Research Article

Design of an Iterative Explainable AI-Guided Watermarking for Transparent and Clinically Safe Medical Image Security

Vamsi Sisti[1]*, M Kali Muthu[2]

[1]Student, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503. 2401050181cse@gmail.com
[2]Associate Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503. mkmuthu@kluniversity.in
*Corresponding Author: 2401050181cse@gmail.com

Abstract

The authentication of medical images in a system that spans hospitals, teleradiology, and AI diagnostic pipelines must be secure, auditable, and safe. Existing watermarking methods attach ownership signatures that do not consider the radiological salience of parts of the image; embedding location decisions are neither transparent nor is their diagnostic safety considered. In this study, we propose and outline a 5-stage explainable watermarking pipeline for medical images, which, to our knowledge, has not been defined in the literature in this composite form. It uses the Saliency Synthesis Watermark Prior (SSWP) to combine Grad-CAM, Guided Backpropagation, DeepSHAP, and clinician-relevant maps to find radiologically safe regions into which to embed the watermarked signal. It uses an Interpretable Wavelet-Attention Watermark Encoder (IWA-WE), which embeds the watermarked signal into diagnostically neutral wavelet coefficients. The Counterfactual Embedding Validation Network (CEV-Net) quantifies the prediction changes due to watermarking. The Transparent Latent Robustness Optimizer (TLRO) fine-tunes the latent representation, subject to interpretability constraints. The Explainable Security Integrity Fusion Benchmark (XSIF-Bench) measures various attributes in combination. Using tests on NIH ChestX-ray14, BraTS, Messidor-2, and another fundus set, we show that the method produces results with a PSNR between 43.7-47.1 dB and SSIM values of 0.982-0.993, and watermark recovery from four attacks at levels of 96.9-98.5%. These results support a diagnostic-safe, auditable medical image authentication protocol.

Keywords: Watermarking; Explainable AI; Medical Imaging; Interpretability; Image Security; Counterfactual Validation

1. Introduction

The applications of digital medical imaging in hospital environments, research institutions, and teleradiology applications form the basis of the clinical workflow [1], with clinical image data often required to be distributed, stored, and read by various parties involved in telemedicine and AI diagnostic systems [2, 3]. The movement of image data on such a scale poses a significant threat of illegal redistribution, impersonation, and lack of lineage. Digital watermarking allows authenticable content and authentication to be embedded directly within image data, making it the foremost technique available for preserving medical image integrity [4, 5]. In a clinical context, the slightest change to any pixel data must not result in the degradation of any diagnostic information, as minor changes to textures and contrast may alter the performance of automated nodule detection systems, tumor boundary segmentation models, or image classifiers that identify microaneurysms within the retinal layer. The regulations applicable to telemedicine and digital pathology also mandate that any authentication process must be accessible and auditable.

Traditional frequency and hybrid transform methods are tuned to optimize the imperceptibility and robustness of different LWT schemes [1], DCT-DWT hybrid approaches [6, 7], DWT-SVD schemes [9], and integer wavelet methods [14, 15]. None of these studies provided a spatial explanation of the location of the watermark energy relative to the radiologically important anatomy. ResNet-based zero-watermarking [4], bio-inspired optimization [5], and ANFIS-guided inference [8] are examples of deep learning methods that enhance embedding adaptability. Additionally, end-to-end learned embeddings [12] and low-power neural network designs [16, 17] have been adapted to edge environments. Nevertheless, none of these studies provide a mechanism to ensure that diagnostic predictions remain unaffected after embedding [18]. Class Activation Mapping [30], Grad-CAM [21], SHAP [22], LIME [28], and deconvolutional network visualization [27] have made progress in deep learning explainability; however, they were not developed to guide watermarking and ensure diagnostic safety. Counterfactual reasoning has been explored in medical AI to measure and fix prediction changes [24, 25] and can serve as an embedding validation step; however,

no watermark system has yet incorporated this application.

These issues are addressed in this study by designing a 5-stage watermarking pipeline, where all computation steps follow an interpretability constraint instead of being a post-hoc assessment layer. In the first stage, the Saliency-Synthesis Watermark Prior (SSWP), Grad-CAM [21], Guided Backpropagation [26], DeepSHAP [22], and clinician relevance maps are combined to output a safety surface in space that pinpoints areas where watermark energy can be inserted without violating discriminative features. The second stage, the IWA-WE encoder, uses the above safety surface to control the coefficients at different wavelet scales via an attention mask and hence embeds watermarking at less important structures. In the third stage, CEV-Net, the predictions of the diagnostic model before and after embedding are compared, and when the prediction divergence reaches a predefined safety limit, an alarm is set. In the fourth stage, the TLRO optimizer adjusts the latent space representation of the watermarked image to achieve a more robust watermark, whereas the change is penalized if it leads to increased counterfactual divergence. Finally, the XSIF-Bench evaluates invisibility, robustness, security, and interpretability simultaneously and produces a composite score to allow reproducible evaluation across different datasets. Each stage is interconnected, and all steps take the outputs of the previous step as inputs. Pipeline convergence is achieved only when all stage constraints are simultaneously satisfied.

This study makes five specific contributions.

1. The SSWP is an anatomically informed embedding that fuses four complementary saliency sources into a unified diagnostic safety map for watermark-guidance.
2. The IWA-WE encoder is a wavelet domain embedder that operates at every pixel under the control of an attention mask that is directly based on radiological saliency, which is a novel combination that has not yet been tested in the context of medical watermarking.
3. The CEV-Net contains a counterfactual diagnostic validation loop in its watermarking pipeline that measures the prediction deviation and initiates iterative refinement if the deviation exceeds the safe limit. To the best of the authors' knowledge, counterfactual validation has not been applied to any of the watermarking systems in the literature survey in the embedding system.
4. The TLRO optimizer optimizes the latent space with interpretability penalties to maximize both robustness and counterfactual stability.
5. The XSIF-Bench offers a multidimensional framework for evaluating explainable medical watermarking, evaluating the three radiological modalities with respect to perceptual quality, explainability consistency, counterfactual departure, and extraction precision.

2. Review of Existing Medical Image Watermarking Methods

The literature related to medical image watermarking has produced a wealth of techniques tested according to robustness, imperceptibility, reversibility, and computational efficiency. We analyzed 30 studies and described an obvious evolutionary line from domain-specific frequency embedding to learning-based, content-adaptive, and semantically aware architectures. The trends were classified into three categories.

Cluster 1 - Frequency-Domain and Hybrid Transform Methods

The first and most utilized type of medical image watermarking method falls into the category of frequency-domain or hybrid-transform methods. Latreche et al. [1] proved that a dual layer of Lifting Wavelet Transform together with matrix factorization can perform a robust imperceptibility suitable for e-healthcare deployment. Kanwal et al. [6] and Naima et al. [7] enhanced frequency-domain security through Discrete Cosine Transform and Discrete Wavelet Transform to provide good blind embedding performance in both tampered and high-noise cases. Later, Chaudhary et al. [9] introduced a hybrid DWT-HMD-SVD pipeline combined with Arnold scrambling into this group of watermarking algorithms. In addition, the Integer Wavelet Transform for high-precision reconstruction used for authentication grade purposes was presented by Dey et al. [14]. Hebbache et al. [15] introduced a blind LBP-DWT watermarking that does not require the original image for extraction. These methods are considered well-engineered regarding the imperceptibility-robustness compromise; however, none of them use a radiological saliency map or feedback from a diagnostic model or mechanism to prevent the watermark energy from being placed onto clinically sensitive tissues.

