European Journal of Radiology 181 (2024) 111732

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Review

Advancements in supervised deep learning for metal artifact reduction in computed tomography: A systematic review

Cecile E.J. Kleber [a,1], Ramez Karius [a,1], Lucas E. Naessens [a,1], Coen O. Van Toledo [a,1], Jochen A. C. van Osch [b], Martijn F. Boomsma [b], Jan W.T. Heemskerk [c], Aart J. van der Molen [c,*]

[a] Department of Clinical Technology, Faculty of Mechanical Engineering, Delft University of Technology, Delft, the Netherlands
[b] Department of Radiology, Isala Hospital, Zwolle, the Netherlands
[c] Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Tomography, X-ray computed
Metals
Artifacts
Deep Learning
Algorithms

ABSTRACT

Background: Metallic artefacts caused by metal implants, are a common problem in computed tomography (CT) imaging, degrading image quality and diagnostic accuracy. With advancements in artificial intelligence, novel deep learning (DL)-based metal artefact reduction (MAR) algorithms are entering clinical practice.
Objective: This systematic review provides an overview of the performance of the current supervised DL-based MAR algorithms for CT, focusing on three different domains: sinogram, image, and dual domain.
Methods: A literature search was conducted in PubMed, EMBASE, Web of Science, and Scopus. Outcomes were assessed using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) or any other objective measure comparing MAR performance to uncorrected images.
Results: After screening, fourteen studies were selected that compared DL-based MAR-algorithms with uncorrected images. MAR-algorithms were categorised into the three domains. Thirteen MAR-algorithms showed a higher PSNR and SSIM value compared to the uncorrected images and to non-DL MAR-algorithms. One study showed statistically significant better MAR performance on clinical data compared to the uncorrected images and non-DL MAR-algorithms based on Hounsfield unit calculations.
Conclusion: DL MAR-algorithms show promising results in reducing metal artefacts, but standardised methodologies are needed to evaluate DL-based MAR-algorithms on clinical data to improve comparability between algorithms.
Clinical relevance statement:
Recent studies highlight the effectiveness of supervised Deep Learning-based MAR-algorithms in improving CT image quality by reducing metal artefacts in the sinogram, image and dual domain. A systematic review is needed to provide an overview of newly developed algorithms.

Abbreviations: ACDNet, Adaptive Convolutional Dictionary Network; CNN, Convolutional Neural Network/Networks; DAN-Net, Dual-Domain Adaptive-Scaling Non-Local network; DICDNet, Deep Interpretable Convolutional Dictionary Network; DL, Deep Learning; DL-MAR, Deep Learning Metal Artifact Reduction; DuDoNet, Dual Domain Network; DuDoNet++, Dual Domain Network Plus-Plus; FBP, Filtered Back Projection; FSMAR, Frequency-Split Metal Artifact Reduction; IDOL-Net, Interactive Dual Domain Parallel Network; iMAR, Iterative Metal Artifact Reduction; InDuDoNet, Interpretable Dual Domain Network; InDuDoNet+, Interpretable Dual Domain Network Plus; LBRN, Lightweight Block Reconstruction Network; LI, Linear Interpolation; MAR, Metal Artifact Reduction; MARGANVAC, Metal Artefact Reduction Method Based on Generative Adversarial Network with Variable Constraints; MSE, Mean Squared error; MODDNet, Coupling Model and Data Driven Network; NMAR, Normalized Metal Artifact Reduction; O-MAR, Orthopaedic Metal Artifact Reduction; OSCNet, Orientation Shared Convolutional Network; OSCNet+, Orientation Shared Convolutional Network Plus; PSNR, Peak Signal-to-Noise Ratio; SART, Simultaneous Algebraic Reconstruction Technique; SEMAR, Single Energy Metal Artifact Reduction; SSIM, Structural Symmetry Index Measure.

* Corresponding author at: Department of Radiology, C-2S, Leiden University Medical Center, Albinusdreef 2, NL-2333 ZA Leiden, the Netherlands.
E-mail address: A.J.van_der_Molen@lumc.nl (A.J. van der Molen).
[1] Joint first authors.

https://doi.org/10.1016/j.ejrad.2024.111732
Received 1 July 2024; Received in revised form 23 August 2024; Accepted 5 September 2024
Available online 7 September 2024
0720-048X/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computed Tomography (CT) is widely used worldwide and is recognized for its high diagnostic accuracy, largely due to the high resolution and reliable anatomical information provided by CT images [1–3]. This high image quality is the result of continuous advancements over the past decades. At the same time, improvements in healthcare and technology have increased the metal implants, which have a detrimental effect in CT image quality [4,5]. Currently, approximately 21 % of all the CT scans contain these implants. However, these metal implants greatly attenuate or even block X-rays from reaching the detector, resulting in poor projection data. When this projection data is reconstructed into a CT image, it leads to unnatural appearances in the CT image, called artefacts [6]. These metal artefacts are typically caused by several mechanisms such as: beam hardening, photon starvation and scatter, all affecting the image quality [7]. Image degradation is most apparent in tissue adjacent to metal, manifesting as dark and bright streaks around the metal implant. This compromise in image quality lead to unreliable or impractical analysis of tissue near metal objects, depending on the quantity, composition, and size of the metals involved [1].

The presence of these artefacts can pose significant challenges in clinical practice. They can hinder the ability of specialists to draw accurate conclusions, because these artefacts can obscure relevant anatomical structures. This can potentially lead to suboptimal, even incorrect treatment decisions [8].

As CT images are used for diagnosis and treatment planning, and the number of patients with metal implants is increasing, accurate imaging is of paramount importance. Therefore, reducing metal artifacts is essential, not only to minimize errors in CT numbers and enhance delineation accuracy, but also because it has the potential to improve diagnostic accuracy and treatment planning [9].

