Neuroradiology (2023) 65:77–87
https://doi.org/10.1007/s00234-022-03023-7

DIAGNOSTIC NEURORADIOLOGY

Deep multi-task learning and random forest for series classification by pulse sequence type and orientation

Noah Kasmanoff[1] · **Matthew D. Lee**[2] · **Narges Razavian**[1,2,3] · **Yvonne W. Lui**[2]

Received: 28 January 2022 / Accepted: 19 July 2022 / Published online: 30 July 2022

Abstract

Purpose Increasingly complex MRI studies and variable series naming conventions reveal limitations of rule-based image routing, especially in health systems with multiple scanners and sites. Accurate methods to identify series based on image content would aid post-processing and PACS viewing. Recent deep/machine learning efforts classify 5–8 basic brain MR sequences. We present an ensemble model combining a convolutional neural network and a random forest classifier to differentiate 25 brain sequences and image orientation.

Methods Series were grouped by descriptions into 25 sequences and 4 orientations. Dataset A, obtained from our institution, was divided into training (16,828 studies; 48,512 series; 112,028 images), validation (4746 studies; 16,612 series; 26,222 images) and test sets (6348 studies; 58,705 series; 3,314,018 images). Dataset B, obtained from a separate hospital, was used for out-of-domain external validation (1252 studies; 2150 series; 234,944 images). We developed an ensemble model combining a 2D convolutional neural network with a custom multi-task learning architecture and random forest classifier trained on DICOM metadata to classify sequence and orientation by series.

Results The neural network, random forest, and ensemble achieved 95%, 97%, and 98% overall sequence accuracy on dataset A, and 98%, 99%, and 99% accuracy on dataset B, respectively. All models achieved > 99% orientation accuracy on both datasets.

Conclusion The ensemble model for series identification accommodates the complexity of brain MRI studies in state-of-the-art clinical practice. Expanding on previous work demonstrating proof-of-concept, our approach is more comprehensive with greater sequence diversity and orientation classification.

Keywords Deep learning · Machine learning · MRI · Sequence classification

Introduction

Advances in medical imaging have led to increasingly complex studies. A brain MRI can have numerous protocol choices (e.g., pediatric, epilepsy, brain tumor, multiple sclerosis, acute stroke), and a single study may consist of over a thousand images separated into dozens of series. A series is a volume of images obtained using a specific pulse sequence, tissue contrast, and scan orientation. Choice of pulse sequence and scanning plane are important clinical considerations. Adding to this complexity are differences in scanner manufacturer, model, software operating system, pulse sequence design, scan parameters, as well as inconsistent or custom naming conventions. Series labeling for effective routing of images to PACS, which at first would appear to be a straightforward task, becomes increasingly challenging. Compounding these issues is the emergence

Noah Kasmanoff and Matthew D. Lee contributed equally to this work.

✉ Matthew D. Lee
matthew.lee4@nyulangone.org

[1] Center for Data Science, New York University, New York, NY, USA

[2] Department of Radiology, NYU Grossman School of Medicine, New York University, New York, NY 10016, USA

[3] Department of Population Health, NYU Grossman School of Medicine, New York University, New York, NY, USA

of large healthcare systems encompassing multiple imaging sites, often arising from previously independent centers with their own conventions. On the other hand, there has never been more of a need for scalable automated classification of images and series to facilitate the curation of large datasets such as required for machine learning application development and real-world deployment.

The literature shows proof-of-concept identification of 5–8 basic brain MRI sequences using machine learning [1–4]. Standard machine learning approaches such as random forest classifiers (RFC) have demonstrated good accuracy, though these depend on consistency of the DICOM metadata, which is known to be variable [2, 4]. Deep learning has also been applied with convolutional neural networks (CNN) tapping image content to directly classify sequences [1–3]. Translation to clinical application from these initial works is limited as they include only a few basic pulse sequences and lack orientation identification (e.g., axial, coronal, sagittal), which is needed for appropriate series classification. In this work, we present an ensemble approach combining a two-dimensional CNN with an RFC trained on DICOM metadata to differentiate 25 clinical brain MRI sequences and image orientation using real-world clinical imaging datasets.

Methods

Data

This retrospective study was conducted in compliance with the Health Insurance Portability and Accountability Act and approved by the institutional review board. To identify brain MRI studies, we searched a deidentified database of MRIs performed on 1.5-T and 3-T scanners at an academic medical center between November 1997 and December 2018 for studies with the DICOM metadata attribute BodyPartExamined (0018,0015) set as "brain," "head," or "neuro." Ground-truth labels for sequence and orientation were based on the SeriesDescription (0008,103E) attribute. A board-certified neuroradiologist with 14 years of post-fellowship experience classified 349 series descriptions into 25 sequence classes by reviewing series descriptions and visually inspecting sample images (2D unless otherwise noted): pre-contrast T1, post-contrast T1, pre-contrast T1 internal auditory canal (IAC), post-contrast T1 IAC, pre-contrast 3D T1, post-contrast 3D T1, pre-contrast 3D volumetric interpolated brain examination (VIBE)/golden-angle radial sparse parallel (GRASP), post-contrast VIBE/GRASP; T2, HASTE (half-Fourier acquisition single-shot turbo spin echo), 3D T2, 3D CISS (constructive interference in steady state), FLAIR (fluid-attenuated inversion recovery), 3D FLAIR; diffusion, ADC (apparent diffusion coefficient), SWI (susceptibility-weighted imaging), SWI magnitude (magnitude image), SWI phase (phase map), SWI MIP (maximum intensity projection; 3-5 mm slab); perfusion (dynamic-susceptibility contrast), TOF MRA (time-of-flight MR angiogram), TOF MRA MIP (3D), scout, and other. The diffusion class consisted of all diffusion-weighted imaging regardless of B-value or number of shells. The other class included all other sequences such as quantitative susceptibility mapping, arterial spin labeling, and functional MRI. Fat-suppressed sequences were included with non-fat-suppressed sequences under their base sequence class. Images of ambiguous series descriptions were examined and classified accordingly. Series orientation was classified as axial, coronal, sagittal, and oblique/other/unlabeled based on series description.

