An Unsupervised Approach for Artifact Severity Scoring in Multi-Contrast MR Images

Savannah P. Hays[1], Lianrui Zuo[2], Blake E. Dewey[3], Samuel W. Remedios[4], Jinwei Zhang[1], Ellen M. Mowry[3], Scott D. Newsome[3], Aaron Carass[1], Jerry L. Prince[1,4]

Departments of Electrical and Computer Engineering[1], Neurology[3], Computer Science[4], Johns Hopkins University, Baltimore, MD, USA
Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA[2]

MIDL Salt Lake City 2025

Introduction

1
- **Motivation:**
  - There exists a need for robust, interpretable, and scalable quality assurance (QA) in MRI, especially for large, multi-site datasets
- **Challenges with existing methods:**
  - Manual QA is subjective and slow[1]
  - MRIQC[2] takes several minutes to run and struggles with certain modalities
  - Supervised methods require labels and struggle to generalize

Our Contribution

- An unsupervised, interpretable framework that uses contrastive learning and simulated artifacts to assign artifact severity scores to MR images
- Works across T1-w, T2-w, FLAIR, and PD images without the need for preprocessing

Methods: Dataset, Model Training, and Model Testing

2
- **Dataset:**
  - 297 high-quality structural MR volumes from the TRaditional vs. Early Aggressive Therapy for Multiple Sclerosis (TREAT-MS) pragmatic, clinical trial (NCT03500328)
  - Acquired from 7 different imaging sites and included T1-w, T2-w, FLAIR, and proton density images

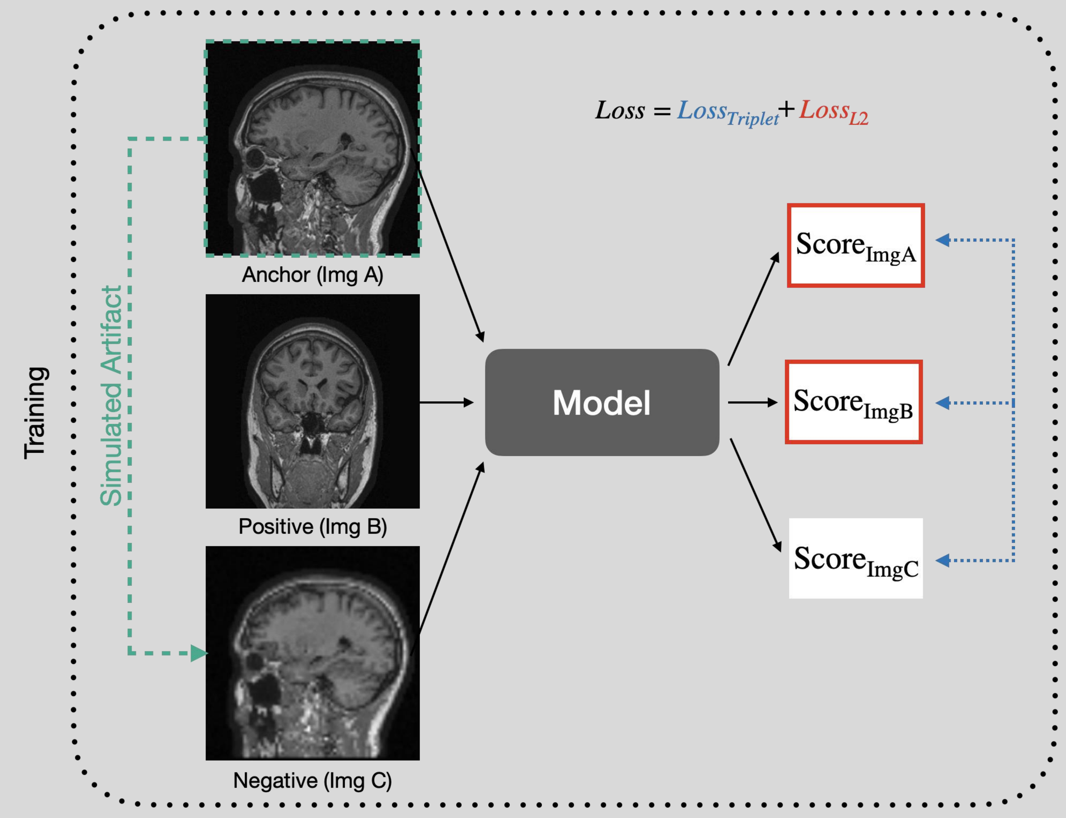

Figure 1: During training, three images are used to calculate the total loss. Img A and Img B are two different, clean image slices, while Img C is Img A with a randomly simulated artifact.

- **Model Training:**
  - Input: 2D slices, which are resized to 224x224
  - Simulated artifacts using the TorchIO library for random bias, random noise, random anisotropy, and random ghosting
  - Triplet loss uses the assigned severity score (SS) as the margin to adapt based on severity level
  - L2 loss anchors clean images to low scores

Table 1: Each artifact and its parameters used in the severity score (SS). The parameters were uniformly sampled in the corresponding range to ensure continuity of the artifact space.