1.1. Metal artefact reduction

Currently, there are several metal artefact reduction (MAR) algorithms (2), ranging from simple imaging protocol adjustments to convolutional neural networks (CNN), all designed to minimise the influence of the artefacts. Simple parameter adjustments, such as increasing the tube current or voltage, can reduce metal artefacts by increasing the number of photons with a higher energy reaching the detector. However, this results in an increased radiation dose to patient or an image with reduced soft tissue contrast, which is undesirable [10]. Alternatively, there are vendor-specific MAR-algorithms such as single-energy MAR (SEMAR) (Canon Medical Systems) [11], MAR for orthopaedic implants (O-MAR) (Philips) [12,13] and iterative MAR (iMAR) (Siemens Healthineers) [14], but these appear to be inadequate and not entirely reliable for complete suppression of metal artefacts [7]. Presently, researchers are actively exploring the potential of deep learning (DL) to improve MAR-algorithms with its rapid advancements, offering a promising solution for MAR.

Selles et al. for instance, developed a DL-based MAR-model (DL-MAR) and compared it with O-MAR (1) and Zhang et al. introduced CNN as an alternative to the filtered back projection (FBP) method [15].

1.2. Metrics

Several metrics assess the reduction of metal artefacts, with Peak-Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) being the most commonly used in studies developing deep learning algorithms for MAR.

PSNR evaluates the mean squared error (MSE) between an image without metal artifacts (Fig. 1a) and an image with artifacts, both before (Fig. 1b) and after correction using a deep learning (DL) algorithm (Fig. 1c). When artifacts resemble noise in a CT image, PSNR is an effective metric for assessing artifact reduction. However, PSNR may not be sufficient to detect reduction of anatomical details, which is important for diagnostic accuracy [8].

SSIM, on the other hand, is specifically designed to evaluate structural information. It considers contrast, luminance and texture. This makes SSIM more capable of assessing images based on how the human visual system would perceive them. Since clinical detail is very important in medical applications, this is a good metric for assessing metal artefact reduction. However, it is much more complex to interpret than PSNR and is not always sensitive to the perception of the amount of noise in the image.

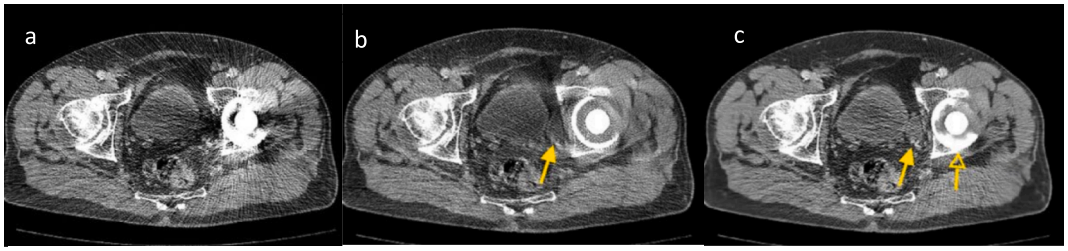
Both are good ways of assessing the image quality of metal artefact reduction algorithms. PSNR and SSIM are often used together to assess the quality of an algorithm. This is because they look at different image properties [8,16].

A common feature of all these MAR-algorithms, whether vendor-specific or DL-based, is that they start by detecting and segmenting the metal that is causing the artefacts. This process can take place in different domains [10]. Most vendor specific MAR-algorithms use image-based or sinogram-based metal segmentation methods [10]. In addition to these domains, it is also possible to detect and segment the metal in the dual domain, which operates in both the image and sinogram domains [17]. However, to date there is no clear consensus on the performance of different supervised DL-algorithms. Therefore, this systematic review provides a first overview of the current supervised DL MAR-algorithms, describing their performances and the domains in which they operate and providing recommendations for further research in this area.

2. Methods

2.1. Literature search

A literature search was performed on April 24, 2024, in the PubMed, EMBASE, World of Science and Scopus databases. The search was based on the keywords:"deep learning","computed tomography","metal artefact" and"reduction", all queries using the same structure of different combinations of related synonyms, MeSH and free text-terms, combined through Boolean operators ('AND','OR'). Searches were limited to



uncorrected image (a), corrected image non-DL MAR algorithm (b), corrected image DL-MAR algorithm (c)

Fig. 1. Visualization of non-DL MAR algorithms versus DL-MAR algorithms.

English and Dutch articles from January 2015 onwards, excluding animal studies (Appendix A). Two teams, formed from the four authors, each screened half of the articles by title, abstract, and, when necessary, full text, adhering predefined selection criteria. Each team member screened the articles assigned to their team individually and tried to reach consensus. Disagreements about the inclusion of articles were resolved by a member of the other team.

2.2. Inclusion and exclusion criteria

Studies were included if they met all of the following criteria: 1. reported a supervised DL-based MAR-algorithm; 2. focused on metal artefacts from implants in patients; 3. measured effectiveness using PSNR, SSIM or other objective numerical outcomes comparing MAR performance to uncorrected images; 4. compared DL-algorithm performance as an outcome to uncorrected images.

Studies were excluded if they met any of the following criteria: 1. used a phantom (physical object designed to simulate human anatomy or tissue properties) for evaluation; 2. used images from cone-beam CT, C-arm CT, spectral CT, micro CT, photoacoustic CT, PET-CT or low dose CT for testing; 3. *meta*-analyses, systematic reviews, editorials, case reports, letters, and abstracts; 4. reported no numerical results in text or tables.

