	0	0	5	0	5
Imaging CNN + metadata RF	*Valid*	0	0	0	16	0	16
	Test	0	0	0	18	0	18

Table 6 Misclassification analysis (MSC) where: (1) Correct indicates cases where the model predicted the correct label for a series that had its ground truth mislabeled; (2) Artifact indicates cases that were misclassified due to severe imaging artifacts; (3) OTHER as original contrast indicates cases that were intentionally labeled OTHER due to severe imaging artifact that we would not want the classifier to deliver downstream to clinicians; (4) Bad slice indicates that the incorrect slice was chosen (top or bottom of head)

		Misclassification analysis: MSC					
		Correct	*Artifact*	*OTHER as original contrast*	*Bad slice*	*Unknown reason*	*Total*
Metadata RF	*Train*	2	0	16	0	18	36
	Valid	5	0	12	0	46	63
	Test	2	0	17	0	32	51
Imaging CNN	*Train*	7	3	0	3	14	27
	Valid	6	1	0	10	10	27
	Test	1	5	0	5	23	34
Imaging CNN + metadata RF	*Train*	6	0	0	0	4	10
	Valid	2	3	0	7	12	24
	Test	2	7	0	4	15	28

for the misclassified MSC data, the accuracy increased to 97.9% and 97.7% for the validation and test set, which was comparable to the imaging-only model. For the combined model, the T1, T1C, T2, and OTHER logits were the most important features, while the importance of "EchoTime" and "ContrastBolusAgentBinarized" was diminished (Table 7). Interestingly, the accuracy dropped significantly for the GR validation and test sets with 32 out of 34 misclassifications associated with high-resolution, high-contrast 3D T2-weighted images classified as OTHER instead of T2 weighted (Fig. 3I).

t-SNE results

Figure 3A depicts the results of the t-SNE analysis. In order to evaluate whether there were visual differences among regions within the same cluster, the coordinates of the t-SNE were used to visualize images from various subregions. Given that a catch-all category of OTHER was used to represent less-common acquisitions, it was of great interest to explore how the subregions of this group are separated by acquisition type. Figure 3B–D depicts representative images corresponding to each subregion. Region B is largely composed of axial gradient-echo echo-planar images with severe spatial distortion artifacts. HARDI diffusion tensor images make up region C, while susceptibility-weighted images (SWI) make up region D. An additional region proximate to region D represented a cluster of high resolution, high contrast T2*-weighted images (not pictured).

T-SNE analysis also allowed the visualization of maximally separated regions in each cluster. Regions G and H both correspond to the cluster of T1-weighted images containing a mixture of sagittal and axial images that had no obvious difference in gray-white matter contrast. Regions I, J, and K corresponded to the T2-weighted cluster. Region I, which protruded from the main cluster shape near the T1C cluster, contained the high-resolution, high-contrast 3D T2-weighted images from the GR cohort that were classified incorrectly in the prior analyses (Table 5). Although regions J and K were overall similar, region K was mostly axial while region J also contained sagittal images. The presence of non-yellow points in region M of the T2-FLAIR cluster (e.g., blue T2, top left of cluster) contained a mix of sagittal, axial, and coronal images—many with extensive pathology, compared to the adjacent highly uniform axial T2-FLAIR images found in region L. On the contrary, regions N and O of the PD cluster had no striking visual differences despite being maximally separated.

To complement the analysis of within-cluster differences, the overlapping region of the T1 and T1C clusters was examined to answer whether the misclassified images had visual similarities to their neighbors. Examples from

