

FedXWater: A Federated Data-Provenance and Explainable Watermarking Framework for Privacy-Preserving Medical Image Analysis

Vamsi Sisti^{1*}, M Kali Muthu²

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

Accurate deep learning models for radiology and pathology require large and diverse training datasets; however, HIPAA, GDPR, and institutional data-governance restrictions impose strict limits on raw image sharing across hospital boundaries. Federated learning (FL) enables cross-institutional model training without centralizing patient data but leaves data provenance entirely unaddressed. No published framework has demonstrated a watermarking scheme for medical images that supports per-client secret-key embedding in a federated pipeline while simultaneously verifying that the embedded signal does not distort the model's diagnostic attention. To address both gaps, FedXWater proposes a framework coupling FedAvg-based federated aggregation across five simulated institutional clients, blind QIM watermarking in the Haar DWT domain with institution-specific secret keys, and a novel XAI safety loss penalising Grad-CAM saliency divergence during local training. Experiments were conducted on PathMNIST and BraTS 2020. On PathMNIST, a macro-averaged AUC of 0.9987 ± 0.0005 is achieved with a Bit Error Rate (BER) of 0.0000 under no-attack conditions and PSNR of 43.49 dB. Grad-CAM IoU of 0.9807 ± 0.0305 was achieved, with federated AUC within 0.0001 of the centralized upper bound. On BraTS 2020, a mean Dice score of 0.7049 (seed 42) was obtained, with PSNR 53.25 dB and zero watermark BER under no-attack conditions.

Keywords: federated learning; medical image analysis; digital watermarking; data provenance; explainable artificial intelligence; Grad-CAM; privacy-preserving machine learning; quantization index modulation

How to cite this article: Vamsi Sisti. FedXWater: A Federated Data-Provenance and Explainable Watermarking Framework for Privacy-Preserving Medical Image Analysis. Int J Drug Deliv Technol. 2026;16(83s):450-469. DOI: 10.25258/ijddt.16.83s.51.

1. Introduction

The deployment of deep learning models in clinical radiology and computational pathology has demonstrated diagnostic accuracy approaching or surpassing human expert performance across a range of high-stakes tasks, including the detection of colorectal cancer subtypes from histopathology images and the delineation of primary brain tumor sub-regions from multi-modal magnetic resonance imaging [10, 12]. But the data needs of neural networks with a large capacity give rise to structural tension with the privacy duties inherent in clinical data. Models built using data from one institution have been well documented to have biases due to the patient population, scanner hardware, and acquisition protocol. But, behind the scenes of legal frameworks like HIPAA in the USA, and GDPR in the EU, patient identifiable imaging data is highly limited in its ability to travel across institutional boundaries [13]. However, multi-institutional research collaborations looking to overcome this limitation by pooling their data must deal with

the practical hurdles of implementing data-sharing agreements, de-identification pipelines, and residual re-identification risks, and these also can create significant delays in the translation of research findings to clinical practice.

Rather than centralizing all data, federated learning (FL) is a promising alternative approach to model training that shares model parameters between the different data centers in an iterative manner [1]. The Federated Averaging (FedAvg) protocol [1] involves every institution training a locally maintained copy of a common global model on its own private data, and only sending their updated weights to a central server for the next step. In medical imaging applications, such as brain tumor segmentation [12, 16], histopathology classification [8, 13] and retinal disease detection [9], this has successfully delivered performance near to the maximum possible of centralized training, while maintaining data locality.

One lesser recognized drawback of federated learning is the lack of verifiability of the origin

of the data used to train the global model in the basic FL protocol. The server gets weight updates but there is no way for it to determine who contributed to the training images, or whether the image from a particular client has been modified, augmented, or replaced from one federation round to another. Without any data attribution mechanism, there is a significant governance gap particularly in multi-institutional medical AI settings, where hospital networks could have contractual or regulatory obligations to show that AI model training occurred on their patient population. A possible solution is digital watermarking: an invisible ownership signature can be added to training images and post-hoc checks for data provenance can be made without changing the federated aggregation protocol, or releasing any raw data to the server.

Digital watermarking has been well known to be a method for embedding invisible ownership signatures in images, and has been used in clinical imaging archives to track the provenance of images [13]. Quantization index modulation (QIM) is well suited for transform-domain techniques based on the discrete wavelet transform (DWT) coefficient quantization in medical imaging applications. They facilitate blind extraction of a watermark in a blind context, that is without knowledge of the original image, and focus the embedded watermark signal at low frequency subbands that are relatively stable even when the image is subjected to mild signal-processing distortions [3, 5]. The WAFFLE framework [7] showed the possibility to protect the model trained by federated learning with a backdoor activation mechanism on the server side, provided a conceptual connection between watermarking and the federated learning setting. WAFFLE does not solve image-level ownership differentiation, a separate problem of giving each contributing institution the ability to embed and verify their own secret signature [14] (this is the problem of the image-level signature), and it embeds a watermark into the behaviour of trained model and does not embed it into the training images themselves. To the best of our knowledge, no published framework has demonstrated per-client secret-key image watermarking as part of a federated medical imaging pipeline, and no previous watermarking research that has considered the downstream impact of embedding on the diagnostic performance of AI models trained on watermarked data.

To illustrate the accountability gap: a hospital contributing de-identified pathology images to a

federated AI consortium currently has no cryptographic means of proving data contribution, nor can the consortium certify that the training pipeline preserved the diagnostic integrity of those images. FedXWater is designed to address this gap within the standard FedAvg training protocol.