2.3. Quality assessment

The quality of all the articles was assessed independently by two reviewers (CK, RK) using a modified version of the Newcastle-Ottawa Quality Assessment Scale (NOS) [18]. If there was a conflict between the assessments of the two reviewers, a third reviewer made the final decision. The following items were assessed: 1.1 Representativeness of cohort; 1.2 Model selection, development and implementation; 1.3 Comparison made; 1.4 Ground truth assessment and data extraction; 2. Applicability and generalizability (data variability, semi-/fully-automatic, different modalities); 3.1 Outcome assessment (clear split, ground truth objectified); 3.2 Outcome reporting (different outcome measures, uncertainty metrics reported); 3.3 Sharing (data or code sharing). Each item could score up to 1 point, except for'Applicability and generalizability,' which could receive a maximum of 2 points, yielding a total of 9 possible points. Studies were categorised into three groups based on their risk of bias: low (7–9 points), moderate (5–6 points), and high (4 or fewer points). Articles were not excluded based on the quality assessment.

2.4. Data extraction

The data extraction was carried out using a structured approach. Firstly, three domains were categorised. Then, the different types of datasets used were defined, specifying the type of metal artefacts used for training purposes. Performance values of each algorithm for MAR and the uncorrected image (images without application of MAR) were extracted from the articles.

The primary outcomes of this systematic review are PSNR, SSIM and other metrics that compare the performance of the DL-algorithm to an uncorrected image. A high PSNR (close to 100) indicates low noise and a high SSIM (close to 1) indicates similarity to artefact-free images.

Additionally, two secondary outcomes were extracted. Both the performance of DL-algorithms on subjective image quality and the performance of non-DL MAR-algorithms for comparison with DL-algorithms were extracted. All results with a p-value < 0.05 were considered significant.

3. Results

3.1. Article selection

The search in PubMed, EMBASE, Web of Science and Scopus yielded 1000 articles. A total of 633 articles were duplicates. 367 articles were screened for eligibility using the title and abstract; 303 were excluded at this stage. The remaining 64 articles were retrieved, but 9 of them had to be excluded because the full text was not accessible. For the analysis, the inclusion and exclusion criteria were applied to the full text of the 55 articles and 43 of these 55 articles were excluded. Two articles were found by reference checking, resulting in the inclusion of 14 articles for the systematic review (Fig. 2). This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

3.2. Quality assessment

Thirteen of fourteen articles received a quality assessment score between 7 and 9, indicating a low risk of bias (Table 1). In Mai et al., minimal information about the dataset was provided and the mechanism of the algorithm for MAR was not clearly explained, resulting in a higher risk of bias [19].

3.3. Study characteristics

Almost all studies used CT scans with artificially generated artefacts superimposed on the original images to train their DL-algorithms with paired data, due to the limited availability of paired clinical data (Table 2). Only in the study by Mai et al. was there ambiguity regarding the source of data with artefacts for both the training and testing datasets [19].

For the test dataset, most articles used 10 different metal artefact masks on 200 different CT scans from the DeepLesion dataset.

Selles et al. (2023) used clinically paired data for testing purposes [20]. All other studies reported outcomes that were evaluated quantitatively on synthetic CT data.

3.4. Sinogram domain

DL-algorithms operating in the sinogram domain take uncorrected sinograms as input and give corrected sinograms as the output, with the aim of reducing artefacts before image reconstruction (Table 3). CNN [19] and LBRN [21], trained and tested on simulated artefact data, have an increased PSNR and SSIM value compared to the uncorrected image. CNN is only compared to the uncorrected image, whereas LBRN is also compared to non-DL MAR-techniques such as FBP and the simultaneous algebraic reconstruction technique (SART). LBRN achieved higher PSNR and SSIM than both FBP and SART.

LBRN was tested on clinical data, while CNN was not. However, there was only a subjective assessment that LBRN performed better than previously developed techniques was stated.

3.5. Image domain

Six algorithms operate in the image domain for MAR in the uncorrected image: DL-MAR [20] and ACDNet [22], DICDNet [23], OSCNet [24], OSCNet+ [25], and MARGANVAC [26] (Table 3). DL-algorithms that operate in the image domain, take an uncorrected CT-image as input and provide a MAR corrected CT-image as output. The performance of the first five algorithms is evaluated quantitatively on the average artefact size using PSNR and SSIM. All these algorithms have shown an increase in the PSNR and SSIM compared to the uncorrected image with artefacts. The first four algorithms are compared in their performance evaluation to the same uncorrected image set; thus, they are directly comparable. OSCNet + achieved the highest PSNR and SSIM