To train and evaluate this work, we used two distinct clinical datasets containing a broad range of normal and abnormal cases. Dataset A was stratified by study and divided into training (16,828 studies; 48,512 series; 112,028 images), validation (4746 studies; 16,612 series; 26,222 images), and test sets (6348 studies; 58,705 series; 3,314,018 images). By stratifying dataset A by study, we ensured that all images in a series and all series in a study were included in only one of the subsets. To balance the variation in the number of series between different sequence classes in the training set, the training data were limited to 4500 images per class. Mean subject age in dataset A was 49.8 years (SD: 21.4 years; range: 1 day–95 years), and 64.6% of studies were of male patients. Distribution of series by scanner manufacturers is provided in Table S1.

Dataset B, obtained from a community hospital prior to its incorporation into the larger health system from which dataset A is drawn, provided out-of-domain external validation with simpler protocols and fewer pulse sequence types (1252 studies; 2150 series; 234,944 images). Mean subject age in dataset B was 56.2 years (SD: 24.9 years; range: 1 day–95 years), and 81.6% of studies were of male patients.

Images were cropped and/or padded to maintain aspect ratio, resized to 256 × 256, and normalized by pixel intensity. Stochastic data augmentations of rotation at a random degree from uniform distribution [− 10°,10°], scaling/cropping [− 5%,5%], and translation [− 20,20 pixels] were applied.

Model development

The RFC implemented in scikit-learn [5] classified images according to DICOM metadata attributes (Table 1). Selected attributes were parameters that often differ between pulse sequences and were often included in prior studies (Table S2) [2–4]. Due to variable institutional conventions, the attribute ContrastBolusAgent (0018,0010) used in these prior studies was not available at either institution. Each attribute was checked individually with missing values

Table 1 DICOM metadata attributes extracted by the random forest classifier

Parameter	DICOM tag
Image type	(0008,0008)
Scanning sequence	(0018,0020)
Sequence variant	(0018,0021)
Scan options	(0018,0022)
MR acquisition type	(0018,0023)
Repetition time	(0018,0080)
Echo time	(0018,0081)
Echo train length	(0018,0091)
Image orientation (patient)	(0020,0037)
Photometric interpretation	(0028,0004)
Pixel spacing	(0028,0030)

imputed as – 1. RFC hyperparameters were tuned, including the number of trees, loss function, and number of random features considered when splitting nodes. The final RFC consisted of 100 trees analyzing 2 to 14 features and otherwise default parameters. The RFC was trained on a CPU, which took approximately 3 h to complete.

The CNN employs multi-task learning (MTL) to learn sequence and orientation from the same input, while a 2D architecture maintains flexibility for clinical practice as such a system is independent of the number of slices per volume. Each slice in the series passes through a custom CNN architecture implemented in PyTorch [6] and trained with PyTorch-Lightning [7] (Fig. 1). After 10 layers of convolution, batch normalization, and rectified linear unit activation, the output is split into two heads for parallel MTL of sequence and orientation. The sequence head outputs a vector of log probabilities of size equal to the number of possible sequences (25). The orientation head produces a vector of log probabilities corresponding to the possible orientations (4). Passing a batch of size M through the network, weighted cross-entropy loss is applied for sequence classification:

$$l_{sequence} = -\frac{1}{M}\sum_{k}^{K}\sum_{m}^{M} w_k y_m^k log\left(\hat{y}_m^k\right) \quad (1)$$

where weights w_k correspond to the relative proportion of images of that class within the training distribution, y_m^k is the ground-truth sequence, and $\hat{y}_m^k$ is the predicted sequence.

The orientation loss function utilizes *masked weighted cross-entropy loss*:

$$l_{orientation} = -\frac{1}{M}\sum_{k}^{K}\sum_{m}^{M} 1_{k\neq unlabeled} w_k y_m^k log(\hat{y}_m^k) \quad (2)$$

where indicator function $1_{\mathrm{k\neq unlabeled}}$ masks instances from the batch where an orientation tag is not provided. This allows the two tasks to train in parallel, albeit with less sampling efficiency for orientation. The contrast and orientation loss functions are summed, $L = l_{sequence} + l_{orientation}$, and the gradient of L is used to update network weights via backpropagation and the Adam optimizer [8].

Many hyperparameter configurations were tuned using the validation set from dataset A, including learning rates between 10^{-5} and 10^{-1}, weight decay factors between 10^{-1} and 10^{-5}, batch sizes between 16 and 256, and different capacity configurations (number of parameters per layer). Final hyperparameter configurations from the most accurate model on the validation set included a learning rate of 10^{-4}, batch size of 8, weight decay factor of 5×10^{-6}, and parameter count of 307,491,484. Learning rate was decreased by a factor of 10 for every 5 epochs in which the validation loss did not improve. The CNN's best performing validation score was achieved in 72 h of training on a 32 GB NVIDIA V100 GPU.

Each image in a series received a prediction from the RFC as well as a prediction from the CNN. For the CNN and RFC, a majority-rules vote over all images in a series resulted in the final prediction of sequence and orientation, except for orientations of scout series as these often contain images of multiple orientations. To produce the ensemble model, a majority-rules vote over all the predictions from the CNN and RFC (two predictions for each image) for the images in the series resulted in a final prediction.

