

Explainable AI-Driven Robustness and Clinical Safety Evaluation of Medical Image Watermarking Schemes

Chukka Harika^{1*}, Dr. N. Raghavendra Sai²

¹ Student, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503. Email: 2401050090cse@gmail.com

² Associate Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503. Email: nallagatlaraghavendra@kluniversity.in

*Corresponding Author: 2401050090cse@gmail.com

Abstract

Medical image watermarking plays a pivotal role in safeguarding clinical imaging processes in Picture Archiving and Communication Systems (PACS) and teleradiology applications. Existing evaluation frameworks rely solely on global signal metrics, such as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which measure average distortion and cannot determine whether the distortion falls within diagnostically critical image regions. This study proposes an explainable artificial intelligence (XAI)-driven framework for evaluating the clinical safety of medical image watermarking schemes across three imaging modalities. Three DenseNet-121 diagnostic models were fine-tuned on publicly available CT, chest X-ray (CXR), and MRI datasets, achieving validation accuracies of 97.5%, 77.5%, and 90.8%, respectively. Grad-CAM saliency maps derived from these models served as the basis for three proposed clinical safety metrics: Saliency Overlap Score (SOS) (Dice/IoU), Diagnostic Probability Shift (ΔP), and Region-specific PSNR/SSIM. A novel XAI-guided watermarking scheme was proposed that restricts embedding to low-saliency background pixels. A total of 7,200 evaluations were conducted across four watermarking schemes, six attack conditions, and three imaging modalities. The XAI-guided scheme achieved the highest SOS Dice on all three modalities (CT: 0.896 ± 0.088 , CXR: 0.855 ± 0.168 , MRI: 0.935 ± 0.105) and the lowest ΔP (CT: 0.008 ± 0.012 , CXR: 0.011 ± 0.009 , MRI: 0.004 ± 0.007). An early training HiDDeN-based distortion-injection baseline, despite achieving an ROI-PSNR of 52–71 dB, satisfied the proposed $\Delta P < 0.02$ reference criterion in only 31–68% of images per modality, suggesting that a high global PSNR does not reliably indicate clinical safety for learned-perturbation watermarking. The proposed framework demonstrates that saliency-aware evaluation provides clinically relevant insights beyond conventional signal fidelity metrics and offers a practical methodology for assessing the clinical safety of medical image watermarking schemes. The findings further suggest that the preservation of diagnostic saliency and model confidence should be considered alongside PSNR, SSIM, and BER when evaluating the watermarking performance in clinical imaging environments.

Keywords: Medical image watermarking, explainable AI, Grad-CAM, saliency overlap score, DenseNet-121, diagnostic probability shift, region-specific PSNR, DWT-SVD, and HiDDeN.

1. Introduction

Digital imaging has become a common part of clinical practice and is creating enormous amounts of medical image data that must be safely stored, communicated, and authenticated within Picture Archiving and Communication Systems (PACS), teleradiology systems, and electronic health record (EHRs) environments [6]. The integrity and provenance of medical images have become issues of clinical governance and patient safety as an increasing amount of clinical AI is involved in the decision-making process relating to diagnosis [14]. To meet this need, watermarking has been introduced into the field of medical image processing and embedded into image data in such a way that it can survive basic clinical processing and be retrieved at any stage along the imaging pathway [4,19]. In telemedicine and PACS workflows, published applications include copyright protection, tamper detection, and chain-of-custody verification [16,20,2].

The usual approach for medical image watermarking evaluation is based on three global signal metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Bit Error Rate (BER) [4,22]. These metrics quantify the average distortion and watermark recoverability but are insensitive to the spatial distribution of distortion relative to diagnostically important structures. A scheme that achieves a PSNR of 52 dB by concentrating distortion within a

pulmonary nodule or brain tumor satisfies every conventional benchmark; however, it may adversely affect the exact image content relied upon by a clinician or diagnostic AI model. Conversely, a scheme that confines the same embedding to a homogeneous background tissue achieves the same global PSNR while leaving the diagnostic region unaffected by design. This distinction is undetectable by PSNR or SSIM but is directly relevant to clinical safety.

Explainable artificial intelligence (XAI) has become an important tool for interpreting deep learning decisions in high-stakes medical domains [21,7]. Gradient-weighted Class Activation Mapping (Grad-CAM) [13] generates spatial saliency maps that highlight the image regions that are most influential to a convolutional neural network's classification outcome. In medical imaging, Grad-CAM activations have been shown to correlate with regions of radiological significance [1,11]: CT activations localize to pulmonary nodule and mass regions; CXR activations localize to the pleural space; and MRI FLAIR activations localize to the tumor core and perilesional edema. These properties make Grad-CAM a candidate instrument for assessing whether the watermark-induced distortion falls within the spatial attention of a diagnostic model, a capability that global PSNR cannot provide.

To the best of our knowledge, no prior study has applied XAI saliency maps as an evaluation instrument for clinical safety in medical image watermarking [4,9,22]. Related work by Khaldi et al. [8] explored the integration of explainability into fragile watermarking for medical deepfake detection, using a multi-scale attention mechanism to modulate the DCT embedding strength in diagnostically relevant regions. The present study proposes a complementary evaluation framework: rather than guiding where to embed, it uses Grad-CAM saliency maps derived from validated diagnostic models to measure the clinical impact of any watermarking scheme on the behavior of diagnostic models.

The specific contributions of this study are as follows: (i) a proposed XAI-driven framework for evaluating the clinical safety of medical image watermarking, grounded in Grad-CAM saliency maps from modality-specific DenseNet-121 diagnostic models; (ii) three proposed clinical safety metrics, the Saliency Overlap Score (SOS Dice/IoU), the Diagnostic Probability Shift (ΔP), and Region-specific PSNR/SSIM, that measure the distortion impact on diagnostic model behavior rather than relying solely on global signal statistics; (iii) a deliberately simple, proof-of-concept XAI-guided watermarking scheme that restricts embedding to low-saliency background pixels, included to demonstrate how saliency information can inform embedding rather than as a competing watermarking algorithm; and (iv) a systematic empirical evaluation comprising 7,200 assessments across three imaging modalities, four watermarking schemes, and six simulated attack conditions. The evaluation framework and the clinical safety perspective it provides are the major methodological novelties of this work, whereas the XAI-Guided scheme is presented as a concrete instantiation of that perspective and not as the key methodological novelty.

Summary of Contributions

Contribution Area	Contribution in This Study
Evaluation perspective	Introduces an XAI-driven framework for assessing the clinical safety of medical image watermarking, rather than relying only on signal fidelity metrics.
Proposed metrics	The SOS (Dice/IoU), Diagnostic Probability Shift (ΔP), and Region-specific PSNR/SSIM were used to quantify the effect of watermarking on the diagnostic model behavior.
Watermarking design	The framework is demonstrated with a deliberately simple, proof-of-concept XAI-Guided scheme that restricts embedding to low-saliency pixels and preserves saliency-defined diagnostic regions by design.
Empirical scope	Evaluated four watermarking schemes under six attack conditions across CT, CXR, and MRI, yielding 7,200 total evaluations.

Clinical relevance	This demonstrates that a high global PSNR alone does not necessarily imply the preservation of diagnostic saliency or classifier confidence.
--------------------	--

The remainder of this paper is organized as follows. The relevant literature on medical image watermarking and XAI in medical imaging is reviewed in Section 2, and the research gap filled in this study is identified. The proposed methodology is described in Section 3. The experimental setup is described in Section 4. The results are reported and discussed in Section 5, respectively. The paper ends with a summary section (Section 6) that summarizes the findings and points out potential avenues for future research.

2. Related Work

2.1 Medical Image Watermarking

Medical image watermarking techniques can be generally categorized into spatial domain and transform domain methods [4,22]. Transform-domain-based methods, which are performed on frequency domain coefficients, are known to yield better imperceptibility and robustness than direct pixel domain modifications. A widely used classical baseline is the discrete wavelet transform with singular value decomposition (DWT-SVD) [15,4], in which a multi-level wavelet decomposition is performed, the singular values in the low-frequency subband are scaled by the watermark payload proportionately, and the image is then reconstructed by applying inverse DWT. This approach normally achieves a PSNR value in the range of 40–50 dB and has moderate robustness when subjected to signal processing attacks. Gull and Parah [4] provide an extensive and thorough review of the current state-of-the-art medical image watermarking techniques and conclude that a major challenge is the lack of clinically based safety evaluation.

Discrete Cosine Transform (DCT)-based methods embed watermark bits in mid-frequency block coefficients using quantization index modulation (QIM), achieving PSNR values of 50–55 dB; however, they are vulnerable to geometric distortions owing to their block-structured nature [15]. To address the problem of distortion within diagnostically important regions, region-of-interest (ROI) and region-of-non-interest (RONI) watermarking methods have been proposed, in which embedding is restricted to non-diagnostic background areas [20,16]. Thanki et al. [20] developed a RONI-based visible watermarking approach for medical image authentication using the Human Visual System (HVS) model to delineate non-diagnostic regions, reporting improved PSNR relative to uniform embedding schemes. Su et al. [16] proposed a reversible data-hiding scheme that uses local rotation within the RONI of medical images to achieve lossless recovery of the original image after the watermark is removed. However, these approaches depend on manually defined or heuristically derived region boundaries, which limit their scalability across different modalities and imaging tasks. The XAI-guided scheme proposed in this study extends the RONI paradigm by replacing manual annotation with data-driven Grad-CAM saliency maps, thereby enabling automatic model-informed region protection.