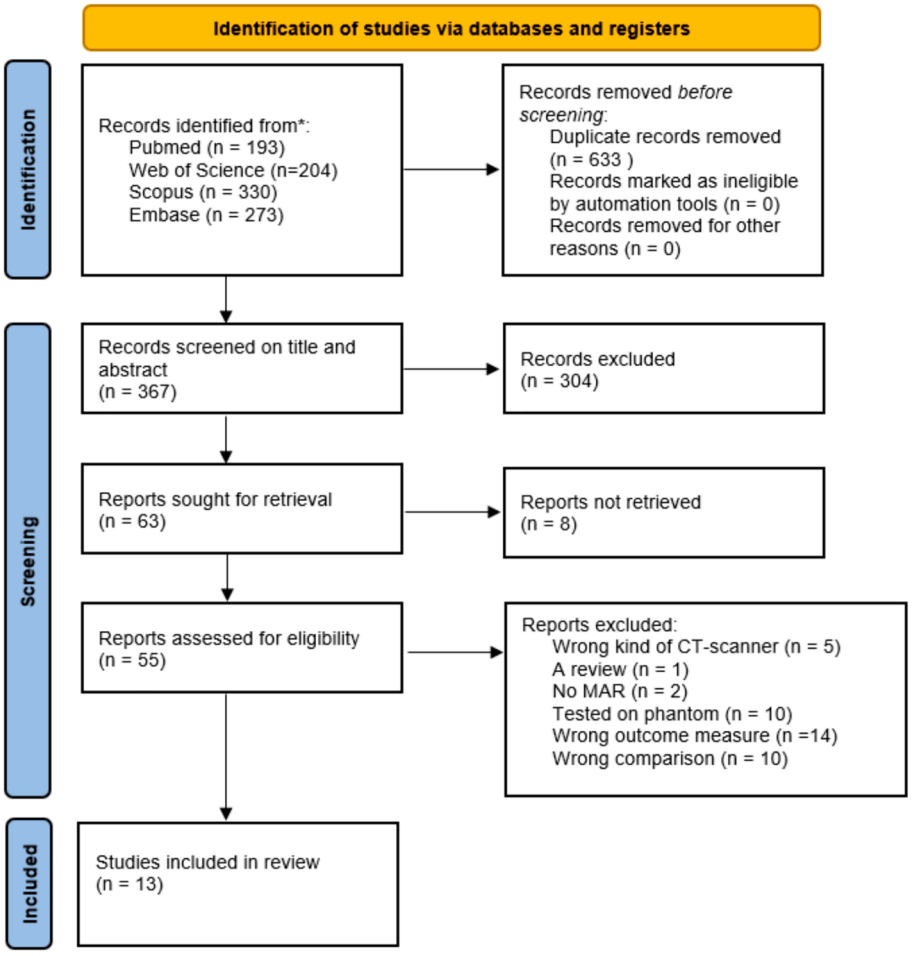


Fig. 2. PRISMA flowchart for the selection of the articles for this systematic review.

Table 1
Results of the modified Newcastle-Ottawa scale (NOS) quality assessment.

Article	1,1	1,2	1,3	1,4	2	3,1	3,2	3,3	Total
Q. Mai et al. (2020) (19)	0	0	1	1	0	1	0	0	3
Y. Lyu et al. (2020) (30)	1	1	1	1	1	1	1	0	7
H. Wang et al. (2021) (31)	1	1	1	1	1	1	0	1	7
T. Wang et al. (2021) (32)	1	1	1	1	1	1	1	1	8
H. Wang et al. (2022) (23)	1	1	1	1	2	1	1	0	8
H. Wang et al. (2022) (22)	1	1	1	1	1	1	0	1	7
H. Wang et al. (2022) (24)	1	1	1	1	1	1	1	1	8
T. Wang et al. (2022) (33)	1	1	1	1	1	1	0	1	7
G. Ma et al. (2022) (21)	1	1	1	1	1	1	1	0	7
M. Selles et al. (2023) (20)	1	1	1	1	1	1	1	1	8
H. Wang et al. (2023) (34)	1	1	1	1	1	1	1	1	8
G. Li et al. (2023) (26)	1	1	1	1	2	1	1	1	9
B.. Shi et al. (2024) (35)	1	1	1	1	2	1	0	1	8
H. Wang et al. (2024) (25)	1	1	1	1	2	1	1	1	9

value of these four algorithms.

In addition, these DL-algorithms were compared to non-DL MAR-algorithms: linear interpolation (LI) [27], normalised MAR (NMAR) [28] or frequency split MAR (FSMAR) [29]. All DL-algorithms outperformed these non-DL MAR-algorithms in terms of PSNR and SSIM, but without proof of statistical significance.

Most of these algorithms have been evaluated on clinical data. However, this evaluation was based on subjective conclusions. The authors found the algorithms in MAR to be sufficient in their opinion.

Other outcomes were used to test DL-MAR [20]. DL-MAR was tested quantitatively on clinical data in six regions of interest (ROIs) around the hip. DL-MAR showed statistically significant noise reduction and percentage MAR in all ROIs compared to the uncorrected image. The percentage of MAR in this article was defined as the percentual deviation of Hounsfield units in the MAR-corrected image compared to the paired image without metal artefacts. DL-MAR significantly outperformed O-MAR in three ROIs for noise reduction and in four ROIs for percentage MAR.

3.6. Dual domain

Dual-domain algorithms are algorithms that operate in both the image and sinogram domains. Six algorithms operated in the dual domain for MAR in the uncorrected images: DuDoNet++ [30], InDuDoNet [31], DAN-Net [32], IDOL-NET [33], InDuDoNet+ [34], and MoDDNet [35]. These were all tested on simulated data with different sizes of artefacts (Table 3). All algorithms provided an increase in PSNR and SSIM value over the uncorrected image. InDuDoNet and InDuDoNet

Table 2
Included studies with corresponding algorithms. The amount of data used for training and testing is provided (All these data are synthesised data except for Selles et al., who tested their algorithm on clinical data).

Sinogram Domain	Q. Mai et al. (2020) (19)	CNN with ground truth elimination	1000 images	102 images
Image domain	M. Selles et al. (2023) (20)	DL-MAR	461 images	25 images (clinical data)
Image domain	H. Wang et al. (2022) (22)	ACDNet	90000 images	2000 images
Image domain	H. Wang et al. (2022) (23)	DICDNet	90000 images	2000 images
Image domain	H. Wang et al. (2022) (24)	OSCNet	90000 images	2000 images
Image domain	H. Wang et al. (2024) (25)	OSCNet+	90000 images	2000 images
Image domain	G. Li et al (2023) (26)	MARGANVAC	360000 images	2000 images
Dual domain	H. Wang et al. (2021) (31)	InDuDoNet	90000 images	2000 images
Dual domain	H. Wang et al. (2023) (34)	InDuDoNet+	90000 images	2000 images
Dual domain	T. Wang et al. (2022) (33)	IDOL-Net	90000 images	2000 images
Dual domain	B. Shi et al. (2024) (35)	MoDDNet	90000 images	2000 images
Dual domain	T. Wang et al. (2021) (32)	DAN-net	90000 images	2000 images
Dual domain	Y. Lyu et al. (2022) (30)	DuDoNet++	360000 images	2000 images

+ were compared to the same uncorrected images.

