Machine Learning Methods for Automated Longitudinal Alignment and Visualization of Clinical MRI Exams

Julia Cluceru[1], Tanya Krishnakumar[3], Riley Bove[3], Atul J. Butte[2] and Jason C. Crane[1]

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA 94158-2330, USA
[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94158-2330, USA
[3]Department of Neurology, University of California, San Francisco, CA 94158-2330, USA

Purpose

- Automate the identification of MR images acquired with similar tissue contrast and align them for longitudinal comparison
- Create a high-throughput system for MR visualization in a research platform launching from the clinical record.

Introduction

- To analyze disease progression and response to therapy with MRI, it is essential to quantify serial changes that occur on images acquired with similar tissue contrast (Figure 1) [1]. However, data in most institutional PACS systems are heterogeneous and unreliably labeled, limiting the ability to automatically retrieve and align images of the same anatomy and contrast. This problem is exacerbated in the context of processing imaging data at scale for population-level analyses. The goal of this study is to create an algorithm that can reliably retrieve images of the same contrast from a cohort of brain exams. We hypothesize that DICOM metadata features, imaging features and a combination of both should yield increasing overall, as well as per-class accuracy. We compare the contribution of radiomics imaging features and those automatically derived through CNNs, as well as model transferability within and between diseases.

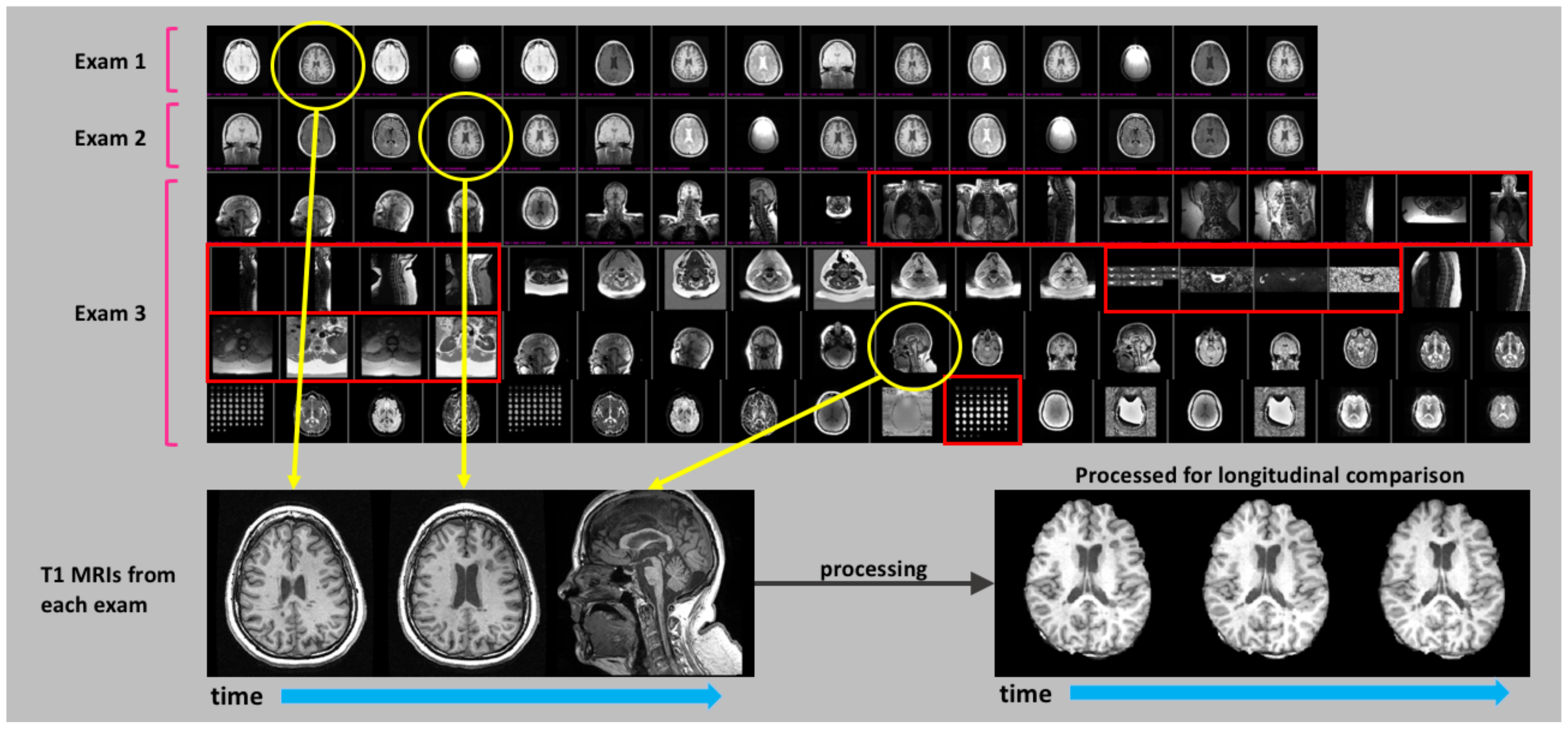


Figure 1. Longitudinal analysis of CNS disease requires retrieval and alignment of images of similar contrast. Three clinical MRI exams from a single patient, each may comprise over 100 series and may include both brain as well as other anatomical regions in a single exam (Exam 3). A T1 weighted volume from each exam is identified (bottom left) and post-processed (bottom right) to facilitate visual or computational comparison of longitudinal changes.

Methods

- Exams and cohorts: Four cohorts (Figure 2) were used. 1) MSR, a highly uniform multiple sclerosis (MS) research cohort[1]; 2) MSC, a heterogeneous MS clinical cohort representative of typical, poorly labeled institutional PACS data and also representing data acquired at external sites; 3) NEWglioma and 4) RECglioma cohorts comprised of newly diagnosed and recurrent research brain tumor exams that are highly uniform in contrast and labeling. Both MSR and MSC have more uniform imaging characteristics than NEW and RECglioma which have on average much more extensive pathology per exam. Each series of MSB cohort visually reviewed + labeled
