Published January 13, 2025 as 10.3174/ajnr.A8512

ORIGINAL RESEARCH

# AI generated synthetic STIR of the lumbar spine from T1 and T2 MRI sequences trained with open-source algorithms

Alice M.L. Santilli*, Mark A. Fontana* PhD, Erwin E. Xia MD, Zenas Igbinoba MD, Ek Tsoon Tan PhD, Darryl B. Sneag MD, J. Levi Chazen MD

## ABSTRACT

**BACKGROUND AND PURPOSE**: To train and evaluate an open-source generative adversarial networks (GANs) to create synthetic lumbar spine MRI STIR volumes from T1 and T2 sequences, providing a proof-of-concept that could allow for faster MRI examinations.

**MATERIALS AND METHODS**: 1817 MRI examinations with sagittal T1, T2, and STIR sequences were accumulated and randomly divided into training, validation, and test sets. GANs were trained to create synthetic STIR volumes using the T1 and T2 volumes as inputs, optimized using the validation set, then applied to the test set. Acquired and synthetic test set volumes were independently evaluated in a blinded, randomized fashion by three radiologists specializing in musculoskeletal imaging and neuroradiology. Readers assessed image quality, motion artifacts, perceived likelihood of the volume being acquired or synthetic, and presence of 7 pathologies.

**RESULTS:** The optimal model leveraged a customized loss function that accentuated foreground pixels, achieving a structural similarity imaging metric (SSIM) of 0.842, mean absolute error (MAE) of 0.028, and peak signal to noise ratio (PSNR) of 26.367. Radiologists could distinguish synthetic from acquired volumes; however, the synthetic volumes were of equal or better quality in 77% of test patients and demonstrated equivalent or decreased motion artifacts in 78% of test patients. For common pathologies, the synthetic volumes had high positive predictive value (75-100%) but lower sensitivity (0-67%).

**CONCLUSIONS:** This work links objective computer vision performance metrics and subject clinical evaluation of synthetic spine MRIs using open-source and reproducible methodologies. High-quality synthetic volumes are generated, reproducing many important pathologies, demonstrating a potential means for expediting imaging protocols.

**ABBREVIATIONS:** AI = Artificial Intelligence; GANs = general adversarial networks; aqSTIR = acquired STIR volume; sSTIR = synthetically generated STIR volume; SSIM = structural similarity imaging metric; PSNR = peak signal to noise ratio; MAE = mean absolute error.

Received June 19 2024; accepted after revision August 22 2024.
From the Orthopedic Data Innovation Lab (ODIL), Hospital for Special Surgery (A.M.L.S., M.A.F.), Department of Radiology and Imaging, Hospital for Special Surgery Centre (E.E.X, Z.I, E.T.T, D.B.S, J.L.C) and Department of Population Health Sciences, Weill Cornell Medicine (M.A.F), New York, New York, USA.

**DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST:**

Darryl Sneag is a consultant for RemedyLogic and GE Healthcare. Ek Tsoon Tan received institutional research support from GE Healthcare, Siemens Healthcare, and AMAG Pharmaceutical Inc. All other authors declare no conflicts of interest with respect to this work.

Please address correspondence to Alice M.L. Santilli, MSc, Orthopedic Data Innovation Lab, Hospital for Special Surgery, 535 E 70th Street, New York, New York, USA; santillia@hss.edu

## SUMMARY SECTION

**KEY FINDINGS:** Generative AI can produce *synthetic* lumbar spine STIR volumes from T1 and T2 volumes using open-source computational packages, achieving comparable image quality and reproducing many underlying spine pathologies as their acquired counterparts.

## INTRODUCTION

Lumbar spine magnetic resonance imaging (MRI) is one of the most common medical imaging examinations. It is the core diagnostic tool for low back pain or lower extremity radiculopathy in patients with persistent or progressive symptoms who have failed 6 weeks of conservative therapy or have clinical “red flags” suspicious for underlying neurological compromise, infection, fracture, or cancer [1]. Thirty minutes of scan (acquisition) time is typically allotted to complete a lumbar spine MRI examination, as multiple sequences are required diagnostically for interpretation [2]. Specifically, these sequences are T1-, T2-weighted, and fluid-sensitive weighted, the latter

Copyright 2025 by American Society of Neuroradiology.

typically involving a short-tau inversion recovery (STIR) technique for fat suppression. However, up to 50% of patients presenting for an MRI report some level of anxiety related to claustrophobia or apprehension and often have difficulty holding still for long periods of time due to pain or discomfort [3], which increases the risk for motion artifacts, repeat scans and overall prolonged exams. These resultantly increase patient stress and healthcare delivery costs. It is estimated that hospitals incur costs of $115,000 per scanner per year in lost revenue from motion artifacts [4].

Artificial intelligence (AI) and generalized adversarial networks (GANs) have the potential to generate synthetic images, which may help reduce scan times and improve patient compliance and comfort. GANs are a class of machine learning algorithms trained to create new data that resemble the training data. They have been previously explored for text and image generation and have demonstrated their potential for synthesizing brain MRI images [5] and cross-modality image-to-image translation [6].

This study aims to create high quality, diagnostically informative synthetic STIR volumes from T1 and T2 volumes using a large dataset of lumbar spine MRIs. The synthetic volumes are evaluated by experienced radiologists in a blinded, randomized fashion. Previous work demonstrates the feasibility of creating synthetic spine STIR volumes but are limited by the smaller size of their datasets and their minimal reporting of technical methodology and objective performance metrics [7,8]. This work builds on these previous investigations by exploring a greater diversity of pathologies and subjective clinical metrics, as well as including detailed reporting of model performance using standard, objective computer vision metrics. Thus, creating one of the first direct comparisons between the two types of evaluation. To increase transparency and reproducibility of GANs in medical imaging, this manuscript also provides thorough technical detail of the selection of loss functions and model other hyperparameters.