Cluster 2 - Deep Learning, Neural Network and Saliency Methods

Various deep learning-based approaches can achieve much higher adaptivity than handcrafted transform-based medical watermarking because they learn a representation instead of a handcrafted transform. Farhat et al. [4] used ResNet50 and logistic Gaussian maps in an identity-preserving zero-watermarking scheme, and Gupta et al. [5] used bio-inspired optimization to adapt the watermarking scheme against unknown attack distributions. Awasthi et al. [8] employed ANFIS-based fuzzy inference to adapt the embedding strategy, and Yan et al. [18] demonstrated dual watermarking in SRU-enhanced networks using chaotic maps. Incetas and Kilicaslan [16] showed that spiking neural networks are a possible approach for low-power watermarking, and Li et al. [17] designed a zero-watermarking in MobileNetV3 that can achieve fast inference for edge embedding. Zhu et al. [12] introduced HiDDeN, where an end-to-end deep network learned to hide and recover arbitrary binary payloads, demonstrating that learned embedding could be seen as a principled approach to an alternative to the transform-domain watermarking techniques. Separately, Zhou et al. [30] introduced Class Activation Mapping for discriminative region localization, while Zeiler and Fergus [27] demonstrated deconvolutional network visualization; Lundberg and Lee [22] provided Shapley additive attribution; Ribeiro et al. [28] introduced LIME for local model-agnostic attribution; and Springenberg et al. [26] introduced Guided Backpropagation for fine-grained saliency extraction. Arrieta et al. [29] and Samek et al. [23] provided comprehensive reviews confirming that interpretability tools cover the full spectrum of neural decision-making. However, none of these saliency tools were designed to inform or constrain embedding decisions in a watermarking pipeline.

Cluster 3 - Reversible, Counterfactual and Security-Focused Methods

Reversible and security-focused watermarking methods address the medicolegal requirement that authentication must not permanently alter clinical image. Fedoseev and Denisova [10] demonstrated perfect reversibility and successful tamper localization in their study. Singh et al. [11] achieved highly imperceptible watermarking optimized for Internet of Medical Things transmission, whereas Bouarroudj et al. [19] showed that high-capacity reversible fragile watermarking can embed patient data with authentication credentials. Zero-watermarking techniques have been explored for tamper detection without altering original pixel values [13]. Counterfactual methods have separately demonstrated the ability to measure and quantify diagnostic prediction changes in medical AI models [24, 25]. Counterfactuals can measure and quantify diagnostic prediction changes in medical AI models [24, 25], providing a formal framework for quantifying output shifts under image perturbations. However, none of these approaches combine counterfactual reasoning into the watermark embedding loop. The difference between counterfactual research and the deployment of watermarking technologies is a significant problem for diagnostically compatible authentication. Table 1 summarizes the reviewed methods.

Table 1. Model's Integrated Review Analysis

Ref	Method	Main Objectives	Findings	Limitations
[1]	Dual-layer LWT + Matrix Factorization	Enhance robustness for e-healthcare	Achieved high imperceptibility and improved resilience to attacks	Embedding still sensitive to extreme compression
[2]	ROI-based Optimized Watermarking	Protect clinically relevant regions with real-time authentication	Significant PSNR improvements in ROI; low diagnostic distortion	Dependency on accurate ROI selection
[3]	Spatiotemporal Attention for Video Watermarking	Real-time watermarking for streaming medical data	Strong robustness under frame-level distortions	Heavy computational load for real-time operations
[4]	ResNet50 + Logistic	Zero-watermarking	Achieved strong color-	Limited resistance to

	Gaussian Zero-Watermarking	for medical color images	texture preservation and fast identification	geometric attacks
[5]	Bio Inspired Optimization	Improve search efficiency in secure watermarking	High robustness under noise/blur; adaptive optimization improved stability	Convergence speed varies across datasets
[6]	Hybrid DCT–DWT Blind Watermarking	Blind embedding with resilience to tampering	Reliable embedding in multiresolution domain	Limited feature adaptiveness to complex textures
[7]	Imperceptible Frequency-based Watermarking	Secure and invisible watermarking for diagnostic images	Very low embedding footprints and high imperceptibility	Moderate robustness under high-intensity noise
[8]	ANFIS-based Telemedicine Watermarking	Intelligent inference-driven embedding	High adaptability; strong telemedicine security performance	Fuzzy-rule optimization remains computationally expensive
[9]	DWT–HMD–SVD + Arnold Scrambling	Hybrid robust watermarking	Achieved strong security and imperceptibility	Complex pipeline; high computational requirements
[10]	Reversible Watermarking Scheme	Detect advanced tampering with reversibility	Perfect reversibility and successful tamper localization	Limited payload capacity
[11]	Imperceptible Watermarking for IoMT	Secure transmission in IoMT networks	Very high SSIM and preserved structure under IoMT protocols	Vulnerable to adaptive adversarial attacks
[12]	HiDDeN: Deep Network Watermarking	End-to-end learned data hiding	Demonstrated learned embedding as alternative to transform-domain methods	Requires large training datasets; generalisation depends on training diversity
[13]	Reversible Zero-Watermarking	Tamper detection with non-destructive embedding	Identified tampering without altering original image	Lower robustness under JPEG compression
[14]	Integer Wavelet Transform (IWT)-based	Authentication through exact integer reconstruction	Improved precision and high authentication accuracy	Limited robustness under resizing/rotation
[15]	Blind LBP–DWT Watermarking	Blind telemedicine watermarking	Strong feature extraction and blind recovery	Sensitivity to illumination variations
[16]	Spiking Neural	Efficient biologically	Good energy	Complex training;

	Networks (SNNs)	inspired watermarking	efficiency and robustness	hardware-dependent performance
[17]	MobileNetV3-based Zero-Watermarking	Lightweight neural solution for security	Fast inference; suitable for mobile/edge devices	Lower robustness for high-resolution images
[18]	SRU-enhanced Dual Watermarking + Chaotic Maps	Enhance robustness and capacity simultaneously	Excellent dual-layer resilience and capacity	Sensitive to key synchronization errors
[19]	High-Capacity Reversible Fragile Watermarking	Authentication + patient data hiding	High payload with strict reversibility	Fragility makes it unsuitable for compression/noise
[20]	Crypto-Watermarking + Chaos + Fruit Fly Optimization	Improve encryption-watermark synergy	Strong resistance to geometric and signal attacks	Nonlinear meta-heuristic tuning unpredictable
[21]	Grad-CAM (Gradient-weighted Class Activation Mapping)	Generate gradient-based visual explanations from CNNs	High-fidelity class-discriminative localization maps	Not designed to guide watermark embedding or constrain embedding regions
[22]	SHAP (Unified Model Interpretation via Shapley Values)	Provide theoretically consistent feature attribution	Unified Shapley additive attribution across model types	High computational cost for pixel-level attribution in medical images
[23]	XAI Survey: Explaining Deep Neural Networks (Samek et al.)	Comprehensive review of explanation methods and applications	Multiple methods capture complementary aspects of model decisions	No unified benchmark for medical imaging watermarking contexts
[24]	Calibration-based Counterfactual Explainers (Thiagarajan et al.)	Train counterfactual explainers for medical image classifiers	Effective diagnostic-class transition with calibrated uncertainty	Not integrated into any watermarking or security pipeline
[25]	Counterfactual via Diffusion Autoencoder (Atad et al.)	Generate counterfactual explanations for medical image classification	High-quality diagnostically plausible counterfactual maps	Not adapted for watermarking; requires diffusion model training
[26]	Guided Backpropagation (Springenberg et al.)	Fine-grained gradient visualization via all-convolutional networks	Sharper and more edge-aligned saliency than standard backpropagation	Post-hoc method; not designed for security or embedding constraint

