OREGON HEALTH & SCIENCE UNIVERSITY
SCHOOL OF MEDICINE – GRADUATE STUDIES

Radiomic Analysis and Comparison of Metal Artifact Reduction Algorithms in Computed Tomography

By

Alex Lindgren-Ruby

A Thesis

Presented to the Department of Medical Physics
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Masters of Science

June 2024

Table of Contents

Abbreviations

CT	Computed Tomography
FOV	Field of View
FWER	Family-Wise Error Rate
MAR	Metal Artifact Reduction
ROI	Region of Interest
RT	Radiation Therapy

Acknowledgements

I would like to thank the technologists of the OHSU radiation medicine department for their kind help in acquiring the images which formed the basis of this analysis.

Dr. Lindsay DeWeese provided helpful guidance towards the direction of this project. I would like to thank my mentor, Dr. Joseph Foy, and well as the members of my advisory committee, Dr. Anna Mench and Dr. Thomas Griglock, without whom this work would not have been possible.

I'm indebted to all the staff of the OHSU medical physics program, the faculty of the math and physics departments of Seattle University, my classmates, and all my friends and loved ones.

Abstract

Purpose: Computed Tomography (CT) images have various applications in radiation oncology, from diagnosis to staging, simulation, and treatment planning. However, the utility of CT images can be compromised by the presence of image artifacts. This work investigates two metal artifact reduction (MAR) algorithms by measuring quantitative features of phantom images. Radiomics-based evaluation of MAR methods have been performed in other works but have limited their analysis to first-order features measuring fidelity of MAR scans to ground truth. Recent advances in the application of texture features to machine learning (ML) models for classification and prognostic evaluation of malignancies have highlighted the relevance of higher-order features to oncological imaging. Therefore, this work presents a texture-based evaluation of two manufacturer MAR algorithms currently in clinical use.

Methods: Images were obtained using a Solid Water™ phantom and six inserts made of Solid Water™, aluminum, titanium, brass, copper, and steel. Each phantom configuration was imaged with and without MAR on two CT scanners used clinically for radiation therapy simulation: the Siemens SOMATOM® X.ceed and General Electric® (GE) Discovery MI, referred to as the "Siemens" and "GE" scanners, respectively. Two regions of interest (ROIs) were manually outlined for each scan—one inner ROI representing the insert and another outer ROI over the nearby Solid Water™ material. 92 2D radiomic features were extracted from each of the 41 image slices per scan in both ROIs. Specific features of interest were analyzed to assess preservation of texture in Solid Water™ with and without MAR enabled in the same phantom configuration, and between Solid Water™ texture surrounding metal inserts with MAR enabled and the ground truth texture of Solid Water™. Other features were analyzed to assess preservation of voxel intensity within the insert ROI with and without MAR enabled. All comparisons were made between scans obtained on the same machine and using a two-sided Kolmogorov-Smirnov (KS) test with a Bonferroni-corrected significance threshold of $p=0.0054$.

Results: Between scans reconstructed with and without MAR, significant differences were observed between outer ROI texture distributions for the titanium, brass, copper, and steel inserts on both machines, with no significant difference for Solid Water™ and aluminum.

Comparing outer ROI texture features of metal insert scans with those of Solid Water™, the GE scanner failed to find significant differences for all insert materials except titanium, indicating that artifacts from titanium inserts could not be entirely removed using the GE MAR algorithm. The Siemens unit found significant differences for all inserts. Lastly, significant differences in the distribution of voxel intensity of the inner ROI with and without MAR were found only in the titanium and steel scans on the GE unit and titanium on the Siemens unit.

Conclusions: This investigation found meaningful differences in the output of the GE and Siemens MAR algorithms. This outcome reinforces the need for standardized choices of reconstruction parameters, even those that one might assume are inconsequential, and highlights the potential of future investigation into the application of radiomic models to images using artifact reduction methods where certain feature groups are well-preserved.

1 Introduction

Computed Tomography (CT) images are used for a variety of diagnostic and screening procedures, making them relevant to multiple disciplines in medicine[1–3]. This especially is true in radiation oncology, where CT images form the basis for clinical evaluation, staging, and radiation therapy (RT) treatment planning for a wide range of pathologies. However, the utility of CT exams for these purposes is limited by the fidelity of voxel intensity, also referred to as Hounsfield Unit (HU), values to the true density or x-ray attenuation of the imaged material[4,5].

For this reason, the application of CT images in radiation oncology may be challenged by the presence of image artifacts which cause large deviations in the reconstructed HU values. Among these are metal artifacts, characterized by high frequency intensity bands or streaks which obscure the internal anatomy, often organized radially around highly attenuating objects. These artifacts are frequently encountered when imaging metal, such as orthopedic implants, and their intensity is affected by the atomic number, density, and size/shape of objects within the image field of view (FOV). Objects with high atomic number and density present physical challenges for CT image reconstruction due to an increase in beam hardening, scatter, noise, partial volume effects, and a lack of suitable projection data from which to reconstruct the CT image[6,7].

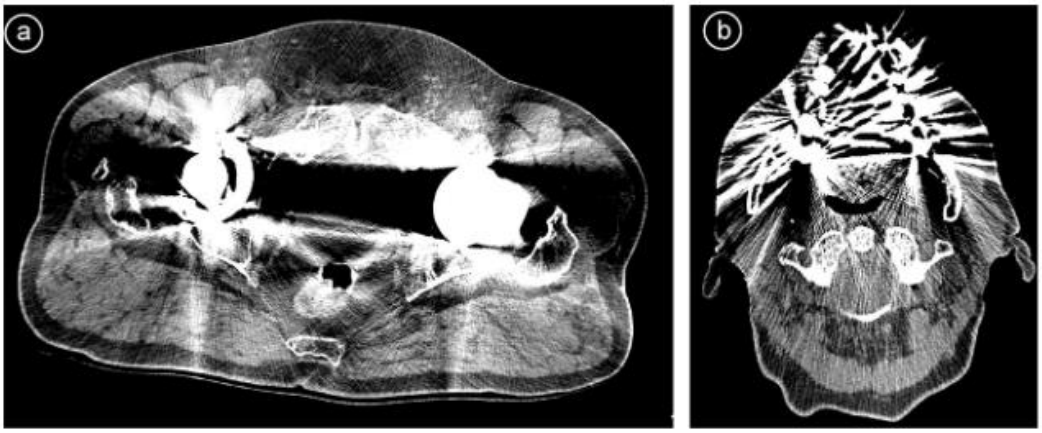

Figure 1: Examples of CT metal artifacts caused by hip prostheses (a) and dental implants (b). Figure from Wllenberg et al., reproduced with permission in this work.

As metal artifacts are the result of physical processes which occur during image acquisition, their intensity and effect on the final image depends greatly on several CT acquisition parameters and the amount of metal in the scan FOV. Altering the FOV and these parameters or using alternative imaging methods, such as dual-energy and photon counting CT techniques, can help to limit this effect[8,9].

Also among the options available to physicists and CT technologists looking to minimize the effect of metal artifacts are metal artifact reduction (MAR) algorithms, a family of post-processing algorithms which can improve qualitative image features while also attempting to reduce the error in HU values[4]. Individual vendors offer proprietary MAR algorithms for use in reconstructing CT images acquired on their equipment. Many of these algorithms use variations of sinogram inpainting in which missing or corrupted sinogram data is recovered via interpolation of nearby valid projections[10]. Other, more novel techniques include iterative reconstruction, where the error between an acquired scan and a reference image is minimized[11,12].

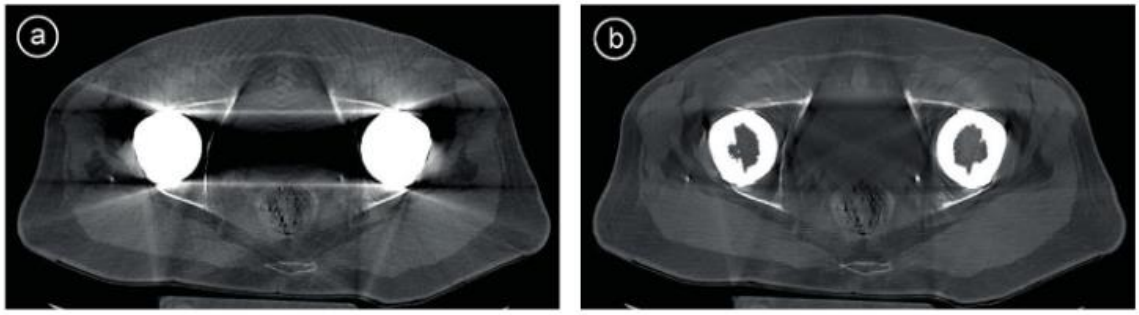

Figure 2: Example of CT metal artifacts from hip prostheses (a) and a corresponding image with metal artifact reduction algorithm applied. Figure from Wllenberg et al., reproduced with permission in this work.