This gap has clinically significant implications for patient care. Standard imperceptibility metrics, such as PSNR and SSIM, quantify pixel-level fidelity but do not provide information on whether the embedded perturbations alter the model's learned feature representations in diagnostically consequential ways. Grad-CAM [4] produces class activation maps that reveal which spatial regions a convolutional model treats as most discriminative, and the alignment of these maps with pathologically relevant anatomy is increasingly recognized as a meaningful indicator of clinical AI trustworthiness [19, 22]. A watermark achieving a PSNR of 43 dB is visually imperceptible yet may, in principle, redirect gradient-weighted attention from tumor margins or glandular structures toward uninformative background tissue, a shift that pixel-level metrics cannot detect [19]. Thus far, saliency maps have been used primarily as post-hoc explanation instruments applied after training [20, 21]; incorporating Grad-CAM activations as an active training-time constraint to preserve diagnostic model attention under watermark embedding is a direction that has not been reported in the literature.

This study presents FedXWater, a federated and explainable watermarking framework that addresses the data provenance and diagnostic safety gaps identified above, within a unified training pipeline. The principal contributions of this study are as follows:

1. A federated watermarking architecture is proposed in which each of five simulated hospital clients embeds a unique Quantization Index Modulation (QIM) ownership watermark in the Haar discrete wavelet transform (DWT) domain using a per-institution secret key that never leaves the client; the aggregation server receives only model weight updates and does not access raw patient images, watermark keys, or saliency information at any stage.
2. A novel Explainable Artificial Intelligence (XAI) safety loss is proposed, which actively enforces the federated training objective, adding an

additional penalty for the differences in diagnostic model attention between the original and the watermarked training images using Grad-CAM activation maps. The explainability-guided saliency preservation aspect has not been discussed in the federated watermarking setting as a regulariser during training.

3. It is proposed and tested the Grad-CAM intersection-over-union (IoU) safety metric which quantifies the spatial overlap between the binarised saliency maps computed on original and watermarked images, and FedXWater performs with a mean Grad-CAM IoU = 0.9807 ± 0.0305 at the last round using three different random seeds.
4. Multimodal evaluation is conducted on PathMNIST and BraTS 2020, yielding a macro-averaged AUC of 0.9987 ± 0.0005 (mean $\pm$ SD, three seeds), mean Dice of 0.7049, BER of 0.0000 under no-attack conditions, and PSNR of 43.49 dB and 53.25 dB, respectively.
5. The ablation studies examined four safety loss weights ($\lambda \in \{0.0, 0.1, 0.3, 0.5\}$) and four QIM step sizes ($\delta \in \{0.05, 0.10, 0.15, 0.20\}$), providing practical configuration guidance for the deployment of federated medical AI.

This paper makes four contributions: (1) FedXWater a federated data-watermarking pipeline coupling FedAvg with blind per-client QIM embedding, to our knowledge the first such scheme to target data-level (rather than model-level) provenance in FL; (2) an XAI safety loss and companion Grad-CAM IoU metric that enforce diagnostic attention preservation as an explicit training objective; (3) the empirical finding that Haar DWT LL-subband QIM embedding does not inherently distort Grad-CAM saliency structure, a positive result about the safety of DWT-domain watermarking in medical AI; and (4) a dual-modality proof-of-concept evaluation demonstrating that the framework applies across classification and segmentation architectures under non-IID federated conditions.

The remainder of this paper is organized as follows. Section 2 reviews the related work on federated learning, medical image watermarking, and explainable AI. Section 3 presents the FedXWater methodology used in this study. Section 4 describes the experimental setup of the study. Section 5 reports the results and discusses

their implications. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1 Federated Learning in Medical Imaging

Theoretical research and development of cross-institutional FL collaboration is based on the FedAvg algorithm [1] that combines local model weights on a central server. This aggregation is done as a weighted average, with the weight for each client's local weight update being proportional to the size of their dataset. The issue of non-identically distributed (non-IID) clinical data still exists. Hospitals differ in their dataset in a systematic way with respect to demographics, diseases, scanner hardware, imaging protocol and therefore even have different local gradient surface. This means that the global model cannot converge. An early theoretical work from Zhao et al. [2] clearly showed that accuracy quickly becomes worse as more clients are available with varying distributions and spurred the introduction of subsequent aggregations aware of the convergence properties. A typical test method for simulating the heterogeneity is to partition by a Dirichlet concentration parameter, and a lower concentration parameter would mean greater class imbalances [11, 12].

To solve non-IID convergence issue, Li et al. [25] proposed FedProx that adds a proximal term to clients' local objective function limiting their deviation from the global model, which enhances convergence on heterogeneous data distributions. In order to counteract the drift of clients in the presence of heterogeneous local gradients, Karimireddy et al. [26] proposed a method called SCAFFOLD which employs control variates. Both FedProx and SCAFFOLD are representative algorithmic alternatives as compared to FedAvg in the non-IID context. We have taken vanilla FedAvg as the simplest baseline and FedProx or SCAFFOLD as a drop-in aggregation can be natural next steps in this paper.