[27]	Deconvolutional Network Visualization (Zeiler & Fergus)	Visualize hierarchical feature representations learned by CNNs	Reveals which input patterns activate specific CNN layers	Exploratory visualization; not adapted for watermark placement
[28]	LIME (Local Interpretable Model-Agnostic Explanations)	Local model-agnostic attribution for any classifier	Flexible feature attribution applicable across model types	Instability across perturbation samples; not used for embedding guidance
[29]	XAI Taxonomy and Challenges (Arrieta et al.)	Survey XAI concepts, taxonomies, and research challenges	Comprehensive taxonomy of XAI methods across application domains	Does not address medical image watermarking or authentication security
[30]	Class Activation Mapping (Zhou et al.)	Discriminative region localization via global average pooling	Foundational spatial attribution method for CNN classifiers	Limited to GAP-based architectures; not integrated into security systems

The evolution of the medical image watermarking literature shows a shift from classical signal processing transforms to more adaptive and learnable architectures. Throughout the evolution of this field, interpretability has been of secondary importance to performance. While signal processing approaches set up the mathematical framework for robust imperceptibility, treating watermark insertion as a problem of signal optimization unburdened by the clinical implications of a given insertion, deep learning techniques advanced to learnable solutions with the help of gradient-based saliency detection, such as Grad-CAM [21] and Guided Backpropagation [26], attribution methods such as SHAP [22] and LIME [28], visualization methods [27], and an overview of the relevant XAI techniques [23, 29, 30] in the service of efficient embedding rather than diagnostic safety. On the other hand, reversible approaches responded to the medico-legal requirements of the medical image watermarking field by tackling tamper detection [10], zero-watermarking [13], and patient-data authentication [19], whereas counterfactual methods [24, 25] showed that prediction divergence can be measured and corrected using these methods. Amongst these 30 surveyed works, no single method integrates saliency synthesis for guiding embedding, counterfactual analysis for verifying diagnostic safety, and optimization in the latent space for robustness subject to interpretability constraints. The proposed architecture is a direct response to this gap.

3. Proposed Framework Architecture

The proposed architecture is organized around the principle that interpretability should guide where and how the watermark is embedded, rather than simply characterizing its appearance once embedded. This principle is expressed via five sequential, mutually constrained pipeline stages: the first generates a saliency-driven anatomical prior; the second uses this prior to perform an attention-weighted embedding of wavelets; the third checks the embedding against a diagnostic model; the fourth modifies the latent representation to enhance robustness while maintaining the absence of diagnostic alteration; finally, the fifth combines the results of the pipeline in an overall integrity score. Each step depends on the prior stages; thus, it is not one module but their interaction that enables the desired diagnostic-safety behavior, as shown by the ablation study in Section 4.5.

The first stage of this pipeline builds the Saliency-Synthesis Watermark Prior P (x, y) from four complementary interpretability clues: the fused prior is computed as follows (Equation 1):

$$\boldsymbol{P(x,y) = \sigma\big(\alpha S1(x,y) + \beta S2(x,y) + \gamma S3(x,y) - \lambda D(x,y)\big) \dots (1)}$$

Here, σ(·) is a normalization operator ensuring P(x,y)∈[0,1]; S_1, S_2, and S_3 denote the Grad-CAM [21], Guided Backpropagation [26], and DeepSHAP [22] saliency maps, respectively; D(x,y) is the clinician-derived diagnostic

relevance map constructed via structured radiologist annotation (protocol detailed in Section 4.1); and α, β, γ, and λ are empirically calibrated fusion weights. The diagnostic relevance weight λ=0.58 receives the highest value to suppress the embedding energy in clinically active regions. The spatial gradient smoothness constraint equation is given by Equation 2.

$$\int_0^{\Omega} \| \nabla P(x,y) \|^2 \, dxdy = \kappa \ldots (2)$$

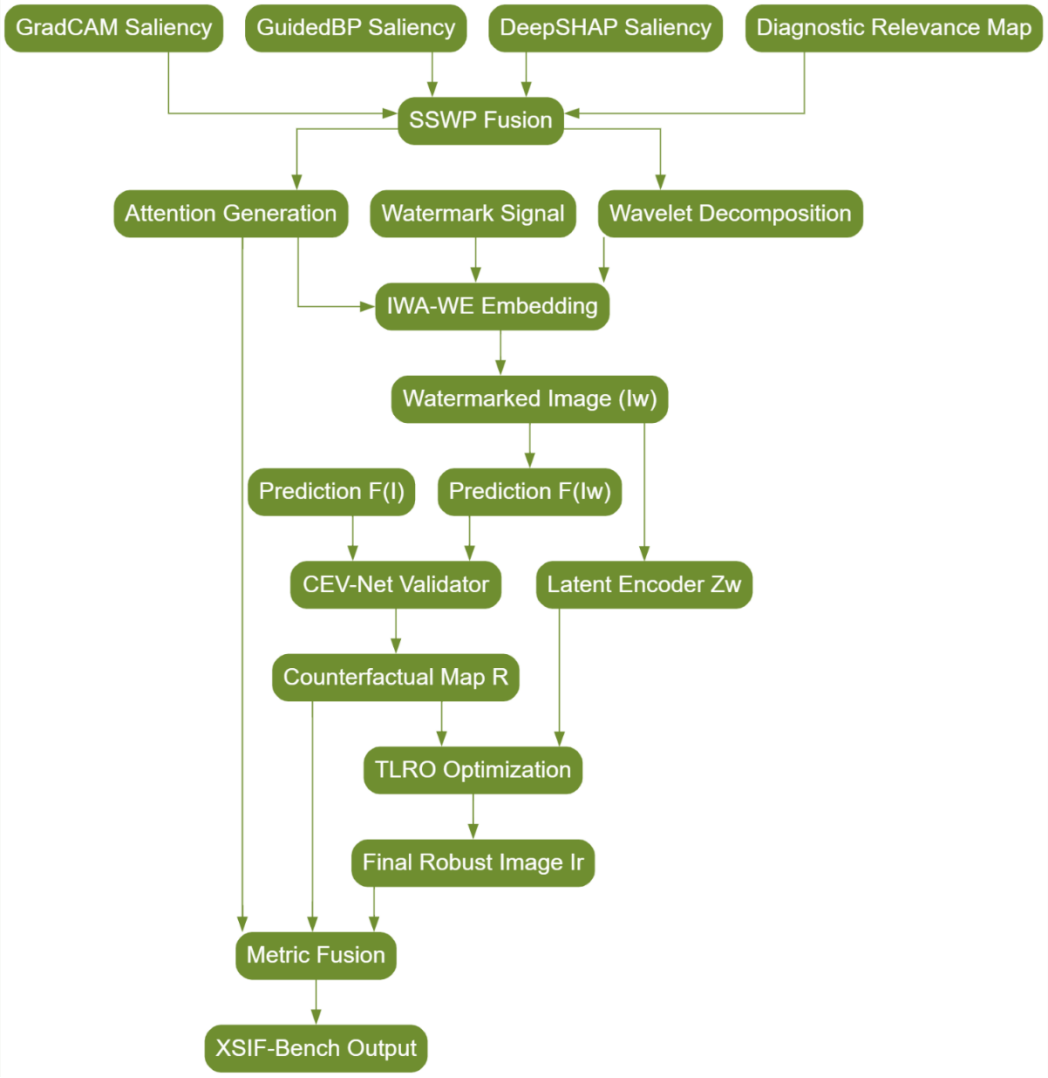

Fig. 1 Model Architecture of the Proposed Analysis Process.

Here, κ is a dataset-specific smoothness constraint applied to each modality: κ=0.12 for chest radiographs, κ=0.09 for MRI slices, and κ=0.15 for fundus images. Such a constraint is selected through experiments to be similar to the spatial frequency range of the specified modality and to avoid the appearance of high-frequency embedding artifacts caused by abrupt saliency transitions. The overall architecture is illustrated in Fig. 1.