Furthermore, all articles include a comparison with non-DL MAR-techniques. The PSNR and SSIM values for LI and NMAR were lower than those of the dual DL-algorithms. The IDOL-Net [33] was only compared with LI, not with NMAR. Although tested on clinical data, this clinical evaluation was based solely on subjective observations. The authors concluded that, in their opinion, the algorithms effectively reduced metal artefacts in the patient CT images.

3.7. Metal artefact sizes

Eight studies provided results on different sizes of simulated artefacts to evaluate their algorithms. Each study tested five sizes, ranging from small to large (Table 4). Six studies [22–25,31,34] compared the results to the same uncorrected image. OSCNet + showed higher PSNR and SSIM values for all sizes of artefacts compared to the other five algorithms. DuDoNet++ and MoDDNet used different datasets resulting in different PSNR and SSIM values for the uncorrected image. In general, smaller metal artefacts result in higher PSNR and SSIM values compared to larger ones after MAR.

4. Discussion

The aim of this systematic review was to provide an overview of the performance of the current supervised DL-based MAR algorithms for CT, focusing on three different domains: sinogram, image, and dual domain. Almost all the included articles used the PSNR and SSIM metrics to assess the quality of the images, with PSNR assessing image noise and SSIM assessing contrast values and structural information. The PSNR is

Table 3
Reported MAR results (PSNR & SSIM) of DL-algorithms from the included studies operating in the sinogram, image and dual domain.

Sinogram domain	Algorithm	Uncorrected image		Deep learning algorithm	
		PSNR	*SSIM*	*PSNR*	*SSIM*
Q. Mai et al. (2020) (19)	CNN with ground truth elimination	21,438	0,4329	36,915	0,9094
G. Ma et al. (2022) (21)	LBRN	30,38	0,9655	42,046	0,9921
Image domain	Algorithm	Uncorrected image		Deep learning algorithm	
		PSNR	*SSIM*	*PSNR*	*SSIM*
H. Wang et al. (2022) (22)	ACDNet	27,06	0,7586	40,68	0,9933
H. Wang et al. (2022) (23)	DICDNet	27,06	0,7586	41,83	0,9923
H. Wang et al. (2022) (24)	OSCNet	27,06	0,7586	42,19	0,9931
H. Wang et al. (2024) (25)	OSCNet+	27,06	0,7586	42,93	0,9943
G. Li et al. (2023) (26)	MARGANVAC	25,88	0,685	39,66	0,9752
M. Selles et al. (2023) (20)	DL-MAR	*Difference in noise between DL and uncorrected image (HU) [95 % CI]*		*Metal artifact reduction of DL-MAR compared to uncorrected image (%)*	
Bone		−80 [-111,5,-48,4], $p < 0.001$		92 %, $p < 0.001$	
Bone contralateral		−9.1 [-10.8,-7.5], $p < 0.001$		57 %, $p < 0.001$	
Gluteus medius		−168 [-196.4,-139.7], $p < 0.001$		86 %, $p < 0.001$	
Gluteus medius contralateral		−12.4[-13.0, −11.7], $p < 0.001$		127 %, $p < 0.001$	
Iliacus		−20.2 [-27.0,-13.3], $p < 0.001$		87 %, $p < 0.001$	
Iliacus contralateral		−14 [-15.4,-12.6], $p < 0.001$		94 %, $p < 0.001$	
Dual domain	Algorithm	Uncorrected image		Deep learning algorithm	
		PSNR	*SSIM*	*PSNR*	*SSIM*
Y. Lyu et al. (2020) (30)	DuDoNet ++	19,42	0,869	37,2	0,971
H.Wang et al. (2021) (31)	InDuDoNet	27,06	0,7586	41,48	0,9904
T. Wang et al. (2021) (32)	DAN-Net	15,33	0,6673	40,61	0,9872
T. Wang et al. (2022) (33)	IDOL-Net	18,93	0,7935	41,57	0,988
H. Wang et al. (2023) (34)	InDuDoNet+	27,06	0,7586	41,5	0,9891
B. Shi et al. (2024) (35)	MoDDNet	27,47	0,7777	44,82	0,9961

widely used, because it is easy to calculate, has a clear physical meaning, and is mathematically convenient in the context of optimisation. However, the PSNR does not correspond very well to the visual quality perceived by humans, so the SSIM is used to compensate for this [16]. Overall, all DL-based MAR algorithms showed an improvement in these metrics compared to uncorrected images.

Furthermore, the sinogram domain results indicate improved image quality in terms of PSNR and SSIM, suggesting effective MAR. However, Mai et al. received a low quality assessment score [19], raising concerns in terms of bias and accuracy. Compared to other domains, there are fewer articles discussing algorithms in the sinogram domain. H. Wang et al. proposed four DL MAR-algorithms operating in the image domain [22–25]. OSCNet + showed superior performance in terms of PSNR and SSIM, efficiently reducing noise levels and closely resembling artefact-free images. Although OSCNet + was the most effective MAR algorithm, its effectiveness remained statistically unproven [25]. In contrast, Selles et al. conducted a comprehensive validation using Philips CT data

Table 4
Reported results (PSNR/SSIM) of the distinct sizes of the metals causing the artefacts in different studies. The final column represents the average of PSNR and SSIM values.