Analysis

Accuracy (number of correctly classified series divided by the total number of series), precision (positive predictive value), recall (sensitivity), and F1-score (harmonic mean of precision and recall) were calculated for each sequence:

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = \frac{2}{precision^{-1} + recall^{-1}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (5)$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives. For each model, average F1-score, weighted by the number of series for each class, was calculated. Accuracy was calculated over all sequences, weighted by the number of series for each class, as well as over all orientations. Confusion matrices were generated to visualize concordance between ground-truth and predicted classes. To test if performance differed between classifiers, Friedman's test based on the F distribution was conducted

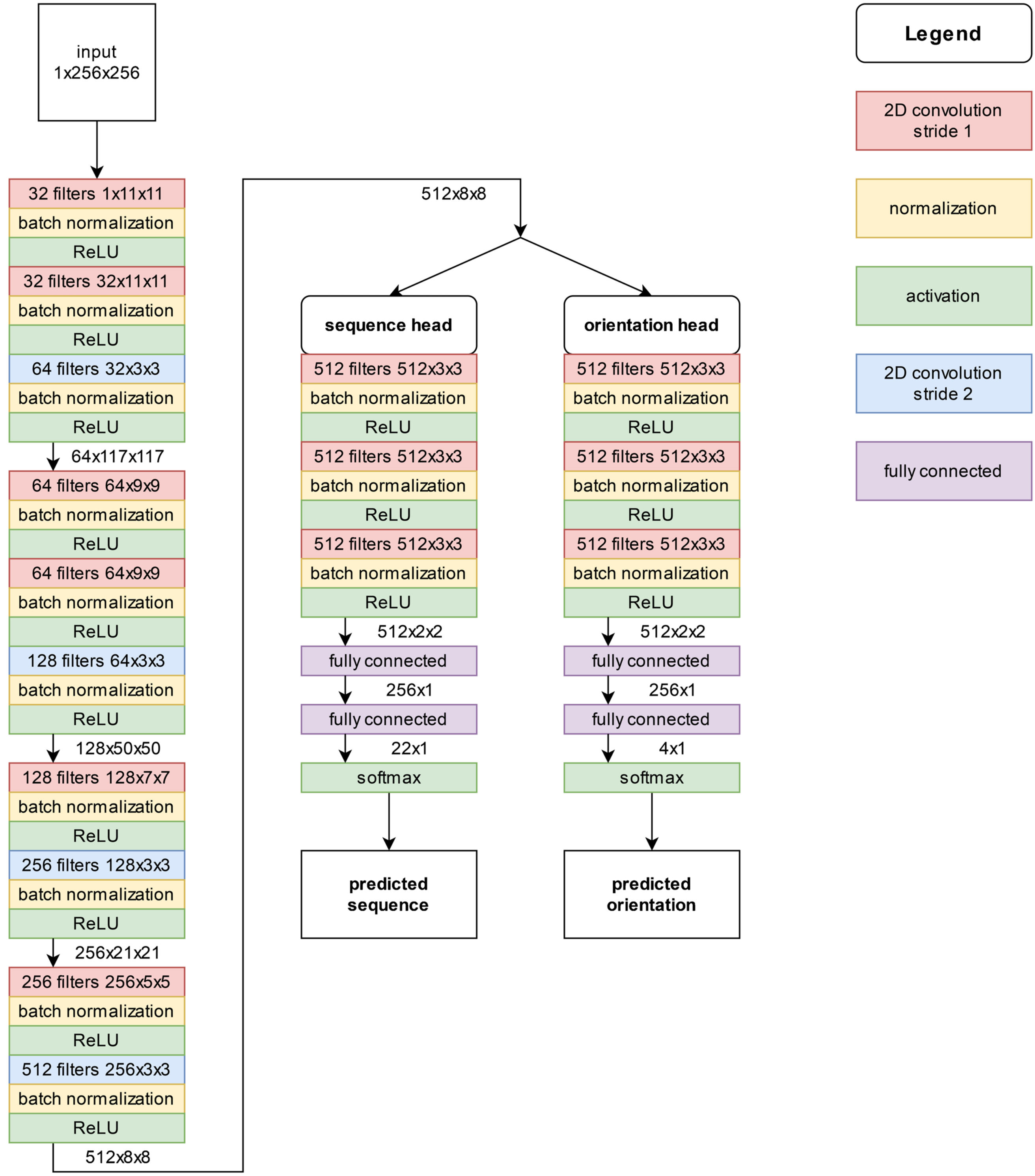

Fig. 1 Convolutional neural network architecture. After 10 layers of convolution, batch normalization, and rectified linear unit (ReLU) activation, the output is split into two heads for parallel multi-task learning of sequence and orientation. Multi-task learning allows a network to optimize more than one loss function to learn related tasks from the same input data

on F1-scores, followed by Nemenyi's post hoc test for pairwise classifier comparisons, with significance level $p < 0.05$. For the most prevalent sequence classification discrepancies with 20 or more series, five randomly selected example series were reviewed by imaging experts to determine the major types of discrepancies. To visualize the portions of the input image that contributed most to the CNN's prediction, saliency maps were generated by calculating the gradient of the predicted class with respect to each input pixel. To determine which DICOM attribute features contributed most to the RFC predictions, we utilized SHAP (Shapley Additive exPlanations), which employs a game theoretic approach to quantify a feature's contribution to the model's prediction of a given data point [9].