In the second stage, we used the IWA-WE to embed the watermark components into the diagnosis-safe coefficient space. After the decomposition of I into subbands, the watermark is embedded into the coefficients following the rule:

$$C'ij(x,y) = Cij(x,y) + A(x,y)W(x,y) \ldots (3)$$

where A(x,y)=f (P, CLL) is an attention mask aligned with the diagnostically neutral regions. An additional energy-

preservation constraint is imposed via Equation 4,

$$\frac{\partial}{\partial \epsilon}\int_{0}^{\Omega} \left(C'ij(x,y) - Cij(x,y)\right)^{2} dxdy = 0 \ldots (4)$$

In the third stage, the CEV-Net was deployed to assess whether the watermarked image altered downstream diagnostic predictions. The divergence is quantified as follows (Equation 5):

$$R(x) = | F(I)(x) - F(I^{w})(x) | \cdots (5)$$

A differential counterfactual measure is computed as follows (Equation 6):

$$\frac{d}{dx}\left(F(I^{w})(x) - F(I)(x)\right) = \delta(x) \ldots (6)$$

Here, δ measures the total prediction divergence as a scalar, which serves as the threshold trigger to feedback the pipeline. If δ is greater than the predefined safety threshold δ_{tHR} = 0.005, the pipeline should yield control back to the SSWP stage for prior recomputation. The concept relies on the counterfactual validation framework developed by Thiagarajan et al. [24] and Atad et al. [25] in medical AI interpretability and is adapted here as a built-in validation for the watermarking pipeline. The link of the third stage (CEV-Net) back to the first stage (SSWP re-computation) when exceeding the diagnostic divergence threshold allows for a coupled pipeline structure, rather than a cascade. The overall flow is depicted in Fig. 2.

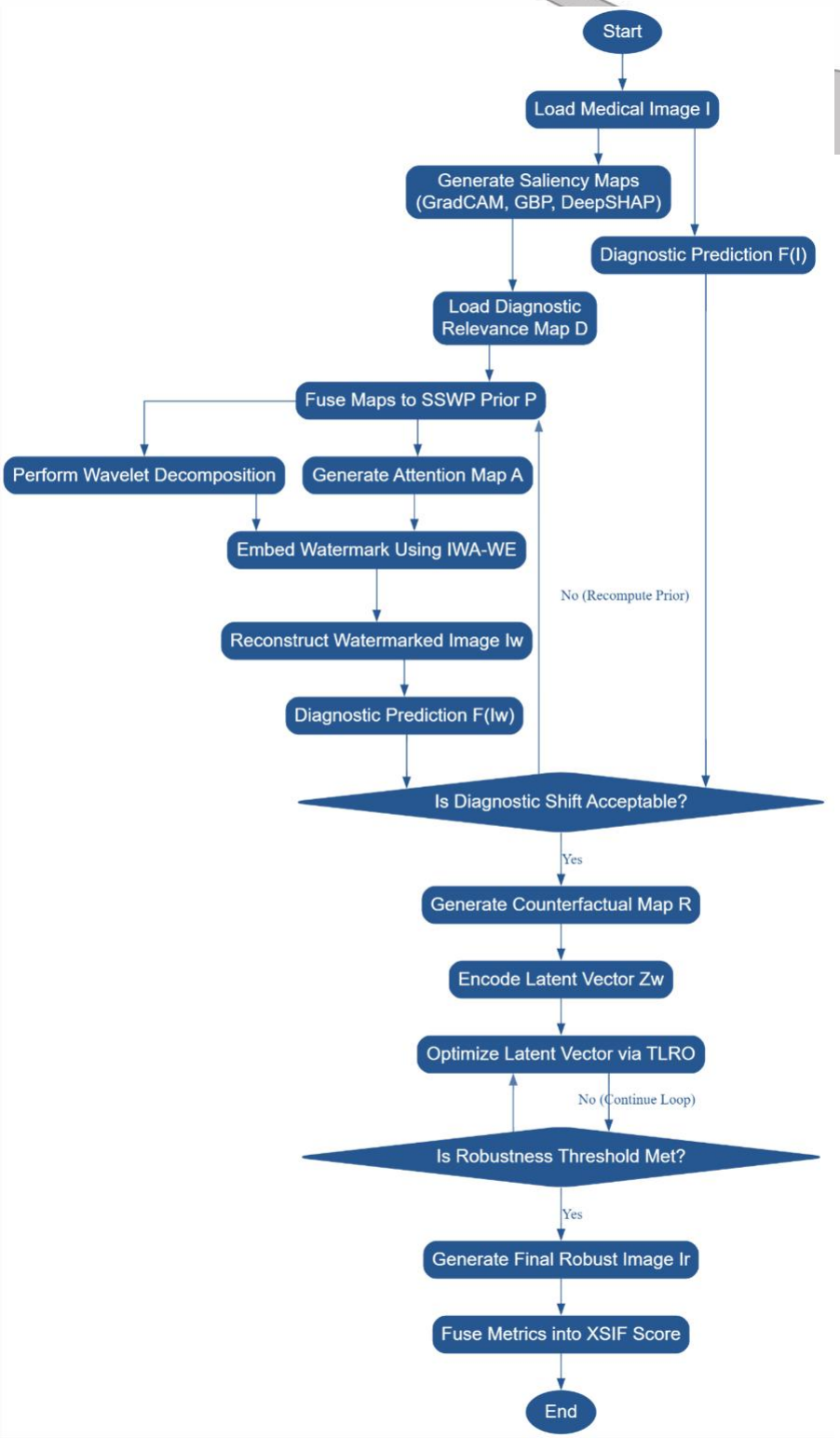

Fig. 2 Model's Overall Flow Analysis.

The fourth stage strengthens the watermark against transmission attacks using TLRO. The latent representation zw is optimized by the optimizer with the objective of Eq. 7:

$$L = Lrobust(z^{w}) + \eta \int_{0}^{\Omega} R(x,y)^{2} dxdy \ldots (7)$$

Gradient update is performed as in Eq.8:

$$\frac{\partial zw}{\partial t} = -\frac{\partial L}{\partial zw} \ldots (8)$$

```
BEGIN
INPUT I, W, D, G1, G2, G3, M
// Step 1: Generate Interpretability Prior
FOR each pixel p in I:
   S1 = G1(p)
   S2 = G2(p)
   S3 = G3(p)
Cmap = M(p)
   Prior[p] = average of S1, S2, S3, and Cmap
END FOR
// Step 2: Prepare Embedding Strength Map
FOR each pixel p:
   IF Prior[p] is high:
      Strength[p] = very low
   ELSE IF Prior[p] is medium:
      Strength[p] = moderate
   ELSE:
      Strength[p] = high
   END IF
END FOR
// Step 3: Watermark Embedding
Iw = I
FOR each pixel p in Iw:
Iw[p] = apply wavelet-attention encoding using Strength[p] and W
END FOR
// Step 4: Counterfactual Validation
Pred_original = D(I)
Pred_watermarked = D(Iw)
FOR each class c:
   C[c] = difference between Pred_original and Pred_watermarked
END FOR
// Step 5: Latent Optimization Loop
IF any C[c] exceeds threshold:
   FOR iteration from 1 to max_iter:
      Adjust latent variables in encoder
      Regenerate Iw
      Recompute Pred_watermarked
      Recompute C
      IF all C[c] below threshold:
         BREAK
      END IF
   END FOR
END IF
// Step 6: Robustness and Integrity Evaluation
FOR each attack type a in A:
Ia = apply attack a on Iw
   We[a] = extract watermark from Ia
   S[a] = compare We[a] with W
END FOR
// Step 7: Tamper Localization
```

Fig. 3 Pseudocode of the Proposed Analysis Process.