## MATERIALS AND METHODS

### Image Selection and Preprocessing

All lumbar spine MRIs performed at a single institution on either 1.5- or 3-Tesla MRI (GE Healthcare, Waukesha, WI) between August and October 2023 were collected, yielding 1817 examinations. Associated T1, T2, and STIR imaging sequences were extracted. Each volume, which we use to refer to a set of images from a sequence, had between 21 and 35 slices (mean 24 slices), a slice thickness of 3.5mm, and in-plane image dimensions of either 256, 512 or 1024 pixels. Within a set of T1, T2 and STIR acquisitions, the spatial resolutions were identical. This continuously collected sample is representative of the distribution of patients and conditions requiring spine imaging at our hospital during this period (with no exclusions based on pathology, presence of spinal instrumentation hardware, or clinical history). Table 1 provides a demographics summary of included patients. Studies were randomly divided on a patient level into training (n=1633), validation (n=84), and test sets (n=100). Additional details on image preprocessing can be found in the supplemental materials.

### Model Architecture

Figure 1 presents an overview of our AI methodology. Three acquired MRI series (T1, T2, STIR) were used to train the GAN to generate synthetic STIR volumes. GANs are machine learning algorithms where two models, a generator and a discriminator, compete to improve their performances by, respectively, generating synthetic STIR (sSTIR) images, and discriminating between the synthetic and acquired STIR (aqSTIR). This competition pushes the generator model to "learn" to generate more realistic images with the goal of "fooling" the discriminator. The open-source conditional GAN model, Pix2Pix [9], was used as a base architecture. The generator is a U-net with 8 downsampling convolutions and 7 upsampling convolution layers. The discriminator model is a convolutional PatchGAN classifier. Both models utilized the Adam optimizer. All models were run for 60 epochs, beyond which overfitting occurred.

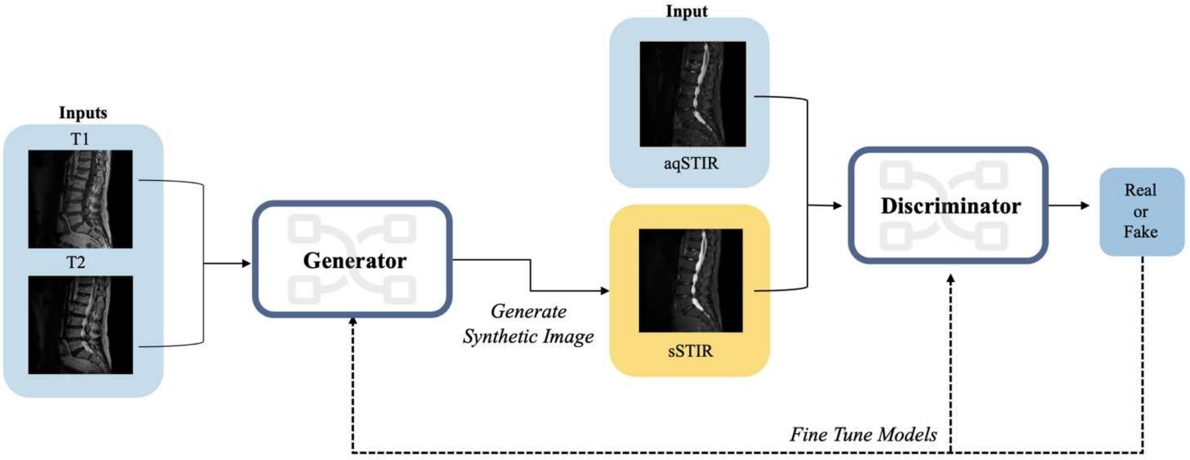

**FIG 1.** Overview of Generative adversarial network (GAN) for producing synthetic STIR volumes. Each patient's three acquired MRI volumes (T1, T2, STIR) are utilized during training of a GAN. The two models, generator and discriminator, compete to improve their own performance by either generating (more realistic) synthetic STIR (sSTIR) images or discriminating between synthetic and acquired STIR (aqSTIR).

### *Model Optimization: Loss Function and Hyperparameter Tuning*

For model optimization, the model performance was assessed using 3 objective metrics: structural-similarity-index (SSIM), mean absolute error (MAE), and peak-signal-to-noise ratio (PSNR).

The generator and discriminator models each have their own loss functions and hyperparameters. The loss function is an equation that calculates the difference, minimized through training, between the output of a model and the "true" target outcome. The learning rate and momentum values are model hyperparameters that influence the rate at which newly acquired information is "learned" to minimize the loss function.

To optimize model performance, a variety of loss functions were evaluated, details of which can be found in the supplemental materials. From preliminary results, a customized loss function was developed where training was focused on the MAE + Sobel loss on foreground pixels (deemed "*Foreground-MAE-loss")*. The Sobel loss uses an edge detection filter to direct the models focus on the edges of spinal structures [10]. All pixels in the image were first binned based on their intensity values, where the lowest bin identified the lowest intensity pixels, i.e. "background" pixels, which are generally of no interest. The MAE was then calculated using the remaining pixels, i.e., foreground pixels, which contain the image-specific spine structures and soft tissue information. After the loss function was selected, a grid search was performed, using varying discriminator and generator learning rates (0.00004 to 0.0005 in increments of 0.0001) and momentum values (0.4 to 0.8 in increments of 0.1). The final loss function and hyperparameters were selected based on the highest performance on the validation set.

### *Radiologist Evaluation*

It is imperative in a clinical environment that abnormalities present in the acquired volumes also appear in the synthetic volumes, while also avoiding the insertion of "fake" pathologies or "hallucinations" [11]. To assess the diagnostic equivalence of the aqSTIR and sSTIR volumes, 3 board-certified radiologists with fellowship training in either musculoskeletal imaging or neuroradiology completed blinded, randomized evaluations of the volumes in the test set. The test set comprised 100 subjects, each of which had one aqSTIR volume and one GAN-generated sSTIR volume. These 200 volumes were then anonymized and uploaded to a picture archiving communication system (PACS) (Sectra IDS, Linköping, Sweden) for radiologic grading. Each radiologist saw all 200 volumes, divided into two sessions, with a two week wash out period between sessions. Each volume (aqSTIR and sSTIR) for a given patient was randomly assigned to either the first or second session, independently for each radiologist, ensuring no overlap between sessions. The order of the volumes was also independently randomized. Radiologists were unable to view the image evaluations of the other radiologists at any time during the study.