Deep-learning-based watermarking methods have emerged as a distinct research area. Zhu et al. [23] introduced HiDDeN (Hiding Data with Deep Networks), an encoder–decoder convolutional neural network trained to embed a fixed-length binary payload as a learned spatial residual, with robustness improved through differentiable attack simulation during training. Subsequent architectures refined HiDDeN's approach by introducing adversarial training [2] and attention-guided embedding [8]. Taj et al. [18] proposed a reversible-zero watermarking scheme that combines VGG19-based feature extraction with integer wavelet transforms for medical image authentication. Taj et al. [19] further proposed a lightweight dual-watermarking framework combining DWT-DCT frequency-domain encoding with a CNN decoder, reporting PSNR values of 64.87–68.75 dB on medical images. Dong et al. [3] proposed a zero-watermarking scheme using NasNet-Mobile and DCT for medical images. Reviews by Bistróń et al. [2] and Zhong et al. [22] confirm that all published medical watermarking studies report only PSNR, SSIM, and BER as measures of clinical acceptability, with no consideration of how distortion is distributed relative to the diagnostic structures.

2.2 Explainable AI in Medical Imaging

XAI methods for deep neural networks encompass gradient-based attribution [13,17], perturbation-based methods, and game-theoretic approaches such as SHapley Additive exPlanations (SHAP) [10]. Grad-CAM [13] generates a

spatial activation map from the gradients of a target class score with respect to the final convolutional layer feature maps, providing a coarse but clinically interpretable visualization of the model's attention. Integrated Gradients [17] offers a complementary attribution method with proven sensitivity and implementation invariance properties. Grad-CAM has been used in medical imaging for the interpretation of models in a variety of modalities [11,9]. Rajpurkar et al. [11] showed DenseNet-121 Grad-CAM activation on chest radiograms correlate well with radiologist attention, especially for pneumonia localisation. Arun et al. [1] systematically assessed eight different saliency map techniques on two large radiology datasets, concluding that Grad-CAM is a reliable method for extracting abnormality localization utility compared to experts' ground truth. Clinically, there are formal guidelines for evaluating clinical XAI in medical imaging by Jin et al. [7] and comprehensive surveys of XAI approaches to deep learning-based medical image analysis by van der Velden et al. [21] considering anatomy, pathology, and modality-specific evaluation criteria. Samek et al. [12] surveyed the general field of visualisation and interpretation techniques for deep networks.

2.3 Research Gap

The above review makes two parallel observations. First, published medical image watermarking studies assess clinical acceptability using global image distortion measures (PSNR, SSIM, and BER) irrespective of the spatial distribution of the distortion [4,15,19,22]. Second, although XAI research has confirmed that Grad-CAM saliency maps are clinically meaningful proxies for model attention in medical imaging cases [1,21], they have not yet been used in watermarking evaluation. Although Khaldi et al. [8] proved that XAI ideas can enhance watermark design for deepfake detection, they did not provide clinical safety evaluation metrics based on saliency.

Research Gap in Medical Image Watermarking and XAI-Based Evaluation:

Study group in reviewed literature	Main evaluation focus	Clinical-safety interpretation
Classical medical watermarking methods (e.g., DWT-SVD, DCT, ROI/RONI schemes)	PSNR, SSIM, BER, robustness under attacks	Distortion is assessed globally, and spatial relevance to diagnostic regions is not directly measured.
Deep learning watermarking methods discussed in the review	PSNR, SSIM, BER, decoder recovery performance	Learned perturbations may appear acceptable under signal metrics, even when diagnostic behavior changes.
Present study	SOS Dice/IoU, ΔP , Region-specific PSNR/SSIM, alongside conventional metrics	Distortion is evaluated with respect to saliency-defined diagnostic regions and classifier behavior.

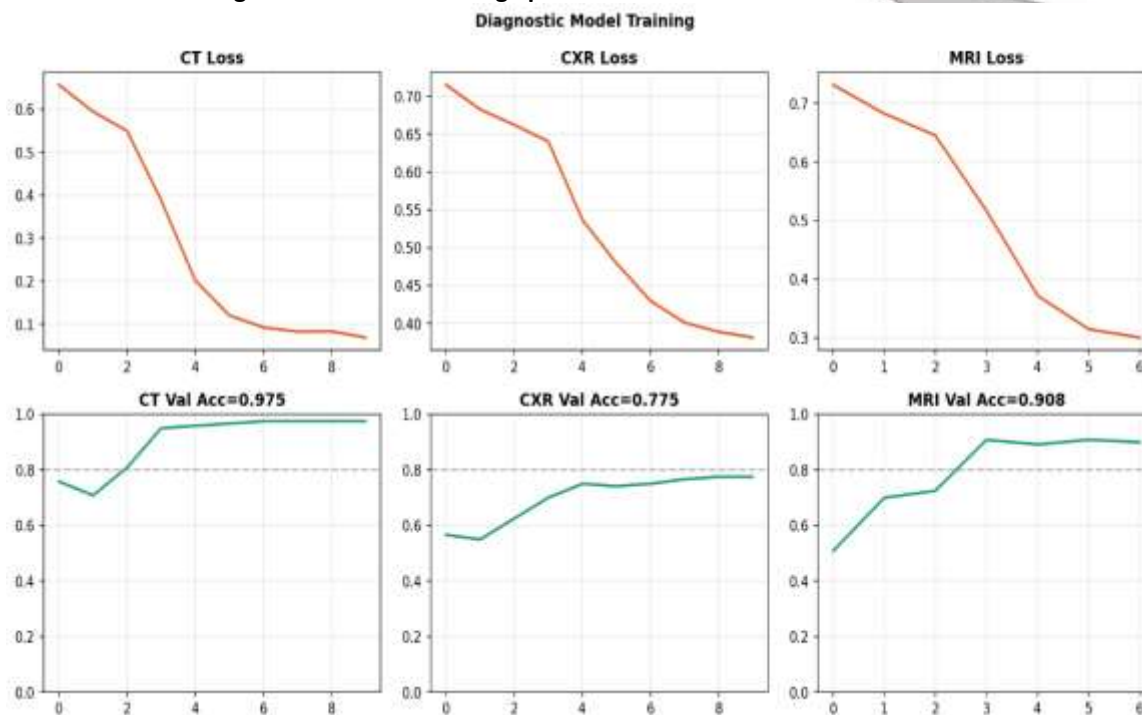
3. Methodology

3.1 Framework Overview

The suggested assessment course of action comprises five stages. In Stage 1, a dedicated DenseNet-121 diagnostic classifier was trained for each imaging modality. In Stage 2, baseline Grad-CAM saliency maps were computed for each of the clean test images to set a spatial attention baseline. Stage 3 injects a watermark into one image of each set of clean images for each of the four schemes tested. To simulate the signal processing operations that are typical in PACS storage, compression, and transmission, six attack conditions were applied to each watermarked image in stage 4. Stage 5 calculates three suggested clinical safety measures based on the Grad-CAM saliency maps and the probabilities of the three classifiers for the clean image and each processed image. The training histories of the three diagnostic models in Stage 1 are plotted in Figure 1, with the quality of these models directly affecting the reliability of the saliency maps they produced in all subsequent stages.

Figure 1. DenseNet-121 training curves for the CT (val_acc = 97.5%), CXR (val_acc = 77.5%), and MRI (val_acc = 90.8%) diagnostic models. Solid curves: training loss (left axis); dashed curves: validation accuracy (right axis). All

three models converge within ten fine-tuning epochs.



3.2 Diagnostic Model and Grad-CAM Saliency Maps

The three main reasons for choosing DenseNet-121 as the diagnostic backbone were as follows: (i) DenseNet-121's dense connectivity architecture maximizes feature reuse across layers, enabling it to empirically generate highly localized Grad-CAM activations of medical structures; (ii) DenseNet-121 has been evaluated as a chest radiograph diagnostic backbone in a number of published studies [11] with radiologist-level performance that gave confidence that features learned from ImageNet were clinically meaningful; and (iii) ImageNet pre-training was required to enable effective fine-tuning from limited medical datasets, with only 480 images used here for training. To avoid overfitting, fine-tuning was performed in two steps. In phase 1 (epochs 1-3) the classifier heads a single fully connected network with sigmoid activation was trained with the convolutional backbone frozen, while using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. In Phase 2 (epochs 4-10), the entire network was unfrozen and trained at a smaller learning rate of 1×10^{-5} using cosine annealing scheduling and early stopping with a patience of three epochs. All input images were resized to 224×224 and then normalized to ImageNet channel statistics. Grad-CAM saliency maps were generated using the Captum v0.7 library, specifically on top of the features. `denseblock4.denselayer16.conv2` layer. This specific layer was selected as the last convolutional layer in DenseNet-121, before global average pooling, because it was the deepest available semantic representation and also had a spatial feature map, which is critical for saliency maps to be concentrated on the anatomical subregions and not diffuse whole-image activations. Raw activation maps (7×7 pixels) were upsampled to 224×224 using bilinear interpolation and normalized to the range of $[0, 1]$. The resulting saliency map $C(x)$ assigns values close to 1 to pixels that most strongly influence the model's classification decision and values close to 0 to background pixels. Although Grad-CAM does not constitute a direct measure of radiologist attention, it provides a reproducible and widely adopted approximation of model attention that is suitable for large-scale automated evaluations.

3.3 Watermarking Schemes

S1 - DWT-SVD: A two-level Haar DWT is applied to each color channel of the input image, producing LL (approximation), LH (horizontal), HL (vertical), and HH (diagonal) sub-band images. Singular value decomposition (SVD) is applied to the LL subband as follows: The singular value matrix is modified as:

$$\Sigma' = \Sigma + \alpha \cdot w \cdot \Sigma \quad (1)$$

where $\alpha = 0.06$ is the embedding strength parameter, and $w \in \{-1, +1\}$ is the binarized watermark pattern derived from the 64-bit payload. The embedding strength $\alpha = 0.06$ was determined through preliminary experiments: values above 0.10 produced visually perceptible ringing artifacts (PSNR < 38 dB), whereas values below 0.04 resulted in insufficient robustness under JPEG Q50 compression. The modified matrix Σ' is used to reconstruct the LL subband ($LL' = U\Sigma'V^T$), and the inverse Haar DWT restores the spatial-domain watermarked image. The watermark is extracted by recomputing the SVD of the LL subband from the received image and comparing the singular values.