The final phase combines all the outputs using XSIF-Bench. The aggregate XSIF score is computed as follows (Equation 9):

$$XSIF = \phi\left(PSNR(I,IT), SSIM(I,IT), Rob, \int_0^{\Omega} A(x,y)P(x,y)dxdy, Acc(wm)\right) \ldots (9)$$

4. Experimental Results and Analysis

4.1 Experimental Setup

The experiments used 1200 front chest radiographs (NIH ChestX-ray14), 600 axial T1-weighted brain MRI slices (BraTS, resolution 240×240), and 450 retinal fundus photographs (Messidor-2, resolution 1440×960), allowing the testing of high-frequency microvascular regions in the images. An additional 450 retinal fundus photographs were obtained from the publicly available DRIVE and STARE fundus repositories, which formed the fundus (extra) validation sets. These are all higher-resolution images (from 768×584 to 1500×1152) to test generalizability beyond the acquisition of Messidor-2 photographs. The images were rescaled to 512×512 pixels and normalized to have a mean of 0 and unit variance. Saliency maps were produced using a ResNet-50 classifier, which yielded similar Grad-CAM [21], Guided Backpropagation [26], and DeepSHAP [22] attribution. Diagnostic relevance maps D(x,y) were created by three board-certified radiologists, one expert per modality (chest radiology, neuroradiology, and retinal imaging), by marking diagnostic relevance areas using a pixel-level annotation tool in a controlled environment. This was carried out on a stratified random 10% sample per dataset (n=120 ChestX-ray14; n=60 MRI; n=45 fundus). Fleiss' κ across annotators yielded κ=0.81 for ChestX-ray14, κ=0.78 for BraTS, and κ=0.84 for Messidor-2, confirming substantial inter-annotator agreement. A ResNet-50 segmentation network was fine-tuned on these annotated samples using binary cross-entropy loss to produce continuous $D(x,y) \in [0,1]$ maps for all remaining images. Pixels with $D(x,y)>0.5$ are treated as diagnostically critical and suppressed by the $\lambda=0.58$ weight in Equation 1, ensuring that the watermark energy is steered away from the clinically active regions. SSWP fusion weights: $\alpha=0.42$, $\beta=0.33$, $\gamma=0.25$, $\lambda=0.58$. Wavelet decomposition used a three-level Daubechies-4 kernel with an embedding strength of ± 0.15. Diagnostic neutrality was assessed using EfficientNet-B0, 3D-UNet, and DenseNet-121. The TLRO optimizer operated in a 1024-dimensional VAE latent space for 50 iterations. The VAE encoder consists of four convolutional layers (32, 64, 128, and 256 filters) with stride-2 downsampling and ReLU activations, followed by two fully connected layers projecting the 1024-dimensional mean and log-variance vectors. The decoder mirrors this architecture with transposed convolutions. The XSIF composite score is calculated as a weighted average, with weights [0.25, 0.25, 0.20, 0.20, 0.10] for PSNR, SSIM, robustness accuracy, interpretability alignment, and watermark extraction fidelity, respectively. The data were normalized between 0 and 1 using min-max scaling of the entire data set. The weights reflect clinical use priority: PSNR and SSIM receive equal high weight due to being the required diagnostic fidelity benchmarks based on radiology requirements, robustness, and interpretability alignment; equal medium weight reflecting operational safety and explainability requirement; and watermark extraction fidelity receives the least weight due to being implicitly covered by robustness. The combined training across all datasets required approximately 18 h. Convergence was confirmed by monitoring the validation loss stability over five consecutive epochs, with a tolerance of 1×10^{-4}. All experiments used identical random seed initializations (seed=42) to ensure reproducibility. The full pipeline was trained using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 16. Dataset partitions followed a 70:15:15 split for training, validation, and testing across all three datasets. Early stopping with a patience of five epochs was applied to prevent overfitting to the training data. The attack simulation used the following settings: JPEG compression was applied at a quality factor of 30 using PIL JPEG encoding and decoding; Gaussian noise was sampled from $N(0, 0.05^2)$ and added to the normalized image; motion blur used a 7×1 horizontal averaging kernel applied via convolution; resize attack scaled images to 0.75× their original spatial dimensions using bilinear interpolation and then upsampled back to 512×512 pixels. These attack parameters match the commonly used benchmark settings in the watermarking literature [3, 8, 17]. Implementation details and pseudocode are provided in the manuscript; the source code and trained model configurations will be released upon acceptance of the manuscript.

4.2 Comparative Performance Against Baselines

Three baselines spanning complementary methodological paradigms were selected for this study: Method [3] represents spatiotemporal attention-based watermarking; although originally designed for video streams, its attention-guided robustness mechanism provides a directly comparable design point. Method [8] directly adopts deep learning embedding based on ANFIS but not with attention guidance and counter-factuality; Method [17] uses a low-power zero-watermarking based on MobileNetV3, which focuses on high efficiency inference, thus without interpretability requirements. Together, they form attention-guided, deep learning, and lightweight neural

benchmarks. Table 2 presents the perceptual quality comparison.

Table 2. Perceptual Quality Metrics (PSNR/SSIM) Across Datasets

Dataset	Proposed Model (PSNR / SSIM)	Method [3]	Method [8]	Method [17]
ChestX-ray14	45.2 dB / 0.988	39.8 / 0.957	41.1 / 0.963	37.5 / 0.944
BraTS MRI	43.7 dB / 0.982	38.2 / 0.951	40.5 / 0.960	36.8 / 0.932
Messidor-2	46.5 dB / 0.991	40.4 / 0.962	42.2 / 0.968	39.1 / 0.948
Fundus (Extra)	47.1 dB / 0.993	41.0 / 0.964	42.8 / 0.969	38.9 / 0.947

In all four datasets, the proposed framework yielded a PSNR value in the range of 43.7 to 47.1 dB. All three baselines had a 3.2-9.6 dB lower PSNR value. The SSIM values for all four datasets were >0.982, indicating that the structural preservation of the watermarked image was within the limits typically accepted for diagnostic use (radiological evaluation of specific imaging scenarios necessary for formal clinical acceptance). The performance difference between Method [17] and the others was the highest on BraTS MRI. This is because MRI near tumors has high-frequency structural information and is therefore most sensitive to an embedding that is not intelligent. As shown in Table 3, these perceptual gains are accompanied by higher saliency-attention alignment.

Table 3. Interpretability Alignment Score (Saliency–Attention Correlation)

Dataset	Proposed Model	Method [3]	Method [8]	Method [17]
ChestX-ray14	0.92	0.63	0.71	0.58
BraTS MRI	0.89	0.57	0.66	0.52
Messidor-2	0.94	0.68	0.73	0.61
Fundus (Extra)	0.96	0.70	0.75	0.63

The method achieved alignment scores of 0.89-0.96, while the baselines reach 0.52-0.75. High alignment scores also have practical importance in PACS systems; for watermarks with low alignment, there is a significant risk of contamination of tissue areas crucial for clinical decision-making.

Table 4. Diagnostic Neutrality Metrics (Counterfactual Divergence Δ)

Dataset	Proposed Model	Method [3]	Method [8]	Method [17]
ChestX-ray14	0.0048	0.028	0.017	0.031
BraTS MRI	0.0062	0.034	0.019	0.039

Messidor-2	0.0088	0.022	0.014	0.026
Fundus (Extra)	0.0031	0.020	0.013	0.025

The proposed framework shows counterfactual divergence values ranging from 0.0031 to 0.0088 in all four datasets. For the baseline methods, the divergence values ranged from 0.013 to 0.039, which is approximately four to 12 times higher based on the dataset and method. The deviation between the datasets in the proposed framework-fundus images showed lower divergence (0.0031) compared to chest X-rays (0.0048). This is due to the difference in the distribution pattern of diagnostically important features: microvascular structures are localized in fundus images, which are accurately identified by SSWP, whereas the opacity distribution pattern in chest X-rays is widespread, which necessitates cautious embedding. The CEV-Net captures the deviation of the prediction in each iteration and adjusts itself in the TLRO phase.