For each volume, the radiologists evaluated overall image quality  (1 = unacceptable, 2 = poor, 3 = acceptable, 4 = good, 5 = excellent), degree of motion artifacts (0 = absent, 1 = mild, 2 = moderate, 3 = severe), whether the volume was acquired or synthetic (1 = definitely synthetic, 2 = probably synthetic, 3 = unsure, 4 = probably acquired, 5 = definitively acquired), and if fat suppression was homogeneous (0/1). Descriptive statistics on these evaluations for each reviewer were calculated and reported.

To assess clinical performance, radiologists were asked whether each volume contained one or more of 6 common lumbar spine pathologies *typically apparent* on STIR images, listed in the results section, or "None". To directly compare the acquired and synthetic results from the radiologists' evaluations, majority-rule was applied to their findings to create the ground truths (from aqSTIR) and the predictions (from sSTIR) for each pathology and patient in the test set. A pathology was considered present in each acquired STIR volume if 2 or 3 reviewers identified it as present, and similarly for each synthetic volume. For each pathology, predictive performance metrics were reported (positive predictive value, PPV; sensitivity; negative predictive value, NPV; and specificity). CLAIM checklist methodology was followed where applicable.

## RESULTS

### *Model Performance and Selection of Loss Functions and Hyperparameters*

All three evaluated metrics plateaued in terms of performance with the traditional binary cross entropy (BCE) and MAE loss functions. Results comparing model performance on the validation set given different training loss functions are provided in Supplemental Table S1. *Foreground-MAE-loss* with a Sobel filter achieved the best performance (SSIM=0.842, MAE=0.028 and PSNR=26.367). The model took ~2.3 seconds to generate a full synthetic STIR volume per patient.

Model selection was primarily based on SSIM performance. From visual inspection, the SSIM metric was determined to be the best proxy for image sharpness, producing the best numerical representation of "image quality" on which to select our models. Figure 2 illustrates one patient example, where the overall sharpness and specific quality/definition of the basivertebral venous plexus was highest with our preferred loss function. Following selection of the loss function, grid search yielded optimal model performance with a discriminator learning rate of 0.00004. discriminator momentum of 0.7, generator learning rate of 0.00008, and generator momentum of 0.6.

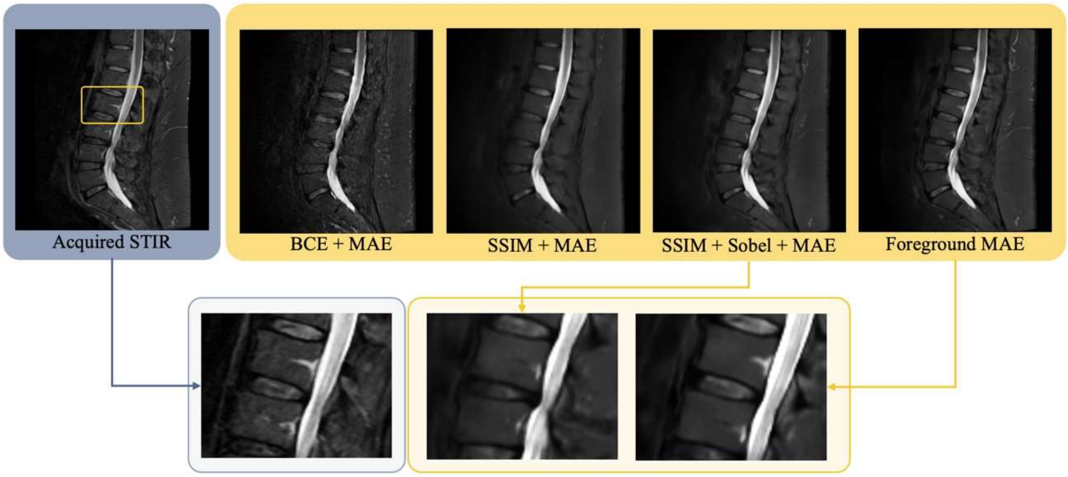

**FIG 2.** (Above) Example of an acquired STIR volume slice from one patient (blue) and the same patient's equivalent synthetic volume slices generated from GANs with 4 different loss functions (yellow). (Below) Three zoomed images allowing comparison of the quality of basi vertebral venous plexus between the acquired volume slice and two synthetic volume slices (SSIM+Sobel+MAE and Foreground-MAE).

## *Radiologist Evaluations*

### *Image Quality, Motion, Fat Suppression*

Table 2 displays summary statistics from radiologists' assessments of the acquired and synthetic STIR volumes from the test set. The reviewers tended to agree with each other, although reviewer 2 consistently assigned worse metrics to both volume types. On average, reviewers could distinguish between the two, rating the acquired volumes correctly as likely acquired and the synthetic volumes as likely synthetic. However, the intra-rater agreement for the quality metrics was high, with raters assigning similar metrics to the synthetic and acquired volumes. For 77% of patients in the test set, the majority of reviewers graded the synthetic volume with equivalent or better quality than the acquired volume. Motion artifacts were also comparable, with 78% of synthetic volumes demonstrating equivalent or decreased motion artifacts. Finally, 96% of the synthetic volumes were graded with homogenous fat suppression (Supplemental Table S2). Figure 3 shows an example of an acquired and synthetic slice from a patient in the test set with evaluated high quality.

### *Identification of Pathologies*

Table 3 details the percentage of patients in the test set with each pathology according to 2 or 3 radiologists. The results in both the aqSTIR and sSTIR volumes demonstrates the substantial variation and reviewer disagreement even among the acquired volumes. Paraspinous muscle edema, Modic type 1 change (M1C) and "None", were the most common pathologies assigned. While edema and M1C were the most common pathologies in both acquired and synthetic volumes, they were less common among the synthetic volumes. Conversely, "None" was more prevalent in the synthetic volumes compared to the acquired volumes.