- Ground truth determination: Each series used for the MSC cohort was visually reviewed and labeled; MSR, RECglioma and NEWglioma were acquired with uniform imaging protocols as part of clinical trials and already well labeled.
- Training and testing splits: 1) Within-disease: training/testing splits used all of the MSR and 60% of the MSC cohort for training, while 20% MSC was used for validation and 20% was used for testing. 2) Between-diseases: training used MSR and MSC cohorts, and testing comprised all available RECglioma and NEWglioma images. In both 1) and 2), splits were randomized and stratified by outcome, with no patient leakage.
- Metadata Models: Metadata were extracted from DICOM headers, hashed, and normalized to create numeric features. Series descriptions were tokenized and binarized using the top 40 most frequent words. The heuristic model was used a logical decision tree branching on a priori assumptions about acquisition parameters. SVMs were built using metadata and series descriptions features using scikit-learn's SVC [2]
- Image Processing: The center slice was used to represent each MR volume. All images normalized by subtracting the mean and dividing by the standard deviation of pixel intensity values. For radiomic analysis images were resampled to 1x1x1 mm resolution and binned into 64 intervals according to IBSI guidelines [3]. All images for CNN resized to 256 x 256 pixels regardless of the FOV.
- Radiomics: First and second order features were extracted using the Python pyradiomics toolkit [4]. SVMs were then built with the radiomics features alone and in combination with metadata features (Figure 3A).
- CNN: Pre-trained ResNet-18, ResNet-34, and ResNet-50[5] first convolutional layers were replaced with a 1 x 64 conv layer (for grayscale DICOM images); their last layers were replaced with 512 (2048) x 6 fully connected layer (Figure 3B). A cosine differential learning rate with 3 max values (0.01, 0.001, 0.0001) was used. All model building, training, and experiments were implemented using PyTorch 1.0.0 on a Tesla V100-PCIE-32GB (NVIDIA) GPU. Metadata features were processed as described above and concatenated with logit values. A final classification was created using scikit-learn's RFC[4].

	MSR	MSC	NEWglioma	RECglioma
Disease	MS	MS	Glioma	Glioma
Exams	1739	1994	88	105
T1	3609	622	65	85
T1C	1686	536	80	96
T2	1372	332	70	66
T2 FLAIR	895	508	77	97
PD	1439	66	0	0
OTHER	6131	1084	0	0

Table 1. Distribution of exams and images among the 4 cohorts used in this analysis. The "Other" category consisted of diffusion images, perfusion images, and other categories that have not yet been manually labeled in the MSC data.