Clinically, metal artifacts are most frequently encountered when imaging dental implants and fillings in the head and neck, hip prosthesis in the pelvis, or spinal fusions[13–15]. In addition to limiting the diagnostic utility of CT images, the presence of implants and associated image artifacts nearby treatment sites poses a significant challenge for treatment planning systems; achieving successful RT outcomes often depends on delivering planned does within 3-4%[16], and local dose calculation errors due to metal artifacts can greatly exceed this margin. Studying effects of metal artifacts on phantom dose calculation, Huang et al. found systemic errors higher than 30% downstream of Cerrobend inserts[17], an alloy

commonly used in radiation therapy applications. Because TPS calculate relative electron densities using HU values, artifacts affecting those values result in inaccurate TPS output. Traditionally, this is overcome by manually overriding regions affected by metal artifacts with suitable densities. However, this technique requires time and labor in the form of manual contouring. Further, the accuracy of the correction is itself limited as it applies a homogenous density to the affected area[18].

Even where density overrides are employed, MAR algorithms present an opportunity to further improve dose calculation and RT treatment delivery. Multiple articles have demonstrated the ability of MAR algorithms to preserve the accuracy of TPS outputs and their potential to improve clinical outcomes. In their literature review, Puvanasunthararajah et al. describe several contexts in which MAR algorithms provide significant benefit for reducing discrepancies in dose planning and preserving ideal dose distributions to targets and healthy tissues. Of note, they highlight the potential for deep-learning-based MAR algorithms used with clinical head and neck CT scans containing dental implants to create IMRT plans with smaller dose errors than those produced with simple density correction using water density (1.0 g cm^{-3})[19].

The relevance of MAR methods to RT treatment outcomes incentivizes physicists to evaluate their relative performance. Differences between MAR algorithms can lead to variations in image quality and the extent to which metal artifacts are reduced with a specific MAR algorithm. Vaishnav et al. describe multiple methods which have been proposed to evaluate and compare the effectiveness of different MAR algorithms, with the majority involving qualitative analysis of image features based on a standardized system of ranking "diagnostic utility"[8]. Other publications have used quantitative and statistical information extracted using radiomic analysis to assess the extent of the metal artifacts with and without the use of MAR algorithms[20,21].

Radiomics is an area of translational research in imaging physics, in which quantitative features are derived or "extracted" from the analysis of pixel or voxel intensity values. These features include descriptors of the shape and geometry of regions of interest (ROIs), and first-order statistical features such as the mean or standard deviation pixel value within an ROI. Additional higher-order features can be measured as well, which consider the

spatial organization of voxel intensities or “texture” of the image. Whereas traditional evaluation of diagnostic images relies on expert identification of qualitative features, radiomic analysis yields image information which would be otherwise inaccessible to the human observer.

Radiomic analysis of oncologic CT images is frequently used in combination with machine learning (ML) algorithms, which automate the selection of stable image features from which a model can be trained and validated (see Figure 3). For example, Garau et al. employed two ML models in identification of lung nodule malignancies and assessed their performance using receiver operating characteristic (ROC) analysis. They found the area under the ROC curve to be between 0.82 – 0.86 during validation[2]. An examination of review literature quickly demonstrates the flexibility of radiomic analysis, with models being developed for multiple sites and validated for use in a variety of clinical tasks from diagnosis and staging to the prediction of therapeutic outcomes [22–24].

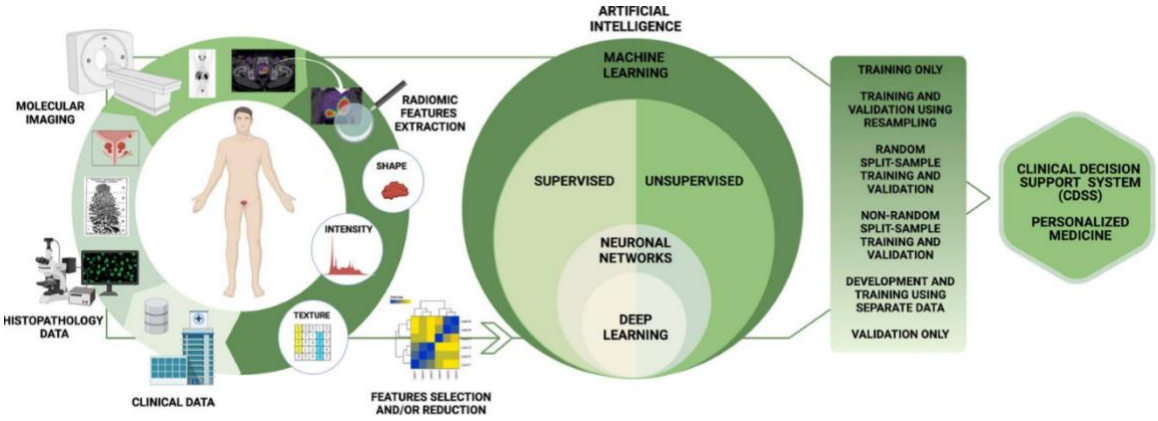

Figure 3: Illustration demonstrating the integration of radiomic analysis into machine learning models. Figure from Liberini et al., reproduced under the creative commons attribution 4.0 international license.

Critically for the application of radiomics to CT images, the reproducibility or stability of radiomic features depends greatly on the acquisition parameters and reconstruction techniques used[25,26]. This dependence makes these features well-suited for the characterization of MAR algorithms and the evaluation of their performance. While radiomics has been previously employed to describe and compare the behavior of different MAR methods, such as in the work of Huang et al.[20] and Lemmens et al.[21] previous analysis has been limited to the use of shape and first-order features such as mean HU value or similar “fidelity metrics”.

Conversely, we use radiomics analysis to assess the changes in images when MAR techniques are and are not applied. In this work, we characterize how the use of two commercially available MAR algorithms affect 2nd order radiomic features in CT scans of a customized image phantom containing a range of high-density materials. By comparing textural features inside and outside metal inserts of a varying atomic number, we assess the ability of each algorithm to restore quantitative image features affected by metal artifacts.

2 Materials and Methods

The CT images used in this analysis were taken using the Siemens SOMATOM® X.ceed™ and General Electric® (GE) Discovery MI™, referred to as "Siemens" and "GE" scanners, respectively . To obtain these images, a roughly 10 cm x 10 cm x 5 cm rectangular RMI® Solid Water™ phantom was scanned on both machines (see Figure 4). This phantom material was chosen for its optical similarity to soft tissue in the keV photon range; soft tissue ($C_5H_{40}O_{18}N$) has an effective atomic (Z) number of 7.22[27] whereas that of the RMI® phantom is 5.96[28]. A 1.27 cm (0.5") diameter cylindrical section of the phantom was removed from the center of the phantom such that the axis of the cylinder was parallel with the scan direction, allowing the phantom to be scanned with Solid Water™, aluminum, titanium, stainless-steel, brass, and copper inserts (see Table 1).

Insert	Material	Z	Density (g mL^{-1})	Implant Application
1	Solid Water™	5.96*	1.03	Soft-tissue analogue
2	Aluminum	13	2.70	Bioactive ceramic composite implants
3	Titanium	22	4.54	Bioinert Ti and Ti-based alloy implants
4	Steel	25.2*	8.00	Bio-tolerant SS implants
5	Brass	29-30*	8.50	NA
6	Copper	29	8.96	NA

Table 1: List of insert materials and their associated atomic (Z) numbers, densities (g mL^{-1}), and applications for use in medical implants[29]. Metal and metal alloy data from NIST[30,31], Solid Water™ data from Hill et al.[28] The Z numbers of Solid Water™, Steel, and Brass are marked with an asterisk (*) to indicate effective Z number, as composite materials have no single Z number.

Scans were taken using the adult head protocol used clinically for stereotactic radiosurgery simulation on both scanners, which use acquisition techniques of 120 kV and 190 mA on the GE scanner and 120 kV and 320 mA on the Siemens. Additional protocol details are provided in Table 2. Although the differences between these techniques (particularly mAs) can be expected to influence radiomic features, comparisons were made between protocols as optimized for clinical use on both scanners in order to better understand the performance of MAR algorithms under typical operating conditions.