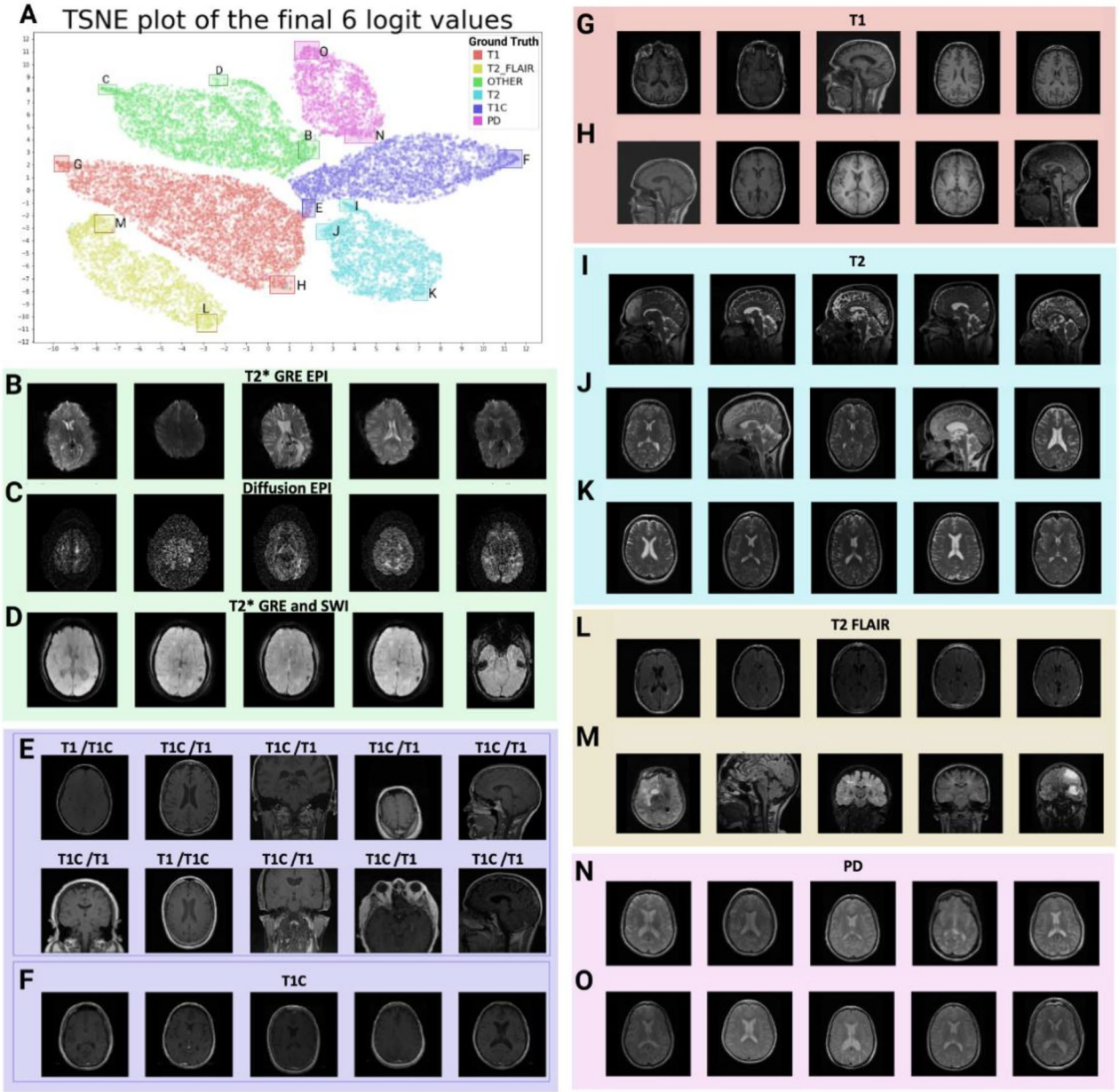

Fig. 3 t-SNE of the 6 logits derived from the final layer of the convolutional neural network. **A** t-Stochastic neighbor embedding of the logits output from the CNN. **B** Examples of images and series descriptions that correspond with the region B on the t-SNE plot, represented largely by gradient-echo echo-planar images. **C** HARDI diffusion volumes largely representing region C on the t-SNE plot. **D** Susceptibility-weighted images corresponding to region D on the t-SNE plot. **E** Misclassified T1 and T1C images in region E that appear similar to the contrast of the other labeled as truth/prediction. **F** In contrast to region E, region F contains highly uniform axial T1C images. **G–H** images corresponding to regions G and H, respectively. Though maximally separated, these regions both contain axial and sagittal images that appear similar in contrast. **I** Region I contains high-contrast, high-resolution 3D T2-weighted images that have their own distinct area within the T2-weighted cluster. **J**, **K** Region J contains both axial and sagittal images with varied contrast compared with region K which appears more uniform. **L** Examples of images located in region L depicting uniform, axial T2-FLAIR images with little evidence of pathology. **M** Examples of images located in region M of the yellow cluster, representing coronal, sagittal, and axial T2-FLAIR images with extensive pathology. **N**, **O** This cluster of PD images does not appear different when comparing maximally separated regions N and O

region E demonstrate that many of the T1 images predicted as T1C had low gray-white matter contrast, typical of T1C images. In addition, T1C images misclassified as T1 had very little or no evidence of contrast enhancement in the center slice, and therefore retained their gray-white matter contrast even in the post-contrast setting.