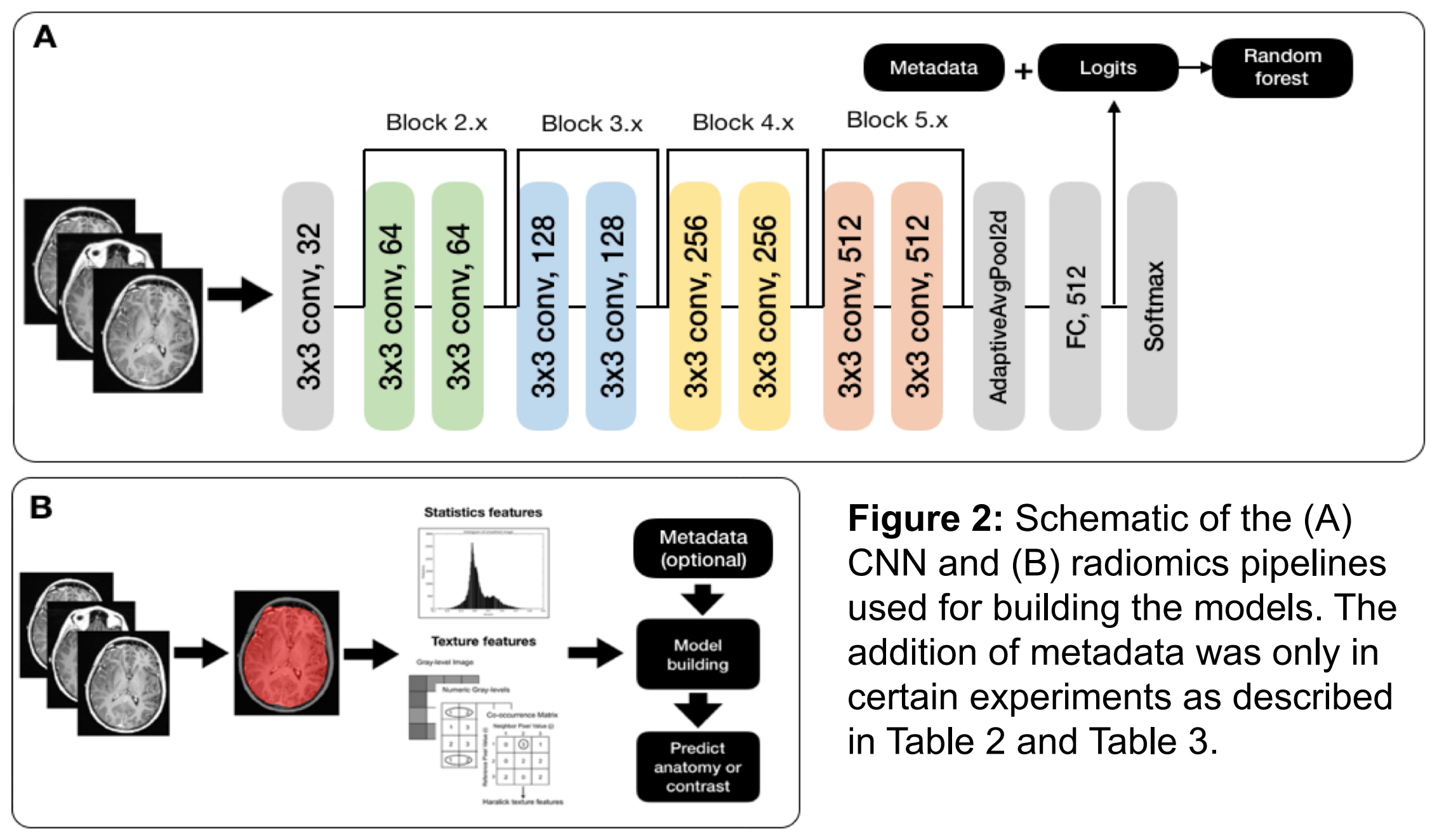