S2 - DCT: The image is divided into non-overlapping 8×8 -pixel blocks, and a 2D Discrete Cosine Transform is applied to each block. One watermark bit is embedded per block by modifying the coefficient at the mid-frequency position (4, 3) using quantization index modulation (QIM):

$$c'(u, v) = \Delta \cdot \text{round}(c(u, v)/\Delta + b/2), \quad b \in \{0, 1\} \quad (2)$$

where $\Delta = 8.0$ is the quantization step size, and $b \in \{0, 1\}$ is the watermark bit. Extraction applies the same QIM rule to the received coefficients, rounding to determine the nearest quantization level. This scheme achieves a higher global PSNR than DWT-SVD owing to smaller per-pixel perturbations; however, its block-structured distortion pattern can produce artifacts that coincide with structured anatomical boundaries.

S3-HiDDeN (Distortion-Injection Baseline): Zhu et al. [23] introduced HiDDeN, an encoder-decoder CNN that learns, through end-to-end training, to embed a fixed-length binary watermark as a spatial residual added to a cover image. The encoder output was scaled by a factor of 0.3 and added to the input. Training jointly minimizes the composite loss:

$$\mathcal{L} = 0.5 \cdot \mathcal{L}_{img} + 0.5 \cdot \mathcal{L}_{wm} \quad (3)$$

where $\mathcal{L}_{img}$ is the mean-squared-error image reconstruction loss and $\mathcal{L}_{wm}$ is the binary cross-entropy watermark recovery loss, weighted equally at 0.5. Due to GPU resource constraints on the Kaggle T4 platform used for this study, HiDDeN was trained for 50 epochs only, yielding a decoder bit accuracy of approximately 0.50. As this value is indistinguishable from random chance, the HiDDeN encoder in this study acts as a distortion-injection baseline: it applies learned spatial perturbations to the image without recovering meaningful watermark bits. It is therefore used as a distortion-injection baseline that allows the clinical impact of deep-learned spatial residuals to be characterised independently of watermark recoverability, rather than as a converged deep watermarking system, and HiDDeN results are interpreted accordingly.

S4 - XAI-Guided (Proposed): As a deliberately simple, proof-of-concept demonstration of the framework rather than an optimised watermarking algorithm, this scheme uses the Grad-CAM saliency map $C(x)$ of the clean image to define a binary protection mask:

$$M = \mathbb{1}[C(x) > \tau] \quad (4)$$

where $\tau \in (0, 1)$ is the saliency threshold. The XAI-Guided watermarked image x^* is constructed by blending the original image and the DWT-SVD watermarked image according to M :

$$x^* = x \cdot M + x_{wm} \cdot (1 - M) \quad (5)$$

Pixels where $M = 1$ (normalized saliency above τ) are retained exactly from the original; only pixels where $M = 0$ receive the DWT-SVD watermark. This design ensures that the ROI-PSNR is structurally equal to ∞ dB, reported as 100.0 dB in the tables for every image, regardless of the embedding strength, because no modification is applied to any pixel within the saliency-defined region of interest. The saliency threshold $\tau = 0.40$ was selected based on the ablation study reported in Section 5.4; it is the lowest value that maintains the ROI-PSNR guarantee across all three modalities while providing greater embedding capacity than the more conservative $\tau = 0.30$ setting.

3.4 Attack Simulation

Six attack conditions were applied to every watermarked image to assess robustness under signal processing operations typical of clinical imaging workflows: (1) clean no attack, providing a baseline under ideal conditions; (2) JPEG Q50 lossy JPEG compression at quality factor 50, representative of aggressive PACS image storage; (3) JPEG Q30 JPEG compression at quality factor 30, representing an upper bound of lossy compression encountered in practice; (4) Gaussian Noise ($\sigma = 0.03$) additive white Gaussian noise simulating sensor acquisition artifacts; (5) Gaussian Blur (5×5 kernel, $\sigma = 1.0$) low-pass spatial filtering representing resampling and PACS viewer processing; and (6) scaling ($0.5 \times$ downscale followed by upscale to original resolution) geometric resampling simulating PACS

thumbnail generation. These conditions are representative of the robustness evaluation protocols adopted in previous watermarking studies [4,19,23].

3.5 Proposed Clinical Safety Metrics

Three metrics were proposed to characterize the clinical safety of a watermarking scheme. The associated reference thresholds are analytically motivated starting points, not established clinical standards; empirical validation through prospective radiologist observer studies is required to establish evidence-based thresholds.

M1 - Saliency Overlap Score (SOS): Binary activation regions A and B are derived from the Grad-CAM maps of the clean image and the processed image respectively $A = \mathbb{1}[C(x) > 0.5]$ and $B = \mathbb{1}[C(x^*) > 0.5]$ where x^* denotes the watermarked or attacked image. The SOS Dice coefficient and intersection-over-union (IoU) are defined in Equations (6) and (7), respectively:

$$SOS_{Dice} = \frac{2|A \cap B|}{|A| + |B|} \quad (6)$$

$$SOS_{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

A reference threshold of $SOS_{Dice} \geq 0.85$ was used in this study. This value is motivated by the reasoning that a 15% reduction in the spatial overlap of model attention may constitute a clinically meaningful redirection of focus for localized pathologies such as pulmonary nodules or brain tumors. Empirically, this threshold would need to be validated using human observer studies comparing detection performance at different SOS Dice levels.

M2 - Diagnostic Probability Shift (ΔP): Let $\hat{y} = \operatorname{argmax}_y P(y | x)$ be the class predicted on the clean image. The absolute change in the predicted probability for class $\hat{y}$ after watermarking is defined as:

$$\Delta P = |P(\hat{y} | x^*) - P(\hat{y} | x)| \quad (8)$$

The proposed reference threshold of $\Delta P < 0.02$, which is a 2 percentage points change in the model output confidence of the predicted class. This value is within the range of typical clinical AI operating margins at fixed decision thresholds (which can vary by several percentage points); however, it is provided as an analytical baseline, not a clinical standard.

M3 - Region-specific PSNR/SSIM: The ROI is defined by the saliency mask $R = \mathbb{1}[C(x) > 0.4]$. ROI-PSNR is computed over pixels inside R, and BG-PSNR is computed over pixels outside R. The PSNR is defined as:

$$PSNR = 10 \cdot \log_{10} \frac{MAX^2}{MSE} \quad (9)$$

where $MAX = 255$ for 8-bit images, and the MSE is calculated for the specified subset of pixels. If the ROI-PSNR is significantly lower than the BG-PSNR, it means that the distortion is concentrated in the diagnostic region, which represents a clinically concerning pattern. This spatial asymmetry is not measurable by the global PSNR and is considered one of the reasons for the present research.

3.6 Statistical Analysis

The Wilcoxon signed-rank test was performed to compare the per-image SOS Dice values between the proposed XAI-guided scheme and the DWT-SVD baseline per modality ($n = 100$ images/modality). As the SOS Dice distributions are not assumed to be Gaussian, this is especially important considering the bimodalities that are likely to occur when images include highly localized focal pathologies as well as diffuse findings. Therefore, a non-parametric test was chosen. A two-sided alpha of 0.05 was used, and each modality was tested separately. The SciPy library was used to perform all statistical tests.

4. Experimental Setup

4.1 Datasets

Data from three publicly available datasets were used, one for each imaging modality. Table 1 summarizes each dataset, its diagnostic task, class definitions, and data allocation. This was a computational study using three publicly available datasets that were de-identified and did not require ethics committee approval for their use.

Table 1. Experimental Datasets Used for Diagnostic Model Training, Validation, and Watermarking Evaluation

Dataset	Modality	Diagnostic Task	Class 0	Class 1	Train	Validation	Source
IQ-OTH/NCCD	CT	Lung Cancer Detection	Normal	Malignant	480	120	Kaggle
CheXpert-v1.0-small	CXR	Pleural Effusion	No Finding	Effusion	480	120	Kaggle
BraTS 2021	MRI (FLAIR)	Brain Tumour	Healthy Slice	Tumour Slice	480	120	Kaggle

Note: 480 images per modality were used for fine-tuning; 120 for validation, of which 100 were used for watermarking evaluation. Kaggle dataset IDs — CT: adityamahimkar/iqothnccd; CXR: ashery/chexpert; MRI: dschettler8845/brats-2021-task1.

The IQ-OTH/NCCD Lung Cancer CT Dataset contains axial CT slices classified as malignant or normal. Malignant slice features confirmed lung cancer tissue characterized by irregular margins, spiculation, and density heterogeneity [9]. The CheXpert-v1.0-small dataset provides chest radiographs with automated labels for 14 radiological findings. The binary pleural effusion versus no-finding task was selected because pleural effusion produces spatially diffuse Grad-CAM activations distinct from the compact activations observed in CT and MRI, enabling modality-specific evaluation of the proposed metrics [6]. The BraTS 2021 dataset provides multiparametric MRI scans of patients with glioma; 2D axial slices from the FLAIR sequence were used to classify tumor-containing slices against healthy tissue slices based on visible FLAIR signal abnormality [14].

4.2 Implementation Details

All experiments were implemented in Python 3.10 using PyTorch 2.1 with DenseNet-121 initialized from ImageNet pre-trained weights. The computations were performed on a Kaggle Notebook using a T4 GPU (16 GB VRAM). Grad-CAM saliency maps were generated using the Captum v0.7 LayerGradCam class, targeting the features. `denseblock4.denselayer16.conv2` layer. DWT computations used PyWavelets (two-level Haar decomposition); region-specific SSIM used scikit-image v0.21; DCT watermarking used OpenCV-Python for block-wise 2D DCT operations. A fixed 64-bit watermark payload was used in all experiments (NumPy random seed 42). The total number of watermarking evaluations is:

$$N_{total} = 100 \times 4 \times 6 \times 3 = 7,200$$

comprising 100 test images $\times$ 4 schemes $\times$ 6 attack conditions $\times$ 3 modalities. An additional 900 evaluations (3 thresholds $\times$ 100 images $\times$ 3 modalities) were conducted for the threshold ablation study described in Section 5.4. All reported values represent the mean $\pm$ standard deviation of $n = 100$ images per modality.