Table 7 Combined model (metadata + imaging) feature importances for training, validation, and test sets

Metadata + imaging combined model					
Rank	***DICOM tag***	***Name***	***Feature importance—train***	***Feature importance—valid***	***Feature importance—test***
0	None	T1_logit	0.0020	0.0311	0.0214
1	None	T1C_logit	0.0003	0.0141	0.0137
2	(0018,0082)	InversionTime	0.0000	0.0088	0.0119
3	(0018,0081)	EchoTime	0.0003	0.0066	0.0089
4	None	OTHER_logit	0.0013	0.0091	0.0044
5	(0018,0080)	RepetitionTime	0.0002	0.0065	0.0044
6	(0018,0010)	ContrastBolusAgentBinarized	0.0002	0.0012	0.0044
7	None	T2_logit	0.0003	0.0121	0.0037
8	(0018,0095)	PixelBandwidth	0.0002	0.0012	0.0026
9	(0018,0091)	EchoTrainLength	0.0000	0.0031	0.0021
10	None	T2FLAIR_logit	0.0000	0.0022	0.0020
11	(0028,0010)	Rows	0.0000	0.0003	0.0020
12	None	PD_logit	0.0005	0.0010	0.0018
13	(0018,0020)	ScanningSequence	0.0000	0.0022	0.0018
14	(0018,1314)	FlipAngle	0.0000	0.0028	0.0016
15	(0018,1312)	InPlanePhaseEncodingDirection	0.0000	0.0007	0.0011
16	(0028,0011)	Columns	0.0001	0.0003	0.0006
17	(None)	NumberOfImagePositions	0.0000	0.0004	0.0004
18	(0018,0023)	MRAcquisitionType	0.0000	0.0001	0.0004
19	(0018,0021)	SequenceVariant	0.0001	0.0006	0.0003
20	(None)	NumberOfVolumes	0.0000	0.0000	0.0003
21	(0018,0094)	PercentPhaseFieldOfView	0.0000	0.0003	0.0001
22	(0018,0086)	EchoNumbers	0.0000	0.0000	0.0001
23	(0018,0050)	SliceThickness	0.0001	0.0003	0.0000
24	(0018,0015)	BodyPartExamined	0.0001	0.0000	0.0000
25	(0018,0089)	NumberOfPhaseEncodingSteps	0.0001	0.0000	0.0000
26	(0018,0024)	SequenceName	0.0001	0.0000	0.0000
27	(0020,0011)	SeriesNumber	0.0001	0.0000	0.0000
28	(0018,0084)	ImagingFrequency	0.0001	0.0000	0.0000
29	(0018,0087)	MagneticFieldStrength	0.0001	0.0000	0.0000
30	(0018,1310)	AcquisitionMatrix	0.0000	0.0000	0.0000
31	(0018,0025)	AngioFlag	0.0000	0.0000	0.0000
32	(0018,0085)	ImagedNucleusQuantized	0.0000	0.0000	0.0000
33	(0020,1002)	ImagesInAcquisition	0.0000	0.0000	0.0000