The image reconstruction process was performed with and without Siemens Iterative Metal Artifact Reduction (iMAR)[32] and GE Smart Metal Artifact Reduction (MAR) algorithms[33], producing a total of 6 phantom configurations for each insert (12 image sets

in total). Each image set yielded 41 axial CT image slices from which features could be extracted.

Scanner	kVp	mA	mAs	Exposure	Slice Thickness	Protocol	Kernel	Filter
GE	120	190	152	800	1	1.5 CT STEREOTACTIC HEAD	STANDARD	BODY FILTER
Siemens	120	320	376	1176	1	CTSIM SRS BRAIN(Adult)	Hr40u	W1

Table 2: Acquisition and reconstruction parameters for the CT images taken on the GE and Siemens scanners. Exposure is measured in ms, and Slice Thickness is measured in mm.

Figure 4: (left) Solid Water™ image phantom (right) Solid Water™ image phantom and implants made of (left to right) Solid Water™, aluminum, titanium, brass, copper, and stainless steel.

ImageJ v1.51i[34] was used to manually segment regions of interest (ROIs) inside and outside each insert in these images, and the resulting ROIs were converted to binary masks using MATLAB R2023a[35]. The mean ROI size was 35.51 ± 1.84 mm^2 and 8131.39 ± 292.27 mm^2 for the inner and outer ROIs, respectively.

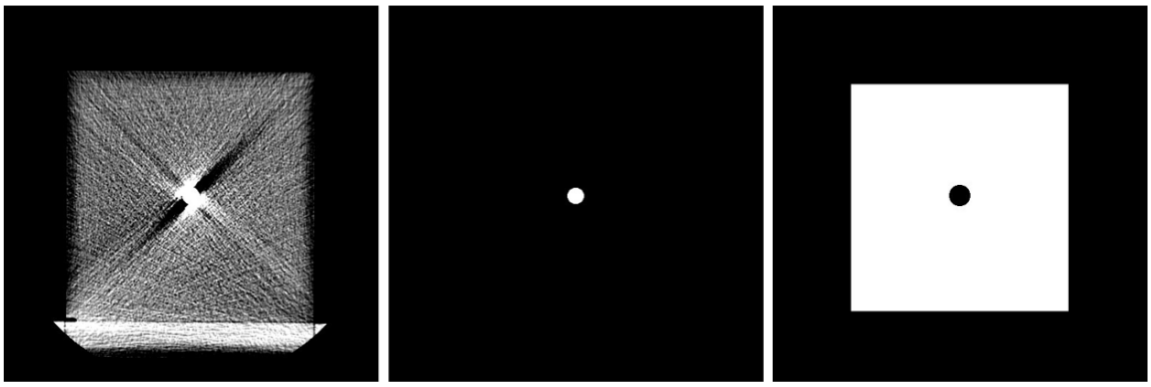

Figure 5: (left) CT image of Solid Water™ image phantom and aluminum insert (middle) binary mask of the inner ROI (right) binary mask of the outer ROI.

The image segmentation described above, visible in Figure 5, allowed us to analyze the performance of the MAR algorithms through comparison of radiomic feature values obtained from scans of different phantom configurations and reconstructions. For each insert material, we performed three comparisons: (1) between features of the outer ROI of each phantom configuration with and without MAR applied, (2) between features of the outer ROI of each metal insert configuration with MAR applied and a reference scan of the Solid Water™ configuration obtained without MAR, and (3) between features of the inner ROI of each phantom configuration with and without MAR applied. Through these feature comparisons we aimed to investigate how the GE MAR and Siemens iMAR algorithms affect features with and without nearby high-attenuation material (1), how well these algorithms replicate the "ground truth" features of Solid Water™ when nearby high-attenuation material is introduced (2), and lastly how MAR algorithms affect the features of highly attenuating objects themselves (3). These comparisons are illustrated in **Error! Reference source not found.**.

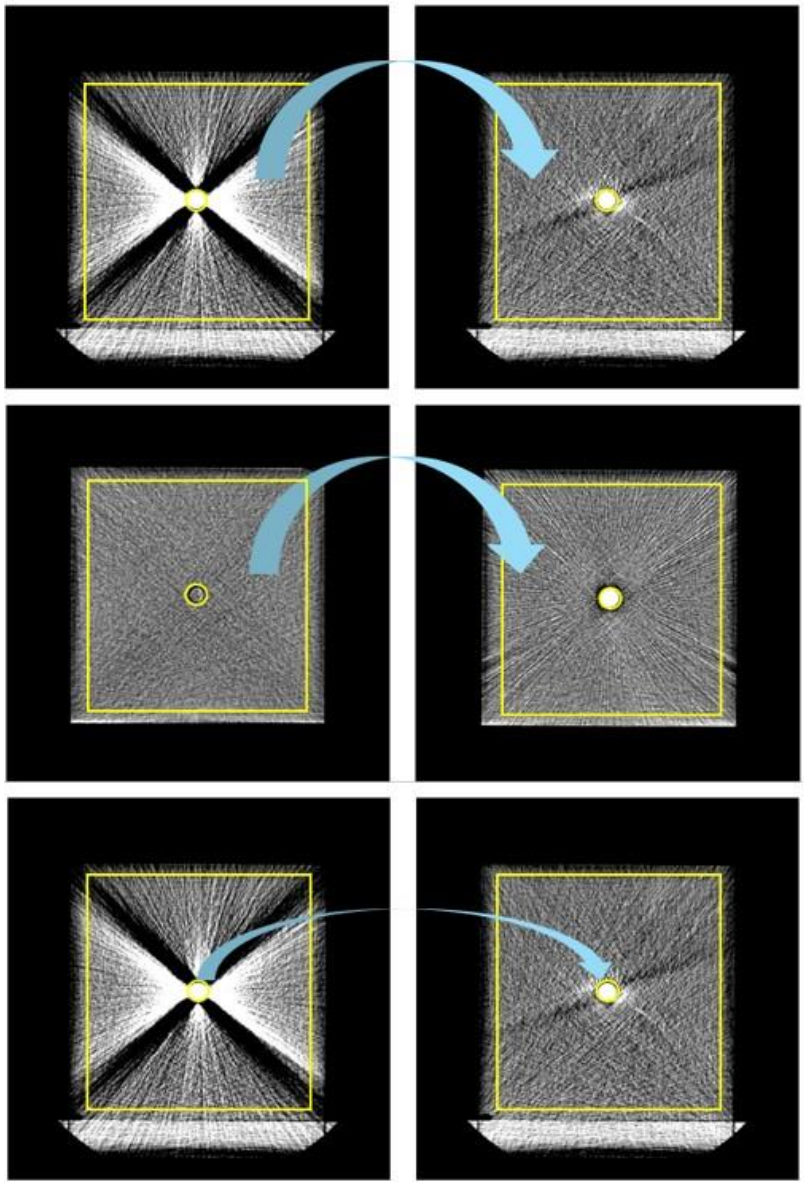

Figure 6: Illustration of the three types of comparisons described in this section (top) comparison 1, of features of the outer ROI between inserts of the same material with and without MAR (middle) comparison 2, of features of the outer ROI between a Solid Water™ without MAR and a metal insert with MAR (bottom) comparison 3 of inner ROI between inserts of the same material with and without MAR.

A total of 92 2D radiomics features were extracted from each image in Python v3.7.12[36], using image analysis library SimpleITK v2.2.124–26[37] and radiomics feature extraction package Pyradiomics v3.0.127[38]. When testing the difference between radiomic features of two image sets we employed a two-sample Kolmogorov-Smirnov (KS) test of the relevant feature sets using SciPy v1.7.3 statistics module[39,40]. The two-sample KS test returns a p-value and test statistic measuring the likelihood that two empirical cumulative

distribution functions calculated from sample data are drawn from the same parent distribution. For each feature comparison a null hypothesis of equal feature value distribution was assumed, and statistical power was adjusted using Bonferroni multiple comparison correction methods[41]. Using an uncorrected threshold of $p = 0.05$ and with 92 relevant features and associated hypothesis tests performed, excluding shape features and features related to global image attributes outside of selected ROIs, this correction resulted in a significance threshold of .05/92 or approximately 0.0054. This correction was made to minimize the family-wise error rate (FWER), or the inflated probability of type 1 error when performing multiple hypotheses tests simultaneously.