5. Results and Discussion

5.1 Diagnostic Model Performance

The training curves for the three DenseNet-121 models are presented in Figure 1. The final validation accuracies were 97.5% (CT), 77.5% (CXR), and 90.8% (MRI). The high CT accuracy reflects the marked visual contrast between malignant slices characterized by irregular, high-density lesions and the normal lung parenchyma in the IQ-OTH/NCCD dataset. The CXR accuracy of 77.5% is consistent with published DenseNet-121 results on CheXpert at comparable training data volumes [11], where label noise introduced by the automated labelling procedure and class imbalance in pleural effusion versus no-finding tasks limited performance. The MRI accuracy of 90.8% is attributable to the strong FLAIR signal contrast between the tumor tissue and healthy brain parenchyma in the BraTS 2021 dataset [14]. Qualitative inspection suggested that all three models produced clinically plausible Grad-CAM activation patterns: CT models localized to pulmonary nodule and mass regions; CXR models localized to the pleural space and lower lung zones; and MRI models localized to the tumor core and surrounding perilesional edema. These

patterns support the use of the resulting saliency maps as proxies for diagnostic model attention in subsequent watermarking safety evaluations.

5.2 Clinical Safety Under Clean Watermarking

Table 2 reports the proposed clinical safety metrics for all four watermarking schemes under clean (no attack) conditions. Figure 2 (CT), 3 (CXR), and 4 (MRI) show representative Grad-CAM saliency maps for a clean image and each of the four watermarked versions, illustrating the qualitative activation changes quantified by the numerical metrics in Table 2.

Table 2. Proposed Clinical Safety Metrics — Clean Watermarking (mean $\pm$ std, n = 100 per modality)

Mod.	Scheme	SOS Dice	SOS IoU	ΔP	Safe%	ROI-PSNR (dB)	BG-PSNR (dB)	BER
CT	DWT-SVD	0.865 $\pm$ 0.110	0.775 $\pm$ 0.143	0.007 $\pm$ 0.009	95%	40.6 $\pm$ 1.2	46.2 $\pm$ 3.3	0.354 $\pm$ 0.035
CT	DCT	0.901 $\pm$ 0.113	0.833 $\pm$ 0.132	0.018 $\pm$ 0.025	75%	50.1 $\pm$ 0.2	53.4 $\pm$ 1.7	0.412 $\pm$ 0.102
CT	HiDDeN $\ddagger$	0.774 $\pm$ 0.156	0.652 $\pm$ 0.167	0.046 $\pm$ 0.041	31%	52.4 $\pm$ 0.2	56.4 $\pm$ 1.8	0.484 $\pm$ 0.000
CT	XAI-Guided	0.896$\pm$0.088	0.821$\pm$0.123	0.008$\pm$0.012	92%	100.0$\pm$0.0	46.2$\pm$3.3	0.511$\pm$0.081
CXR	DWT-SVD	0.832 $\pm$ 0.197	0.748 $\pm$ 0.215	0.011 $\pm$ 0.010	81%	40.4 $\pm$ 2.0	42.4 $\pm$ 1.7	0.345 $\pm$ 0.016
CXR	DCT	0.807 $\pm$ 0.211	0.714 $\pm$ 0.219	0.014 $\pm$ 0.012	75%	50.2 $\pm$ 0.2	52.0 $\pm$ 1.2	0.537 $\pm$ 0.061
CXR	HiDDeN $\ddagger$	0.617 $\pm$ 0.235	0.482 $\pm$ 0.218	0.043 $\pm$ 0.037	34%	50.8 $\pm$ 0.3	52.7 $\pm$ 1.3	0.625 $\pm$ 0.000
CXR	XAI-Guided	0.855$\pm$0.168	0.774$\pm$0.187	0.011$\pm$0.009	82%	100.0$\pm$0.0	42.4$\pm$1.7	0.406$\pm$0.076
MRI	DWT-SVD	0.918 $\pm$ 0.110	0.863 $\pm$ 0.146	0.006 $\pm$ 0.008	95%	64.4 $\pm$ 24.7	67.9 $\pm$ 22.5	0.432 $\pm$ 0.085
MRI	DCT	0.485 $\pm$ 0.285	0.366 $\pm$ 0.248	0.053 $\pm$ 0.039	19%	54.9 $\pm$ 1.8	57.0 $\pm$ 1.0	0.562 $\pm$ 0.000
MRI	HiDDeN $\ddagger$	0.809 $\pm$ 0.182	0.712 $\pm$ 0.225	0.015 $\pm$ 0.018	68%	71.2 $\pm$ 20.3	75.2 $\pm$ 18.0	0.516 $\pm$ 0.000
MRI	XAI-Guided	0.935$\pm$0.105	0.892$\pm$0.140	0.004$\pm$0.007	97%	100.0$\pm$0.0	67.9$\pm$22.5	0.519$\pm$0.072

Note: XAI-Guided rows (in bold) denote the proposed scheme (Equations 4–5). ROI-PSNR = 100.0 dB denotes no modification of the ROI pixels by design. $\ddagger$ HiDDeN used at 50-epoch early training state (decoder bit accuracy $\approx$ 0.50); results characterize a distortion-injection baseline, not a converged deep watermarking system. Safe% = percentage of images meeting $\Delta P < 0.02$ (proposed reference criterion).

Figure 2. Grad-CAM saliency maps for a representative CT image — clean (row 1) versus four watermarking conditions: DWT-SVD (row 2), DCT (row 3), HiDDeN $\ddagger$ (row 4), and XAI-guided (row 5). The columns show the input image, Grad-CAM overlay, and binary activation mask (binarization threshold = 0.5). XAI-Guided produced near-identical activation regions to the clean image (mean SOS Dice = 0.896 $\pm$ 0.088), whereas HiDDeN $\ddagger$ showed the greatest activation shift (SOS Dice = 0.774 $\pm$ 0.156).

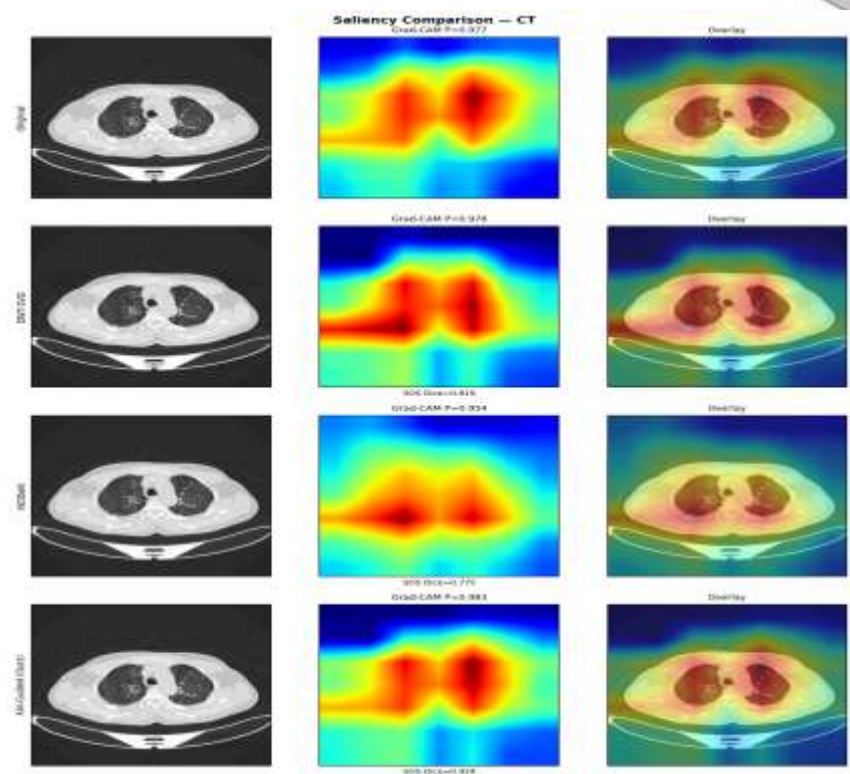


Figure 3. Grad-CAM saliency maps for a representative CXR image — clean versus four watermarking conditions (same layout as in Figure 2). Pleural effusion activations are inherently diffuse, which reduces the sensitivity of the SOS Dice to localized perturbations. Both DWT-SVD (SOS Dice = 0.832 ± 0.197) and XAI-Guided (SOS Dice = 0.855 ± 0.168) preserved activations well for this modality.