Table 7 (continued)

Metadata + imaging combined model					
Rank	*DICOM tag*	*Name*	*Feature importance—train*	*Feature importance—valid*	*Feature importance—test*
34	(0008,0016)	SOPClassUID	0.0000	0.0000	0.0000
35	(0018,1251)	TransmitCoilName	0.0000	0.0000	0.0000
36	(0018,0083)	NumberOfAverages	0.0001	−0.0006	0.0000
37	(None)	NumberOfFiles	0.0001	0.0001	0.0000
38	(0018,0022)	ScanOptions	0.0000	0.0003	-0.0001
39	(0018,0093)	PercentSampling	0.0001	−0.0006	-0.0001
40	(0018,0088)	SpacingBetweenSlices	0.0001	0.0003	-0.0004
41	(0028,0030)	PixelSpacing	0.0000	0.0001	-0.0006

Compared with region E, images from region F were more uniform in appearance with similar contrast and axial orientation.

Discussion

Being able to automatically identify and quantify heterogeneously acquired MRI images will be central to delivering imaging-based precision medicine to the point of care in the real world. In this study, multiple algorithms were developed to predict the contrast mechanism of diverse MRI series with > 97% accuracy. These tools were developed to satisfy a real institutional clinical objective for the UCSF *Bridge* application [42]. *Bridg*e is a web-based application that displays longitudinally aligned MR images acquired with similar contrast to clinicians alongside other pertinent clinical metrics, biomarkers, and therapy history [43] (Fig. 4B). The MRI series contrast classifiers presented in this study represent an important step in the image analysis pipeline, after classifying the MRI series by the anatomical region and before longitudinal image registration [44, 45]. By streamlining the presentation of longitudinal imaging and presenting it in a comprehensive personalized dashboard with statistical context, the tool aims to facilitate the patient-clinician consultation and discussion of treatment options. Misclassification of the contrast of a particular MRI series within an exam or selection of a low-quality image can lead to inaccurate interpretation of the imaging data. Retrieving high-quality images of the brain that were acquired with the contrast of interest is crucial for achieving accurate, high-throughput data preparation.

This work overcomes some of the limitations in prior work through requiring 1) only a single slice from each patient which can be acquired in any orientation: coronal, sagittal, axial, or oblique; 2) the center for image-based computation rather than the entire 3D volume, yielding computational advantages over models requiring 3D or multi-slice data; 3) ground-truth classification determined by a process of automatic rule-based derivation of weak labels, followed by visual review and correction by multiple reviewers; 4) inclusion of an additional "OTHER" category that does not force the algorithm to return the fixed set of contrast types. The performance of the following models was compared to 1) a rule-based classifier, 2) a metadata-only RF classifier, 3) an imaging-only convolutional neural network, and 4) a combined model that uses both imaging data outputs from (3) and metadata in a RF. The primary goal was to obtain high accuracy on the MSC dataset, a heterogeneous clinical cohort, while the secondary goal was to obtain good classification accuracy on the GR dataset, containing images with more extensive pathological burden. All three trained algorithms vastly improved the classification performance over the rule-based approach, with models including imaging