However, the percentages of each pathology in the aqSTIR and sSTIR volumes alone do not suggest whether the pathologies reported on the sSTIR volumes are true positives or incorrectly generated. The aqSTIR and sSTIR volumes were therefore directly compared to produce the model's predictive performance statistics, reported in Table 4, using a majority-rule aggregator as described in the methodology section above. Some pathologies were not present in the test set. Overall high PPVs (75%-100%) indicate that when a pathology was detected in the synthetic volume, it was typically also found in the associated acquired volume. However, lower sensitivities (0%-67%) suggests that the synthetic volumes occasionally "missed" pathologies that were in fact found in the aqSTIR counterpart. Overall, the GAN-generated synthetic volumes were conservative in their inclusion of pathologies. This is also shown by the high proportion of cases (50%) that were deemed to have no pathologies ("None") according to the synthetic volumes compared to only 25% among the acquired volumes (Table 3).

Of note, two common image types that are sometimes removed from datasets in previous studies [12,13] were retained in our dataset: instrumented spines (due to susceptibility effect from metallic hardware) and scoliosis (due to lack of confidence in assessing stenosis due to spinal curvatures). Figure 4 displays four examples of cases selected from the test set with different pathologies and imaging features, including instrumented spines and scoliosis. There were 15 patients in the test set with instrumentation and another 15 with scoliosis. Their synthetic volumes were on average assigned a quality metrics of 3.3 and 3.4, respectively.

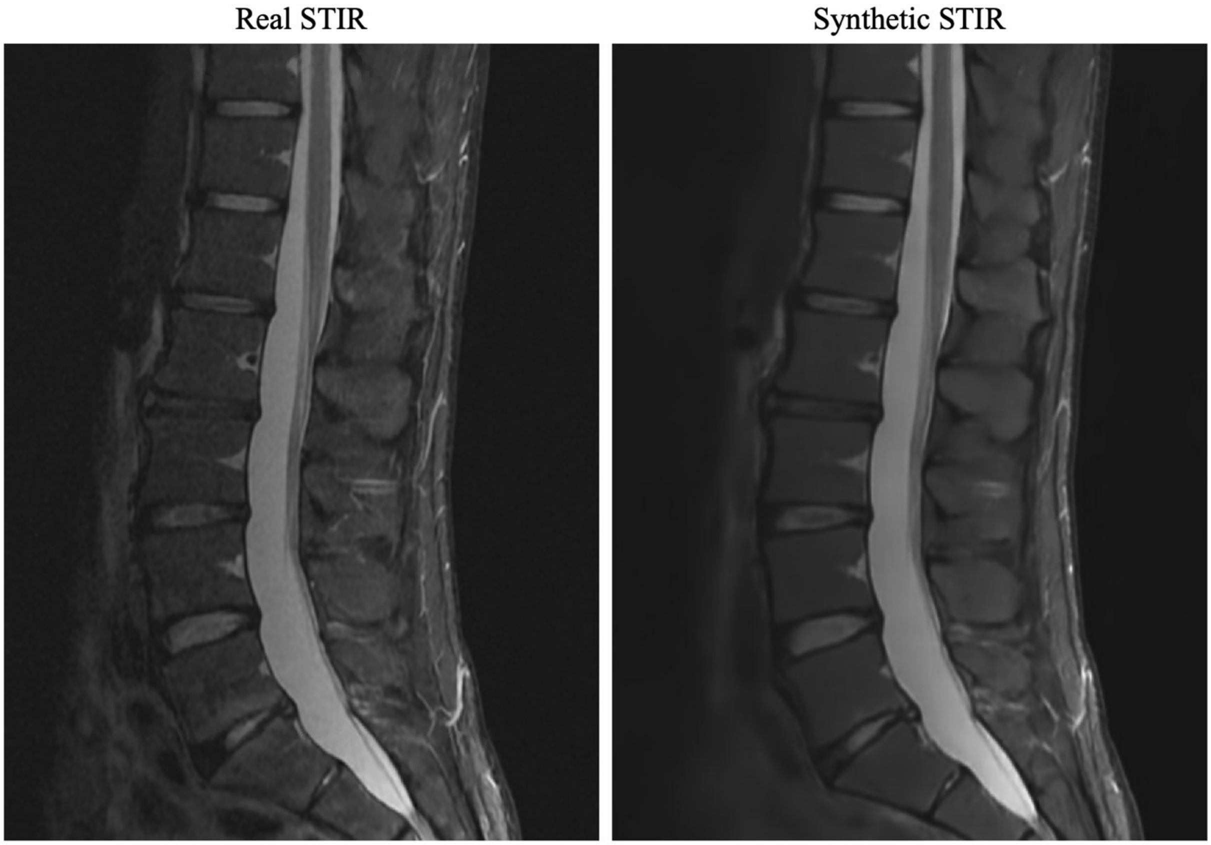

**FIG 3**. *A slice from a randomly selected volume from the test set which was assigned a quality metric of 5/5 from the three reviewers. Left: slice from the acquired STIR volume. Right: same slice from the synthetic STIR volume*