Method	Small metal	→	Medium metal	→	Large metal	Average
Uncorrected image H. Wang et al.	28,78/ 0,8076	28,53/ 0,7964	27,75/ 0,7659	26,13/ 0,7471	24,12/ 0,6761	27,06/ 0,7586
ACDNet (22)	42,64/ 0,9965	42,43/ 0,9961	41,14/ 0,9949	39,30/ 0,9920	37,91/ 0,9872	40,68/ 0,9933
InDuDoNet (31)	45,01/ 0,9948	44,47/ 0,9942	41,86/ 0,9931	39,32/ 0,9896	36,74/ 0,9801	41,48/ 0,9904
InDuDoNet+ (34)	45,15/ 0,9959	45,03/ 0,9952	41,81/ 0,9937	39,23/ 0,9872	36,28/ 0,9736	41,50/ 0,9891
DICDNet(23)	45,27/ 0,9958	44,91/ 0,9953	42,25/ 0,9941	39,53/ 0,9908	37,19/ 0,9853	41,83/ 0,9923
OSCNet (24)	45,45/ 0,9962	45,04/ 0,9958	42,92/ 0,9950	39,88/ 0,9902	37,70/ 0,9883	42,19/ 0,9931
OSCNet + (25)	45,99/ 0,9968	45,51/ 0,9965	43,46/ 0,9956	40,72/ 0,9930	38,98/ 0,9897	42,93/ 0,9943
Uncorrected image Y. Lyu et al.	27,64/ 0,899	26,60/ 0,893	26,12/ 0,887	23,07/ 0,854	19,42/ 0,811	24,58/ 0,869
DuDoNet ++ (30)	38,38/ 0,975	38,34/ 0,974	37,84/ 0,974	36,84/ 0,970	34,60/ 0,962	37,20/ 0,971
Uncorrected image B. Shi et al.	29,32/ 0,8275	28,77/ 0,8108	28,66/ 0,8014	26,35/ 0,7612	24,27/ 0,6878	27,47/ 0,7777
MoDDNet (35)	47,36/ 0,9979	47,30/ 0,9977	45,48/ 0,9973	42,35/ 0,9948	41,59/ 0,9930	44,82/ 0,9961

by comparing paired data from clinical patients and demonstrated statistically significant improvements in MAR compared to both uncorrected images and the non-deep learning O-MAR algorithm. Their comprehensive validation serves to reinforce the reliability and efficacy of their methodology and provides compelling statistical evidence in support of its superiority in reducing artefacts in clinical contexts using paired data, in contrast to the other algorithms, for which statistical clinical validation has not been performed [14].

In the dual domain, all algorithms [30–35] effectively reduce metal artefacts, based on the PSNRs and SSIMs. InDuDoNet+ [34] seems to have the best MAR performance compared to InDuDoNet [31].

Although image domain DL-algorithms are user-friendly in clinical settings, most DL-algorithms in this systematic review operate in the dual domain. A possible reason for this is that the dual domain potentially provides the most information to preserve important features in the CT-image while removing artefacts. Nevertheless, the OSCNet + image domain by H. Wang et al. [25] outperformed the dual domain algorithms by the same author [24,31,34] in terms of PSNR and SSIM, indicating superior MAR performance.

4.1. Limitations

Objective comparisons of the performance of all DL algorithms described are challenging because they are assessed using different uncorrected images. This variation makes it impossible to conclude which algorithm performs best. In contrast, the studies by H. Wang et al. compare DL-algorithms on the same uncorrected image [22–24,31,34], which may introduce authorship bias. Additionally, there may be publication bias as all articles report positive results.

Although addressed in some articles, comparisons between DL MAR-algorithms and non-DL MAR-algorithms are limited, and most articles do not include commercial vendor-specific MAR-algorithms that are widely used in healthcare, such as O-MAR [12] and SEMAR [36].

Due to the lack of standardised evaluation methods in most of the included articles, the differences in performance between the three different domains could not be adequately compared. For this reason, and also because no standard deviations are known, it was not possible to perform a *meta*-analysis. Moreover, Selles et al. [20] were the only authors to present their results with statistical evidence.

In this systematic review, only Selles et al. reported the performance of the algorithm quantitatively on clinical data. Although several articles used subjective approaches to assess performance on clinical data [21–25,30–35], no statistical evidence was provided.

Articles evaluating MAR-performance on different sizes of artefacts clearly reported higher PSNR and SSIM values for the reduction of smaller artefacts [22–25,30,31,34,35]. However, there is no statistical significance for the difference in MAR-performance of DL-algorithms for different artefact sizes. Given that all included articles have a positive outcome, it can be argued that publication bias is a factor that should be considered when drawing conclusions.

4.2. Recommendations for further research

For further research on supervised DL MAR algorithms, the following key areas highlighted in this systematic review should be addressed. Currently, almost no studies use an objective quantitative assessment using the same metric on comparable clinical data with statistical testing. More research has been conducted on metrics to measure metal artefacts by Chammin et al. [37]. This article introduces the ρ-index and shows that this new metric is significantly more robust than both the contrast-to-noise ratio and the artifact index.

The only study in this review that validates their DL-algorithm on clinical data is the DL-algorithm of Selles et al. [20]. This study used noise to assess the severity of metal artefacts. Therefore, based on Chammin [37], it cannot be conclusively stated whether the ρ-index would be a better metric than the noise calculation used in the study of Selles et al. [20]. However, it is significantly better than the artefact index. The artefact index is also largely determined by the noise in a region of interest. A standardized metric for evaluating paired clinical data is needed, but a definitive gold standard metric has not yet been established and further research is required.

Furthermore, subjective image quality evaluations were not based on subjective image quality scales, but rather on conclusions drawn from individual opinions. Subjective image quality scales could be used for conclusions based on statistical evidence, which is more favourable.

Objective quantitative performance evaluation and subjective image quality scales should be used for standardised evaluation method to improve comparability of performance between algorithms, allowing future systematic reviews to perform *meta*-analyses. The approach of Selles et al. [20] could serve as a standardised objective evaluation method for clinical data. This would allow researchers to investigate which operating domain for DL MAR-algorithms provides optimal performance and usability.

Finally, given the trend towards improved performance of DL MAR-algorithms on smaller metal implants, future research should investigate the impact of artefact sizes on the performance of DL-algorithms with statistical power.