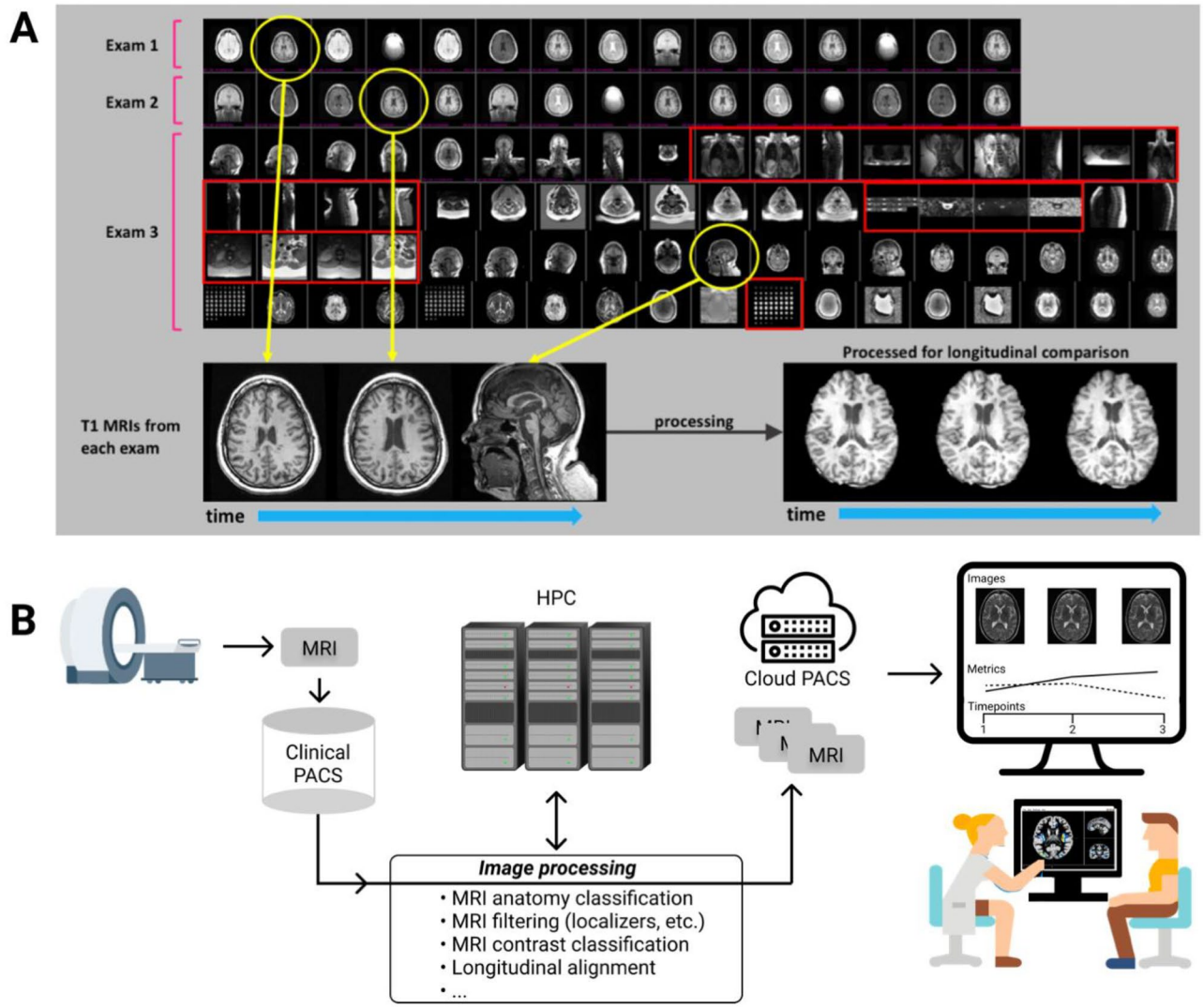

Fig. 4 The application of contrast classification in MRI. **A**, top: Representative longitudinal exams from a single patient retrieved from the clinical PACS. Exams 1–3 consist of different numbers and types of acquisition sequences and even different anatomical regions. Acquisitions indicated in red are spine images despite being labeled as brain in the DICOM headers. Typical downstream applications require identification of input images acquired with specific tissue contrast. Images circled in yellow represent T1-weighted images from each exam used as input to a downstream application. **B**, bottom: Representative downstream application to align longitudinal MRI exams for visual review

(imaging-only and combined models) performing the best on the MSC validation and test sets.

The imaging-only and combined model were comparable in their performance on the MSC dataset, with the imaging-only model achieving 98.0% validation and 97.1% test set accuracy, and the combined model achieving 97.9% validation and 97.7% test set accuracy, respectively (Table 3). The combined model improved the classification of T2-FLAIR images in the test set and decreased the number of T1 pre- and post-contrast mistakes made by the algorithm. The imaging-only CNN improved upon the classification of the OTHER category compared with the combined model for the MSC dataset. With the perspective of delivering the image contrast of interest for the MSC dataset, the combined model outperformed the imaging-only model.