Artifact	Input	Parameters	Severity Score (SS)
Noise	std	$\mathcal{U}[0.005, 0.2]$	$\frac{\text{std} - 0.005}{0.2 - 0.005}$
Ghosting	num_ghosts intensity	$\mathcal{U}\{2, \dots, 10\}$ $\mathcal{U}\{0.2, 1.5\}$	$\frac{(\text{intensity} - 0.2) + \frac{\text{num\_ghosts}}{10}}{(1.5 - 0.2) + 1}$
Bias Field	coefficients	$\mathcal{U}[0.01, 0.3]$	$\frac{\text{coefficients} - 0.01}{0.3 - 0.01}$
Anisotropy	scale	$\mathcal{U}[1, 4]$	$\frac{\text{scale} - 1}{4 - 1}$

- **Model Testing:**
  - During inference, a 3D volume can be input as 2D slices
  - Volumetric score is computed from the average of the middle 60% of slices

Methods: Simulated Artifacts

3

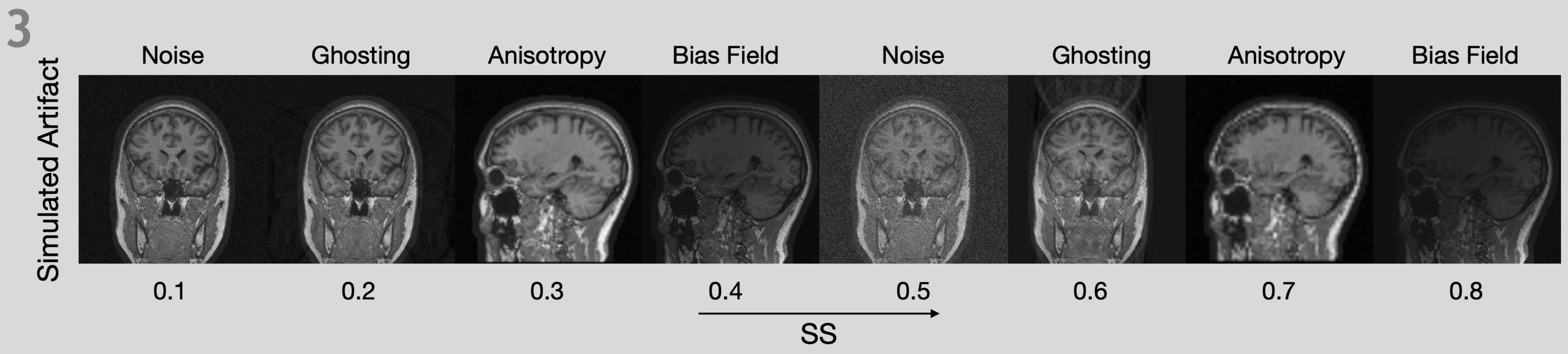

Figure 2: Increasing SS in the range [0, 1] (left to right) of selected simulated artifacts seen during training. During training, each artifact type is simulated on a continuous [0,1] scale.

Results

4

Public Dataset (OASIS3, N=20)

- MRIQC often failed, especially on T2w images
- Our model: ~1 second per volume (MRIQC: 7-9 minutes per volume)

Table 2: Simulated artifact type and SS with MRIQC and our model results. We report the Pearson coefficient between the SS and result.

Artifact Type	SS	MRIQC				Ours↓
		CJV↓	CNR↑	EFC↓	FBER↑	
None	0.0	0.77 ± 0.13	1.14 ± 0.25	0.49 ± 0.05	6962 ± 2097	0.17 ± 0.47
Bias	0.1	0.80 ± 0.19	1.08 ± 0.27	0.51 ± 0.60	2367 ± 1572	0.01 ± 0.03
Motion	0.3	0.87 ± 0.19	1.05 ± 0.27	0.51 ± 0.05	5501 ± 2010	1.74 ± 0.49
Anisotropy	0.6	0.78 ± 0.05	1.32 ± 0.23	0.52 ± 0.05	9876 ± 3326	2.32 ± 0.48
Ghosting	0.8	1.03 ± 0.18	0.82 ± 0.12	0.53 ± 0.06	4161 ± 1852	2.34 ± 0.60
Noise	0.9	–	–	–	–	3.54 ± 0.06
Pearson		**0.36**	**-0.16**	**0.13**	**0.35**	**0.92**

Clinical Dataset (TREAT-MS, N=124)

- High-quality images score near 0, poor-quality images score > 1
- 83% of clinical scans flagged as low quality (expected due to 2D acquisitions)