5. Conclusion

DL MAR algorithms in one or dual domains seem promising solutions for MAR and appear to be the way forward for acquiring CT images that are as free from metal artefacts as possible. However, a standardised method to evaluate these algorithms on clinical data is needed to statistically compare the different algorithms and to determine what further steps need to be taken for clinical implementation.

CRediT authorship contribution statement

Cecile E.J. Kleber: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis,

Conceptualization. **Ramez Karius:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Lucas E. Naessens:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Coen O. Van Toledo:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Jochen A. C. van Osch:** . **Martijn F. Boomsma:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jan W. T. Heemskerk:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Aart J. van der Molen:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Cecile Kleber, Ramez Karius, Lucas Naessens, and Ramez Karius report writing assistance was provided by Isala Zwolle. The Department of Radiology, Isala, Zwolle, the Netherlands, has established a research collaboration with Philips Healthcare regarding metal artifact reduction. This work was supported by a research exhibit with Philips Healthcare (Exhibit B-12: Photon Counting in Metal Artifact Reduction). However, the subsiding party had no decisive role in deep learning model development, data collection, data analysis, nor data - interpretation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper].

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.ejrad.2024.111732.

References

[1] M. Selles, R.H.H. Wellenberg, D.J. Slotman, I.M. Nijholt, J.A.C. Van Osch, K.F. Van Dijke, et al., Image quality and metal artifact reduction in total hip arthroplasty CT: deep learning-based algorithm versus virtual monoenergetic imaging and orthopedic metal artifact reduction, Eur. Radiol. Exp. 8 (2024) 31, https://doi.org/10.1186/s41747-024-00427-3.

[2] D. Kumar, B. Pratap, N. Boora, R. Kumar, N.K. Sah, A comparative study of medical imaging modalities, Int. J. Radiol. Sci. 3 (2021) 9–16.

[3] M.M. Njiti, N.D. Osman, M.S. Mansor, N.A. Rabaiee, M.Z. Abdul Aziz, Potential of metal artifact reduction (MAR) and Deep Learning-based reconstruction (DLR) algorithms integration in CT metal artifact correction: A review, Radiat. Phys. Chem. 218 (2024) 111541, https://doi.org/10.1016/j.radphyschem.2024.111541.

[4] R. Davis, A. Singh, M.J. Jackson, R.T. Coelho, D. Prakash, C.P. Charalambous, et al., A comprehensive review on metallic implant biomaterials and their subtractive manufacturing, Int. J. Adv. Manuf. Technol. 120 (2022) 1473–1530, https://doi.org/10.1007/s00170-022-08770-8.

[5] A. Blum, J.B. Meyer, A. Raymond, M. Louis, O. Bakour, R. Kechidi, et al., CT of hip prosthesis: New techniques and new paradigms, Diagn. Interv. Imaging. 97 (2016) 725–733, https://doi.org/10.1016/j.diii.2016.07.002.

[6] L. Gjesteby Q. Yang Y. Xi H. Shan B. Claus Y. Jin et al., Deep learning methods for CT image-domain metal artifact reduction, Proc. SPIE 10391, Developments in X-Ray Tomography XI, 103910W (25 September 2017); Doi: 10.1117/12.2274427.

[7] F.E. Boas, D. Fleischmann, Evaluation of two iterative techniques for reducing metal artifacts in computed tomography, Radiology. 259 (2011) 894–902, https://doi.org/10.1148/radiol.11101782.

[8] L. Gjesteby, B. De Man, Y. Jin, H. Paganetti, J. Verburg, D. Giantsoudi, et al., Metal artifact reduction in CT: where are we after four decades? IEEE Access 4 (2016) 5826–5849, https://doi.org/10.1109/ACCESS.2016.2608621.

[9] J. King, S. Whittam, D. Smith, B. Al-Qaisieh, The impact of a metal artefact reduction algorithm on treatment planning for patients undergoing radiotherapy of the pelvis, Phys. Imaging. Radiat. Oncol. 24 (2022) 138–143, https://doi.org/10.1016/j.phro.2022.11.007.

[10] M. Katsura, J. Sato, M. Akahane, A. Kunimatsu, O. Abe, Current and novel techniques for metal artifact reduction at CT: Practical guide for radiologists, Radiographics. 38 (2018) 450–461, https://doi.org/10.1148/rg.2018170102.

[11] Y.B. Chang, D. Xu, A.A. Zamyatin, Metal artifact reduction algorithm for single energy and dual energy CT scans, IEEE, Washington DC, 2012, pp. 3426–3429, https://doi.org/10.1109/NSSMIC.2012.6551781.

[12] Philips. Metal Artifact Reduction for Orthopedic Implants (O-MAR). https://www.philips.co.uk/c-dam/b2bhc/master/sites/hotspot/omar-metal-artifact-reduction/O-MAR%20whitepaper_CT.pdf/ 2012 (Accessed 22, August 2024).

[13] H. Li, C. Noel, H. Chen, H.H. Li, D. Low, K. Moore, et al., Clinical evaluation of a commercial orthopedic metal artifact reduction tool for CT simulations in radiation therapy, Med. Phys. 39 (2012) 7507–7517, https://doi.org/10.1118/1.4762814.

[14] M. Axente, A. Paidi, R. Von Eyben, C. Zeng, A. Bani-Hashemi, A. Krauss, et al., Clinical evaluation of the iterative metal artifact reduction algorithm for CT simulation in radiotherapy, Med. Phys. 42 (2015) 1170–1183, https://doi.org/10.1118/1.4906245.

[15] Y. Zhang, H. Yu, Convolutional neural network based metal artifact reduction in X-Ray computed tomography, IEEE Trans. Med. Imaging. 37 (2018) 1370–1381, https://doi.org/10.1109/TMI.2018.2823083.

[16] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image. Process. 13 (2004) 600–612, https://doi.org/10.1109/TIP.2003.819861.