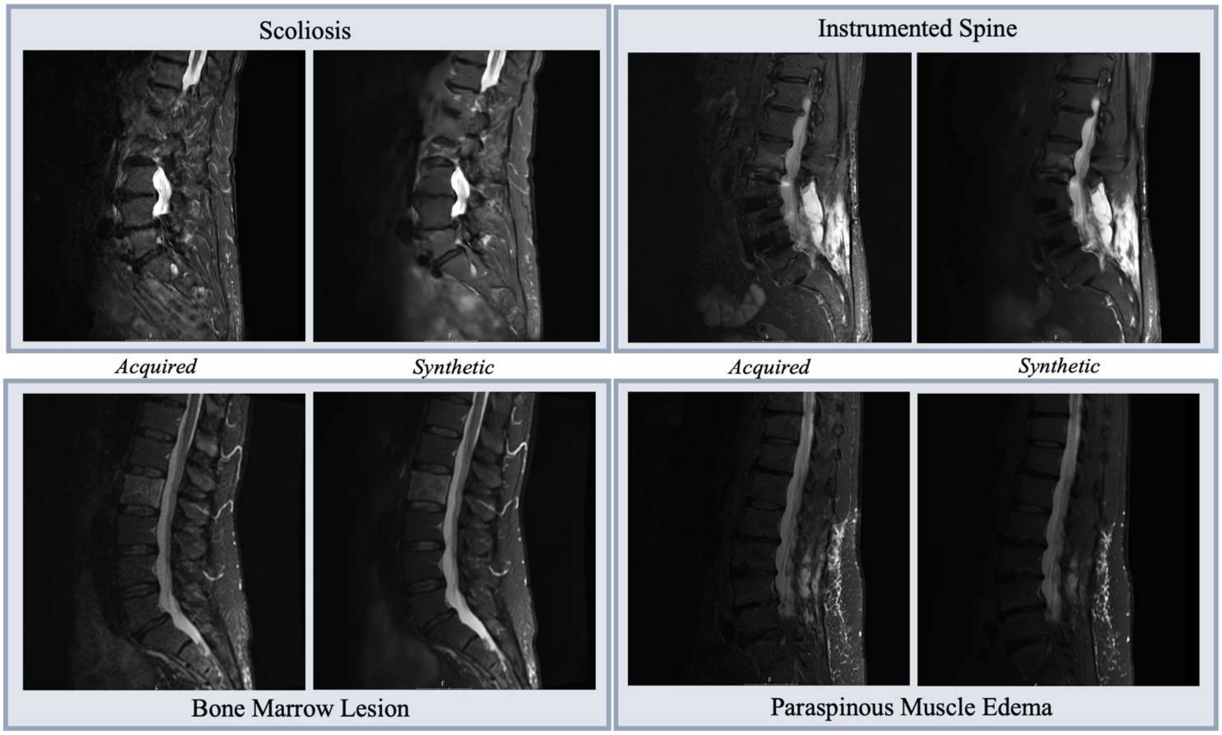

**FIG 4.** *Examples of a variety of imaging features. Each pair shows a sample sagittal slice from the acquired STIR volume with its corresponding model-generated synthetic STIR volume.*

## DISCUSSION

Fat-suppressed, fluid-sensitive imaging using STIR is a critical part of routine spine imaging protocols given its ability to depict marrow and soft tissue lesions with uniform fat suppression. However, with prolonged scan times, anxiety regarding time spent in the MRI scanner, and costs of repeated imaging due to motion artifacts, the potential impact of synthetically-generated STIR imaging in alleviating these bottlenecks is high. This study describes the AI generation of synthetic STIR sequences from non-contrast T1 and T2-weighted sagittal lumbar spine series and validates them in a blinded, randomized evaluation comparing synthetic and acquired STIR series by expert radiologists.

Previous work on synthetic MRI sequence generation [8, 14] has been performed but generally using proprietary commercial software lacking methodological detail and none have reported objective model performances metrics common in computer vision research against which we could compare. This study demonstrated performance improvements based on hyperparameter tuning and loss function customization. The combined loss function of SSIM, Sobel filter loss, and foreground MAE loss proved successful at overcoming high proportion of "blank" background pixels, the desire for precise image details, and the presence of original motion artifacts. No single set of parameters will be optimal for all medical applications of GANs, but providing these technical and methodological detail is a positive step towards greater reproducibility of generating clinically robust synthetic MRI images.

In terms of radiologist-assessed quality, using the same Likert scale as Tannenbaum et al. [8], we found average quality metrics of 3.61 (std 0.22) for the aqSTIR and 3.38 (std 0.33) for the sSTIR; theirs were 3.21 (std 1.08) for their aqSTIR and 3.71 (std 1.14) sSTIR. Tannenbaum found the average quality metric for their sSTIR higher than the aqSTIR, while we found the opposite. In discussing their results, they noted their sSTIR's appeared relatively smooth and interpreted this smoothness as representing lower degree of motion artifacts, while our radiologists interpreted this same smoothing in the sSTIR as a reduction in image quality. In addition, the standard deviations of our quality measurements are lower than those in Tannenbaum et al., consistent with greater reviewer agreement. They also evaluated a different list of pathologies, finding no difference between the acquired and synthetic STIRs, but did not calculate predictive performance statistic (e.g., PPV, sensitivity) as we did, making direct comparison difficult.

The data selected for this study were from patients scanned with either one of two commonly used magnetic field strengths (1.5T or 3T) over a 3-month period at our institution. The pathologies present in this dataset were not specifically selected for, nor were certain images with troublesome features removed. Given this, the model presents encouraging clinical validity in reproducing image artifacts relevant for diagnosis, some of which were rare. It also demonstrates that our sampling and computation methodology led to a robust model able to generate different patient pathologies without explicit dataset curation. One of the limitations of assessing synthetic image quality is the lack of a consensus on the ground truth acquired imaging [15]. The pathologies evaluated in our study varied largely across reviewers (even among the acquired images), which renders it difficult to confirm if something "should" be present in the synthetic image. Our model produces synthetic images with high PPV and relatively lower sensitivity and is hence conservative in its diagnostic representations. It is important to note that no systematic evidence of the GAN producing major "hallucination" artifacts were found.

Given the centrality of STIR imaging for patient care and diagnosis, synthetic STIR image generation still requires more work before its clinical implementation. Although the results presented here were trained with a large dataset, the work is still limited by imaging from one scanner type, focused on the lumbar region, from one institution. Future work should prioritize augmenting training with a more comprehensive list of specific or rare pathologies and adding images from different scanners. It may also be of interest to measure the effect of acquisition order between the T1, T2 and STIR series on the severity of motion artifacts found in the images. Nonetheless, this exploratory investigation supports the assertion that synthetic STIR imaging could be a viable option in lieu of a directly acquired STIR sequence.