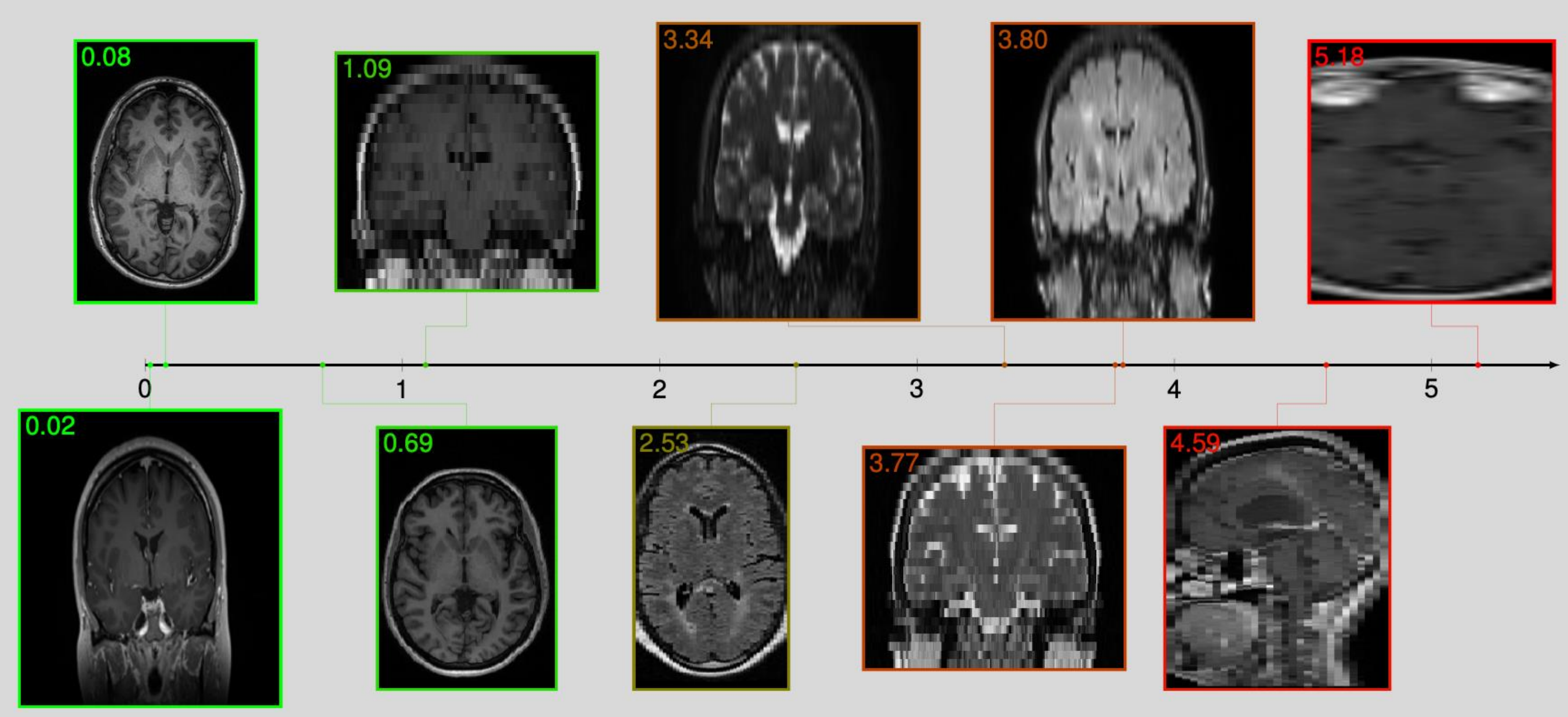

Figure 3: Clinically acquired images from the TREAT-MS dataset with our model scores. Although images are acquired at various sites following a standardized protocol, several low resolution 2D acquisitions are observed.

Discussion and Conclusion

5
- **Key Takeaways:**
  - Scalable and unsupervised QA model for MR images.
  - Higher Pearson coefficient and faster runtime than MRIQC
  - Handles diverse contrasts and artifacts without preprocessing
  - Threshold of 1 can be used to separate acceptable from poor-quality images.
- **Future Work:**
  - Simulate combinations of artifacts
  - Multi-dimensional score representing the level of each artifact type

References

6
[1]Alfaro-Almagro, et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *NeuroImage*, 166:400-424, 2018.
[2]Esteban et al. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9):1-21, 09 2017.

Acknowledgements: This material is partially supported by the Johns Hopkins University Percy Pierre Fellowship (Hays) and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139757 (Hays) and Grant No. DGE-1746891 (Remedios). Development is partially supported by FG-2008-36966 (Dewey), CDMRP W81XWH2010912 (Prince), NIH R01 CA253923 (Landman), NIH R01 CA275015 (Landman), the National MS Society grant RG-1507-05243 (Pham) and Patient-Centered Outcomes Research Institute (PCORI) grant MS-1610-37115 (Newsome and Mowry). The statements in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

shays6@jhu.edu