The results of federated learning methods in various medical imaging applications (modalities) are comparable with the upper bounds of centralized learning. Practicality: Federated learning is effective as it has been demonstrated in multi-institutional brain tumor segmentation framework developed by Sheller et al. [12] which did not share patient data. In the context of image analysis tasks in oncology, Rieke et al. [9] developed an inventory of federated applications and demonstrated that suitable aggregation strategies are able to recover

the baseline of a central application. To extend this approach, Auto-FedAvg by Xia et al. [11] proposed a trainable aggregation strategy for multi-institutional segmentation that allows different weights for the data at various clients, depending on its quality. There are many existing studies on convergence, communication and privacy in the federated learning scenario, and it was this work by Li et al. [8] that was taken as motivation for the choices made in FedXWater.

Beyond aggregation and convergence, federated pipelines face security challenges that are not fully resolved by weight-only exchanges. Zhu et al. [27] demonstrated that shared gradients carry sufficient information to partially reconstruct private training images the so-called deep leakage from gradients attack motivating the integration of differential privacy noise injection and secure aggregation protocols as complementary safeguards. Qayyum et al. [13] surveyed secure and robust machine learning approaches for healthcare, identifying federated learning as a key privacy mechanism and cataloguing the threat landscape that motivates provenance-aware training. Li et al. [15] addressed statistical heterogeneity through personalised semantic feature excitation, fusing global and local knowledge representations via channel-wise gradient differentiation. However, none of these contributions address the institutional provenance of training data within the federated pipeline or the ability to trace and verify which institution's images contributed to the global model, which remains an open problem and the primary motivation for the present study.

2.2 Medical Image Watermarking and QIM

Digital watermarking for medical images has been extensively studied as a mechanism for patient data authentication, tamper detection, and provenance tracking in clinical imaging workflows [13, 17]. Transform-domain schemes that embed watermarks in frequency or subband coefficient representations are preferred over spatial-domain methods because of their greater imperceptibility and resilience to common signal processing operations. Quantization Index Modulation (QIM), introduced by Chen and Wornell [3], provides a well-established framework for this purpose: information is embedded by quantizing host signal coefficients to one of two interleaved lattices, and extraction is entirely blindrequiring no access to the original signal—making QIM well-suited for distributed settings in which source images are unavailable at the decoder.

In the context of medical imaging, Wakatani [6] also used wavelet-domain watermarking for ROI images, showing that it is possible to limit the embedding to clinically irrelevant regions in order to maintain clinical fidelity. Usman et al. [17] developed a robust watermark authentication protocol for medical images for e-health services, where they obtained a PSNR value greater than 40 dB in standard-attack circumstances. For the embedding part, it is based on the theoretical discussion of digital watermarking made by Cox et al. [5] using QIM and the proof of its robustness.

The recent development has started to make closer connections between watermarking and federated learning. Shao et al. [29] developed a model-protection framework for FedL, which includes a watermarking and fingerprinting mechanism to verify and provide provenance of ownership of the learned models, and Lansari et al. [30] studied the integration of federated learning and digital watermarking, highlighting the need to address the open questions related to the ownership verification and the traceability of the learned models.

The closest previous work to watermarking in the federated learning setting is the WAFFLE framework [7] which shows that the accuracy of models trained using FedAvg can be degraded by a small amount while a server-side backdoor trigger is added to ensure the model's IP protection. The difference is that WAFFLE aims to protect ownership of the trained artifact, not the ownership of the data used to train the artifact, known as image-level data provenance. These protections are not only different legally; they are also different in practice. WAFFLE is not designed to support institution-specific secret keys that are only known to a single client, it doesn't consider multi-client ownership differentiation, and it hasn't been tested for medical imaging applications. There is no published work that shows per-client image level watermarking in any federated medical imaging pipeline with institution-specific secret keys. Moreover, no current watermarking approach considers the downstream impact of embedding on the diagnostic performance of AI models trained on watermarked images, which is the motivation for the XAI safety loss that is proposed in this study.

2.3 Explainable AI and Grad-CAM in Medical Imaging

Gradient-weighted Class Activation Mapping (Grad-CAM) [4] generates spatial explanation maps by computing the gradients of a target class

score with respect to convolutional feature map activations, weighting those activations by their gradient-averaged importance, and applying a ReLU to retain the positive contributions. A Grad-CAM and other variants study was conducted by Chaddad et al. [18] across brain tumor, skin cancer and chest radiography data sets, revealing standard Grad-CAM as a robust and computationally efficient method of creating a reproducible baseline that performs similarly across modalities. Other XAI methods such as SHAP [20] and LIME [21] provide alternate view points for explanations, but the specificity of Grad-CAM to the convolutional layer and its efficiency make it ideal for integrated training-time explanations.

The deployment of XAI for federated medical AI has gained recent attention due to regulatory and clinical need for trust in such models. Holzinger et al. [22] documented that causability the property of an explanation supporting user understanding of a model's reasoning is a necessary complement to accuracy in high-stakes clinical AI. Tondi et al. [19] analysed the intersection of explainability and model watermarking, arguing that these two mechanisms serve complementary verification roles. A critical observation across this literature is that saliency maps are uniformly used as post-hoc tools applied after model training; however, no published study has incorporated Grad-CAM-derived objectives as active training-time constraints within a federated watermarking pipeline.