## CONCLUSIONS

It is feasible to generate STIR volumes using generative AI algorithms and T1 and T2 volumes as inputs. Based on a large training set, high quality synthetic images were produced that accurately represent many important pathologies seen in acquired STIR volumes. The transparent reporting of model building and optimization parameters in this work, as well as leveraging open-source, non-commercial computational packages help support the standard for reproducibility and evaluation of medical generative imaging models. Moreover, our reporting of computer vision performance metrics alongside clinical performance metrics creates a key, novel link between objective and subjective reporting criteria in the realm of medical imaging. The diagnostically conservative nature of our model requires further refinement to achieve diagnostic equivalency and suitability for clinical use, however this project provides a promising first step toward expediting imaging protocols and unlocking the potential of generative AI in medical spine imaging.

**Table 1: Patient Cohort and Imaging Descriptions**

Patient Cohort and Imaging Descriptions. Number of patients (including sex and age information), number of image volumes, and number of slices in each randomized data set. Note: each volume has between 21-35 slices (mean of 24), and the T1 and T2 slices are stacked together before being inputted into the model.

| Variables | Data Set |
|---|---|

| | Train | Validation | Test |
|---|---|---|---|
| **Patients** | 1633 | 84 | 100 |
| **Sex (% Female)** | 53% | 54% | 47% |
| **Age (Mean ± SD)** | 57.66 ± 17.7 | 55.05 ± 19.5 | 61.7 ± 17.8 |
| **Volumes** | 4899 | 252 | 300 |
| **T1/T2 Slices** | 42688 | 2121 | 2702 |

**Table 2. Summary Statistics of Radiologists' Assessments of Acquired and Synthetic Volumes**

Summary statistics of radiologists' assessments of acquired and synthetic STIR volumes from test set of 100 patients. Three radiologists (R1, R2, R3) each reviewed in a blinded fashion 1 synthetic STIR volume and 1 acquired STIR volume from 100 patients and evaluated each volume's quality (1 = Unacceptable, 2= Poor, 3= Acceptable, 4= Good, 5 = Excellent), degree of motion artifacts (0 = absent, 1=mild, 2=moderate, 3=severe), whether they believed the volume to be acquired or synthetic (1= Definitely Synthetic, 2= Probably Synthetic, 3=Unsure, 4=Probably Acquired, 5= Definitively Acquired) and whether or not the fat suppression is homogenous (Y/N) in the volume.

| | | Acquired STIR Volumes | | | Synthetic STIR Volumes | | |
|---|---|---|---|---|---|---|---|
| | | ***R1*** | ***R2*** | ***R3*** | ***R1*** | ***R2*** | ***R3*** |
| **Quality (1-5)** | *Average* | 3.92 | 3.38 | 3.53 | 3.83 | 3.01 | 3.31 |
| | *STD* | 0.84 | 1.17 | 0.79 | 1.18 | 1.27 | 1.08 |
| | *Median* | 4 | 3 | 4 | 4 | 3 | 4 |
| | *Min-Max* | 1-5 | 1-5 | 1-5 | 1-5 | 1-5 | 1-5 |
| **Motion Artifact (0-3)** | *Average* | 0.69 | 0.74 | 1.2 | 0.68 | 1.14 | 1.4 |
| | *STD* | 0.81 | 0.88 | 0.653 | 0.89 | 1.04 | 0.8 |
| | *Median* | 1 | 0.5 | 1 | 0 | 1 | 1 |
| | *Min-Max* | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 | 0-3 |
| **Acquired vs Synthetic (1-5)** | *Average* | 4.16 | 3.85 | 3.48 | 1.11 | 1.11 | 1.63 |
| | *STD* | 0.73 | 1.14 | 1.05 | 0.34 | 0.34 | 0.89 |
| | *Median* | 4 | 4 | 3 | 1 | 1 | 1 |
| | *Min-Max* | 2-5 | 1-5 | 1-5 | 1-4 | 1-3 | 1-5 |
| **Fat Suppression Y/N** | *Yes %* | 100% | 93% | 100% | 98% | 81% | 96% |

**Table 3. Radiologist-Assessed Occurrence of Pathologies Present in Acquired and Synthetic Volumes**

Percentage of acquired and synthetic volumes in the test set (n=100) with each pathology (or none) as determined by the three radiologists during the blinded review. Rows detail pathologies available to the 3 radiologist reviewers (including none). Columns detail the percentage of the 100 patients in the test set with each pathology according to 2/3 reviewers or 3/3 reviewers.

| | Acquired STIR | | Synthetic STIR | |
|---|---|---|---|---|
| **Pathologies** | *2/3 Reviewers* | *3/3 Reviewers* | *2/3 Reviewers* | *3/3 Reviewers* |
| **Bone Marrow Lesion** | 3% | 2% | 2% | 2% |
| **Spinal Cord Signal Change** | 0% | 0% | 0% | 0% |
| **Paraspinous Muscle Edema** | 24% | 9% | 9% | 3% |
| **Facet Arthritis with Synovitis** | 13% | 2% | 4% | 1% |
| **Fracture with Edema** | 2% | 1% | 0% | 0% |
| **Modic Type 1 Change** | 41% | 21% | 19% | 4% |
| **Synovial Cyst** | 0% | 0% | 0% | 0% |
| **None** | 25% | 9% | 50% | 21% |

**Table 4. Majority-Rule Pathology Predictive Performance**

Predictive performance statistics of pathologies from synthetic STIR volumes versus acquired STIR volumes in the test set of 100 patients. Three radiologists each reviewed in a blinded fashion the synthetic and acquired STIR volumes and evaluated the pathologies present in each of them. The rows detail the pathologies available to the 3 radiologist reviewers (including none). Their evaluations of presence of pathologies were aggregated using majority-rule; a pathology was considered present in each acquired STIR volume if 2 or 3 reviewers identified it as present, and similarly for each synthetic volume. The columns detail predictive performance metrics (numerator, denominator, and percentage), comparing the prediction (from sSTIR) to the ground truth (from aqSTIR) for each pathology. PPV = positive predictive value, NPV = negative predictive value. 95% confidence intervals for the percentage values are available in supplemental table S3.