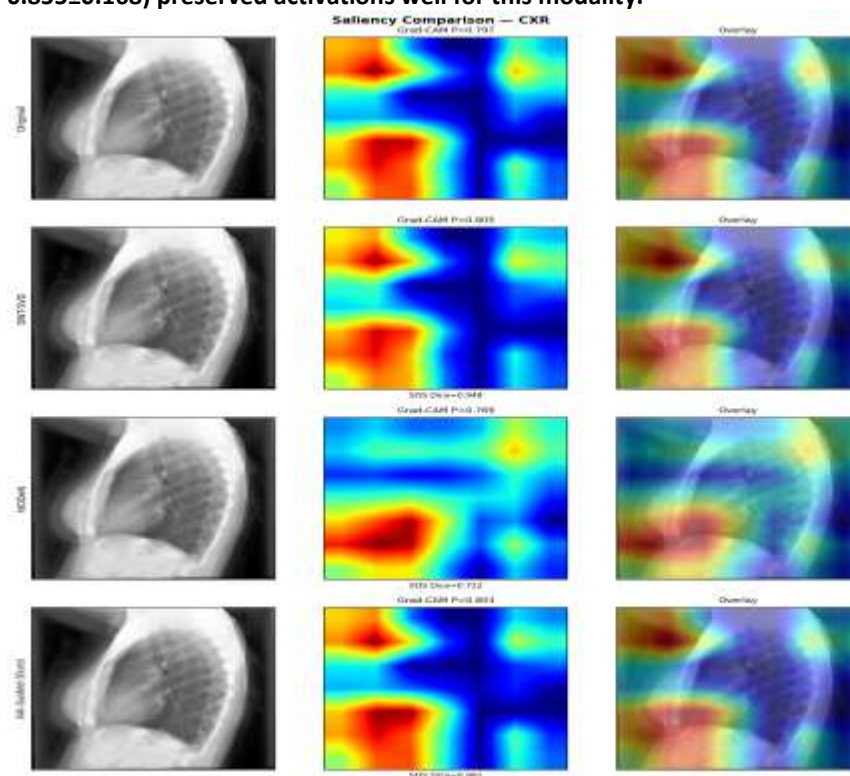
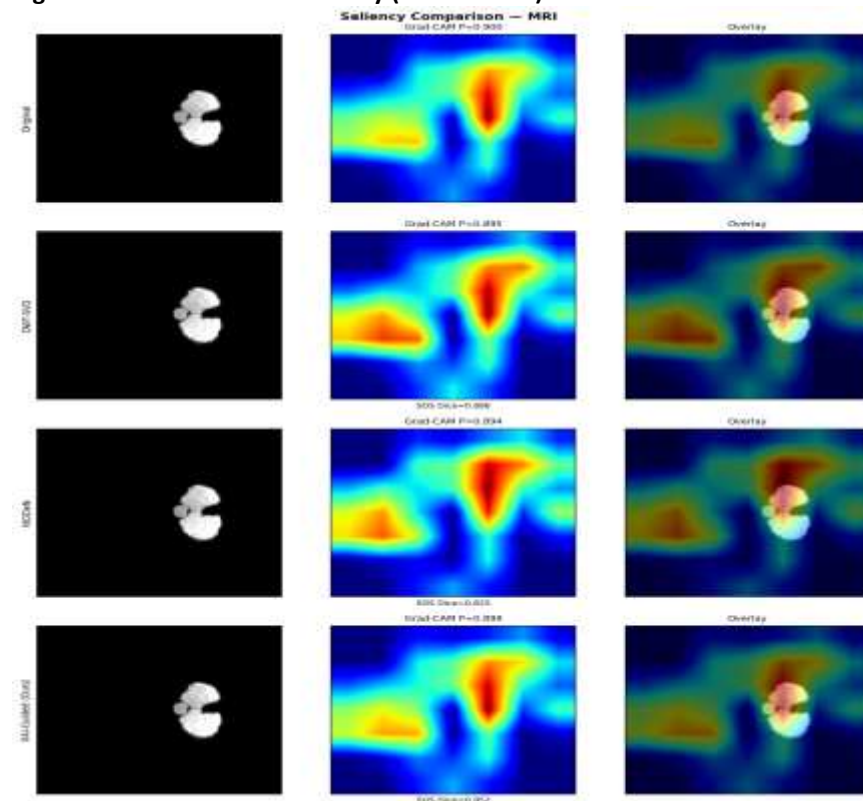


Figure 4. Grad-CAM saliency maps for a representative MRI FLAIR image — clean versus four watermarking conditions (same layout as in Figure 2). DCT produced the most pronounced activation disruption (SOS Dice = 0.485 ± 0.285) owing to block-DCT artifacts coinciding with FLAIR tumor boundaries. XAI-Guided achieved the highest SOS Dice for this modality (0.935 ± 0.105).

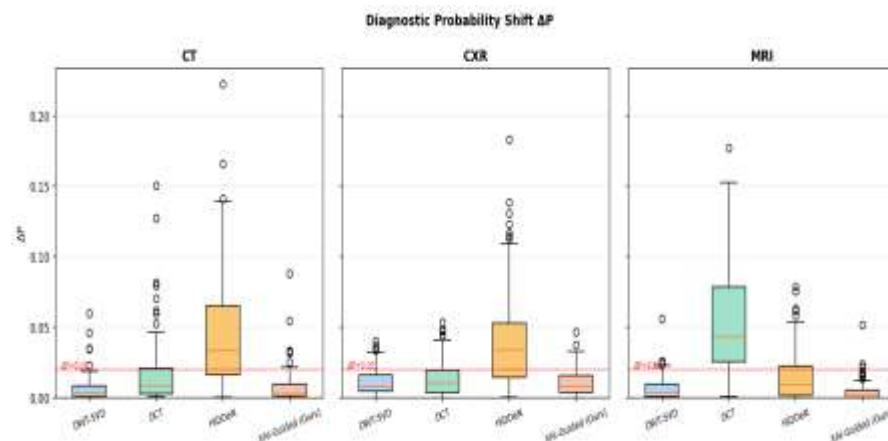


As shown in Table 2, the XAI-guided scheme achieved the highest SOS Dice on all three modalities (CT: 0.896 ± 0.088 , CXR: 0.855 ± 0.168 , MRI: 0.935 ± 0.105) and the lowest ΔP (CT: 0.008 ± 0.012 , CXR: 0.011 ± 0.009 , MRI: 0.004 ± 0.007). The relatively narrow standard deviations for CT and MRI reflect the design invariant that all high-saliency pixels are excluded from embedding, regardless of the image content, making the results consistent across diverse pathological appearances. DWT-SVD performed well with respect to SOS Dice on both CT (0.865 ± 0.110) and MRI (0.918 ± 0.110), and it met the proposed ΔP reference criterion in 95%, 81%, and 95% of images on CT, CXR, and MRI, respectively, confirming that classical wavelet-domain embedding at $\alpha = 0.06$ is broadly compatible with diagnostic model attention for these modalities. DCT achieved the second-highest SOS Dice on CT (0.901 ± 0.113) but degraded severely on MRI (0.485 ± 0.285), where block-structured artifacts coincided with FLAIR tumor boundaries, a modality-specific hazard that the global PSNR (54.9 dB) did not indicate.

Figure 5 shows the per-image ΔP distributions across all four schemes aggregated over the three modalities. Despite achieving ROI-PSNR values of 52.4 dB (CT), 50.8 dB (CXR), and 71.2 dB (MRI) values that comfortably exceed the commonly cited 40 dB acceptability threshold [4], the HiDDen $\ddagger$ distortion-injection baseline met the proposed $\Delta P < 0.02$ reference criterion in only 31%, 34%, and 68% of images on CT, CXR, and MRI, respectively. This finding suggests that a high global PSNR does not reliably indicate clinical safety for learned-perturbation watermarking; a scheme can achieve a PSNR above 50 dB while still substantially shifting the diagnostic classifier's output probability for a significant fraction of images. This observation is consistent with concerns raised in the recent watermarking literature [4,8] regarding the adequacy of signal metrics as safety proxies, and is a central contribution of the present evaluation framework.

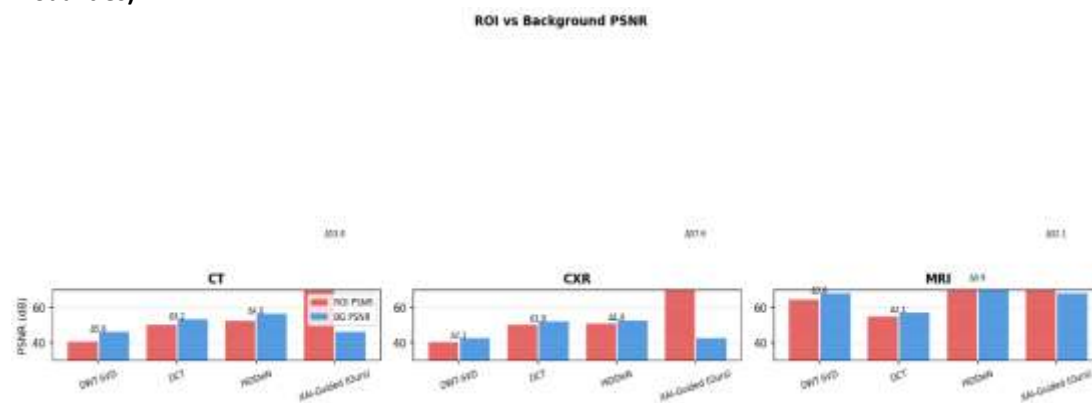
Figure 5. Diagnostic probability shift (ΔP) distributions per watermarking scheme, aggregated across all three modalities ($n = 300$ images per scheme). The proposed reference criterion ($\Delta P = 0.02$) is shown as a horizontal

dashed line. XAI-guided and DWT-SVD remained predominantly below this criterion. The HiDDeN[‡] early training baseline showed the widest distribution and highest proportion of exceedances across all three modalities.



The region-specific PSNR analysis illustrated in Figure 6 reveals further limitations of the global metrics. For MRI DWT-SVD, ROI-PSNR = 64.4±24.7 dB compared with BG-PSNR = 67.9±22.5 dB, a 3.5 dB gap indicating that the wavelet-domain perturbation concentrates slightly more distortion within the tumour region than in the background. This spatial asymmetry is invisible to the global PSNR and would not be detected by any currently published evaluation protocol. The XAI-Guided scheme shows ROI-PSNR = 100.0 dB by design and BG-PSNR = 67.9±22.5 dB for MRI, because the non-protected pixels receive the same DWT-SVD transformation as in S1, which is consistent with the mask-blending operation defined in Equation (5).