Results

Dataset A

Sequence classification performance on the holdout test set is summarized in Table 2. Confusion matrices show concordance between predicted and ground-truth labels (Fig. 2). Orientation accuracy was > 99% for the CNN, RFC, and ensemble model.

Overall RFC sequence classification accuracy was 97% with weighted average F1-score of 0.97. F1-score was 1.00 for diffusion, ADC, FLAIR, 3D FLAIR, MRA, MRA MIP, SWI, and SWI magnitude/MIP/phase. F1-score was lowest for VIBE/GRASP post-contrast (0.69; 51 series).

Overall CNN sequence classification accuracy was 95% with weighted average F1-score of 0.95. F1-score was 1.00 for diffusion, MRA MIP, SWI MIP, and SWI phase. F1-score was lowest for VIBE/GRASP post-contrast (0.42; 51 series).

Overall ensemble sequence classification accuracy was 98% with weighted average F1-score of 0.98. F1-score was 1.00 for diffusion, ADC, FLAIR, 3D FLAIR, MRA MIP, perfusion, SWI, SWI magnitude/MIP/phase, and 3D T1 pre-contrast. F1-score was lowest for VIBE/GRASP post-contrast (0.64; 51 series).

According to aggregate F1-scores, RFC sequence classification significantly outperformed the CNN ($p = 0.02$), as did the ensemble ($p = 0.01$). RFC and ensemble performance did not significantly differ ($p > 0.05$).

Dataset B

Sequence classification performance is summarized in Table 3 and confusion matrix (Fig. 3). Orientation accuracy was > 99% for the CNN, RFC, and ensemble.

The RFC and ensemble achieved 99% overall sequence classification accuracy with weighted average F1-score of 1.00. F1-score was 1.00 for other, perfusion, scout, 3D T1 pre-contrast, and T2. F1 was lowest for T1 pre-contrast (0.61; 16 series).

Overall CNN sequence classification accuracy was 98% with weighted average F1-score of 0.98. F1-score was 1.00 for other and perfusion. F1 was lowest for scout (0.00; 9 series).

According to aggregate F1-scores, sequence classification performance of the RFC, CNN, and ensemble did not significantly differ ($p > 0.05$).

Interpretability analysis

Saliency maps demonstrated high gradients around head contours, which may be useful to the CNN for both orientation and sequence classification, as well as high-contrast boundaries such as along the ventricular surface, which may be useful for sequence classification (Fig. 4).

SHAP analysis of the RFC is summarized in Fig. S1. ImageType was the most impactful feature, especially when set to TRACEW, DIFFUSION, or ADC, as these corresponded directly to diffusion, diffusion, and ADC sequence classes respectively.

Discussion

We developed an ensemble model combining a CNN with a novel MTL architecture and RFC trained on DICOM metadata to accurately classify up to 25 brain MRI sequences and identify image orientation. The model was trained on a large real-world clinical dataset encompassing a broad range of scanners, protocols, and normal and abnormal cases. The model performed well on the holdout test set and external data, indicating good generalization across imaging sites, scanners, protocols, and pathologies.

We extend work from previous studies providing proof-of-concept classification of 5–8 basic brain MRI sequences using deep learning and conventional machine learning [1–4]. Remedios et al. developed PhiNet, a cascaded 3D CNN that first classifies series as T1, T2, or FLAIR, then classifies T1 and FLAIR as pre-/post-contrast [1]. Pizarro et al. developed a CNN to classify series as T1, T1 post-contrast, T2, FLAIR, proton density (PD), high-resolution T1, magnetic-transfer-on, and magnetic-transfer-off [2]. Pizarro et al. compared 2D and 3D convolution but did not find a significant improvement with 3D convolution, which required excessive computational memory. DeepDicomSort, a 2D CNN developed by van der Voort et al., classifies slices as T1, T1 post-contrast, T2, PD, FLAIR, diffusion, perfusion, or derived (e.g. ADC, CBF maps) [3]. Though initially trained on brain tumor studies, DeepDicomSort achieved excellent performance on a test set of Alzheimer's disease patients. Gauriau et al. developed an RFC that classifies series as T1, T2, FLAIR, diffusion, susceptibility, angiography, scout, or other (e.g., screenshots, perfusion, spectroscopy) with performance varying by scanner manufacturer [4]. Like DeepDicomSort, our models consider pre- and post-contrast T1-weighted sequences as distinct initial classes.

Table 2 Sequence classification performance on dataset A's holdout test set. N refers to the number of series of a specific sequence class. The weighted average is weighted by the proportion of series per sequence

		CNN			RFC			Ensemble		
Sequence	*N*	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
CISS[a]	425	0.90	0.97	0.94	0.73	0.99	0.84	0.80	0.99	0.89
Diffusion	4390	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Diffusion ADC	4201	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
FLAIR	3940	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
3D FLAIR	1072	0.98	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00
MRA	236	0.96	0.99	0.97	1.00	1.00	1.00	1.00	0.99	0.99
MRA MIP	55	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Other[b]	200	0.87	0.96	0.91	1.00	0.96	0.98	0.98	0.96	0.97
Perfusion	321	0.95	1.00	0.98	1.00	1.00	1.00	1.00	0.99	0.99
Scout	7949	0.99	0.93	0.96	0.99	0.99	0.99	0.99	0.99	0.99
SWI[b]	3458	0.76	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00
SWI Magnitude[b]	3301	1.00	0.67	0.80	1.00	1.00	1.00	1.00	1.00	1.00
SWI MIP	3456	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SWI Phase	3364	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
T1 IAC post[b]	420	0.78	0.93	0.85	0.90	0.98	0.94	0.95	0.97	0.96
T1 IAC pre[a]	532	0.87	0.86	0.88	0.65	0.84	0.74	0.79	0.88	0.83
T1 post[a]	2294	0.82	0.96	0.88	0.78	0.84	0.80	0.82	0.85	0.83
T1 pre[a]	4985	0.99	0.87	0.92	0.92	0.84	0.87	0.92	0.88	0.90
3D T1 pre	2304	0.98	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00
T2	5960	0.96	0.99	0.97	1.00	1.00	1.00	0.99	1.00	0.99
T2 SPACE	451	0.91	0.99	0.95	0.98	0.99	0.98	0.98	0.99	0.98
HASTE[b]	2172	0.84	0.95	0.89	0.97	0.98	0.97	0.96	0.99	0.98
VIBE/GRASP post[b]	51	0.27	0.98	0.42	0.53	0.96	0.69	0.48	0.96	0.64
VIBE/GRASP pre	1157	0.99	0.90	0.95	1.00	0.96	0.98	0.99	0.95	0.97
Weighted avg	56,694	0.96	0.95	0.95	0.97	0.97	0.97	0.98	0.98	0.98