Conducting a t-SNE analysis (Fig. 3) helped shed light on why certain groups of images were misclassified by the imaging-only CNN. The between-cluster t-SNE analysis of the T1 and T1C images in region E of Fig. 2 confirmed that the CNN misclassified T1 and T1C images that were visually similar to one another. The canonical features that visually differentiated T1C from T1 images were regions of bright contrast enhancement and lower gray-white matter contrast. Visual observation of misclassified cases suggested that the T1C images that were misclassified as T1 by the CNN were those that retained their dynamic range

and gray-white matter contrast due to little or no enhancement. Within-cluster t-SNE analysis of the OTHER cluster of images (Fig. 3B–D) demonstrated that the resulting logit values from the imaging-only CNN model contained a rich diversity of imaging features that were able to numerically differentiate different kinds of images within the same label without explicitly defining them during training. The observation that the imaging-only CNN resulted in spatial clustering of the HARDI, SWI, and GRE EPI image types within the t-SNE analysis suggests the feasibility of classifying these additional acquisition contrasts that would need to be investigated in further studies.

The generalizability of the models to images from a different disease domain which contained a more extensive disease burden was examined by testing performance of the present models on a glioma research imaging dataset. The best performing algorithm was the metadata model achieving 99.7% on both validation and test GR datasets, classifying only one image incorrectly in each of the GR validation and test sets. This is likely due to the strict research protocol that resulted in both homogeneous acquisition of the images and consistent labeling of metadata. Contrary to expectations, the combined model had lower accuracy on the GR dataset due to the classification of a set of high-resolution 3D T2-weighted images acquired with a BrainLab protocol as OTHER (Fig. 3I). These high-resolution T2-weighted images with much longer TR than the T2-weighted images seen in training (2912.5 ms vs. 2370.9 ms) which drastically changed the image contrast were not included in the training dataset. The imaging-only CNN segregates these images into their own small, separated section within the T2-weighted cluster. Taken together, the information stored in the imaging-based logits, the TR difference contained in the metadata, and the lack of similar acquisitions in the training data likely all contributed to the combined RF failing to segregate these specific images from the rest of the T2-weighted images. This underscores the risks of inferring on data with characteristics not present in the training set. Before deploying these models for clinical imaging of patients with glioma, the model would need to be retrained with examples of these high-resolution, T2-weighted 3D images and all three algorithms tested on clinical glioma data stored in an institutional PACS, where it is expected that imaging-based models would perform at least comparably with metadata-only models, similar to the MSC data.

Though the combined model only slightly outperformed the imaging-only CNN on the MSC dataset, it was clear that the metadata features that describe the acquisition parameters were highly valuable for classifying the acquisition and contrast. However, there also were advantages to using an imaging-only model in some contexts. Compared with the combined model, the imaging-only model performed comparably on the MSC data while increasing the accuracy on the GR dataset, due to generalizing better to the high-contrast, high-resolution 3D T2-weighted images that it had not seen before. The number of misclassified images in the training data that were, in fact, labeled incorrectly was also increased for the imaging-only model compared with the others, indicating it was more robust to overfitting to the training set. Compared with the metadata-only model, the imaging-only CNN was able to identify series that were impacted by severe artifacts as OTHER instead of the contrast that it was acquired with, which is beneficial in the target MS pipeline. However, it may be the case that in other applications, a T1 image delivered with substantial artifacts is preferable to no T1 image selected at all for a timepoint; in that case, this would be a disadvantage. Though using imaging data is more computationally expensive compared with metadata alone, the inference time difference is nominal in the context of the Bridge application since the only preprocessing step required is calculating the center of the image volume to identify the slice most representative of the image contrast. Though metadata-only models shorten inference time at scale, they have lower overall accuracy and will never be able to filter out MRI sequences with artifacts, which is essential for preventing downstream issues with alignment and display. Therefore, for use cases that require delivering high-quality images to an application, an imaging-only CNN model is advantageous for contrast classification.