Figure 6. ROI-PSNR versus BG-PSNR scatter plot for all four watermarking schemes across the three modalities. Points on the diagonal indicate spatially uniform distortion, and points below the diagonal indicate disproportionate distortion within the diagnostic region. The MRI DWT-SVD falls 3.5 dB below the diagonal (ROI = 64.4 dB, BG = 67.9 dB). XAI-Guided is plotted at ROI-PSNR = 100.0 dB by design (off the main axis for all three modalities).



5.3 Robustness Under Attack Conditions

Table 3 presents the SOS Dice, ΔP , and BER across all four schemes and six attack conditions, aggregated over the three modalities. The attack-condition trends are illustrated graphically in Figure 7, which plots the mean SOS Dice with ± 1 standard deviation bands.

Table 3. Proposed Clinical Safety Metrics Under Six Attack Conditions (mean $\pm$ std, aggregated across CT, CXR, and MRI modalities; $n = 300$ per scheme-condition pair)

Scheme	Attack	SOS Dice	ΔP	BER
--------	--------	----------	------------	-----

	Condition			
DWT-SVD	Clean	0.872±0.149	0.008±0.009	0.377±0.066
DWT-SVD	JPEG Q50	0.540±0.272	0.117±0.114	0.405±0.087
DWT-SVD	JPEG Q30	0.466±0.282	0.141±0.131	0.408±0.089
DWT-SVD	Gaussian Noise	0.316±0.257	0.161±0.095	0.397±0.075
DWT-SVD	Gaussian Blur	0.314±0.288	0.137±0.117	0.499±0.059
DWT-SVD	Scaling 0.5×	0.362±0.281	0.145±0.135	0.494±0.062
DCT	Clean	0.731±0.279	0.028±0.033	0.504±0.095
DCT	JPEG Q50	0.531±0.282	0.123±0.120	0.537±0.047
DCT	JPEG Q30	0.450±0.285	0.143±0.128	0.546±0.043
DCT	Gaussian Noise	0.326±0.265	0.162±0.097	0.508±0.063
DCT	Gaussian Blur	0.309±0.288	0.130±0.109	0.604±0.062
DCT	Scaling 0.5×	0.360±0.281	0.139±0.127	0.518±0.059
HiDDeN‡	Clean	0.733±0.211	0.035±0.036	0.542±0.060
HiDDeN‡	JPEG Q50	0.528±0.282	0.122±0.119	0.542±0.060
HiDDeN‡	JPEG Q30	0.457±0.283	0.144±0.132	0.542±0.060
HiDDeN‡	Gaussian Noise	0.338±0.270	0.160±0.099	0.542±0.060
HiDDeN‡	Gaussian Blur	0.303±0.281	0.146±0.124	0.542±0.060
HiDDeN‡	Scaling 0.5×	0.363±0.278	0.152±0.141	0.542±0.060
XAI-Guided	Clean	0.896±0.129	0.008±0.010	0.479±0.092
XAI-Guided	JPEG Q50	0.546±0.268	0.122±0.117	0.491±0.089
XAI-Guided	JPEG Q30	0.460±0.277	0.142±0.131	0.493±0.088
XAI-Guided	Gaussian Noise	0.333±0.261	0.157±0.095	0.478±0.087
XAI-Guided	Gaussian Blur	0.310±0.287	0.136±0.116	0.538±0.045
XAI-Guided	Scaling 0.5×	0.362±0.278	0.144±0.133	0.537±0.046

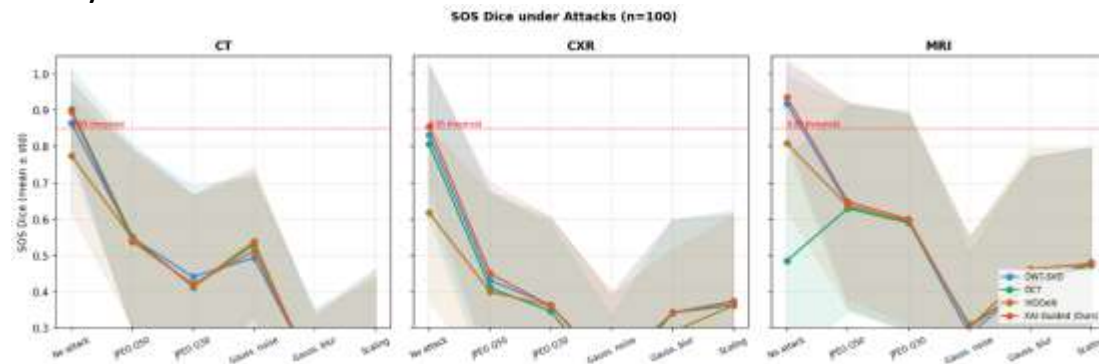
Note: Proposed XAI-guided scheme. ‡ HiDDeN early training distortion injection baseline. HiDDeN‡ BER = 0.542 across all conditions owing to the random-chance decoder output at the training state used.

Under clean conditions, the XAI-guided scheme obtained the highest SOS Dice (0.896±0.129), which is similar to the modality-level result shown in Table 2. The advantage of the new methods over the classical baselines is most significant when the signals are clean and the attacks are mild, such as JPEG Q50 and Gaussian noise, which are the most common signal processing operations in the regular implementation of PACS. With JPEG Q50, XAI-Guided outperforms DWT-SVD (0.540), DCT (0.531), and HiDDeN‡ (0.528) with a consistent but small margin of 0.546.

Under severe attack conditions, JPEG Q30 and geometric scaling all four schemes converged toward similar low SOS Dice values (0.30–0.47), as is evident from Table 3 and Figure 7. At this level of signal degradation, the dominant driver of the saliency shift is the attack itself rather than the watermark; global image statistics change substantially, causing Grad-CAM activations to shift even on unmodified images. Under these conditions, no watermarking scheme evaluated here fully preserves model saliency, which represents an open challenge for the field. The XAI-Guided design advantage is therefore most meaningful in the clean-to-mild-attack regime that characterises routine PACS operation.

The HiDDeN $\dagger$ baseline shows a constant BER of 0.542 ± 0.060 across all six attack conditions, confirming that the output distribution of the decoder is independent of the received image and that the random-chance bit accuracy persists regardless of the applied attack. This invariance validates the characterization of this implementation as a distortion-injection baseline: the per-attack ΔP values in Table 3 characterize the clinical impact of the encoder's fixed spatial perturbation pattern under each attack, not the behavior of a functional-watermarking system.

Figure 7. Mean SOS Dice (± 1 standard deviation, shaded) under six attack conditions for all four watermarking schemes, aggregated across the three modalities. The XAI-Guided advantage over the baselines is most pronounced under clean and JPEG Q50 conditions. Under JPEG Q30, Gaussian noise, Gaussian blur, and scaling, all schemes converge, indicating that the attack, rather than the watermark, becomes the dominant source of the saliency shift.



5.4 Threshold Ablation Study

Table 4 presents the ablation results for the XAI-guided scheme at $\tau \in \{0.30, 0.40, 0.50\}$ across all three modalities. Figure 8 illustrates the resulting trade-offs among the four key metrics.

Table 4. Threshold Ablation Study — XAI-Guided Scheme at $\tau \in \{0.30, 0.40, 0.50\}$ (mean $\pm$ std, $n = 100$ per modality)

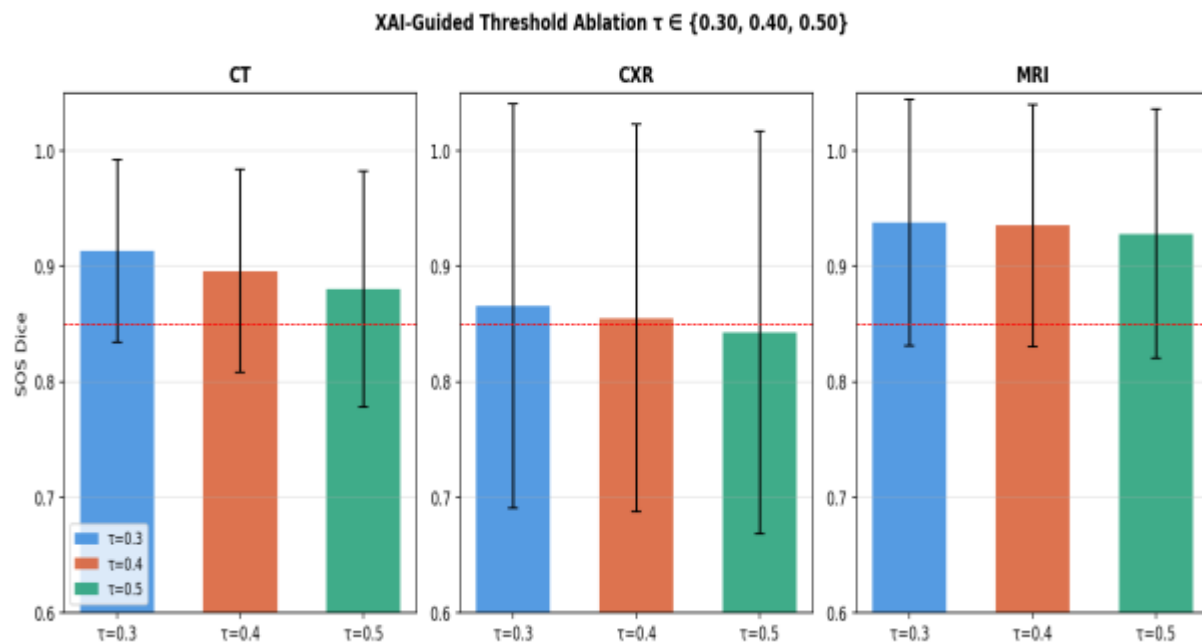
Mod.	τ	Protected%	SOS Dice	ΔP	ROI-PSNR (dB)	BER
CT	0.30	30%	0.914 $\pm$ 0.079	0.007 $\pm$ 0.012	100.0 $\pm$ 0.0	0.539 $\pm$ 0.059
CT	0.40	40%	0.896$\pm$0.088	0.008$\pm$0.012	100.0$\pm$0.0	0.511$\pm$0.081
CT	0.50	50%	0.881 $\pm$ 0.102	0.007 $\pm$ 0.011	47.1 $\pm$ 1.9	0.455 $\pm$ 0.087
CXR	0.30	30%	0.866 $\pm$ 0.175	0.011 $\pm$ 0.008	100.0 $\pm$ 0.0	0.444 $\pm$ 0.087
CXR	0.40	40%	0.855$\pm$0.168	0.011$\pm$0.009	100.0$\pm$0.0	0.406$\pm$0.076
CXR	0.50	50%	0.843 $\pm$ 0.174	0.010 $\pm$ 0.010	46.0 $\pm$ 2.8	0.370 $\pm$ 0.053
MRI	0.30	30%	0.938 $\pm$ 0.106	0.005 $\pm$ 0.008	100.0 $\pm$ 0.0	0.523 $\pm$ 0.071
MRI	0.40	40%	0.935$\pm$0.105	0.004$\pm$0.007	100.0$\pm$0.0	0.519$\pm$0.072
MRI	0.50	50%	0.928 $\pm$ 0.108	0.005 $\pm$ 0.007	72.8 $\pm$ 23.4	0.507 $\pm$ 0.076