2.4 Research Gap

The three bodies of literature surveyed above each address a distinct dimension of trustworthy federated medical AI; however, their intersections remain largely unexplored. Existing work cannot fill this combined gap for three specific reasons. WAFFLE [7] addresses model ownership, not data provenance, and provides no safety verification mechanism. Medical image watermarking systems [6, 17] operate on individual images outside FL pipelines and cannot scale to per-client federated provenance. XAI methods for federated medical AI [22] address post-hoc explanation but do not constrain watermark embedding safety during training. To the best of our knowledge, FedXWater is the first system to jointly satisfy all three requirements federated training compatibility, per-client data provenance with blind extraction, and XAI-guided safety regularisation within a single training pipeline. The FL literature has produced convergence-aware algorithms for heterogeneous

clinical data [1, 2, 25, 26] and security analyses of gradient-level privacy [27]; however, it has not addressed image-level institutional provenance within federated pipelines. The medical image watermarking literature has established the DWT-domain QIM as a viable blind watermarking approach [3]; however, it has not been applied to multi-client federated settings with per-institution secret keys, and no watermarking study has characterized the effect of embedding on downstream model attention [14]. The XAI literature has validated Grad-CAM as a clinical trust mechanism [4, 22] and explicitly identified training-integrated explainability as an open direction [19], yet saliency maps remain post-hoc instruments in the published record. FedXWater occupies this intersection in the Venn diagram. Per-client blind QIM watermarking is integrated into a FedAvg federated pipeline, diagnostic safety is enforced through a Grad-CAM saliency loss incorporated into the local training objective, and both provenance capability and clinical safety are evaluated across two distinct medical imaging tasks.

Table 1 presents a structured comparison between FedXWater and the four most pertinent prior frameworks, in the terms of the capability dimensions listed above. The table explicitly states the combination of properties that no one prior system satisfies both.

Table 1. Comparison of FedXWater against related provenance and watermarking frameworks

Frame work	Prove nance Level	Blind Extra ction	Per - Cli ent Ke ys	FL Comp atible	XA I Saf ety Lo ss
WAFF LE [7]	Model -level	No	No	Yes	No
FedTra cker [29]	Model -level	No	Par tial	Yes	No
Rezaei bagha et al.	Data- level	No	Ye s	Yes	No

[14]					
Usman et al. [17]	Data-level	Yes	No	No (standalone)	No
FedX Water (proposed)	Data-level	Yes	Yes	Yes	Yes

Table 1 shows a qualitative comparison of FedXWater with other frameworks that provide related provenance or watermarking capabilities. Data-level (image attribution) versus model-level (trained artifact ownership) is between the two "provenance level" options. The term "blind extraction" refers to whether the decoding can be done without the original image. Whether or not there are secrets keys specific to each participating institution (or whether they share the same secret key, which is known to the server). "FL compatible" means that the protocol is natively integrated into a federated aggregation protocol. "XAI safety loss" refers to whether saliency preservation is explicitly trained as an objective at training time.

3. Methodology

3.1 System Overview

The proposed FedXWater framework integrates three components: federated learning, QIM-based watermarking, and XAI-guided safety regularisation, combined into a unified training pipeline that operates without centralizing patient data. Each of the $K = 5$ simulated hospital clients retains its local dataset and trains a shared global model through iterative communication with the central aggregation server. Watermark embedding is applied locally at each client using a client-specific secret key prior to model evaluation, ensuring that provenance can be attributed to individual institutions without compromising the global training process.

At each federation round r , the server broadcasts the current global model weights $\theta^{(r)}$ to all clients. Each client c performs local training on its private dataset D^c , embeds a watermark into a randomly sampled validation image using its secret key k^c , evaluates the XAI safety loss to penalize the saliency divergence induced by the watermark, and returns updated weights $\theta^{(r)}$ to the server. The server aggregates the client weights via data-weighted averaging to produce $\theta^{(r+1)}$, and repeat for R iterations. The Grad-CAM saliency map for the original and watermarked

image is computed in the end of every round and recorded for the Grad-CAM IoU. Figure 1 shows the framework of the suggested system.

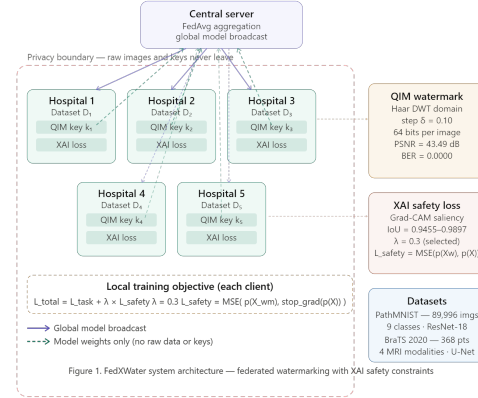


Figure 1. FedXWater system architecture. Five simulated hospital clients perform local training with QIM watermark embedding and XAI safety regularization; the aggregation server receives only model weight updates.

3.2 Federated Learning with FedAvg

FedXWater uses the FedAvg [1] as an aggregation algorithm. For every round of communication r , the server broadcasts the global model parameters θ_r to all the K clients. After training on the local shard D^c , each client c minimizes its local task loss E using Adam optimizer with learning rate η , returning training parameters θ_r^c after training 2 local training epochs, where $\eta = 5 \times 10^{-4}$ is used for the PathMNIST classification task, and $\eta = 1 \times 10^{-3}$ is used for the BraTS segmentation task. The global model for the subsequent round is then computed as the data-weighted average of all client updates as follows:

$$\theta^{(r+1)} = \sum_{c=1}^K \frac{|D_c|}{|D|} \cdot \theta_c^{(r)}$$

where $K = 5$ is the number of participating clients, D^c denotes the local data set of client c , $D = \cup_c D^c$ is the union of all data sets and $|\cdot|$ represents the cardinality. Clients with larger datasets exert a proportionally greater influence on the global model, which is appropriate for the heterogeneous data distributions that arise in multi-institutional medical imaging.