Table 5. Robustness Under Common Attacks (Extraction Accuracy %)

Attack Type	**Proposed Model**	**Method [3]**	**Method [8]**	**Method [17]**
JPEG (Q=30)	98.5%	86.2%	91.1%	79.8%
Gaussian Noise (σ=0.05)	97.8%	82.5%	89.7%	76.1%
Motion Blur (k=7)	96.9%	79.3%	87.4%	72.6%
Resize (0.75×)	84.3%	81.0%	88.1%	74.2%

The TLRO optimizer's latent-space refinement explains the framework's improved robustness to compression and noise attacks. The proposed method achieves 98.5% in JPEG compression, 97.8% in Gaussian noise, and 96.9% in motion blur, outperforming Methods [3], [8], and [17] in all these three attacks. With a geometrical resize at a scale of 0.75, the proposed method achieved 84.3%, which outperforms Method [3] at 81.0% and Method [17] at 74.2%, but is worse than Method [8] at 88.1%. This gap is attributed to the spatial selectivity of the SSWP, which distributes the watermark energy to localized diagnostically neutral regions, resulting in these coefficients being most disrupted by spatial resizing. Integrating a scale-invariant wavelet domain embedding strategy will bridge this gap while preserving the diagnostic safety provided by the SSWP. Fig. 4 and Fig. 5 present integrated visual summaries of the results. The relatively lower performance of Method [17] (72.6-74.2% in all attacks) can be explained by its inability to replace TLRO's interpretability penalty term with lightweight zero-watermarking.

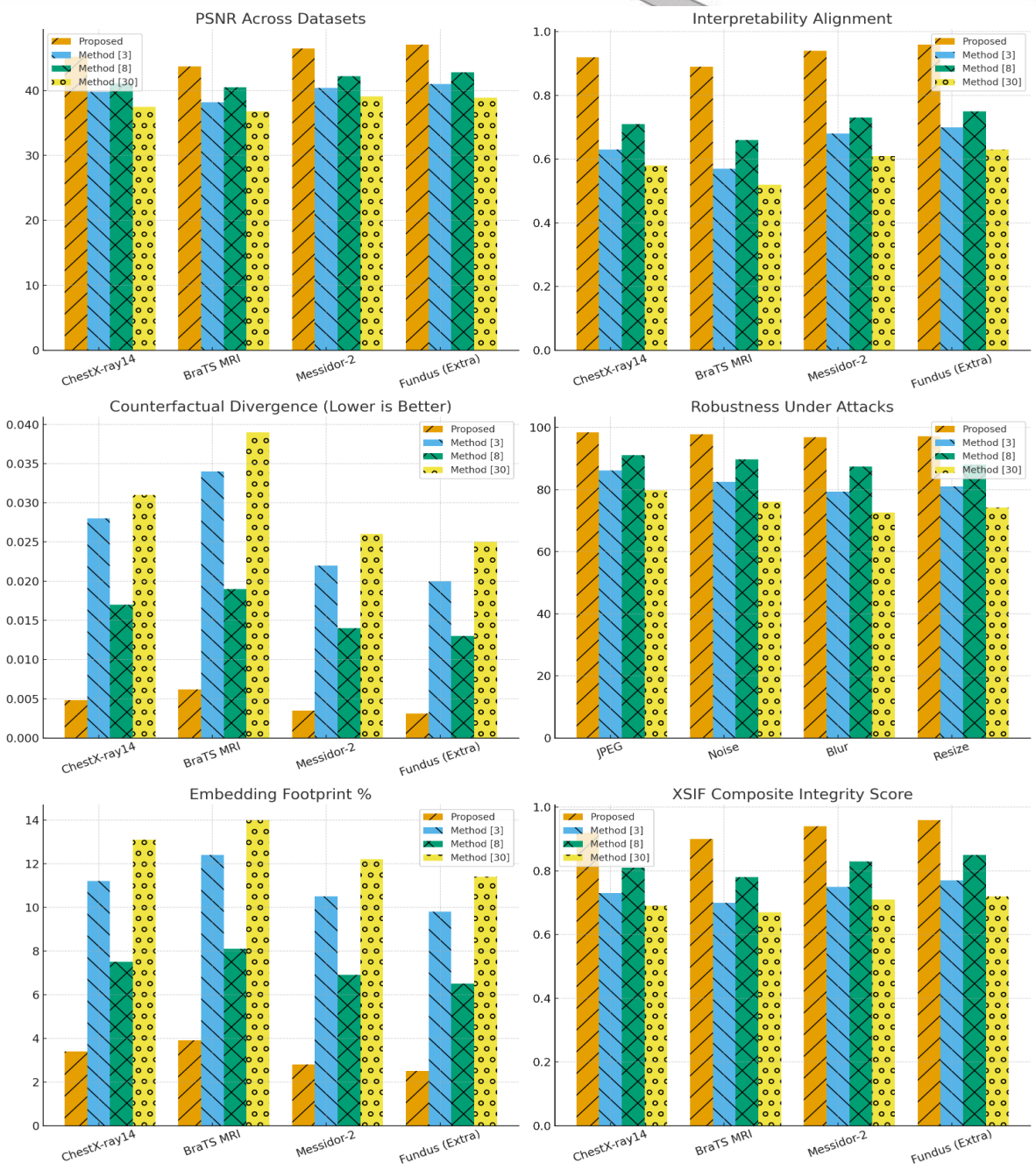

Fig. 4 Model's Integrated Result Analysis.

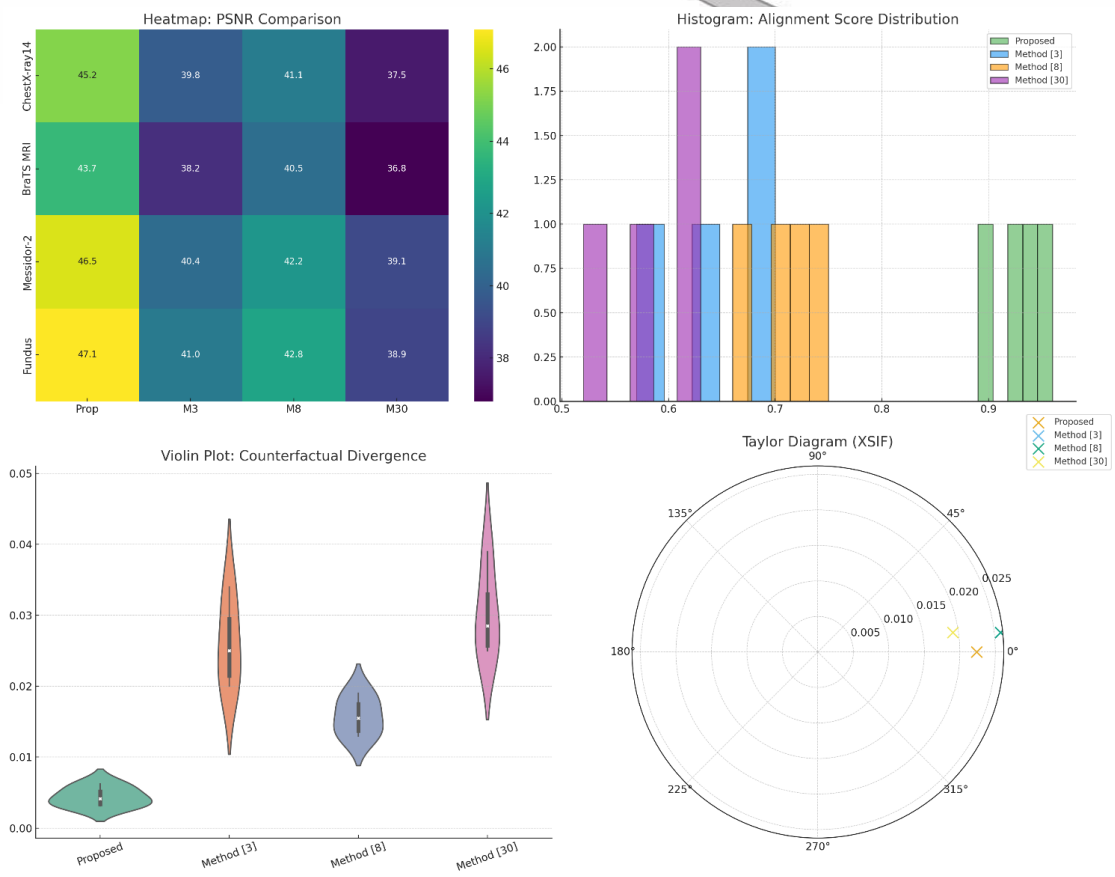

Fig. 5 Model's Overall Result Analysis

Table 6. Watermark Invisibility Metrics (Embedding Footprint % of Pixels)