[a] CNN outperforms RFC (ΔF1-score ≥ 0.05)

[b] RFC outperforms CNN

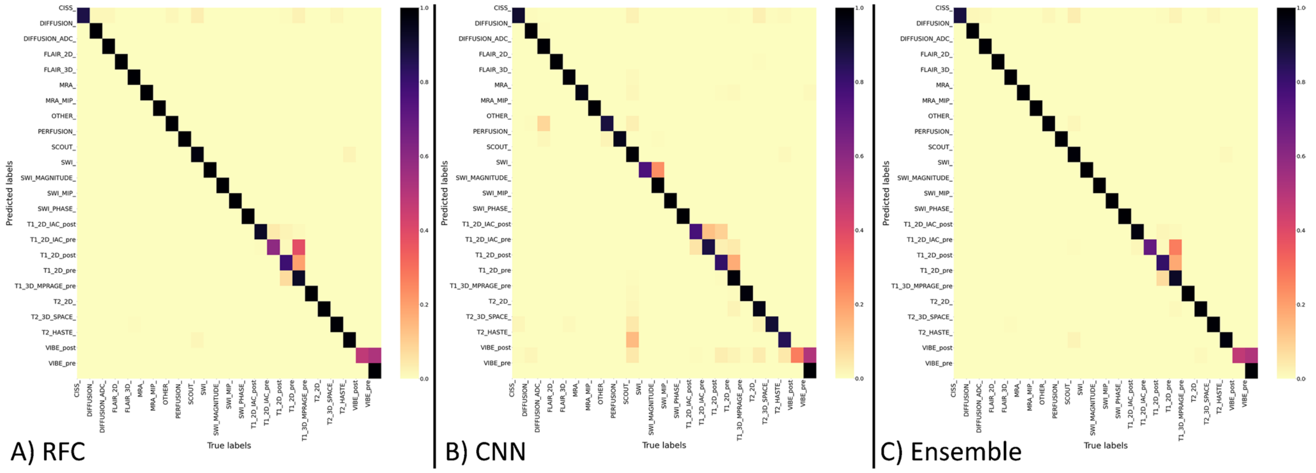

Fig. 2 Confusion matrix of sequence classification on dataset A (holdout test set) for the **A** random forest classifier, **B** convolutional neural network, and **C** ensemble. The color of each square corresponds to the proportion of sequence labels predicted for each ground-truth label

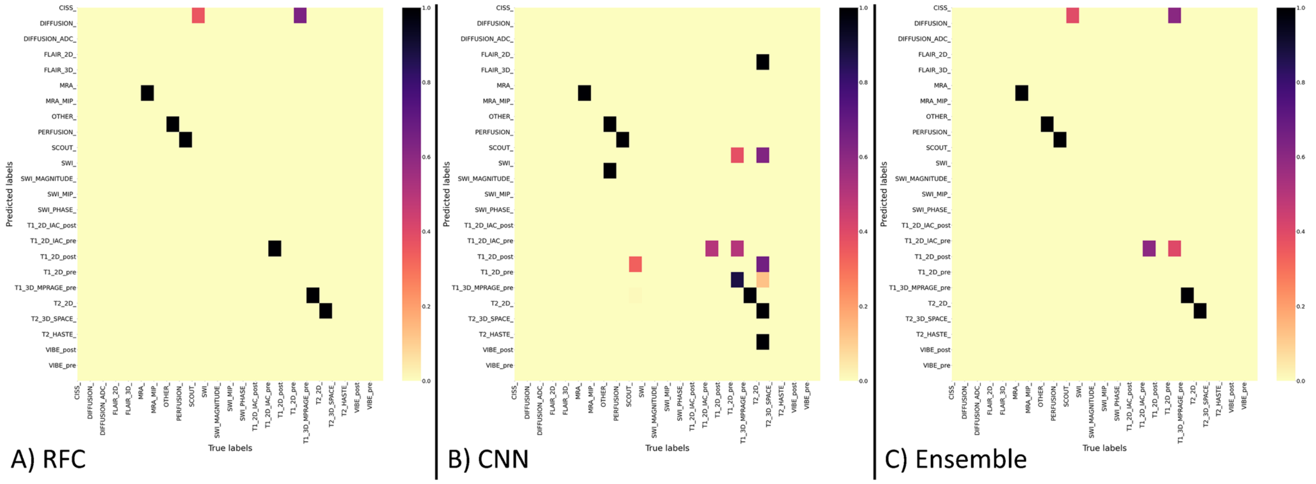

Fig. 3 Confusion matrix of sequence classification on dataset B (out-of-domain external validation) for the **A** random forest classifier, **B** convolutional neural network, and **C** ensemble