Though the heterogeneous clinical data in the MSC dataset was visually reviewed and labeled, it is still subject to human labeling error. The authors acknowledge that multiple reviewers for each sequence would have likely increased data-label fidelity and potentially improved model performance. In addition, though the model was trained on clinical PACS data derived from many institutions, there may still be bias as to where those prior imaging centers were located and the machines those data were acquired on. Therefore, it may be the case that this model would perform worse on data derived from machines or institutions with lesser representation in the MSC dataset.

Conclusion

In this study, different modeling paradigms for classifying MRI series based on their acquisition contrast were developed and evaluated. Classification accuracies sufficient for deployment within the pipeline detailed in Fig. 4B were achieved for the classification of clinical data for both the Bridge application and a separate MRI dataset of patients with glioma with more extensive disease that was not seen in training. The t-SNE analysis suggests that the imaging-based CNN might possess power to differentiate additional contrasts such as ADC maps, HARDI, and T2*-weighted

images. This could suggest generalization power for a broad range of applications. Future directions will aim to explore this question by extending the present models to additional acquisition types (e.g., high-resolution, high-contrast 3D T2-weighted images) and anatomy through retraining and validation with supplementary datasets Given that robust automated data selection is a critical preliminary step in many high-throughput clinically deployed inference pipelines and also for content-based retrieval in large-scale computational health studies, such methods hold promise for multiple applications.