For the first and second comparisons the glcm_JointEntropy feature was used to assess the ability of MAR scans to remove metal artifacts affecting textural features and maintain textural similarity to that of Solid Water™ in the outer ROI. While several features could be selected for this comparison, we chose glcm_JointEntropy for its pronounced role in the literature; one review article found 14 pulmonary nodule classification and 7 pulmonary tumor prognostic models which identified glcm_JointEntropy as a stable feature during their feature selection process[42]. In addition, Khurshid et al. have demonstrated its reliability as a prognostic feature in theranostics involving prostate cancer[43]. Given the proximity of each tumor cite to common orthopedic implant locations (spinal fusions and partial/full hip replacements), this feature selection was deemed most likely to yield clinically informative results.

In the third comparison, where texture within the inserts with and without MAR applied were compared, the firstorder_Mean feature was selected as a measure of the effect of MAR algorithms on the variability of HU values within high-density material. First-order features were preferred in this comparison to best assess the impact on MAR algorithms on treatment planning systems, in which the fidelity of high-density HU values to real-world attenuation properties will affect the accuracy of the system's output. Texture features such as those derived from the GLCM, while useful for the characterization of soft-tissue, are not relevant to the appearance of metal implant materials themselves in CT and were not highlighted in this work.

3 Results

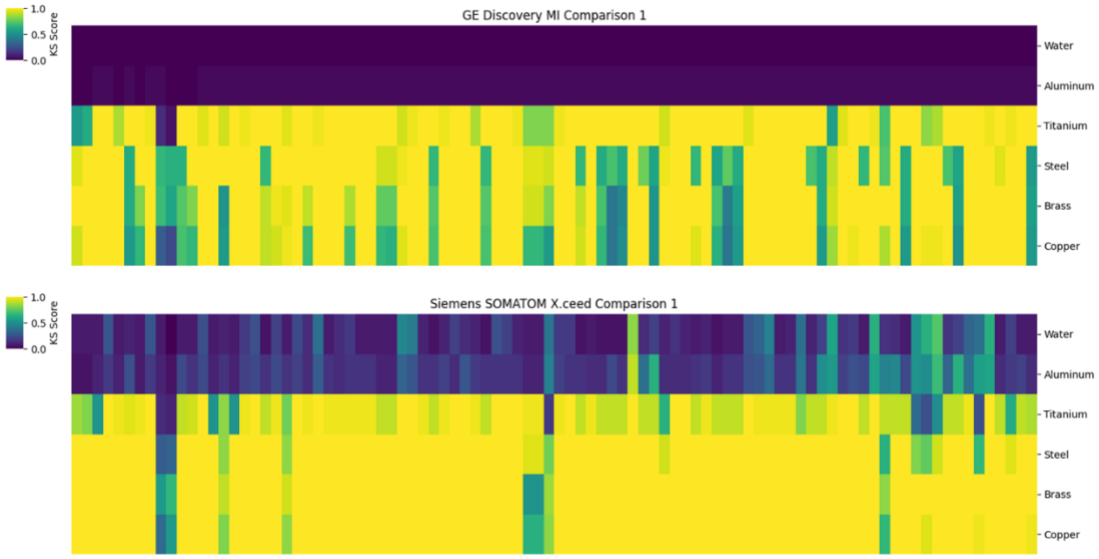

Figure 7: Heatmaps showing Kolmogorov-Smirnov (KS) statistics for each of the 92 radiomics features extracted in the first comparison on the GE (top) and Siemens (bottom) scanners. In comparison 1, CT scans of each insert material were compared with and without MAR applied with features extracted from the outer ROI. Rows represent different metal insert materials in ascending order of density, and columns individual radiomic features.

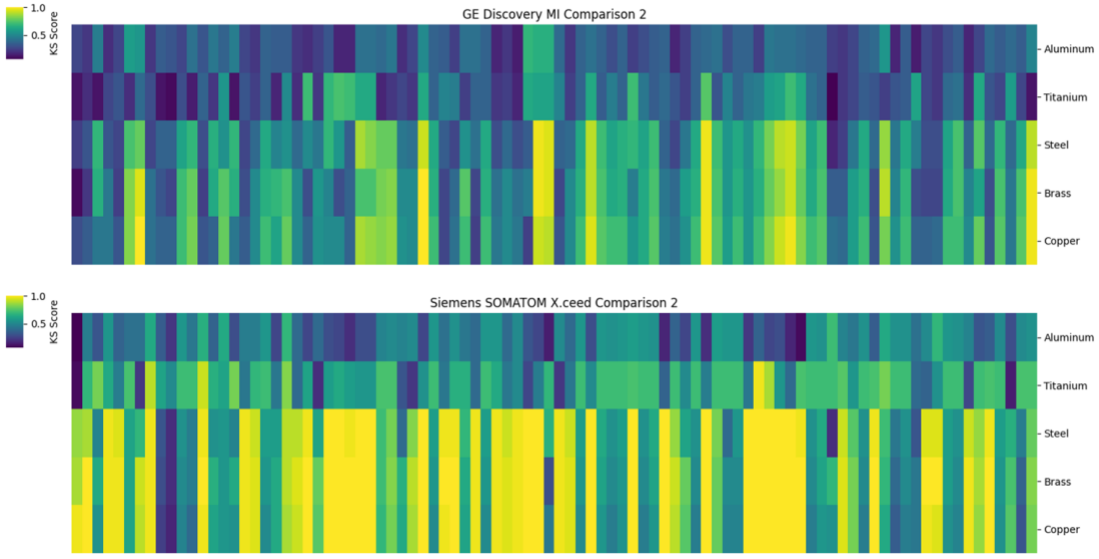

Figure 8: Heatmaps showing Kolmogorov-Smirnov (KS) statistics for each of the 92 radiomics features extracted in the second comparison on the GE (top) and Siemens (bottom) scanners. In comparison 2, CT scans with metal inserts and MAR applied were compared to the reference scan that was void of metal artifacts with features extracted from the outer ROI. Rows represent different metal insert materials in ascending order of density, and columns individual radiomic features.

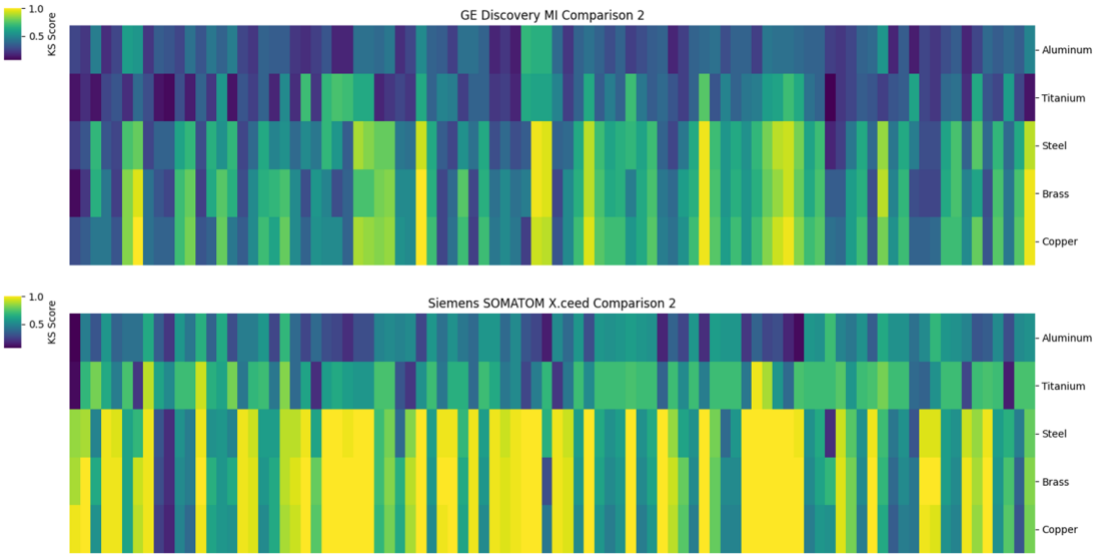

Figure 9: Heatmaps showing Kolmogorov-Smirnov (KS) statistics for each of the 92 radiomics features extracted in the third comparison on the GE (top) and Siemens (bottom) scanners. In comparison 3, CT scans of each insert material were compared with and without MAR applied with features extracted from the inner ROI. Rows represent different metal insert materials in ascending order of density, and columns individual radiomic features.

The KS statistic obtained in each hypothesis test performed in this analysis are plotted in Figures 7-9. These statistics indicate the probability that feature samples were measured from the same theoretical distribution, with a KS-score of 1 indicating high probability that the parent distributions which produced the samples being compared are different and a score of 0 failing to infer the same. A summary of the number of features which saw significant differences for each comparison and insert material are reported in **Error! Reference source not found.**.