Table 3 Sequence classification performance on dataset B (out-of-domain test set). N refers to the number of series of a specific sequence class. The weighted average is weighted by the proportion of series per sequence

		CNN			RFC			Ensemble		
Sequence	*N*	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MRA	376	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Other[a]	7	1.00	0.86	0.92	1.00	1.00	1.00	1.00	1.00	1.00
Perfusion	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Scout	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T1 IAC pre[a]	3	0.50	1.00	0.67	1.00	1.00	1.00	0.60	1.00	0.75
T1 pre[b]	16	0.88	0.44	0.58	1.00	0.06	0.12	1.00	0.06	0.12
3D T1 pre	848	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
T2	888	1.00	0.95	0.97	1.00	1.00	1.00	1.00	1.00	1.00
Weighted avg	2150	0.99	0.97	0.98	1.00	0.99	0.99	1.00	0.99	0.99

[a] RFC outperforms CNN

[b] CNN outperforms RFC (ΔF1-score ≥ 0.05)

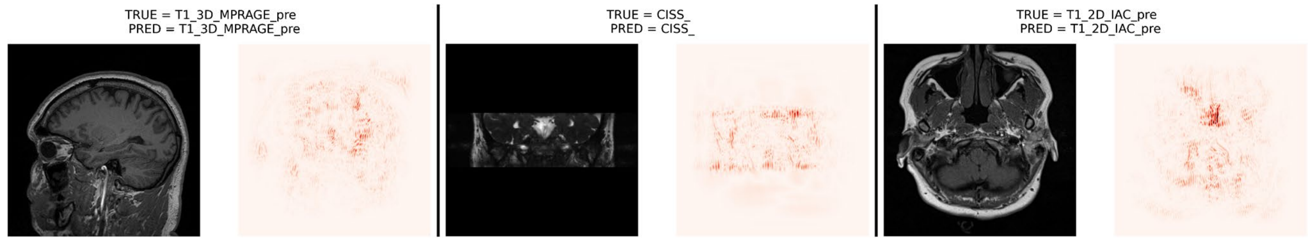

Fig. 4 Saliency maps demonstrate high gradients around head contours, which may be particularly useful for orientation determination, as well as brain parenchyma and ventricles, likely for sequence classification

Beyond including a more complete variety of state-of-the-art brain MRI sequences, our model provides several key advantages and innovations. Our model is the first to combine deep learning and standard machine learning approaches, as well as the first to use MTL to learn MRI sequence and orientation. In MTL, a CNN optimizes more than one loss function to learn related tasks from the same input data [10]. MTL is a well-studied technique that provides multiple advantages by increasing focus on relevant features, allowing classifiers to share features, and reducing overfitting. The literature of deep MTL applied to medical imaging is sparse but growing as CNNs in radiology become increasingly sophisticated to address more complex problems. Most MTL studies involve detection, segmentation, and classification, such as of gliomas [11], breast lesions [12, 13], and COVID-19 pneumonia [14]. Furthermore, we include “other” sequence and orientation classes, giving the model flexibility in accommodating sequences it may not have been trained on. We employ a 2D architecture to accommodate series of variable slice number and avoid excessive computational power or additional preprocessing that 3D convolution requires, such as resampling and volumetric inference.

While all models performed well, the ensemble model had overall highest and near-perfect accuracy. The ensemble model tended to match whichever model was the higher performer for a specific sequence, suggesting that the CNN and RFC may complement each other. The RFC slightly outperformed the CNN on dataset A, in contrast to previous studies that had shown CNNs to outperform RFCs [2, 3]. The small samples of certain sequences may have contributed to the RFC’s edge over the CNN. On dataset A, the CNN outperformed the RFC on CISS, T1 pre-/post-contrast, and T1 IAC pre-contrast, whereas the RFC outperformed the CNN on other, SWI, SWI magnitude, T1 IAC post-contrast, HASTE, and VIBE/GRASP post-contrast. These findings suggest that the two models may complement each other and that an ensemble approach combining both can mitigate differences in imaging and DICOM metadata.