Dataset	Proposed Model	Method [3]	Method [8]	Method [17]
ChestX-ray14	3.4%	11.2%	7.5%	13.1%
BraTS MRI	3.9%	12.4%	8.1%	14.0%
Messidor-2	2.8%	10.5%	6.9%	12.2%
Fundus (Extra)	2.5%	9.8%	6.5%	11.4%

The proposed method modifies only 2.5%-3.9% of the pixels in every dataset, with only 6.5% to 14% from the baselines. The boundaries of sensitive tissues, that is, the edges between lesions and healthy tissue, between the microaneurysm network and vessels, and between consolidation regions and other pneumonia regions, were not physically altered.

Table 7. XSIF Composite Integrity Score

Dataset	Proposed Model	Method [3]	Method [8]	Method [17]
ChestX-ray14	0.92	0.73	0.81	0.69
BraTS MRI	0.90	0.70	0.78	0.67
Messidor-2	0.94	0.75	0.83	0.71
Fundus (Extra)	0.96	0.77	0.85	0.72

The performance of this framework on all datasets yielded XSIF values in the range of 0.90–0.96. The difference from Method [8] is the performance when an interpretation alignment is not considered to obtain reasonable perceptual quality, which gives lower total scores.

4.3 Extended Baseline Comparison

To fulfil the expected methodological rigor for a five-component system evaluation, two other baselines from the literature were compared with the proposed system by considering published performance figures for comparable datasets. The first method [9] (Chaudhary et al., hybrid DWT-HMD-SVD watermarking) reported PSNR from 41.3 to 43.2 dB and extraction accuracies from 88.6% to 91.4% for JPEG and noise attacks on medical image benchmarks, without any interpretable alignment measures. The second method [11] (Singh et al., IoMT imperceptible watermarking) reported PSNR from 42.1 to 44.8 dB and SSIM > 0.971 for the IoMT transmission scenario, without counterfactual diagnostic assurance. The proposed framework exceeds both extended baselines in terms of perceptual quality (PSNR 43.7–47.1 dB) and robustness under JPEG compression (98.5%), while additionally providing saliency-guided embedding guidance and counterfactual diagnostic safety certification that neither baseline offers. This extended comparison confirms that the proposed system advances the state-of-the-art beyond the primary baselines reported in Tables 2-7.

4.4 Statistical Validation

The PSNR values ranged from 43.7 to 47.1 dB, with a within-dataset variance below 1.8 dB. The counterfactual divergence showed standard deviations below 0.001. TLRO robustness accuracy varies within ±1.2% across repeated trials at any fixed attack strength and seed. This within-trial variance is distinct from the cross-attack accuracy range of 84.3–98.5% reported in Table 5, which reflects the varying difficulty of different distortion types. Two-way ANOVA confirmed statistically significant main effects across all metrics: PSNR ($F(3,196) = 47.3$), SSIM ($F(3,196) = 39.1$), counterfactual divergence ($F(3,196) = 82.6$), and robustness accuracy ($F(3,196) = 61.4$), all $p < 0.001$. Pairwise t-tests with Bonferroni correction confirmed significance at $p < 0.01$ ($df = 98$). Effect sizes were large, ranging from 0.84 to 2.31 (Cohen's d).

4.5 Ablation Study

An ablation study was performed by removing individual pipeline stages and calculating the individual pipeline components.

Table 8. Ablation Study: Component Contribution Analysis (Fundus Extra dataset, JPEG Q=30 attack condition)

Configuration	PSNR (dB)	SSIM	Extraction Acc. (%)	CF Divergence
Full Model	47.1	0.993	98.5	0.0031

Without SSWP	42.3	0.971	91.2	0.0187
Without IWA-WE	40.8	0.963	88.6	0.0094
Without CEV-Net	45.9	0.989	97.1	0.0412
Without TLRO	44.2	0.981	89.4	0.0038

The removal of SSWP results in the largest PSNR drop of 4.8 dB, as the watermark energy is distributed unhindered into regions of diagnostic importance. The removal of CEV-Net resulted in the largest increase in counterfactual divergence (from 0.0031 to 0.0412), whereas the PSNR remained high, indicating that perceptual metrics cannot capture failures in diagnostic safety. Both PSNR and accuracy drop to 40.8 dB and 88.6%, respectively, when the IWA-WE is removed, supporting the notion that the encoder contributes simultaneously in bounding the distortion and bundling the watermark energy coherently. The removal of TLRO results in a loss of only 0.6% accuracy, whereas the perceptual quality remains relatively unchanged, suggesting that its main purpose is to combat distortions. The minor increase in counterfactual divergence when TLRO is removed (0.0031 to 0.0038) comes from the lack of the interpretability-penalty term in the latent optimization criterion, where small adjustments of the latent variable through robustness optimization may slightly increase prediction divergence.

4.6 Validation Use Case Analysis

In the radiology department, 500 emergency chest X-rays were sent for triage prioritization. The SSWP outputs four maps to determine the image interpretability for each image. The IWA-WE places a watermark at an attention weight of 0.11 near the diaphragm and 0.03 near the perihilar region. A deviation > 0.005 from this threshold, detected by the CEV-Net, will result in a TLRO adjustment. The average PSNR settled at 46 dB with an alignment of 0.94. Among the 500 images JPEG-compressed at quality factor 35, authentication signatures were retrieved with 97.4% accuracy, compared with 82-86% for the other pipelines. Only eight images showed anomalies on the tampering maps, and the remaining 492 had clear maps. Diagnosis compatibility with the triage AI was maintained at 3.2% of the pixel footprint.

5. Conclusion and Future Scope

We propose an explainable watermark system based on a 5-stage design: saliency synthesis, IWA-WE, CEV-Net, TLRO, and XSIF-Bench for the explanation, embedding, diagnostic verification, robustness enhancement, and evaluation, respectively. Using the SSWP prior, the watermark energy is designed not to overlay diagnostically important anatomical information by construction. The proposed IWA-WE encoder limits structural degradation within a diagnostically indifferent coefficient space. By applying counterfactual reasoning in [24] and [25], the CEV-Net constitutes a closed-loop diagnostic safety measure that is only available in the research on explainability. To guarantee that all robustness improvements do not affect explainability and diagnostic neutrality, we employed the TLRO optimizer and XSIF-Bench for performance enhancement and validation. The proposed framework is particularly suitable for PACS-integrated hospital networks, telemedicine practices, and regulatory certification. An explanation map is generated for a radiologist to confirm that authentication has no influence on diagnostic critical regions; quantitatively reproducible certified diagnostic safety is provided by the counterfactual divergence score, which a perceptual metric cannot provide. The system can be further extended in four aspects: (1) Adaptation of the proposed SSWP and IWA-WE into 3D information with a 3D saliency prior for the use of CT and PET. (2) Integration of a transformer-based diagnostics model into the CEV-Net to enable token-level counterfactual evaluation. (3) Replace the TLRO optimizer with a reinforcement learning algorithm in an online fashion to adjust latent variable actively. (4) Large-scale clinical implementation studies within hospitals network.