GE						
Comparison	Solid Water™	Aluminum	Titanium	Stainless Steel	Brass	Copper
1	0.0%	0.0%	97.8%	100.0%	97.8%	95.7%
2	NA	10.9%	30.4%	59.8%	69.6%	67.4%
3	0.0%	0.0%	87.0%	93.5%	41.3%	69.6%
Siemens						
Comparison	Solid Water™	Aluminum	Titanium	Stainless Steel	Brass	Copper
1	8.7%	13.0%	93.5%	97.8%	100.0%	98.9%
2	NA	51.1%	77.2%	92.4%	93.5%	92.4%
3	0.0%	0.0%	58.7%	5.4%	33.7%	14.1%

Table 3: Number of radiomic features, expressed as a percentage of 92 total features, which showed significant differences between the compared ROIs in comparisons 1-3. This result is not reported for Solid Water™ in comparison 2, due to the nature of this comparison.

In the first comparison, no feature reflected significant differences between MAR and non-MAR scans with Solid Water™ and Aluminum inserts on the GE scanner. This trend can be seen in the top heatmap of Figure 7 by the two first rows of uniformly insignificant KS-scores, followed by rows of variably significant scores for other insert materials on the same scanner. This sudden change can be contrasted with lower heatmap of the same figure, representing the Siemens scanner, which displays a variety of KS-score values in each row.

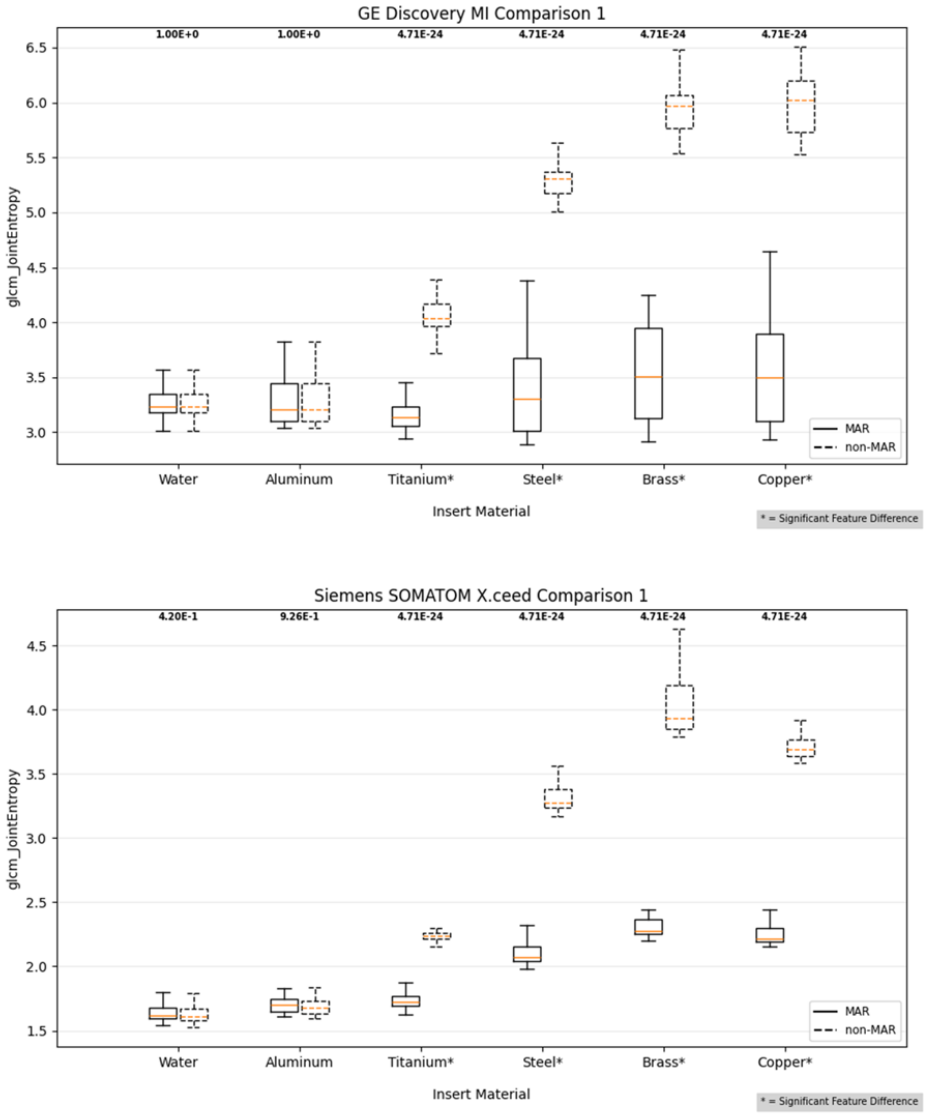

Figure 10: Box plots representing the glcm_JointEntropy feature distribution in the first comparison for scans obtained using the GE (top) and Siemens (bottom) scanners. Insert materials are plotted on the x-axis and feature value is plotted on the y-axis. An asterisk (*) by an insert label is used to indicate significance. P-values of the KS tests performed for each insert are printed above the boxplots which were tested. Scans with MAR algorithms used have box plots printed with solid lines, and those without have dotted lines as illustrated in the plot legend.

Similar behavior can be seen on both scanners in the third comparison, represented numerically in Table 3 and graphically in Figure 9. Here, while low-density inserts resulted in no features reflecting significant differences, inserts made of titanium and denser materials did. In the second comparison, the number of features reflecting significant differences is greater than zero for all insert materials.

In comparison 1, CT scans of each insert material were compared with and without MAR applied. This comparison found significant differences between the glcm_JointEntropy feature samples obtained from the outer ROIs with and without MAR enabled on both the GE and Siemens scanners for the titanium, brass, copper, and steel inserts. The Solid Water™ and aluminum inserts did not yield significant differences in this comparison on either scanner. The observed glcm_JointEntropy samples for each machine and insert combination are represented in Figure 10, along with the associated p-values of the hypothesis tests performed. On both scanners, the titanium, steel, brass, and copper inserts reflected highly significant p-values ($p << 0.0001$), with average differences in glcm_JointEntropy between scans with MAR enabled and those without being -22.06%, -36.05%, -40.64%, and -39.97% respectively on the GE scanner and -22.89%, -36.96%, -42.99%, -40.19% on the Siemens scanner.