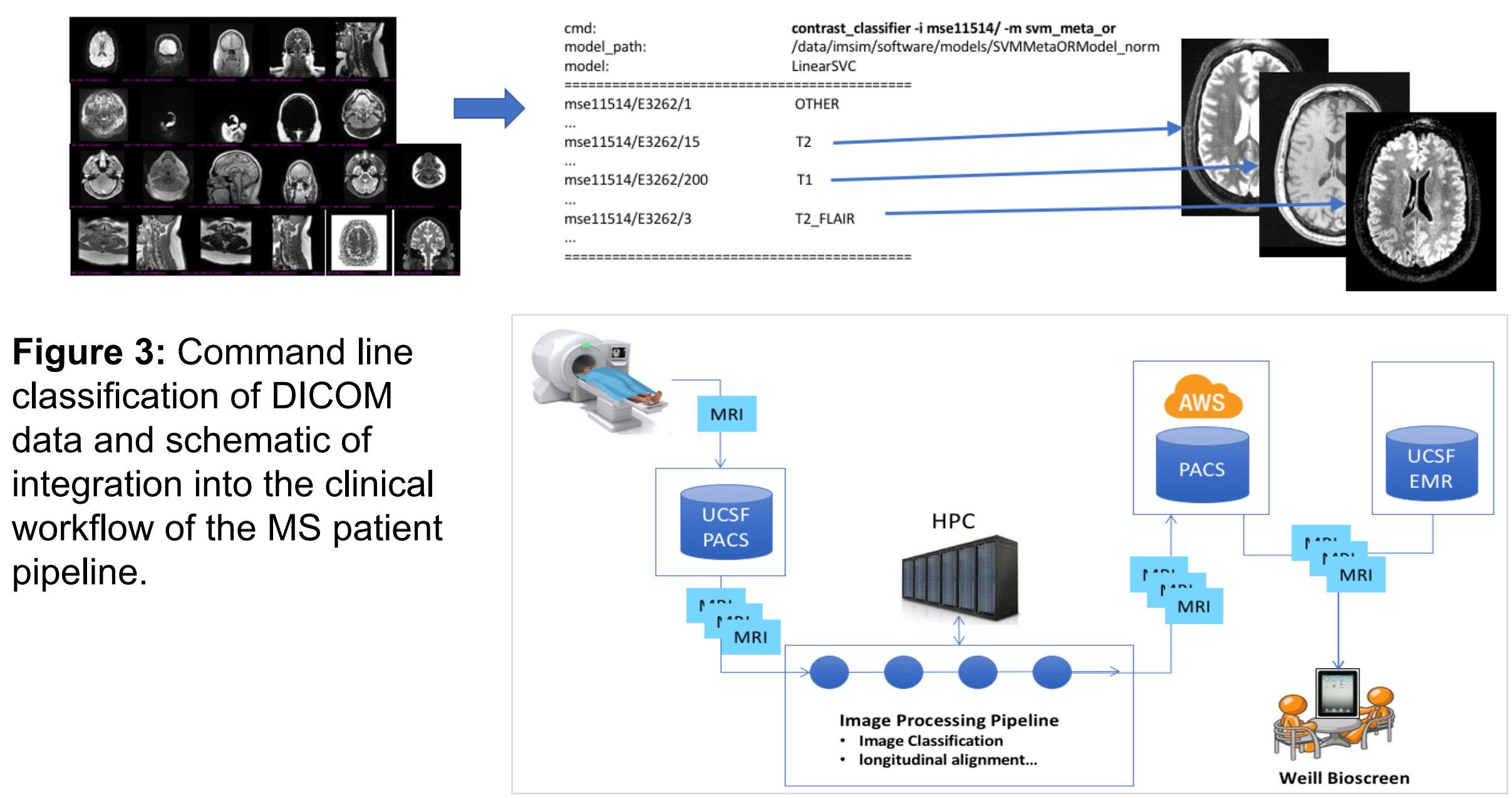