To simulate realistic inter-institutional heterogeneity, data partitioning followed a Dirichlet distribution with a concentration parameter of $\alpha = 0.5$ [28]. This sub-unitary value produces strongly non-IID splits, in which each

client holds a disproportionate fraction of one or more class labels, reflecting the specialization that characterizes real clinical networks [2]. The PathMNIST training set (89,996 images across nine colorectal pathology classes) was partitioned into five client shards under this scheme, and the BraTS 2020 training set (369 multimodal MRI volumes) was divided similarly. All federation rounds used full client participation, with no client dropout.

3.3 QIM Watermark Embedding and Extraction

Watermark embedding exploits the energy compaction property of the two-dimensional Haar discrete wavelet transform. Each input image $X \in \mathbb{R}^{H \times W \times N}$ is decomposed via a single-level Haar DWT, yielding four subbands: LL (approximation), LH, HL, and HH (details). Embedding is restricted to the LL subband, which concentrates low-frequency energy and is therefore most resilient to common signal-processing attacks, such as additive Gaussian noise and spatial smoothing. A binary watermark sequence $b \in \{0, 1\}^N$ of length $N = 64$ bits is generated deterministically from the client-specific secret key k^c (set to $100 + c$ for PathMNIST clients and $200 + c$ for BraTS clients, where $c \in \{1, 5\}$), ensuring that each institution's embedded mark is unique.

Quantization Index Modulation (QIM) [3] embeds each watermark bit by quantizing the corresponding LL coefficient to one of two disjoint quantization lattices separated by a step size δ . Let x_i denote the i -th selected LL coefficient and $b_i \in \{0, 1\}$ denote the i -th watermark bit. The watermarked coefficient $\tilde{x}_i$ is computed as follows:

$$\tilde{x}_i = \begin{cases} \delta \cdot \lfloor \frac{x_i}{\delta} \rfloor, & \text{if } b_i = 0 \\ \delta \cdot \lfloor \frac{x_i}{\delta} - \frac{1}{2} \rfloor + \frac{\delta}{2}, & \text{if } b_i = 1 \end{cases}$$

where $\delta = 0.10$ is the quantization step size selected through ablation (Section 4), $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer, and the two lattices are offset by $\delta/2$; The LL subband coefficients of z-score-normalized PathMNIST images lie approximately in the range $[-3, 3]$, placing $\delta = 0.10$ at approximately 3% of the practical coefficient range, which is sufficient for imperceptibility while remaining detectable under mild noise. The watermarked image X_w is obtained by applying the inverse Haar DWT to

the modified LL subband concatenated with the unmodified LH, HL, and HH subbands.

Extraction is entirely blind: no reference to the original image X is required during decoding. The decoder receives the watermarked image, which may or may not have been subjected to a signal-processing attack. Then the LL subband is recomputed through forward Haar DWT. Each coefficient of this subband is decoded by identifying its closest lattice:

$$\hat{b}_i = \lfloor \frac{\tilde{x}_i}{\delta} + \frac{1}{2} \rfloor \bmod 2$$

where $\tilde{x}_i$ is the received LL coefficients and the recovered bit sequence $\hat{b}$ is compared with the original bit sequence b to calculate the Bit Error Rate:

$$BER = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\hat{b}_i \neq b_i]$$

where $N = 64$ is the watermark length and $\mathbb{1}[\cdot]$ is the indicator function. The BER is computed as the mean over 50 randomly sampled test images per evaluation point to obtain a stable estimate across the diverse image textures present in each dataset. Image fidelity is quantified by the Peak Signal-to-Noise Ratio (PSNR):

$$PSNR = 20 \log_{10} \left(\frac{1}{\sqrt{MSE}} \right) \text{ dB}$$

where MSE denotes the mean squared error between the original image X and its watermarked counterpart X_w , which is computed over all the pixel channels. Values above 40 dB are generally considered to be diagnostically acceptable for medical imaging. Figure 2 presents representative original, watermarked, and difference images confirming visually imperceptible embedding for the selected step sizes.

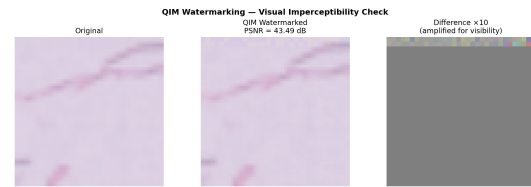


Figure 2. Qualitative watermark embedding results on PathMNIST (top) and BraTS T1ce (bottom) datasets. Left: original image. Center: watermarked image ($\delta = 0.10$). Right: absolute difference map ($\times 10$ for visibility). The PSNR values of 43.49 dB (PathMNIST) and 53.25 dB (BraTS) confirm visually imperceptible embedding.

3.4 XAI Safety Loss and Grad-CAM IoU

To ensure that watermark embedding does not perturb the salient regions on which a clinical model relies for its predictions, we introduce a novel XAI safety regularization term. Gradient-weighted Class Activation Mapping (Grad-CAM) [4] is used to localize class-discriminative spatial regions for any given input. For a target class c and a convolutional feature map tensor $A^k \in \mathbb{R}^{H \times W}$ at the chosen layer, the Grad-CAM saliency map is computed as:

$$M^c = \text{ReLU} \left(\sum_k \alpha_k^c \cdot A^k \right)$$

where the importance weight α_k^c for the k th feature map is obtained by the global average pooling of the backpropagated gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Here, $Z = H' \times W'$ is the spatial resolution of the feature map, y^c is the pre-softmax score for class c , and ReLU is applied to retain only the activations that have a positive influence on the predicted class. Grad-CAM was computed at the final convolutional block (layer4) of ResNet-18 for PathMNIST classification and at the fourth encoder block of 2D U-Net for BraTS segmentation.