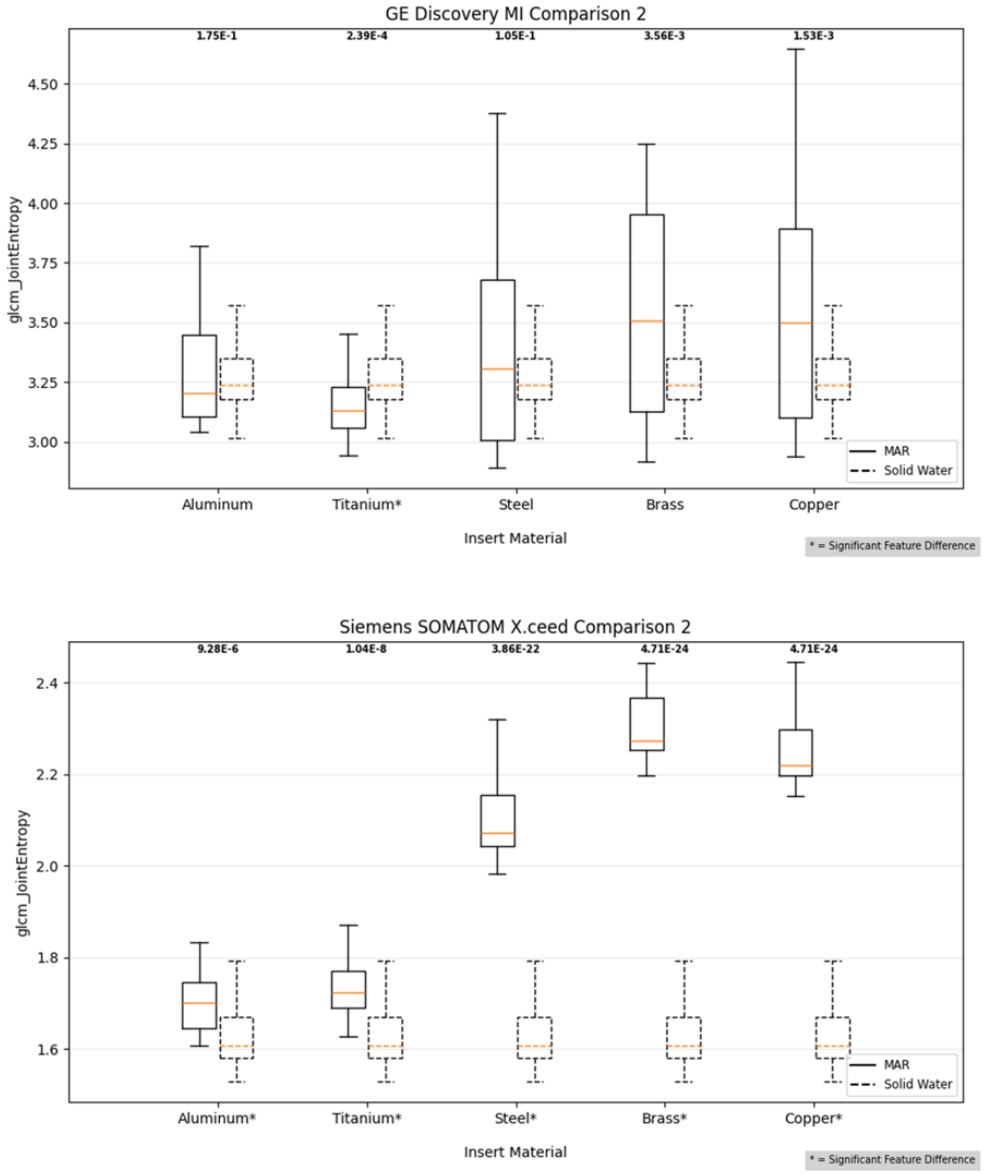

Figure 11: Box plots representing the glcm_JointEntropy feature distribution in the second comparison for scans obtained using the GE (top) and Siemens (bottom) scanners. Insert materials are plotted on the x-axis and feature value is plotted on the y-axis. An asterisk (*) by an insert label is used to indicate significance. P-values of the KS tests performed for each insert are printed above the boxplots which were tested. Scans with MAR algorithms used have box plots printed with solid lines, scans of Solid Water™ have dotted lines as illustrated in the plot legend.

In comparison 2, CT scans with metal inserts and MAR applied were compared to the reference scan that was void of metal artifacts. This comparison found significant differences between the GE unit glcm_JointEntropy feature sample belonging to the outer ROI of the titanium insert and the reference scan. No other insert material on the GE unit

drew a similar inference. While qualitative differences in the mean and interquartile range of this feature's value were observed between metal insert images with MAR enabled and the reference scan and are visible in Figure 11, hypothesis testing did not reveal these differences to be significant. This behavior is contrasted by the Siemens unit, which did find significant differences between the sample glcm_JointEntropy sample distributions for each of the metal inserts in this comparison. The steel, brass, and copper inserts reflected highly significant p-values on this scanner ($p << 0.0001$), with average differences in glcm_JointEntropy between scans with MAR enabled and the reference scan being 28.80%, 40.60%, and 37.92%.

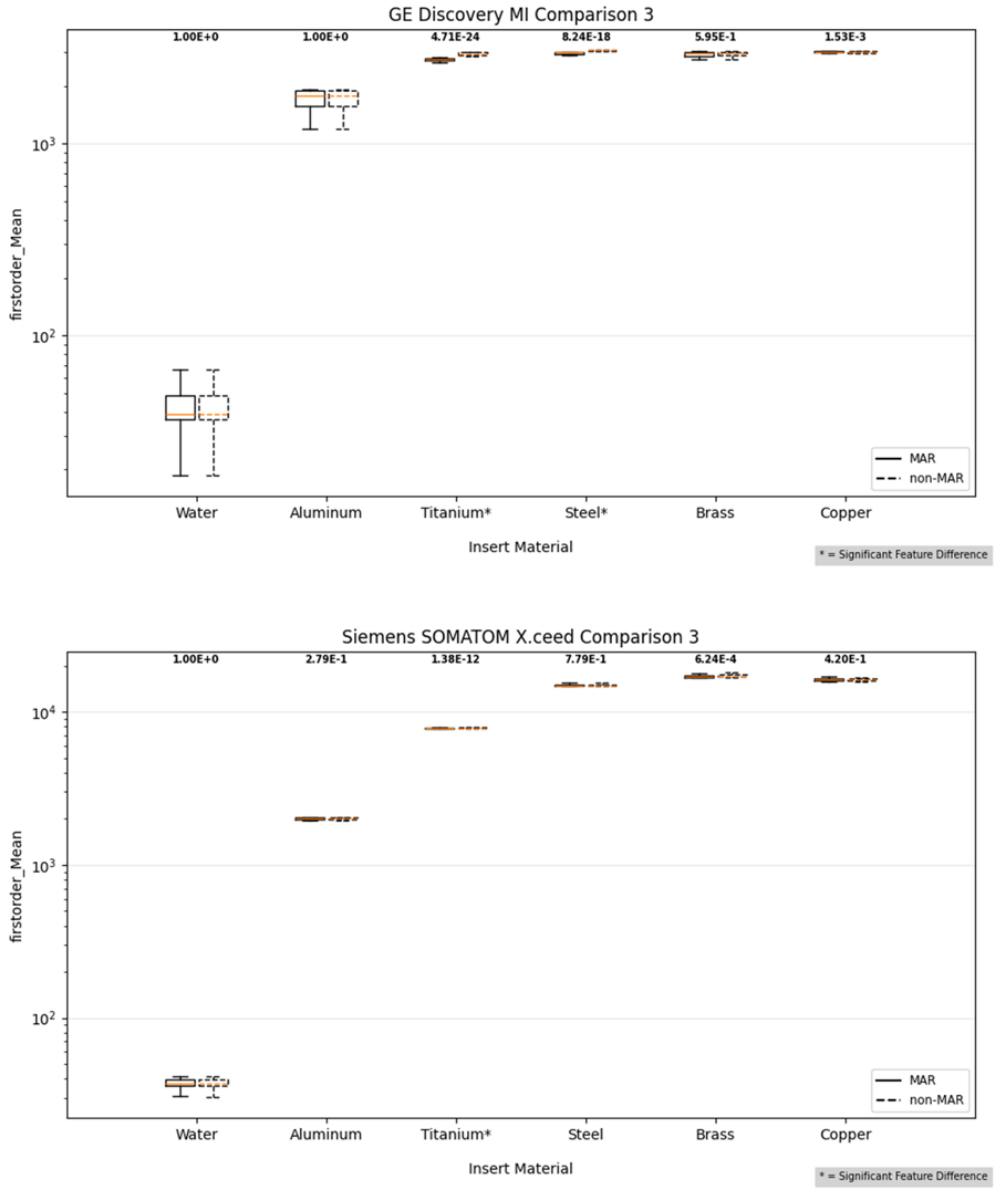

Figure 12: Box plots representing the firstorder_Mean feature distribution in the third comparison for scans obtained using the GE (top) and Siemens (bottom) scanners. Insert materials are plotted on the x-axis and feature value is plotted on the y-axis. An asterisk (*) by an insert label is used to indicate significance. P-values of the KS tests performed for each insert are printed above the boxplots which were tested. Scans with MAR algorithms used have box plots printed with solid lines, and those without have dotted lines as illustrated in the plot legend.

Finally, in comparison 3, CT scans containing all insert materials were compared with and without MAR applied. In this comparison the sample firstorder_Mean distributions were compared between inner ROIs on both units. The GE unit did not find significant differences between the Solid Water™, aluminum, brass, and copper insert samples, while

titanium and steel inserts reflected highly significant p-values ($p << 0.0001$) with associated average differences in firstorder_Mean between scans with MAR enabled and those without of -6.03% and -2.58%, respectively. On the Siemens unit the titanium insert yielded highly significant results ($p << 0.0001$) and an associated average difference in firstorder_Mean value of -0.81%, with all other inserts failing to achieve significance.

4 Discussion

Differences in the behavior of the MAR methods examined in this analysis can be observed beginning in **Error! Reference source not found.**. The first and third comparisons show KS statistics of zero for all 92 feature tests performed with the Solid Water™ and aluminum inserts on the GE unit, while a distribution of non-zero KS statistic values are observed for the same insert materials on the Siemens scanner. This indicates that the GE MAR reconstruction setting does not alter the radiomic features of the image unless material above an attenuation threshold is present in the FOV. The lack of statistical inference drawn in this comparison is supported by the observation that the difference between matching slices taken from the Solid Water™ or aluminum scans with and without MAR was found to be zero. In other words, images taken on the GE scanner with Solid Water or aluminum inserts were identical regardless of the chosen reconstruction setting. In contrast, the Siemens iMAR setting operated independently of the image subject. Changing reconstruction settings on this scanner caused differences in pixel values for each insert material and, correspondingly, significant differences in feature values were observed in this comparison.