Note: Rows in bold denote the selected threshold of $\tau = 0.40$. At $\tau = 0.50$, the ROI-PSNR degrades below 100.0 dB for CT (47.1 dB) and CXR (46.0 dB) because pixels with saliency in the range $[0.40, 0.50]$ are erroneously included in the embedding region. $\tau = 0.40$ provided the best balance of SOS Dice, ROI protection, and BER across all three modalities.

As shown in Table 4, reducing τ to 0.30 maximizes the SOS Dice (CT: 0.914 $\pm$ 0.079, CXR: 0.866 $\pm$ 0.175, MRI: 0.938 $\pm$ 0.106) by protecting a larger fraction of pixels, but this comes at the cost of reduced embedding capacity, as

reflected in the highest BER values at $\tau = 0.30$ (CT: 0.539 ± 0.059). Increasing τ to 0.50 allows pixels with saliency in the range $[0.40, 0.50]$ to receive the watermark, breaking the ROI-PSNR guarantee on CT (47.1 ± 1.9 dB) and CXR (46.0 ± 2.8 dB). MRI is comparatively less affected at $\tau = 0.50$ (ROI-PSNR = 72.8 ± 23.4 dB) because FLAIR tumor activations tend to concentrate at higher peak-saliency values. At $\tau = 0.40$, ROI-PSNR = 100.0 dB was maintained for all three modalities, SOS Dice satisfied the proposed reference criterion (CT: 0.896, CXR: 0.855, MRI: 0.935), and BER was improved relative to $\tau = 0.30$. Across the three evaluation axes, ROI protection, saliency preservation, and embedding capacity, $\tau = 0.40$ represents the most balanced choice for the experimental conditions examined.

Figure 8. Threshold ablation results for the XAI-Guided scheme at $\tau \in \{0.30, 0.40, 0.50\}$ across CT, CXR, and MRI modalities. The selected threshold $\tau = 0.40$ (dashed vertical marker) maintains ROI-PSNR = 100.0 dB for all three modalities while achieving competitive SOS Dice and BER values. At $\tau = 0.50$, the ROI-PSNR guarantee breaks down for CT and CXR.



5.5 Statistical Significance

Wilcoxon signed-rank tests comparing per-image SOS Dice between XAI-Guided and DWT-SVD ($n = 100$ per modality, $\alpha = 0.05$) yielded the following results: For CT, the test statistic W was 1003.0, and p was < 0.0001 (statistically significant). For MRI: $W = 392.0$, $p < 0.0001$ (statistically significant). For CXR: $W = 2158.5$, $p = 0.269$ (not statistically significant). The CXR results indicate that XAI-guided and DWT-SVD perform comparably on this modality in terms of saliency overlap, which is consistent with the diffuse nature of pleural effusion activations, making CXR saliency maps less sensitive to localized wavelet-domain perturbations. Despite the non-significant SOS Dice difference, XAI-Guided maintains strict superiority over DWT-SVD on CXR in terms of ROI-PSNR (100.0 dB vs. 40.4 ± 2.0 dB). The statistically significant improvements on CT and MRI are consistent with the hypothesis that compact, high-intensity focal-pathology pulmonary nodules and brain tumors are more sensitive to wavelet-domain perturbations within their saliency-defined regions than diffuse pleural effusion findings on CXR.

5.6 Discussion

The present findings highlight the distinction between conventional watermarking imperceptibility and clinically meaningful safety assessments. Global metrics such as PSNR and SSIM remain useful because they provide established summaries of overall signal preservation and facilitate comparisons with prior watermarking literature; however, they cannot reveal whether the observed distortion is concentrated within diagnostically critical regions [4]. The current results highlight this limitation in the baseline HiDDeN early training result, where the ROI-PSNR values are clustered in the 52-71 dB range, with significant shifts in probability and low Safe% values for both CT and

CXR.

This pattern has several important implications. First, although it may seem that the degradation is global and high-fidelity, it can be locally variable for clinical relevance [4,8]. Second, a watermarking scheme should not only be evaluated by image similarity or a high PSNR (average) but also by the diagnostic model attention and confidence about the parts that are significant to image interpretation. The framework proposed here is different from all of the classical robustness and fidelity measures; it is used in conjunction with those measures to provide a clinically anchored point of view on the proposed framework. The area-specific analysis confirms this. For instance, ROI-PSNR = 64.4 ± 24.7 dB compared to BG-PSNR = 67.9 ± 22.5 dB, where the area between them is 3.5 dB, which is not identified by global PSNR. Such spatial asymmetry is a type of behavior that could be relevant in medical imaging, where errors in the vicinity of the lesion can be more significant than errors in more remote regions, irrelevant to the anatomical context. All the above observations suggest a more conservative and clinically realistic assessment of watermark quality. While signal metrics are still needed for compatibility with previous studies and for measuring overall imperceptibility and robustness, they are insufficient for clinical-safety-oriented evaluation when diagnostic AI systems or saliency-sensitive reading tasks are concerned [4]. Saliency-aware evaluation provides an engineering-to-clinically meaningful model behavior analysis link. The present evaluation is model-based, not observer-based, and is best viewed as a scalable and automated initial first-stage screening mechanism that complements and is designed to precede, not replace, prospective radiologist observer evaluations.

5.7 Practical Implications for Clinical Deployment

The reported results are particularly relevant to the modern clinical imaging environment where images can be stored in PACS, compressed in a compression pipeline, transferred via teleradiology platform, archived in hospital archival platform and then viewed by a clinician or downstream AI application [6]. In such workflows, watermarking is attractive for authentication, provenance, and tamper detection, but the results of this manuscript suggest that a watermarking method should ideally be evaluated not only for robustness and imperceptibility, but also for whether it perturbs diagnostically important regions or destabilises model predictions.

For hospital image archives and authentication workflows, the proposed framework provides a way to examine whether embedding strategies preserve diagnostic content in saliency-defined regions while maintaining watermark functionality. For AI-assisted diagnosis systems, the use of ΔP and saliency-overlap metrics offers a practical mechanism for screening whether a watermarking pipeline changes the classifier behavior in ways that may not be visible from the PSNR or SSIM alone.

In the PACS and teleradiology contexts, the finding that the XAI-Guided advantage is most meaningful under clean and mild-attack conditions is operationally important because these conditions are often closer to routine deployment than severe synthetic degradations. Simultaneously, the convergence of all schemes under stronger attacks indicates that clinically robust watermarking for harsher post-processing environments remains an open challenge, which future deployment-oriented validation should address before widespread adoption.

5.8 Limitations

Limitation 1—HiDDeN Training State: HiDDeN was trained for 50 epochs and had a decoder bit accuracy of approximately 0.50, which was reported as a distortion-injection baseline and not as a converged model. Different SOS Dice, ΔP , and ROI-PSNR values may be obtained for a fully trained HiDDeN.

Limitation 2 - Proposed Threshold Calibration: The reference thresholds (SOS Dice ≥ 0.85 and $\Delta P < 0.02$) are analytically motivated but not validated with prospective studies on human observers. Experiments by radiologists are required to provide empirical thresholds for the diagnostic task performance on clean images vs images marked with watermarks at different levels of the watermark. In this study, the thresholds are analytically motivated and do not serve as clinical cut-offs; rather, they are taken to be transparent, reproducible reference thresholds, which are the logical next step in validation, following the natural studies identified here.

Limitation 3: Two-dimensional Slice Evaluation: All experiments were conducted on 2D axial slices taken from 3D volumetric data. This is a proof-of-concept for slices. However, for realistic clinical deployment scenarios, where interslice consistency and volumetric embedding capacity are important, an extension to full volumetric watermarking and 3D Grad-CAM evaluation is necessary.

Limitation 4 - Public Benchmark Datasets: All three datasets are publicly available benchmark collections with binary diagnostic labels. The generalization of the proposed metrics and scheme to heterogeneous multi-class tasks, diverse scanner types, imaging protocols, and real-world PACS environments remains to be demonstrated.

Limitation 5 - CT Task Granularity: The CT classification task (malignant versus normal) was relatively coarse-grained. Evaluation of nodule detection, sizing, or multi-class staging tasks would provide a more clinically demanding test of CT safety claims.

Limitation 6 - Model Dependency: The proposed framework was evaluated using DenseNet-121 as the diagnostic backbone. Although DenseNet-121 is widely used in medical imaging research and provides reliable Grad-CAM visualizations, the behavior of the proposed saliency preservation metrics may vary across alternative architectures. Future studies should investigate the consistency of the SOS Dice, ΔP , and region-specific PSNR/SSIM using other deep learning models, including EfficientNet, Vision Transformers, and foundation-model-based medical imaging systems.