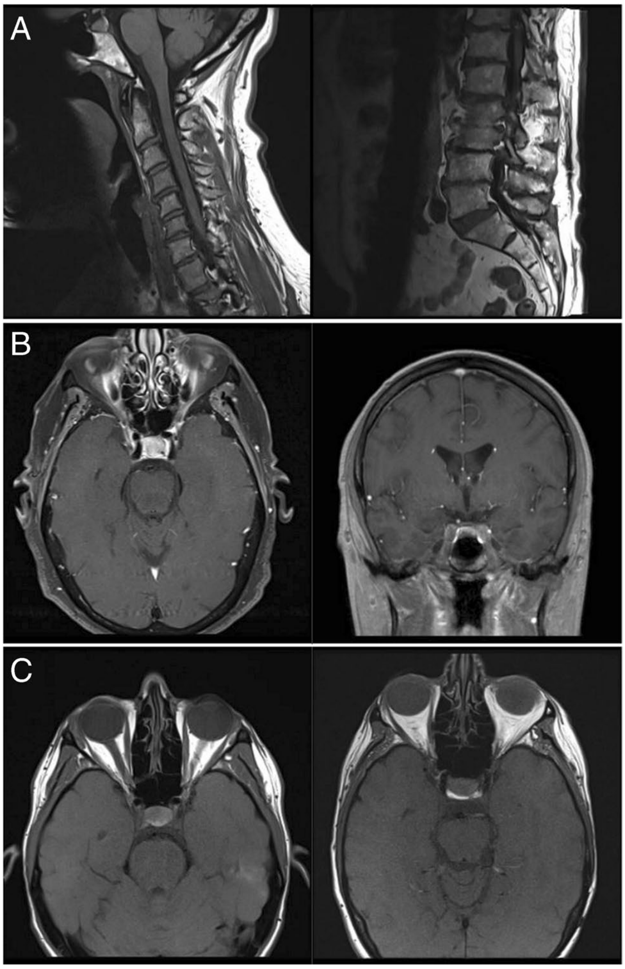

Fig. 5 Examples of sequence classification discrepancies by the convolutional neural network. **A** T1 pre-contrast series classified as T2 were non-brain studies inadvertently included due to erroneous DICOM metadata. **B** T1 internal auditory canal pre-contrast series classified as T1 internal auditory canal post-contrast were indeed post-contrast. **C** T1 pre-contrast series classified as T1 internal auditory canal pre-contrast had limited fields-of-view through the orbits, resembling internal auditory canal studies

Error/discrepancy analysis

The following major types of classification discrepancies were identified: (1) inputs that were initially mislabeled and not actually brain imaging, (2) incorrect ground-truth labels, and (3) incorrect predictions. Some of the misclassifications had clear explanations after review; others did not.

Review of discrepancies in dataset A's test set revealed non-brain images included in the data. For example, all reviewed examples of T1 pre-contrast that were classified as T2 by the CNN (138 series, 2.8% of T1 pre-contrast) were non-brain studies (e.g., spine) (Fig. 5A). Non-brain studies occurred in dataset A, even after efforts to exclude non-brain studies manually and by filtering out studies using DICOM metadata. One purpose of our study was to develop models using large real-world clinical datasets as opposed to small, perfectly curated datasets. In any clinical dataset, it is more likely than not that there will be some mislabeled images. For example, one analysis of computer vision datasets estimates that up to 6% of images in ImageNet (the most popular benchmark dataset containing more than 14 million images) are mislabeled [15]. One of the major strengths of machine learning methods is that even somewhat imperfect data can be extremely useful, yielding highly accurate models given large enough datasets [16, 17]. Of note, no non-brain studies were found in dataset B.

Some discrepancies between ground-truth and model labeling were due to either ambiguous or incorrect series descriptions. The CNN and ensemble were able to correctly classify many of them, more than the RFC. For example, all inspected examples of T1 pre-contrast classified as T1 post-contrast by the CNN (469 series, 9% of all T1 pre-contrast) were deemed to be T1 post-contrast images by consensus two-expert review, though some were fat-suppressed (Fig. 5B). Similarly, all examples of T1 IAC pre-contrast classified as IAC post-contrast (65 series, 12% of all T1 IAC pre-contrast) were in fact IAC post-contrast. The CNN's robustness to label noise is advantageous and consistent with prior studies demonstrating that deep learning models form robust representations of classes instead of memorizing specific examples [16, 17].

Some discrepancies were true misclassifications, though many were explainable upon further review. Review of T1 pre-contrast series classified as T1 IAC pre-contrast by the CNN (23 series, 0.5% of all T1 pre-contrast) revealed several studies which were T1 pre-contrast through the orbit with a smaller field-of-view than typical of routine brain imaging, resembling IAC studies (Fig. 5C). Some examples of VIBE/GRASP pre-contrast misclassified as VIBE/GRASP post-contrast by the CNN (95 series, 8% of VIBE/GRASP pre-contrast) and RFC (43 series, 4% of VIBE/GRASP pre-contrast) were thin-section source images from pre-/non-contrast head/neck MR angiograms with hyperintense vessels.

True misclassifications without readily available explanations tended to occur more frequently with the RFC than the CNN. The RFC demonstrated certain types of true misclassifications that did not occur with the CNN: T1 post-contrast classified as T1 pre-contrast (338 series, 15%), T1 IAC pre-contrast classified as T1 pre-contrast (49 series, 9%), T1 post-contrast classified as CISS (24 series, 5%).

Some scout images posed problems for both CNN and RFC. Discrepancies between CNN and ground-truth labels included classifying scout series as HASTE (358 series, 5% of scout), T2 (94 series, 1%), or 3D T1 pre-contrast (24 series, 0.3%). RFC discrepancies included classifying scouts as CISS (30 series, 0.4%) or HASTE (75 series, 0.9%). In some regards, those classified as HASTE were not technically incorrect, as most scout series performed at our institution are HASTE (a T2-weighted sequence), though with lower resolution than the clinical HASTE series. Further, non-HASTE scouts are occasionally obtained and underrepresented in the training data, making it understandable that T1-weighted scouts were misclassified as a T1-weighted sequence.

From dataset B, all examples of T2 classified as HASTE by the CNN and ensemble (27 series, 3% of T2) were in fact HASTE of pediatric patients. T2-weighted series classified as FLAIR by the ensemble (7 series, 11% of FLAIR) were T2 with significant motion artifact.

Review of discrepancies between ground-truth labels and model predictions revealed the following study limitations: our method of identifying brain MRIs was imperfect and introduced a small proportion of non-brain studies. Another limitation is determining ground-truth sequence and orientation labels from series description, the fallibility of which motivates our work in the first place. Nevertheless, the size of the datasets allowed the CNN to overcome many instances of errant labeling; the discrepancy analysis demonstrates several instances where series descriptions were ambiguous or incorrect, but the CNN was in fact correct. Finally, some sequences contained significantly fewer examples due to their relative rarity. We attempted to mitigate this by capping the number of examples in each class and applying weighted cross-entropy loss. The generalizability of our study is limited by the fact that our training data comes from a single institution with conventions and protocols that may differ from other institutions. Also, the external validation dataset B included few sequence types compared to dataset A, though dataset B exclusively consisted of studies from the scanner manufacturer that was least represented in the training and validation datasets. Our goal was to explore the potential of machine learning models in improving routing and image display tasks. For generalizability, future studies will need to fine-tune on local data or re-train using multi-institutional datasets.