In the first comparison both MAR methods found significant test results for the four densest inserts, including titanium, brass, copper, and steel, but not in the case of Solid Water™ or aluminum. This finding is intuitive given the qualitative differences between images with metal artifacts and those reconstructed using MAR algorithms. However, this comparison alone does not evaluate the degree to which those algorithms "restore" quantitative features in the former to match those present in unaffected Solid Water™.

Instead, we look to the results of the second comparison. Here the two machines perform differently from one another; the GE unit being more successful in replicating the glcm_JointEntropy distribution found in the Solid Water™ sample when removing metal artifacts from scans with metal objects inside the FOV than the Siemens unit—failing to find significant differences between distributions in 4 out of 5 inserts where all 5 were significantly different in the latter. More broadly, the GE unit was more successful in replicating the overall radiomic features of Soldi Water™ when removing artifacts, with

fewer features reporting significant differences when compared with the reference scan for each insert material than on the Siemens unit (see Table 3).

The third comparison showed that most tests of the firstorder_Mean feature distribution belonging to inner ROIs with and without MAR enabled did not find significant differences on either scanner. Titanium and steel yielded significant differences on the GE scanner, and only titanium found significant differences on the Siemens scanner. More broadly, the number of features showing significant differences was consistently lower on the Siemens unit than on the GE unit. However, the lack of statistical inference in the case of firstorder_Mean and other features is a poor indicator of the clinical utility of MAR methods when analyzing ROIs representing high density material itself; even when MAR reconstruction is in use, the density of these regions is typically overwritten manually during the dose planning process as discussed previously. While this investigation did not independently analyze the effects of metal artifacts on treatment planning systems, this outcome suggests that the benefits of MAR methods for dose accuracy in treatment planning observed by Huang et al. and others can likely be attributed to the effect which MAR algorithms have in restoring features outside metal objects, rather than within them[17].

These behaviors could be explained by the difference in approach taken by the MAR methods. As suggested by its name, the Siemens iMAR algorithm uses an iterative reconstruction approach[32] whereas the GE MAR algorithm uses an implementation of sinogram inpainting or an interpolation-based approach[33]. These results highlight the impact which a manufacturer's choice of CT image reconstruction can have—not only on the quality of images under normal acquisition conditions but also on the artifact correction options which are available to the operator and the affects they have on image features as a result.

In this case, the observation that images are altered by the iMAR algorithm even in the absence of high attenuation material is relevant to researchers looking to develop radiomic models or clinical tests on the Siemens SOMATOM X.ceed. This finding reinforces the need for standardized and deliberate choices of reconstruction parameters even in the case of those an operator might assume are inconsequential.

The relative success in preservation of textural features by the GE MAR algorithm has less obvious application in research or clinical practice. Searches in the journals Medical Physics and the Journal of Applied Clinical Medical Physics at the time of writing this work did not return any published research which applies or develops radiomic models using CT images with metal implants or prosthetics inside the FOV, either for classification or prognosis. This gap in the literature is easily attributed to the clear detrimental effects of metal artifacts on image quality; any model trained using images with such artifacts is likely to generalize poorly, being adapted to recognize trends caused by photon starvation and beam hardening rather than underlying physiology or pathology. However, our results point to the potential for models trained on non-artifact images to be applied in analysis of images where artifacts have been suitably removed. In cases where specific features or future groups are shown to be well-preserved by a MAR method, we propose that radiomic models which show promising clinical impacts and which use those feature groups could be evaluated for generalization to images which have been reconstructed using that method.

A potential area of future work would include the comparative validation of radiomic ML models on sets of clinical CT images with and without MAR methods applied during reconstruction. Additionally, analysis of radiomic features could be used to compare novel deep learning methods of metal artifact reduction with traditional MAR methods. These avenues would build on the findings presented in this work and further deepen the understanding of CT reconstruction methods as applied to oncologic imaging.

These results, while informative of the differences between the GE MAR and Siemens iMAR algorithms, have limited clinical application because of the poor resemblance of the image phantom to real anatomical structures. While Solid Water™ has attenuation properties which are roughly analogous to soft tissue, the lack of tissue heterogeneity and total absence of denser structures such as bone prevent us from drawing conclusions about the behavior of these MAR methods in vivo. In addition, while the selection of insert materials represented a range of effective densities sufficient to explore the threshold behavior discussed above, it did not represent many implant materials in common use. Some examples of relevant materials which were omitted include metal alloys containing cobalt and magnesium, as well as ceramic, polymer, and composite implants[29]. Finally, this

phantom is small relative to most patient anatomy and, as result, fails to faithfully model the scatter and attenuation one would expect from even a head and neck scan.

While Solid Water™ does not contain the same detail and heterogeneity inherent in human anatomy, future work could assess how radiomic features vary in CT images of human subjects. With a sufficiently large patient database, variability in patient anatomy and clinical scan protocols could be overcome, allowing one to meaningfully compare reconstruction methods such as those analyzed in this work as they function in vivo.

Another limitation of this analysis can be found in the treatment of statistical inference when comparing radiomic feature distributions between samples. Because 92 separate hypotheses tests were performed for each comparison of two scans, we applied a Bonferroni correction factor of 1/92 to our significance threshold to account for the increased FWER that occurs when performing multiple hypotheses tests simultaneously. The universal application of Bonferroni methods in medical science has been challenged, however, as it can lead to different interpretations of a test's result depending on the number of other tests that were performed in the same work[44]. We acknowledge, for example, that in our second comparison the failure of the Discovery MI to find significant differences in glcm_JointEntropy distributions of Brass and of Copper with and without MAR enabled (see: Figure 11) was influenced by this stricter threshold and that the samples would have been shown to be significantly different with a traditional threshold of p=0.05.

It is arguable that by seeking to minimize FWER in our analysis we have increased the likelihood of type 2 error, which is especially relevant to our conclusions regarding differences between the MAR algorithms examined. While this challenges the relevance of tests comparing samples of individual features, the choice of threshold would not affect the broader trends observed between scanners and inserts of different densities.

Ultimately, we agree with Nichols et al. that "[T]he strength of Bonferroni and related methods are their lack of assumptions"[45], and we find utility in the ability to perform multiple comparisons using the same threshold for significance. Further, we observe that the difference in behavior between GE and Siemens units in the second comparison would still hold for the aluminum and steel inserts with an uncorrected significance threshold.

As tests and models which rely on radiomic measurements continue to proliferate the importance of standardization in design, testing, and implementation of those tools will only grow. Among the factors which researchers must consider when creating clinically useful tools with radiomics are the image acquisition protocols and procedures used to correct artifacts which affect feature measurement[46]. This work builds on the understanding of one of those procedures by shedding light on inner workings of two MAR algorithms and illustrating how they impact the radiomic features which may be observed with their use.

5 Summary

CT images are widely used throughout radiation oncology, playing an essential role in diagnosis, staging, and treatment planning, and radiomics presents a powerful suite of tools with which researchers and clinicians can utilize quantitative features of CT images. Recent advances in machine learning have drawn increased attention to radiomic analysis as feature groups have found use in the classification of malignancies and the prediction of therapeutic outcomes. However, both the general utility of CT images and the extraction of reliable radiomic features are challenged by common metal implants and the image artifacts they create. Although proprietary artifact reduction algorithms are widely available on modern commercial CT scanners, the finer details of their effect on reconstructed image features are not always known to operators. In this work two metal artifact reduction algorithms were evaluated and found to exhibit different behaviors. The Smart Metal Artifact Reduction algorithm used with the GE Discovery MI was shown to only alter image features when sufficiently attenuating material was present in the FOV, and to preserve a chosen 2D radiomic texture feature in most phantom configurations where metal was introduced. In contrast, the Iterative Metal Artifact Reduction algorithm used with the Siemens SOMATOM X.ceed was found to alter image features absent of highly attenuating material in the FOV and showed significant differences in texture feature distributions from Solid Water™ when removing artifacts. These differences emphasize the importance of standardization of reconstruction parameters for radiomic study.