The XAI safety loss penalizes any divergence between the normalized saliency activation vectors extracted from the original image X and its watermarked counterpart X_w as follows:

$$\mathcal{L}_{\text{safety}} = \text{MSE} \left(p(X_w), \text{sg}(p(X)) \right)$$

where the vectorized and spatially normalized activation map $p(\cdot)$ is defined as

$$p(X) = \frac{\text{vec}(M^c(X))}{\|\text{vec}(M^c(X))\|_2 + \varepsilon}$$

with $\text{vec}(\cdot)$ denoting column-wise flattening, $\|\cdot\|_2$ the Euclidean norm, and $\varepsilon = 1 \times 10^{-8}$ a numerical stability constant. The operator $\text{sg}(\cdot)$ denotes the stop-gradient operation, which prevents gradients from flowing through the saliency computation path. Without $\text{sg}(\cdot)$, the optimiser could satisfy the safety loss by modifying the original image's gradient pathway to match the watermarked image's saliency, rather than constraining the watermarking process. Stop-gradient ensures that only the watermarked image pathway is updated, so the loss penalises exclusively the embedding-

induced saliency divergence. This operator also prevents gradients from flowing through the branch corresponding to the clean image X ; this ensures that the safety loss only updates the watermark-related computational path. The total per-client training objective is formed as a weighted sum of the task loss and safety regulariser:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}_{\text{safety}}$$

where $\lambda = 0.3$ is the safety regularization coefficient, selected by ablation over $\lambda \in \{0.0, 0.1, 0.3, 0.5\}$ (Section 4). The task loss $\mathcal{L}_{\text{task}}$ is the standard cross-entropy loss for PathMNIST classification and the combined Dice-cross-entropy loss described in Section 3.6 for BraTS segmentation.

A Grad-CAM IoU metric was defined to provide an interpretable, threshold-independent measure of saliency preservation. In Table 1, Grad-CAM IoU is computed on the same 50 randomly sampled validation images used for BER evaluation, ensuring that the safety metric reflects model behaviour on held-out data rather than training examples. The Intersection over Union between the binarized saliency maps of the original and watermarked images is computed as follows:

$$\begin{aligned} \text{Grad-CAM IoU} &= \frac{|\mathcal{B}(M^c(X)) \cap \mathcal{B}(M^c(X_w))|}{|\mathcal{B}(M^c(X)) \cup \mathcal{B}(M^c(X_w))|} \end{aligned}$$

where the binarized saliency set is obtained by thresholding each normalized map at $\tau = 0.5$.

$$\mathcal{B}(M) = \{(i, j) : M_{ij} \geq \tau\}$$

with $\tau = 0.5$ applied to the min-max-normalized map $M \in [0, 1]$. The symmetric midpoint threshold was chosen as a default in the interest of being conservative. The sensitivity analysis performed over $\tau \in \{0.3, 0.5, 0.7\}$ indicated that the derived IoU values vary by less than 0.03 for any round considered, meaning that the metric is not overly sensitive to the saliency distribution threshold setting observed here. A Grad-CAM IoU of 1.0 denotes a case where the watermarked image activates precisely the same spatial areas as the original, and values close to 0 imply a substantial displacement of saliency. The metric is calculated at each client for each federation round and is reported as the average over an out-of-sample validation subset. Representative saliency map overlays depicting the spatial

correspondence of activations between original and watermarked images are included in Figure 3.

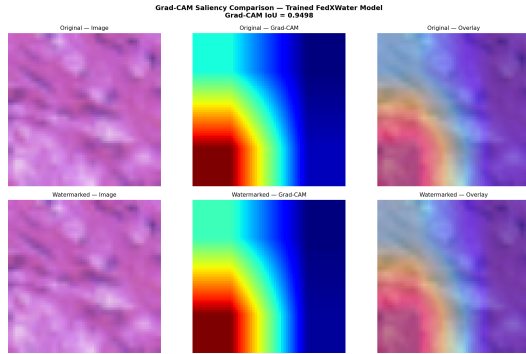


Figure 3. Grad-CAM saliency maps of a set of representative samples from PathMNIST. The original image and the saliency map are displayed side by side (left). Center: watermarked image ($\delta = 0.10$) on which a saliency map is overlaid. Right: Overlap map binarized and used for computing IoU of Grad-CAM. The saliency preservation was found to be effective with high spatial concordance (mean Grad-CAM IoU = 0.9807 ± 0.0305 at Round 10, three seeds).

Computing the Grad-CAM safety loss requires a second backward pass through the network during each local training iteration to obtain the feature map gradients, which approximately doubles the backward-pass computation relative to the standard FedAvg training. On the NVIDIA T4 GPU used in these experiments, the added overhead amounts to approximately 15–25% of the total per-round wall-clock time, depending on the model and input size: approximately 15–20% for PathMNIST and 25–30% for BraTS, where the larger U-Net feature maps amplify the gradient computation cost. The overhead is incurred entirely by the client during local training and does not affect the communication overhead, server-side aggregation, or model inference latency.