Figure 2: Schematic of the (A) CNN and (B) radiomics pipelines used for building the models. The addition of metadata was only in certain experiments as described in Table 2 and Table 3.

Clinical Integration

- Command line MRI contrast classifier was developed: reads DICOM exam and reports the predicted classes (Figure 3)
- Automated pipeline:
  - Delivers longitudinally registered T1, T2, T1+contrast, PD and T2 FLAIR images
  - Web-based MS application that launches from the UCSF EMR
  - Clinical and research data (biomarker, genetics and imaging data) in a contextualized display to facilitate patient/clinician consultations (Figure 4).
- Can serve 7000 patients and 15 clinicians with ~1000 projected patient visits per year (processing ~10,000 MRI exams/year)
- Image classifier and registration pipeline is integrated into the MS application



```
cmd:           contrast_classifier -i mse11514/ -m svm_meta_or
model_path:    /data/imsim/software/models/SVMMetaORModel_norm
model:         LinearSVC
============================================
mse11514/E3262/1          OTHER
...
mse11514/E3262/15         T2
...
mse11514/E3262/200        T1
...
mse11514/E3262/3          T2_FLAIR
...
============================================
```

Figure 3: Command line classification of DICOM data and schematic of integration into the clinical workflow of the MS patient pipeline.

Results + Discussion

- Tables 2 and 3 describe the overall accuracy and per-class accuracy for each experiment from the first (MSC) and second (glioma) train/test splits.
- ResNet-34 + metadata trained on MSR + MSC and tested on MSC resulted in 95.6% overall accuracy and >95% per-class accuracy
- ResNet-50+metadata trained on MSR + MSC and tested on glioma data resulted in 99.6% accuracy >97% per-class accuracy.
- Suggest that imaging features and metadata work synergistically to represent the contrast of brain MR images.
- Imaging features automatically generated through training a CNN are overall more valuable for classifying image contrast than standard IBSI features.
- Despite the distinct imaging features introduced by glioma presence, the glioma data were acquired with uniform imaging protocols compared with the MSC cohort which could explain why our between-disease results are why our between-disease results are superior than our within-disease results.
- Manually labeling the MSC dataset introduces some error into the ground truth labels.
- Next, expand this to more contrast types (e.g. diffusion, perfusion) and test on a cohort from a different disease from a different institution.

		Overall Accuracy	T1	T1C	T2	T2-FLAIR	PD	OTHER
Baselines	Heuristic	0.752	0.52	n/a	0.23	0.99	0.43	0.96
	SVM+Metadata	0.82	0.86	0.62	0.66	1	1	0.85
CNNs	ResNet-18	0.892	0.87	0.81	0.9	0.93	0.9	0.93
	ResNet-18+Metadata	0.941	0.98	0.95	0.9	0.98	0.9	0.95
	ResNet-34	0.886	0.86	0.82	0.88	0.92	0.9	0.92
	ResNet-34+Metadata	0.956	0.95	0.98	0.9	0.99	0.9	0.95
	ResNet-50	0.866	0.95	0.98	0.9	0.98	0.9	0.95
	ResNet-50+Metadata	0.953	0.95	0.89	0.9	0.95	0.9	0.92
Radiomics	IBSI	0.528	0.5	0.27	0.66	0.2	0	0.74
	IBSI + Metadata	0.876	0.89	0.74	0.95	1	0	0.86

Table 2. Overall and per-class accuracy of algorithms that were trained on MSR + 60% of the MSC cohort, validated on 20% of the MSC cohort. The metadata features were concatenated with the 6 ResNet logits and a random forest was used to predict the final overall accuracy.

		Overall Accuracy	T1	T1C	T2	T2-FLAIR	PD	OTHER
Metadata	SVM+Metadata	0.76	0.67	0.56	0.9	0.93	n/a	n/a
CNNs	ResNet-18	0.98	0.98	0.99	0.97	1	n/a	n/a
	ResNet-18+Metadata	0.99	1	0.99	0.99	1	n/a	n/a
	ResNet-34	0.98	1	0.91	0.89	0.92	n/a	n/a
	ResNet-34+Metadata	0.99	1	0.97	0.99	1	n/a	n/a
	ResNet-50	0.99	1	0.97	0.99	1	n/a	n/a
	ResNet-50+Metadata	**0.99**	**1**	**0.97**	**0.99**	**1**	**n/a**	**n/a**
Radiomics	IBSI	0.49	0.79	0.84	0.27	0.1	n/a	n/a
	IBSI + Metadata	0.83	0.53	0.93	0.91	0.93	n/a	n/a

Table 3. Overall and per-class accuracy of algorithms that were trained on MSR + MSC cohort, validated on 50% of the RECglioma cohort and 50% of the NEWglioma cohort. The metadata features were concatenated with the 6 ResNet logits and a random forest was used to predict the final overall accuracy.

Conclusions

- We compare the results of using DICOM metadata, IBSI radiomics imaging features, and CNN imaging features to classify the contrast of brain MRI images.
- We conclude that for both within-disease and between-disease prediction, deep ResNets for imaging feature extraction combined in a random forest with DICOM metadata performs best and can be deployed to reliably deliver images of specific contrasts for clinical use or large imaging analyses.

Acknowledgements

This work was funded by PCORI, the Hilton Foundation, the Weill Institute for the Neurosciences, the Valhalla Foundation, gifts from Friends of the Multiple Sclerosis Research Group at UCSF, and a grant from an anonymous donor.

References

[1] University of California, San Francisco MS-EPIC Team:, Cree BAC, Gourraud P-A, et al. Long-term evolution of multiple sclerosis disability in the treatment era. Ann Neurol 2016;80(4):499–510. [2] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research2011; 12, p 2825-2830; [3] Hatt M, Vallieres M, Visvikis D, Zwanenburg A. IBSI: an international community radiomics standardization initiative. Journal of Nuclear Medicine 2018; [4] van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7. [5] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 770–8.