5.9 Threats to Validity

The first threat to validity arises from the dependence on Grad-CAM quality. Although Grad-CAM is justified here through prior medical-imaging literature and the resulting activation maps appear clinically plausible, any saliency-based evaluation is only as reliable as the attribution method used to define diagnostically important regions [1]. If the saliency map under-localizes or over-localizes relevant anatomy, the resulting SOS and region-specific distortion metrics may misrepresent the true clinical importance of the perturbation. Accordingly, the reliability of Grad-CAM as a localization proxy is claimed only for the specific datasets and models examined here, each trained on 480 images per modality, and is not asserted as a general property across modalities, architectures or clinical populations.

The second threat is the model-specific saliency behavior. Saliency patterns can vary by alternative architecture, training regime, or calibration features of the framework and can be different if DenseNet-121 classifiers are replaced with other architectures. This implies that some of the abovementioned safety relationships are partly model-dependent and not necessarily architecture-invariant.

The third threat relates to the dataset selection. The study adopts relatively standard task sets and binary classification problems from public benchmark datasets, thereby improving the level of reproducibility and limiting the potential for generalization to more heterogeneous clinical populations, scanners, and multi-center workflows. The CT task is also relatively coarse and may not exert as much pressure on the watermarking framework as one of the more difficult lesion detection or staging tasks.

The fourth is threshold sensitivity. The thresholds for the SOS Dice and ΔP are analytically suggested reference values, and the embedding mask XAI also relies on the saliency threshold τ . While the results of the ablation study suggest $\tau = 0.40$ for the current experimental conditions, the choice of thresholds may vary from task to task, modality to modality, and model to model and should, therefore, not be considered as yet as universal clinical thresholds.

The fifth threat is the generalizability across modality and architecture. The present work is more than a one-modality proof of concept, as it is broader than that, including CT, CXR, and MRI; however, it remains within the realm of 2D slices and one primary diagnostic family of the backbone. Expansion to other modalities, complete volumetric pipelines, and transformer-based topologies must be performed for it to be generally deployed.

5.10 Future Work

Several future directions follow naturally from the current results and limitations. The extension from 2D slice-level experiments to 3D volumetric CT and MRI analysis is the first priority, in which watermarking decisions and saliency preservation can be evaluated not only based on the slice-level continuity but also on the volumetric context. The second approach is to extend DenseNet-121 to newer diagnostic backbones, such as the Vision Transformer models,

to determine whether the current relationships between saliency and safety hold for alternative representation learning methods. The third direction is the methodological extension of Grad-CAM to other methods of attribution, such as SHAP and Integrated Gradients, which may yield similar attribution behaviour and can be used to assess the stability of the proposed framework across different attribution methods.

The fourth dimension is the evaluation of multicenter and more heterogeneous clinical datasets, such as a wider class structure, protocol variety, and scanner variability. The fifth direction is prospective radiologist observer studies to empirically validate clinically meaningful thresholds for SOS Dice and ΔP and to test whether saliency-aware watermark evaluation correlates with human diagnostic performance.

The final direction concerns regulatory and deployment considerations. If watermarking is to be used in AI-integrated imaging workflows, future research must consider validation frameworks that are acceptable for hospital governance, clinical quality assurance, and eventual regulatory scrutiny, especially where watermarking interacts with diagnostic decision support.

6. Conclusion

This study proposes an XAI-driven evaluation framework for the clinical safety assessment of medical image watermarking, motivated by the fundamental limitation that the global PSNR and SSIM do not indicate whether the watermark-induced distortion falls within diagnostically important image regions. Three proposed clinical safety metrics — Saliency Overlap Score, Diagnostic Probability Shift, and Region-specific PSNR — were applied across 7,200 evaluations on CT, CXR, and MRI datasets, using Grad-CAM saliency maps from fine-tuned DenseNet-121 diagnostic models as the evaluation instrument. The central and most transferable contribution of this study is the evaluation framework and the saliency-based clinical safety metrics it defines. The proposed watermarking scheme primarily serves as a demonstration of how these metrics can guide the embedding design.

The proposed XAI-guided watermarking scheme, which confines embedding to low-saliency background pixels, achieved superior SOS Dice and lower ΔP compared with classical transform-domain baselines across all three modalities. Improvements over DWT-SVD were statistically significant on CT ($W = 1003.0$, $p < 0.0001$) and MRI ($W = 392.0$, $p < 0.0001$), with the performance advantage concentrated under clean and mild-attack conditions representative of routine PACS operation. Under severe attacks, all current schemes converge to similar SOS Dice values, indicating that the attack, rather than the watermarking scheme, becomes the dominant source of saliency change under such conditions.

The results for the early training HiDDeN distortion-injection baseline indicate that a high global PSNR does not reliably predict a low ΔP , suggesting that a signal-metric-only evaluation may be insufficient for characterizing the clinical impact of learned-perturbation watermarking. These findings demonstrate that saliency-aware evaluation provides clinically relevant information beyond conventional signal-fidelity metrics and may contribute to the development of more clinically grounded evaluation standards for medical image watermarking in AI-integrated imaging workflows [4,8]. Directions for future research include prospective radiologist observer studies for empirical threshold validation, extension to 3D volumetric watermarking and Grad-CAM evaluation, evaluation of fully converged deep watermarking models [23,2], and testing on heterogeneous multi-class clinical tasks and diverse imaging sites.

Acknowledgements

The authors acknowledge the use of three publicly available de-identified datasets: the IQ-OTH/NCCD Lung Cancer CT Dataset, CheXpert-v1.0-small, and BraTS 2021. Computational experiments were performed using the GPU resources provided by the Kaggle Notebooks. The authors also acknowledge the developers of Captum, PyWavelets, and scikit-image open-source libraries used in this study.

Credit Author Contributions Statement

Chukka Harika: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, writing – Original Draft, writing – Review and Editing.

Ethics Statement

The datasets used in this study were publicly available, de-identified datasets (IQ-OTH/NCCD, CheXpert-v1.0-small, and BraTS 2021). No human subjects were recruited, nor was patient information other than what is publicly available. Institutional ethics committee approval was not required.

AI Assistance Disclosure

The manuscript was prepared using AI-assisted writing tools for grammar checking and language refinement. All scientific content, experimental design, analysis, results, and conclusions are the sole responsibility of the author(s).

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

1. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Hoebel K, Gupta S, Patel J, Gidwani M, Adebayo J, Li MD, Kalpathy-Cramer J. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3(6):e200267. doi:10.1148/ryai.2021200267
2. Bistrón M, Žurada JM, Piotrowski Z. Deep learning for image watermarking: a comprehensive review and analysis of techniques, challenges, and applications. *Sensors*. 2026;26(2):444. doi:10.3390/s26020444
3. Dong F, Li J, Bhatti UA, Liu J, Chen YW, Li D. Robust zero watermarking algorithm for medical images based on improved NasNet-Mobile and DCT. *Electronics*. 2023;12(16):3444. doi:10.3390/electronics12163444
4. Gull S, Parah SA. Advances in medical image watermarking: a state of the art review. *Multimed Tools Appl*. 2023. doi:10.1007/s11042-023-15396-9
5. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*. 2017:4700–4708. doi:10.1109/CVPR.2017.243
6. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–510. doi:10.1038/s41568-018-0016-5
7. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *arXiv preprint*. 2022. arXiv:2202.10553.
8. Khaldi A, Zermi N, Moad MS, Boukhamla A, Kafi MR, Sahu AK. Explainable and proactive fragile watermarking for medical deepfake detection. *Intell Syst Appl*. 2026;31:200684. doi:10.1016/j.iswa.2026.200684
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. doi:10.1016/j.media.2017.07.005
10. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
11. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Ball RL, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint*. 2017. arXiv:1711.05225.
12. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv preprint*. 2017. arXiv:1708.08296.
13. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proc IEEE Int Conf Comput Vis (ICCV)*. 2017:618–626. doi:10.1109/ICCV.2017.74
14. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng*. 2017;19:221–248. doi:10.1146/annurev-bioeng-071516-044442
15. Singh P, Devi KJ, Thakkar HK, Santamaría J. Blind and secured adaptive digital image watermarking approach for high imperceptibility and robustness. *Entropy*. 2021;23(12):1650. doi:10.3390/e23121650
16. Su GD, Lin CC, Chang CC. Privacy-preserving reversible data hiding for medical images employing local rotation. *J Healthc Eng*. 2021;2021:5709513. doi:10.1155/2021/5709513
17. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proc Int Conf Mach Learn (ICML)*. 2017;70:3319–3328.

18. Taj R, Tao F, Kanwal S, Almogren A, Altameem A, Rehman AU. A reversible-zero watermarking scheme for medical images. *Sci Rep.* 2024;14:16891. doi:10.1038/s41598-024-67672-9
19. Taj R, Kanwal S, Alsirhani A, Alserhani F, Alanazi R, Ali A. Lightweight dual-watermarking framework for medical image authentication and integrity preservation. *Sci Rep.* 2025;15:10583. doi:10.1038/s41598-025-30615-z
20. Thanki R, Borra S, Dwivedi V, Borisagar K. A RONI based visible watermarking approach for medical image authentication. *J Med Syst.* 2017;41(9):143. doi:10.1007/s10916-017-0795-3
21. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* 2022;79:102470. doi:10.1016/j.media.2022.102470
22. Zhong X, Das A, Alrasheedi F, Tanvir A. A brief yet in-depth survey of deep learning-based image watermarking. *arXiv preprint.* 2023. arXiv:2308.04603.
23. Zhu J, Kaplan R, Johnson J, Fei-Fei L. HiDDeN: hiding data with deep networks. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018. Lecture Notes in Computer Science*, vol 11219. Springer; 2018:682–697. doi:10.1007/978-3-030-01267-0_40