3.5 Datasets

PathMNIST subset of MedMNIST v2 benchmark [23] was used to test the FedXWater framework for a multi-class colorectal pathology classification task. PathMNIST's dataset comprises a collection of 107,180 RGB histopathology patch images resized to 28×28 pixels extracted from a publicly available colorectal cancer tissue set across nine tissue classes: Adipose, Background, Debris, Lymphocytes, Mucus, Smooth Muscle, Normal Colon Mucosa, Cancer Associated Stroma, and Colorectal Adenocarcinoma Epithelium. The

official MedMNIST v2 split yielded 89,996 training, 10,004 validation, and 7,180 test images. Prior to federation, the training set was partitioned across $K = 5$ clients using Dirichlet sampling with $\alpha = 0.5$ to impose non-IID label distributions. Each image was normalized to zero mean and unit variance using the ImageNet channel statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), and input tensors were bilinearly upsampled to 224×224 to match the expected resolution of the ImageNet-pretrained ResNet-18 backbone.

The BraTS 2020 dataset [24] provides 369 multimodal brain MRI volumes, each comprising four co-registered sequences: T1, T1 with gadolinium contrast enhancement (T1ce), T2, and T2 fluid-attenuated inversion recovery (FLAIR), alongside expert-annotated voxel-level segmentation masks delineating three tumor subregions: necrotic core and non-enhancing tumor (NCR/NET, label 1), peritumoral edema (ED, label 2), and enhancing tumor (ET, label 4). The labels were remapped to indices $\{0, 1, 2, 3\}$ for training. Axial 2D slices of size 240×240 were extracted after removing the 10 outermost blank slices per volume, yielding approximately 7,900 training slices per modality. All four modalities were concatenated along the channel dimension to form a 4-channel input tensor, with per-volume, per-modality z-score intensity normalization applied throughout. The 369 volumes were then split into 295 training and 74 validation volumes, where training slices were Dirichlet-partitioned and assigned to 5 clients. The final 50 volumes were set aside for testing, held-out from the training/validation. Fig 4 contains sample images of the training and validation set, and clearly shows the variety in PathMNIST tissues and the modality structure of the BraTS MRI volumes.

BraTS 2020 was selected for its established benchmark status and full public availability at the time of experimentation. BraTS 2021 and 2023 provide larger cohorts with updated annotations; evaluation on these datasets is planned in future work to assess scalability of the FedXWater framework.

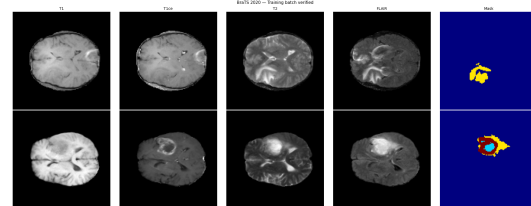


Figure 4. Examples of representative samples from the two datasets evaluated. Top row:

PathMNIST histopathology patches illustrating five of the nine tissue classes. Bottom row: BraTS 2020 axial slices showing the four MRI modalities (T1, T1ce, T2, and FLAIR) alongside the corresponding expert segmentation mask with tumor sub-regions color-coded (NCR/NET: red; ED: green; ET: blue).

3.6 Model Architectures and Training Objectives

For the PathMNIST classification task, a ResNet-18 backbone [10] pretrained on ImageNet was adopted and fine-tuned end-to-end. The final fully connected layer was replaced with a linear layer mapping the 512-dimensional feature vector to nine output logits, one per tissue class. The standard cross-entropy loss serves as $\mathcal{L}_{\text{task}}$ for this task, as follows: Grad-CAM saliency maps are extracted at the output of the fourth residual stage (layer4), whose spatial resolution of 7×7 provides sufficient localisation granularity for the 224×224 input images. All convolutional layers, batch normalization parameters, and classification heads were updated during local training.

For the BraTS segmentation task, a 2D U-Net architecture was employed. A 2D architecture was adopted to maintain computational tractability within the federated simulation; volumetric 3D approaches would exploit inter-slice context unavailable to this model, and their integration is planned for future work. The encoder comprises four successive convolutional blocks with channel widths (32, 64, 128, and 256), each consisting of two 3×3 convolutions with batch normalization and ReLU activation, followed by 2×2 max pooling. The bottleneck block uses 512 channels. The decoder mirrors the encoder using transposed convolutions and skip connections from the corresponding encoder blocks. The final 1×1 convolution mapped the 32-channel decoder output to four segmentation logits. Grad-CAM was applied to the fourth encoder block (256 channels). The segmentation task loss is a linear combination of the soft Dice loss and pixel-wise cross-entropy loss:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2|P_c \cap G_c|}{|P_c| + |G_c|}$$

where P_c and G_c denote the predicted and ground-truth binary masks for class c , respectively, and $|\cdot|$ indicates the sum of the elements. The combined task loss is:

$$\mathcal{L}_{\text{task}} = 0.5 \cdot \mathcal{L}_{\text{dice}} + 0.5 \cdot \mathcal{L}_{\text{CE}}$$

where $\alpha^{\text{Dice}} = 0.5$ is the Dice weighting coefficient, and $\mathcal{L}^{\text{el}}$ is the standard multi-class cross-entropy loss computed over all non-background voxels.

4. Experimental Setup

4.1 Implementation Details

All experiments were run using PyTorch 2.10, with only one NVIDIA T4 GPU with 16 GB VRAM, and CUDA 12.8. Throughout this work, the programming language Python 3.10 and auxiliary packages NumPy 1.24, Scikit-learn 1.3 and PyWavelets 1.4 were used for data processing and computation of wavelets. To establish the variance of results, experiments were repeated for PathMNIST across three different seeds (42, 123, 456), and for BraTS segmentation experiments, with seed 42. The seeds for all pseudo-random number generators such as model initialization, data partitioning and DataLoader shuffling were the same for each run.