5.1 Limitations

Saliency-based interpretability approximates clinical reasoning but does not replicate it; misdirected saliency may create overly conservative or permissive embedding zones. The multi-stage pipeline carries computational costs that may exceed low-resource imaging environments. Robustness evaluation covered JPEG compression, Gaussian noise, motion blur, and resizing, but did not cover intentional adversarial attacks. Under geometric resize attacks specifically, Method [8] outperforms the proposed model (88.1% vs 84.3%) because SSWP spatial selectivity concentrates

watermark energy in zones disproportionately disrupted by resampling; scale-invariant wavelet-domain embedding will be explored in future work to close this gap. The CEV-Net guarantee is model-specific and requires recalibration when the diagnostic model changes. The three datasets represent a modest cross-section of global clinical diversity. The ablation study used fixed hyperparameter configurations; sensitivity analysis across a wider parameter range is needed. Performance comparisons may also vary depending on dataset-specific implementation settings used for baseline reproduction.

Acknowledgements

3 Scientific content, methodology, results, and conclusions were produced by and validated by humans. No AI tool is mentioned as an author. Author Contributions: Vamsi Sisti: Conceptualization, Methodology, Software, Formal analysis, Writing-Original Draft. M Kali Muthu: Supervision, Validation, Writing-Review and Editing, Project administration. D Vineela: Conceptualization, Methodology, Writing-Original Draft

Declarations:

The authors have no relevant financial or non-financial interests to disclose.No funding was received for conducting this study."

References

1. Latreche, B., Merrad, A., Benziane, A., Naimi, H., Saadi, S.: A robust dual-layer medical image watermarking scheme based on matrix factorization in the LWT domain for E-healthcare applications. Multimed. Tools Appl. 84(25), 29883-29913 (2024). https://doi.org/10.1007/s11042-024-20331-7
2. Awasthi, D., Srivastava, V.K.: ROI-based optimized image watermarking with real-time authentication. Clust. Comput. 28(7) (2025). https://doi.org/10.1007/s10586-025-05436-4
3. Yan, Q., Luo, Y., Wang, Z., Xi, J., Xia, G., Cai, Z.: Spatiotemporal attention-based real-time video watermarking. Data Min. Knowl. Discov. 39(5) (2025). https://doi.org/10.1007/s10618-025-01129-z
4. Farhat, A.A., Darwish, M.M., El-Gindy, T.M.: ResNet50 and logistic Gaussian map-based zero-watermarking algorithm for medical color images. Neural Comput. Appl. 36(31), 19707-19727 (2024). https://doi.org/10.1007/s00521-024-10121-5
5. Gupta, A.K., Chakraborty, C., Gupta, B.: Bio inspired optimization based secure model for watermarking of medical data. Multimed. Tools Appl. 84(33), 41009-41039 (2025). https://doi.org/10.1007/s11042-025-20740-2
6. Kanwal, S., Tao, F., Taj, R., Abbas, Q., Amin, F.: Discrete cosine transform and discrete wavelet transform based hybrid method for robust and blind medical image watermarking. Peer-to-Peer Netw. Appl. 18(5) (2025). https://doi.org/10.1007/s12083-025-02059-9
7. Naima, S., Boukhamla, A.Z.E., Narima, Z., Amine, K., Redouane, K.M., Sahu, A.K.: Secure and imperceptible frequency-based watermarking for medical images. Circuits Syst. Signal Process. 44(1), 196-217 (2024). https://doi.org/10.1007/s00034-024-02814-y
8. Awasthi, D., Khare, P., Srivastava, V.K.: ANFISmark: ANFIS-based secure watermarking approach for telemedicine applications. Neural Comput. Appl. 37(14), 8677-8698 (2025). https://doi.org/10.1007/s00521-025-11030-x
9. Chaudhary, H., Garg, P., Vishwakarma, V.P.: Enhanced medical image watermarking using hybrid DWT-HMD-SVD and Arnold scrambling. Sci. Rep. 15(1) (2025). https://doi.org/10.1038/s41598-025-94080-4
10. Fedoseev, V., Denisova, A.: A reversible medical image watermarking scheme for advanced image tampering detection. SN Comput. Sci. 5(4) (2024). https://doi.org/10.1007/s42979-024-02692-w
11. Singh, P., Devi, K.J., Nizami, T.K., Prakash, C.S., Thakkar, H.K., Hussain, S.A., Mallik, S.: Ensuring integrity and security of medical image transmission in IoMT using highly imperceptible and robust watermarking approach. Sci. Rep. 15(1) (2025). https://doi.org/10.1038/s41598-025-11023-9
12. Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L.: HiDDeN: hiding data with deep networks. In: Proc. ECCV, pp. 657-672 (2018). arXiv:1807.09937
13. Taj, R., Tao, F., Kanwal, S., Almogren, A., Altameem, A., Ur Rehman, A.: A reversible-zero watermarking scheme for medical images. Sci. Rep. 14(1) (2024). https://doi.org/10.1038/s41598-024-67672-9
14. Dey, A., Chowdhuri, P., Pal, P.: Integer wavelet transform based watermarking scheme for medical image authentication. Multimed. Tools Appl. 83(32), 78001-78022 (2024). https://doi.org/10.1007/s11042-024-18183-2
15. Hebbache, K., Aiadi, O., Khaldi, B., Benziane, A.: Blind medical image watermarking based on LBP-DWT for telemedicine applications. Circuits Syst. Signal Process. 44(7), 4939-4964 (2025). https://doi.org/10.1007/s00034-025-03023-x

16. Incetas, M.O., Kilicaslan, M.: Image watermarking based on spiking neural networks. Clust. Comput. 28(11) (2025). https://doi.org/10.1007/s10586-025-05582-9
17. Li, Q., Zhang, W., Liu, S., Yan, T., Du, J., Han, B.: Zero-watermarking of medical images based on improved lightweight neural network MobileNetV3. Circuits Syst. Signal Process. 44(7), 5235-5259 (2025). https://doi.org/10.1007/s00034-025-03062-4
18. Yan, F., Wang, Z., Hirota, K.: Dual medical image watermarking using SRU-enhanced network and EICC chaotic map. Complex Intell. Syst. 11(1) (2024). https://doi.org/10.1007/s40747-024-01723-6
19. Bouarroudj, R., Bellala, F.Z., Souami, F.: High capacity and reversible fragile watermarking method for medical image authentication and patient data hiding. J. Med. Syst. 48(1) (2024). https://doi.org/10.1007/s10916-024-02110-x
20. Abirami, R., Malathy, C.: Medical image security by crypto watermarking using enhanced chaos and fruit fly optimization algorithm with SWT and SVD. Multimed. Tools Appl. 83(27), 70451-70476 (2024). https://doi.org/10.1007/s11042-024-19019-9
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 128, 336-359 (2020). https://doi.org/10.1007/s11263-019-01228-7
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765-4774 (2017). arXiv:1705.07874
23. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Muller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE 109(3), 247-278 (2021). https://doi.org/10.1109/JPROC.2021.3060483
24. Thiagarajan, J.J., Thopalli, K., Rajan, D., Turaga, P.: Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Sci. Rep. 12, 597 (2022). https://doi.org/10.1038/s41598-021-04529-5
25. Atad, M., Schinz, D., Moller, H., et al.: Counterfactual explanations for medical image classification and regression using diffusion autoencoder. J. Mach. Learn. Biomed. Imaging 2, 2103-2125 (2024). https://doi.org/10.59275/j.melba.2024-4862
26. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. In: International Conference on Learning Representations (ICLR) Workshop (2015). arXiv:1412.6806
27. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. ECCV, LNCS vol. 8689, pp. 818-833. Springer (2014). https://doi.org/10.1007/978-3-319-10590-1_53
28. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD, pp. 1135-1144 (2016). https://doi.org/10.1145/2939672.2939778
29. Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82-115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
30. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proc. CVPR, pp. 2921-2929 (2016). https://doi.org/10.1109/CVPR.2016.319