| | PPV | | Sensitivity | | NPV | | Specificity | |
|---|---|---|---|---|---|---|---|---|
| **Pathologies** | *Ratio* | *%* | *Ratio* | *%* | *Ratio* | *%* | *Ratio* | *%* |
| **Bone Marrow Lesion** | 2/2 | 100% | 2/3 | 66.6% | 97/98 | 98.9% | 97/97 | 100% |
| **Spinal Cord Signal Change** | 0/0 | - | 0/0 | - | 100/100 | 100% | 100/100 | 100% |
| **Paraspinous Muscle Edema** | 9/9 | 100% | 9/24 | 37.5% | 76/91 | 83.5% | 76/76 | 100% |
| **Facet Arthritis with Synovitis** | 3/4 | 75% | 3/13 | 23% | 86/96 | 89.5% | 86/87 | 98.8% |
| **Fracture with Edema** | 0/0 | - | 0/2 | 0% | 98/100 | 98% | 98/98 | 100% |
| **Modic Type 1 Change** | 18/19 | 94.7% | 18/41 | 43.9% | 58/81 | 71.6% | 58/59 | 98.3% |
| **Synovial Cyst** | 0/0 | - | 0/0 | - | 100/100 | 100% | 100/100 | 100% |
| **None** | 30/50 | 60% | 30/34 | 88.2% | 46/50 | 92% | 46/66 | 69.6% |

## REFERENCES

1. Patel ND, Broderick DF, Burns J, et al. ACR Appropriateness Criteria Low Back Pain. *J Am Coll Radiol*. 2016;13(9):1069-78. DOI: 10.1016/j.jacr.2016.06.008.
2. Sartoretti E, Sartoretti T, Binkert C, et al. Reduction of procedure times in routine clinical practice with Compressed SENSE magnetic resonance imaging technique. *PLoS One*. 2019;12;14(4):e0214887. DOI: 10.1371/journal.pone.0214887. PMID: 30978232; PMCID: PMC6461228.
3. Munn Z, Pearson A, Jordan Z, et al. Patient anxiety and satisfaction in a magnetic resonance imaging department: Initial results from an action research study. *Journal of Medical Imaging and Radiation Science*s. 2015;46(1):23–29. DOI: 10.1016/j.jmir.2014.07.006.
4. Andre JB, Bresnahan BW, Mossa-Basha M, et al. Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical mr examinations. *J Am Coll Radiol*. 2015;12(7):689–695. DOI: 10.1016/j.jacr.2015.03.007.
5. Ali H, Biswas R, Mohsen F, et al. The role of generative adversarial networks in brain mri: a scoping review. *Insights into Imaging*, 2022;13(98). DOI: 10.1186/s13244-022-01237-0.
6. Fard A, Reutens D, and Vegh V. From cnns to gans for cross-modality medical image estimation. *Comput Biol Med*. 2022;146(10555). DOI: 10.1016/j.compbiomed.2022.105556.
7. Haubold J, Demircioglu A, Theysohn JM, et al. Generating virtual short tau inversion recovery (stir) images from t1- and t2-weighted images using a conditional generative adversarial network in spine imaging. *Diagnostic*s. 2021;11(9):1542. DOI: 10.3390/diagnostics11091542.
8. Tanenbaum LM, Bash SC, Zaharchuk G, et al. Deep learning–generated synthetic mr imaging stir spine images are superior in image quality and diagnostically equivalent to conventional stir: A multicenter, multireader trial. *American Journal of Neuroradiology*. 2023;44(8):987–993. DOI: 10.3174/ajnr.A7920.
9. Pix2Pix TensorFlow. Last Accessed: 2024-01-17. URL: https://www.tensorflow.org/tutorials/generative/pix2pix.
10. Sobel operator: https://en.wikipedia.org/wiki/Sobel_operator
11. Cohen JP, Luck M, and Honari S. Distribution matching losses can hallucinate features in medical image translation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018. DOI: 10.48550/arXiv.1805.08841.
12. Raudner M, Toth DF, Schreiner MM, et al. Synthetic T2-weighted images of the lumbar spine derived from an accelerated T2 mapping sequence: Comparison to conventional T2 w turbo spin echo. *Magnetic Resonance Imaging*. 2021;84:92-100.
13. Hong KT, Cho Y, Kang CH, et al. Lumbar spine computed tomography to magnetic resonance imaging synthesis using generative adversarial network: visual Turing test. *Diagnostics*. 2022;12(2):530. DOI: 10.3390/diagnostics12020530
14. Yurt M, Uh Dar S, Erdem A, et al. Mustgan: multi-stream generative adversarial networks for mr image synthesis. *Med Image Anal*. 2021;70(101944). DOI:10.1016/j.media.2020.101944.
15. Whiting P, Sutjes AWS, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Annals of Internal Medicine*. 2004;140(3). DOI: 10.7326/0003-4819-140-3-200402030-00010

## SUPPLEMENTAL FILES

### Materials & Methods

### Preprocessing:

The intensity values of the image volumes were normalized between -1 and 1 and then resampled to 512x512 (up-sampled or down sampled as needed depending on their original size). All T1 and T2 image pairs were of the same depth (number of slices). Each slice pair was stacked into a new, 2-channel image used as the input to the model. Therefore, the “T1/T2 slices” row seen in Table 1 refers to the number of stacked image inputs (T1-T2 slice pairs) used in each of, a variety of loss functions these sets.

### Model Optimization Metric Definitions:

*Structural Similarity Index (SSIM)* is a measure of perceived quality and similarity between two images where 1 represents a perfect match between images.

*Mean Absolute Error (MAE)* is a pixel-wise measure of error between pair images where 0 represents no differences between the two images.

*Peak Signal To Noise Ratio (PSNR)* is the ratio between the maximum power of a signal (original image) and the corrupting noise that affects its representations (synthetized image). Higher values are better, with no upper bound on the metric.

### Loss Functions:

To optimize model performance, a variety of loss functions were evaluated (binary cross entropy loss, BCE-loss; structural-similarly-index-loss, SSIM-loss; mean absolute error loss, MAE-loss; and Sobel-loss).