The flexibility and modularity of our approach support many directions for future work. The model's output can be modified to accommodate more classes such as fat-suppressed sequences. The MTL approach can be scaled to learn and predict additional medical image attributes, such as slice thickness and body part examined. The model may be transferred to learn sequences for other anatomy or incorporated into a larger tool that first classifies studies by anatomy. Additional considerations for clinical implementation include integration with PACS for radiologist viewing/hanging protocols and user ability to correct mislabeled series. Automated sequence classification could also serve as the first step in image post-processing pipelines, such as a tool that automatically identifies FLAIR and T1 sequences at acquisition to detect and segment white matter lesions in patients with multiple sclerosis. Certain results stimulate additional possibilities. For example, the model's systematic misclassification of T2 with severe motion artifact suggests an extension to image quality control. While our model's saliency maps suggest that head contours and high-contrast boundaries are useful for classification, further detailed investigation of model explainability may be worthwhile. While our ensemble model combined CNN and RFC results using a majority-rules voting system, which allowed us to assess the relative performance and contribution of images and DICOM metadata to predict sequence and orientation, a model trained on both pixel-level data and DICOM metadata may perform even better.

Conclusions

This work shows that an ensemble approach combining a CNN trained on images and RFC trained on DICOM metadata for series identification accommodates the complexity of brain MRI studies in state-of-the-art clinical practice. Expanding on previous work demonstrating proof-of-concept, our approach is more comprehensive with many more classes, as well as orientation classification, and employs unique methods including MTL to formulate a flexible model. The ensemble model including CNN and RFC had overall high accuracy with weighted mean F1-score of 1.00 on external validation, and results indicate that the two approaches may be complementary.

Author contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Noah Kasmanoff and Matthew Lee. Funding acquisition was performed by Noah Kasmanoff. Supervision and project administration were performed by Narges Razavian and Yvonne Lui. The first draft of the manuscript was written by Matthew Lee, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Partial financial support was received from the Moore-Sloan Data Science Environment Summer Research Initiative.

Data availability The data for this project is not publicly available.

Code availability The code for this project is publicly available at https://github.com/nkasmanoff/mri-content-detection.

Declarations

Ethics approval This study was conducted in compliance with the Health Insurance Portability and Accountability Act and approved by the institutional review board.

Consent to participate Informed consent was waived with approval by the institutional review board.

Consent for publication Informed consent was waived with approval by the institutional review board.

Conflict of interest The authors declare no competing interests.

References

1. Remedios S, Roy S, Pham DL, Butman JA (2018) Classifying magnetic resonance image modalities with convolutional neural networks. Medical Imaging 2018: Computer-Aided Diagnosis 558–63. https://doi.org/10.1117/12.2293943
2. Pizarro R, Assemlal H-E, De Nigris D et al (2019) Using deep learning algorithms to automatically identify the brain MRI contrast: Implications for Managing Large Databases. Neuroinform 17:115–130. https://doi.org/10.1007/s12021-018-9387-8
3. van der Voort SR, Smits M, Klein S, for the Alzheimer's Disease Neuroimaging Initiative (2020) DeepDicomSort: an automatic sorting algorithm for brain magnetic resonance imaging data. Neuroinformhttps://doi.org/10.1007/s12021-020-09475-7
4. Gauriau R, Bridge C, Chen L et al (2020) Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. J Digit Imaging 33:747–762. https://doi.org/10.1007/s10278-019-00308-x
5. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830. https://doi.org/10.5555/1953048.2078195
6. Paszke A, Gross S, Massa F, et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 8024–35
7. Falcon W (2019) PyTorch Lightning. https://github.com/PyTorchLightning/pytorch-lightning
8. Kingma D, Ba J (2015) Adam: a method for stochastic optimization. arXiv:1412.6980 [cs.LG]
9. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions
10. Ruder S (2017) An overview of multi-task learning in deep neural networks. arXiv:1706.05098 [cs.LG]

11. van der Voort SR, Incekara F, Wijnenga MMJ, et al (2020) WHO 2016 subtyping and automated segmentation of glioma using multi-task deep learning. arXiv:2010.04425 [eess.IV]
12. Sainz de Cea MV, Diedrich K, Bakalo R, et al (2020) Multi-task learning for detection and classification of cancer in screening mammography. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 12266:241–250. https://doi.org/10.1007/978-3-030-59725-2_24
13. Gao F, Yoon H, Wu T, Chu X (2020) A feature transfer enabled multi-task deep learning model on medical imaging. Expert Syst Appl 143:112957. https://doi.org/10.1016/j.eswa.2019.112957
14. Amyar A, Modzelewski R, Ruan S (2020) Multi-task deep learning based CT imaging analysis for COVID-19: classification and segmentation. medRxiv. https://doi.org/10.1101/2020.04.16.20064709
15. Northcutt CG, Jiang L, Chuang IL (2021) Confident learning: estimating uncertainty in dataset labels. arXiv:1911.00068 [cs, stat]
16. Rolnick D, Veit A, Belongie S, Shavit N (2018) Deep learning is robust to massive label noise. arXiv:1705.10694 [cs.LG]
17. Tajbakhsh N, Jeyaseelan L, Li Q et al (2020) Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med Image Anal 63:101693. https://doi.org/10.1016/j.media.2020.101693

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