[17] M. Selles, J.A.C. van Osch, M. Maas, M.F. Boomsma, R.H.H. Wellenberg, Advances in metal artifact reduction in CT images: A review of traditional and novel metal artifact reduction techniques, Eur. J. Radiol. 170 (2024) 111276, https://doi.org/10.1016/j.ejrad.2023.111276.

[18] G. A. Wells, B. Shea, D. O'Connell, J. Peterson, V. Welch, M. Losos, et al, The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses, Available at: https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp 2000 (Accessed 22. August 2024).

[19] Q. Mai, J.W.L. Wan, Metal artifact reduction in CT scans using convolutional neural network with ground truth elimination, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (2020) 1319–1322, https://doi.org/10.1109/EMBC44109.2020.9176173.

[20] M. Selles, D.J. Slotman, J.A.C. van Osch, I.M. Nijholt, R.H.H. Wellenberg, M. Maas, et al., Is AI the way forward for reducing metal artifacts in CT? Development of a generic deep learning-based method and initial evaluation in patients with sacroiliac joint implants, Eur. J. Radiol. 163 (2023) 110844, https://doi.org/10.1016/j.ejrad.2023.110844.

[21] G. Ma, X. Zhao, Y. Zhu, H. Zhang, Projection-to-image transform frame: a lightweight block reconstruction network for computed tomography, Phys. Med. Biol. 67 (2022) 035010, https://doi.org/10.1088/1361-6560/ac4122.

[22] H. Wang, Y. Li, D. Meng, Y. Zheng. Adaptive convolutional dictionary network for CT metal artifact reduction. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), 2022, pp. 1401–1407. Available at: https://www.ijcai.org/proceedings/2022/0195.pdf. (Accessed: 22. August 2024).

[23] H. Wang, Y. Li, N. He, K. Ma, D. Meng, Y. Zheng, DICDNet: Deep interpretable convolutional dictionary network for metal artifact reduction in CT images, IEEE Trans. Med. Imaging. 41 (2022) 869–880, https://doi.org/10.1109/TMI.2021.3127074.

[24] H. Wang, Q. Xie, Y. Li, Y. Huang, D. Meng, Y. Zheng. Orientation-shared convolution representation for CT metal artifact learning. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S (eds). Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, 2022; 13436: 665-675. Springer, Cham. Doi: 10.1007/978-3-031-16446-0_63.

[25] H. Wang, Q. Xie, D. Zeng, J. Ma, D. Meng, Y. Zheng, OSCNet: orientation-shared convolutional network for CT metal artifact learning, IEEE Trans. Med. Imaging. 43 (2024) 489–502, https://doi.org/10.1109/TMI.2023.3310987.

[26] G. Li, L. Ji, C. You, S. Gao, L. Zhou, K. Bai, et al., MARGANVAC: metal artifact reduction method based on generative adversarial network with variable constraints, Phys. Med. Biol. 68 (2023) 205005, https://doi.org/10.1088/1361-6560/acf8ac.

[27] W.A. Kalender, R. Hebel, J. Ebersberger, Reduction of CT artifacts caused by metallic implants, Radiology. 164 (1987) 576–577, https://doi.org/10.1148/radiology.164.2.3602406.

[28] E. Meyer, R. Raupach, M. Lell, B. Schmidt, M. Kachelrieß, Normalized metal artifact reduction (NMAR) in computed tomography, Med. Phys. 37 (2010) 5482–5493, https://doi.org/10.1118/1.3484090.

[29] E. Meyer, R. Raupach, M. Lell, B. Schmidt, M. Kachelrieß, Frequency split metal artifact reduction (FSMAR) in computed tomography, Med. Phys. 39 (2012) 1904–1916, https://doi.org/10.1148/radiology.164.2.3602406.

[30] Y. Lyu, W.A. Lin, J. Lu, S.K. Zhou. DuDoNet++ : Encoding metal mask projection for metal artifact reduction in computed tomography. In: Martel AL, et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12262: 147-157. Springer, Cham. Doi: 10.1007/978-3-030-59713-9_15.

[31] H. Wang, Y. Li, H. Zhang, J. Chen, K. Ma, D. Meng, et al. InDuDoNet: An interpretable dual domain network for CT metal artifact reduction. In: de Bruijne, M, et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12906: 107-118. Springer, Cham. Doi: 10.1007/978-3-030-87231-1_11.

[32] T. Wang, W. Xia, Y. Huang, H. Sun, Y. Liu, H. Chen, et al., DAN-Net: Dual-domain adaptive-scaling non-local network for CT metal artifact reduction, Phys. Med. Biol. 66 (2021) 115009, https://doi.org/10.1088/1361-6560/ac1156.

[33] T. Wang, Z.X. Lu, Z.Y. Yang, W.J. Xia, M.Z. Hou, H.Q. Sun, et al., IDOL-Net: An interactive dual-domain parallel network for CT metal artifact reduction, IEEE Trans. Radiat. Plasma. Med. Sci. 6 (2022) 874–885.

[34] H. Wang, Y. Li, H. Zhang, D. Meng, Y. Zheng, InDuDoNet+: A deep unfolding dual domain network for metal artifact reduction in CT images, Med. Image. Anal. 85 (2023) 102729, https://doi.org/10.1016/j.media.2022.102729.

[35] B. Shi, S. Zhang, K. Jiang, Q. Lian, Coupling model- and data-driven networks for CT metal artifact reduction, IEEE Trans. Comput. Imaging. 10 (2024) 415–428, https://doi.org/10.1109/TCI.2024.3369408.

[36] D. Zhang, Single Energy Metal Artifact Reduction, White paper. Available at: https://us.medical.canon/download/ct-aq-one-genesis-wp-semar. 2017 (Accessed: 22. August 2024).

[37] J. Cammin, A robust index for metal artifact quantification in computed tomography, J. Appl. Clin. Med. Phys. 25 (2024) e14453.