*Binary Cross Entropy Loss (BCE-loss):* In GANs, this loss is the probability assigned to the generated image by the discriminator. Given a “perfect” generator which always has a realistic output, the discriminator loss assigned should always be close to 1.

*Structural Similarly Index Loss (SSIM-loss)*: Loss reflecting the difference in quality between the generated and target image. Since the optimal value for SSIM is 1, the loss is defined as 1-the calculated SSIM metric of the observation.

*Mean Absolute Error Loss (MAE-loss):* This loss passes the true MAE metric value to the model as a loss. The MAE metric is reflective of the pixel-wise intensity error between the two images (target and predicted).

*Sobel Filter Loss (Sobel-loss):* We applied the Sobel filter to both the predicted and target image and calculate the mean squared error (MSE) between the two. This MSE value was the loss value passed back to the model.

## Supplemental Tables

| Loss function | SSIM | MAE | PSNR |
|---|---|---|---|
| | *Structural similarity index measure* | *Mean absolute error* | *Peak signal to noise ratio* |
| BCE + MAE | 0.766 | 0.034 | 24.816 |
| SSIM + MAE | 0.840 | 0.029 | **26.367** |
| SSIM + Sobel | 0.840 | **0.028** | 26.360 |
| SSIM + Sobel + MAE | 0.839 | 0.029 | 26.193 |
| SSIM + Sobel + Foreground MAE | **0.842** | 0.032 | 25.969 |
| **Test set:**<br>SSIM + Sobel + Foreground MAE | 0.829 | 0.033 | 25.667 |

**Table S1. Performance Metrics with Various Loss Functions**

Performance metrics (SSIM, structural similarity index; MAE, mean absolute error; PSNR, peak signal to noise ratio) on the validation set (n=84) for models trained with different loss functions (first column). Performance metrics on the final test set (n=100) shown on the last row with our preferred loss function. BCE: Binary Cross Entropy.

| | Quality | Motion Artifact | Acquired vs Synthetic | Fat Suppression |
|---|---|---|---|---|
| | *Synthetic STIR is as good or better quality than the acquired STIR.* | *Synthetic STIR has a much or less motion artifact than the acquired STIR.* | *Patients where reviewer though acquired STIR was more likely to be synthetic than the synthetic STIR* | *Synthetic STIR has the same fat suppression as the acquired.* |
| **Reviewer 1** | 74/100 | 77/100 | 2/100 | 98/100 |
| **Reviewer 2** | 59/100 | 53/100 | 3/100 | 78/100 |
| **Reviewer 3** | 72/100 | 74/100 | 20/100 | 96/100 |
| **2+ reviewers agree** | 77/100 | 78/100 | 3/100 | 96/100 |

**Table S2. Radiologist-Specific Assessments on the Test Set**

Percentage of patients in the n=100 test set for whom each reviewer agreed with the relevant column comparing their responses between acquired versus synthetic STIR volumes. The final row shows the percentage of patients for whom the relevant column was agreed on for 2 or 3 of the reviewers.

**Table S3. Majority-Rule Pathology Predictive Performance *with Confidence intervals***

Majority-rule predictive performance statistics of pathologies from synthetic STIR volumes versus acquired STIR volumes in the test set of 100 patients. Three radiologists each reviewed in a blinded fashion 1 synthetic STIR volume and 1 acquired STIR volume from 100 patients and evaluated each volume's pathologies. The rows detail pathologies available to 3 radiologist reviewers (including none). Columns detail performance metrics for the 100-patient test set for each pathology (numerator, denominator, and percentage), assuming a pathology was present in the acquired STIR volume if 2 or 3 of 3 total reviewers identified it as present, and similarly for the synthetic volume. PPV = positive predictive value, NPV = negative predictive value. 95% confidence intervals were derived via bootstrapping with 1000 iterations. * When the point estimate is 100% or 0%, the confidence intervals via bootstrapping always collapse to [100%, 100%] or [0%, 0%], respectively, and are hence meaningless and not reported. ** For PPV and facet arthritis with synovitis, n=33/1000 bootstrapped iterations did not contain any relevant observations, hence the confidence interval was formed with 977 iterations.

| | PPV | | Sensitivity | | NPV | | Specificity | |
|---|---|---|---|---|---|---|---|---|
| **Diagnoses** | *Ratio* | *%*<br>[95% CI] | *Ratio* | *%*<br>[95% CI] | *Ratio* | *%*<br>[95% CI] | *Ratio* | *%*<br>[95% CI] |
| **Bone Marrow Lesion** | 2/2 | 100%<br>* | 2/3 | 66.6%<br>[0%,100%] | 97/98 | 98.9%<br>[96%,100%] | 97/97 | 100%<br>* |
| **Spinal Cord Signal Change** | 0/0 | - | 0/0 | - | 100/100 | 100%<br>* | 100/100 | 100%<br>* |
| **Paraspinous Muscle Edema** | 9/9 | 100%<br>* | 9/24 | 37.5%<br>[19%,56%] | 76/91 | 83.5%<br>[76%,91%] | 76/76 | 100%<br>* |
| **Facet Arthritis with Synovitis** | 3/4 | 75%**<br>[66%,100%] | 3/13 | 23%<br>[0%,50%] | 86/96 | 89.5%<br>[83%,96%] | 86/87 | 98.8%<br>[96%,100%] |
| **Fracture with Edema** | 0/0 | - | 0/2 | 0%<br>* | 98/100 | 98%<br>[94%.99%] | 98/98 | 100%<br>* |
| **Modic Type 1 Change** | 18/19 | 94.7%<br>[81.2%,100%] | 18/41 | 43.9%<br>[29%,59%] | 58/81 | 71.6%<br>[62%,81%] | 58/59 | 98.3%<br>[95%,100%] |
| **Synovial Cyst** | 0/0 | - | 0/0 | - | 100/100 | 100%<br>* | 100/100 | 100%<br>* |
| **None** | 30/50 | 60%<br>[46%,73%] | 30/34 | 88.2%<br>[76%,97%] | 46/50 | 92%<br>[83%,98%] | 46/66 | 69.6%<br>[57%,81%] |