Acknowledgements Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Funding Julia Cluceru was supported in part by T32 grant, P01CA118816. Riley Bove is supported by a National Multiple Sclerosis Society Harry Weaver Award.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Bitar, R. *et al.* MR pulse sequences: what every radiologist wants to know but is afraid to ask. *Radiographics* **26**, 513–537 (2006).
2. Nishimura, D. Chapter 1. in *Principles of Magnetic Resonance Imaging* (2010).
3. NEMA. DICOM. http://medical.nema.org/. Accessed 30 Nov 2021.
4. Plewes, D. B. The AAPM/RSNA physics tutorial for residents. Contrast mechanisms in spin-echo MR imaging. *Radiographics* **14**, 1389–404; quiz 1405 (1994).
5. Nitz, W. R. & Reimer, P. Contrast mechanisms in MR imaging. *Eur. Radiol.* **9**, 1032–1046 (1999).
6. Wen, P. Y. *et al.* Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* **28**, 1963–1972 (2010).
7. Willemink, M. J. *et al.* Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
8. Gauriau, R. *et al.* A Deep Learning-Based Model for Detecting Abnormalities on Brain MRI for Triaging: Preliminary Results from a Multi-Site Experience. *Radiology: Artificial Intelligence* e200184 (2021). https://doi.org/10.1148/ryai.2021200184.
9. Akgül, C. B. *et al.* Content-based image retrieval in radiology: current status and future directions. *J. Digit. Imaging* **24**, 208–222 (2011).
10. Kumar, A. *et al.* Adapting content-based image retrieval techniques for the semantic annotation of medical images. *Comput. Med. Imaging Graph.* **49**, 37–45 (2016).
11. Kumar, A., Kim, J., Cai, W., Fulham, M. & Feng, D. Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. *J. Digit. Imaging* **26**, 1025–1039 (2013).
12. Gauriau, R. *et al.* Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. *J. Digit. Imaging* **33**, 747–762 (2020).
13. Remedios, S., Roy, S., Pham, D. L. & Butman, J. A. Classifying magnetic resonance image modalities with convolutional neural networks. in *Medical Imaging 2018: Computer-Aided Diagnosis* (eds. Mori, K. & Petrick, N.) 89 (SPIE, 2018). https://doi.org/10.1117/12.2293943.
14. Gai, N. D. Highly Efficient and Accurate Deep Learning-Based Classification of MRI Contrast on a CPU and GPU. *J. Digit. Imaging* **35**, 482–495 (2022).
15. Ranjbar, S. *et al.* A deep convolutional neural network for annotation of magnetic resonance imaging sequence type. *J. Digit. Imaging* **33**, 439–446 (2020).
16. Pizarro, R. *et al.* Using deep learning algorithms to automatically identify the brain MRI contrast: implications for managing large databases. *Neuroinformatics* **17**, 115–130 (2019).
17. University of California, San Francisco MS-EPIC Team: *et al.* Long-term evolution of multiple sclerosis disability in the treatment era. *Ann. Neurol.* **80**, 499–510 (2016).
18. Cluceru, J. *et al.* Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro Oncol.* **24**, 639–652 (2022).
19. Cluceru, J. *et al.* Recurrent tumor and treatment-induced effects have different MR signatures in contrast enhancing and non-enhancing lesions of high-grade gliomas. *Neuro Oncol.* **22**, 1516–1526 (2020).
20. ADNI. http://adni.loni.usc.edu. Accessed 30 Nov 2021.
21. MRI Scanning. http://mriquestions.com/what-are-the-steps.html. Accessed 30 Nov 2021.
22. Jung, B. A. & Weigel, M. Spin echo magnetic resonance imaging. *J. Magn. Reson. Imaging* **37**, 805–817 (2013).
23. mri pulse sequence parameters. https://radiopaedia.org/articles/mri-sequence-parameters. Accessed 30 Nov 2021.
24. Cluceru, J. *et al.* Automatic Classification of MR Image Contrast. in (ISMRM, 2020).
25. Pydicom. https://pydicom.github.io/. Accessed 30 Nov 2021.
26. DICOM. DICOM MR Image Module. http://dicom.nema.org/medical/dicom/current/output/chtml/part03/sect_C.8.3.html#table_C.8-4. Accessed 30 Nov 2021.
27. DICOM. DICOM Secondary Capture. http://dicom.nema.org/dicom/2013/output/chtml/part03/sect_A.8.html. Accessed 30 Nov 2021.
28. Essock-Burns, E. *et al.* Comparison of DSC-MRI post-processing techniques in predicting microvascular histopathology in patients newly diagnosed with GBM. *J. Magn. Reson. Imaging* **38**, 388–400 (2013).
29. Essig, M. *et al.* Perfusion MRI: the five most frequently asked technical questions. *AJR Am J Roentgenol* **200**, 24–34 (2013).
30. hashlib. https://pypi.org/project/hashlib/. Accessed 30 Nov 2021.
31. scikit-learn. scikit-learn Random Forest Classifier. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble. Accessed 30 Nov 2021.
32. scikit-learn. scikit-learn Support Vector Machines. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm. Accessed 30 Nov 2021.
33. scikit-learn. scikit-learn Model Selection. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection. Accessed 30 Nov 2021.
34. scikit-learn. scikit-learn Inspection. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.inspection. Accessed 30 Nov 2021.
35. scikit. Permutation feature importance. *Permutation feature importance* https://scikit-learn.org/stable/modules/permutation_importance.html. Accessed 30 Nov 2021.

36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.90.
37. PyTorch, P. *PyTorch ResNet50*. (PyTorch, 2021).
38. ImageNet. ImageNet. *ImageNet* http://www.image-net.org/. Accessed 30 Nov 2021.
39. PyTorch. PyTorch TorchVision.Models. *Torchvision.Models* https://pytorch.org/vision/stable/models.html. Accessed 30 Nov 2021.
40. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
41. scikit-learn TSNE. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html. Accessed 30 Nov 2021.
42. UCSF Bridge. https://bridge.ucsf.edu/. Accessed 30 Nov 2021.
43. Gourraud, P.-A. *et al.* Precision medicine in chronic disease management: The multiple sclerosis BioScreen. *Ann. Neurol.* **76**, 633–642 (2014).
44. Adeel Azam, M., Bahadar Khan, K., Ahmad, M. & Mazzara, M. Multimodal medical image registration and fusion for quality enhancement. *Computers, Materials & Continua* **68**, 821–840 (2021).
45. Azam, M. A. *et al.* A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **144**, 105253 (2022).