References

1. Gazi PM, Yang K, Burkett GW, Aminololama-Shakeri S, Anthony Seibert J, Boone JM. Evolution of spatial resolution in breast CT at UC Davis. *Med Phys*. 2015;42(4):1973-1981. doi:10.1118/1.4915079

2. Garau N, Paganelli C, Summers P, et al. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis. *Med Phys*. 2020;47(9):4125-4136. doi:10.1002/mp.14308

3. Pfannenberg AC, Aschoff P, Brechtel K, et al. Value of contrast-enhanced multiphase CT in combined PET/CT protocols for oncological imaging. *Br J Radiol*. 2007;80(954):437-445. doi:10.1259/bjr/34082277

4. Ziemann C, Stille M, Cremers F, Rades D, Buzug TM. The effects of metal artifact reduction on the retrieval of attenuation values. *J Appl Clin Med Phys*. 2017;18(1):243-250. doi:10.1002/acm2.12002

5. Andersson KM, Dahlgren CV, Reizenstein J, Cao Y, Ahnesjö A, Thunberg P. Evaluation of two commercial CT metal artifact reduction algorithms for use in proton radiotherapy treatment planning in the head and neck area. *Med Phys*. 2018;45(10):4329-4344. doi:10.1002/mp.13115

6. Bushberg JT, ed. *The Essential Physics of Medical Imaging*. 3rd ed. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2012.

7. De Man B, Nuyts J, Dupont P, Marchal G, Suetens P. Metal streak artifacts in X-ray computed tomography: a simulation study. *IEEE Trans Nucl Sci*. 1999;46(3):691-696. doi:10.1109/23.775600

8. Vaishnav JY, Ghammraoui B, Leifer M, Zeng R, Jiang L, Myers KJ. CT metal artifact reduction algorithms: Toward a framework for objective performance assessment. *Med Phys*. 2020;47(8):3344-3355. doi:10.1002/mp.14231

9. Simard M, Panta RK, Bell ST, Butler APH, Bouchard H. Quantitative imaging performance of MARS spectral photon-counting CT for radiotherapy. *Med Phys*. 2020;47(8):3423-3434. doi:10.1002/mp.14204

10. Meyer E, Raupach R, Lell M, Schmidt B, Kachelrieß M. Normalized metal artifact reduction (NMAR) in computed tomography. *Med Phys*. 2010;37(10):5482-5493. doi:10.1118/1.3484090

11. Gjesteby L, De Man B, Jin Y, et al. Metal Artifact Reduction in CT: Where Are We After Four Decades? *IEEE Access*. 2016;4:5826-5849. doi:10.1109/ACCESS.2016.2608621

12. Desai SD, Kulkarni L. Comprehensive Survey on Metal Artifact Reduction Methods in Computed Tomography Images: *Int J Rough Sets Data Anal*. 2015;2(2):92-114. doi:10.4018/IJRSDA.2015070106

13. Boas FE, Fleischmann D. CT artifacts: causes and reduction techniques. *Imaging Med*. 2012;4(2):229-240. doi:10.2217/iim.12.13

14. Wellenberg RHH, Hakvoort ET, Slump CH, Boomsma MF, Maas M, Streekstra GJ. Metal artifact reduction techniques in musculoskeletal CT-imaging. *Eur J Radiol*. 2018;107:60-69. doi:10.1016/j.ejrad.2018.08.010

15. Katsura M, Sato J, Akahane M, Kunimatsu A, Abe O. Current and Novel Techniques for Metal Artifact Reduction at CT: Practical Guide for Radiologists. *RadioGraphics*. 2018;38(2):450-461. doi:10.1148/rg.2018170102

16. Reft C, Alecu R, Das IJ, et al. Dosimetric considerations for patients with HIP prostheses undergoing pelvic irradiation. Report of the AAPM Radiation Therapy Committee Task Group 63. *Med Phys*. 2003;30(6):1162-1182. doi:10.1118/1.1565113

17. Huang JY, Followill DS, Howell RM, et al. Approaches to reducing photon dose calculation errors near metal implants. *Med Phys*. 2016;43(9):5117-5130. doi:10.1118/1.4960632

18. Ignatius D, Alkhatib Z, Rowshanfarzad P, et al. Radiotherapy planning of spine and pelvis using single-energy metal artifact reduction corrected computed tomography sets. *Phys Imaging Radiat Oncol*. 2023;26:100449. doi:10.1016/j.phro.2023.100449

19. Puvanasunthararajah S, Fontanarosa D, Wille M, Camps SM. The application of metal artifact reduction methods on computed tomography scans for radiotherapy applications: A literature review. *J Appl Clin Med Phys*. 2021;22(6):198-223. doi:10.1002/acm2.13255

20. Huang JY, Kerns JR, Nute JL, et al. An evaluation of three commercially available metal artifact reduction methods for CT imaging. *Phys Med Biol*. 2015;60(3):1047-1067. doi:10.1088/0031-9155/60/3/1047

21. Lemmens C, Faul D, Nuyts J. Suppression of Metal Artifacts in CT Using a Reconstruction Procedure That Combines MAP and Projection Completion. *IEEE Trans Med Imaging*. 2009;28(2):250-260. doi:10.1109/TMI.2008.929103

22. Liberini V, Laudicella R, Balma M, et al. Radiomics and artificial intelligence in prostate cancer: new tools for molecular hybrid imaging and theragnostics. *Eur Radiol Exp*. 2022;6(1):27. doi:10.1186/s41747-022-00282-0

23. Reginelli A, Nardone V, Giacobbe G, et al. Radiomics as a New Frontier of Imaging for Cancer Prognosis: A Narrative Review. *Diagnostics*. 2021;11(10):1796. doi:10.3390/diagnostics11101796

24. Casà C, Piras A, D'Aviero A, et al. The impact of radiomics in diagnosis and staging of pancreatic cancer. *Ther Adv Gastrointest Endosc*. 2022;15:263177452210815. doi:10.1177/26317745221081596

25. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp*. 2018;2(1):36. doi:10.1186/s41747-018-0068-z

26. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol*. 2020;93(1108):20190948. doi:10.1259/bjr.20190948

27. C A Jayachandran. Calculated effective atomic number and Kerma values for tissue-equivalent and dosimetry materials. *Phys Med Biol*. 1971;16(4):617-623. doi:10.1088/0031-9155/16/4/005

28. Hill R, Kuncic Z, Baldock C. The water equivalence of solid phantoms for low energy photon beams. *Med Phys*. 2010;37(8):4355-4363. doi:10.1118/1.3462558

29. Davis R, Singh A, Jackson MJ, et al. A comprehensive review on metallic implant biomaterials and their subtractive manufacturing. *Int J Adv Manuf Technol*. 2022;120(3-4):1473-1530. doi:10.1007/s00170-022-08770-8

30. NIST: X-Ray Mass Attenuation Coefficients - Table 1. Accessed May 17, 2024. https://physics.nist.gov/PhysRefData/XrayMassCoef/tab1.html

31. REFERENCE TABLES | NIST. Accessed May 17, 2024. https://www.nist.gov/ncnr/sample-environment/sample-mounting/reference-tables

32. iMAR - Iterative Metal Artifact Reduction. Accessed May 5, 2024. https://www.siemens-healthineers.com/en-us/molecular-imaging/options-and-upgrades/software-applications/imar

33. Smart Metal Artifact Reduction. Accessed May 5, 2024. https://www.gehealthcare.com/en-sg/products/computed-tomography/radiation-therapy-planning/metal-artifact-reduction

34. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671-675. doi:10.1038/nmeth.2089

35. The MathWorks Inc. MATLAB version: 24.1.0.2537033 (R2024a). Published online 2024. Accessed April 5, 2024. https://www.mathworks.com

36. Python 3.7 documentation — DevDocs. Accessed May 6, 2024. https://devdocs.io/python~3.7/

37. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The Design of SimpleITK. *Front Neuroinformatics*. 2013;7. doi:10.3389/fninf.2013.00045

38. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

39. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

40. Hodges JL. The significance probability of the smirnov two-sample test. *Ark För Mat*. 1958;3(5):469-486. doi:10.1007/BF02589501

41. Pagano M, Gauvreau K, Mattie H. *Principles of Biostatistics*. Third edition. CRC Press; 2022.

42. Ge G, Zhang J. Feature selection methods and predictive models in CT lung cancer radiomics. *J Appl Clin Med Phys*. 2023;24(1):e13869. doi:10.1002/acm2.13869

43. Khurshid Z, Ahmadzadehfar H, Gaertner FC, et al. Role of textural heterogeneity parameters in patient selection for 177Lu-PSMA therapy via response prediction. *Oncotarget*. 2018;9(70):33312-33321. doi:10.18632/oncotarget.26051

44. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316(7139):1236-1238. doi:10.1136/bmj.316.7139.1236

45. Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*. 2003;12(5):419-446. doi:10.1191/0962280203sm341ra

46. Huang EP, O'Connor JPB, McShane LM, et al. Criteria for the translation of radiomics into clinically useful tests. *Nat Rev Clin Oncol*. 2023;20(2):69-82. doi:10.1038/s41571-022-00707-0