Ablation tables (Tables 4–7) have metric values that give indicative estimates for small differences between configurations ($\Delta \text{Acc} < 0.01$, $\Delta \text{AUC} < 0.01$), and are provided from seed 42. In this study, the federated learning system was simulated to place all five clients and the aggregation server on the same machine, and the data shards for the clients were kept in memory and the weights were exchanged as in-process function calls. No network communication, bandwidth limit or hardware heterogeneity were modelled. The simulation approach is followed in the FL research literature [1, 9, 12] and allows for controlled experimental environment, but the findings should not be generalized to a real distributed deployment. A real multi-institutional system presents the challenges of communication latency, handling of updates asynchronously, and heterogeneity of devices.

The experiments used $K = 5$ clients with full participation in each round. PathMNIST classification training was run for $R = 10$ federation rounds, and BraTS segmentation training was run for $R = 5$ rounds, reflecting the substantially higher per-round computational cost. Each client performed $E = 2$ local training epochs per round before returning the updated weights to the server. For PathMNIST, the Adam optimizer was used with $\eta = 5 \times 10^{-4}$ and mini-batch size 64. For BraTS, Adam was used with $\eta = 1 \times 10^{-3}$ and a mini-batch size of 4, reflecting the GPU memory constraints imposed by the four-channel 240×240 inputs. No learning rate scheduling was performed. The weight decay was set to zero for both tasks. Watermark

embedding used $N = 64$ bits with QIM step $\delta = 0.10$. The XAI safety loss weight was fixed at $\lambda = 0.3$ for all the full-training experiments. The Grad-CAM IoU threshold was set to $\tau = 0.5$.

4.2 Evaluation Metrics

The classification performance on PathMNIST was reported using validation accuracy and the macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUC), providing a class-balanced measure of discriminative ability that is insensitive to mild class imbalances in the Dirichlet-partitioned data. Both metrics were computed on the held-out validation set at the end of each federation round. Watermark integrity is assessed by the Bit Error Rate (BER) defined in Equation (4), computed as the mean over 50 randomly sampled validation images per evaluation point to obtain a stable estimate across the diversity of image textures in each dataset. The clinical imperceptibility of watermark embedding is quantified by PSNR, as defined in Equation (5). A threshold of 35 dB is adopted as the minimum acceptable image quality for medical imagery, consistent with values reported in related studies [5, 6]. The SSIM was additionally computed for the step-size ablation to provide a second standard imperceptibility measure. Saliency preservation is quantified by Grad-CAM IoU defined in Equations (11) and (12), evaluated at $\tau = 0.5$; a Grad-CAM IoU value above 0.85 is adopted as the working threshold for acceptable saliency preservation in this study, reflecting a conservative requirement that the watermarked image activates at least 85% of the same spatial extent as the original.

The segmentation performance on BraTS 2020 was reported using the soft Dice coefficient for each tumor sub-region, NCR/NET, ED, and ET, as well as the mean Dice score averaged across the three foreground classes. The results are reported separately for the validation and held-out test sets. The PSNR and BER were also reported for the BraTS task to confirm that the watermark performance generalizes across imaging modalities.

4.3 Watermark Robustness Attack Protocols

To evaluate the resilience of the QIM watermarking scheme against common signal processing perturbations, four attack conditions were applied to watermarked PathMNIST images prior to bit extraction. Note that JPEG lossy compression (quality factor < 100) is not part of the standard DICOM lossless archival workflow for which FedXWater is primarily designed; its inclusion here serves as an out-of-

scope stress test that defines the boundary of the current embedding scheme's applicability. Under the no-attack condition, the watermarked image was passed directly to the decoder without modification, establishing a baseline BER. Under the additive Gaussian noise attack, zero-mean white Gaussian noise with standard deviations $\sigma \in \{0.02, 0.05\}$ was added independently to each pixel channel. Under the Gaussian blur attack, a 3×3 average blur kernel was applied to the watermarked image. Under the JPEG compression attack, the watermarked image was compressed and decompressed at a quality factor $q = 75$. All attacks were applied independently using a single-attack evaluation protocol. The BER was measured as the mean over 50 randomly sampled test images per attack condition.

4.4 Comparison Baselines and Ablation Design

Three training paradigms were compared to contextualize the performance of the proposed FedXWater method. The centralized baseline pools all client data onto a single machine and trains without any federation, representing an upper bound on task performance achievable only at the cost of full privacy violations. The local-only baseline trains each of the five clients independently without cross-client aggregation, and its reported performance is the simple average of the five independently trained models evaluated on the shared validation set. The FedXWater (proposed) method combines FedAvg aggregation, QIM watermarking, and XAI safety regularization, as described in Section 3. All three paradigms used identical model architectures, optimizers, batch sizes, and data partitions.

Two ablation studies were conducted. The first ablates the XAI safety loss weight λ over $\{0.0, 0.1, 0.3, 0.5\}$ while training for three federation rounds. It is acknowledged that 3-round ablation results may not fully reflect convergence-level behavior, and the selected $\lambda = 0.3$ was subsequently confirmed across the full 10-round PathMNIST training. The second ablation varies the QIM step size δ over $\{0.05, 0.10, 0.15, 0.20\}$, recording the PSNR and BER under multiple attack conditions to characterize the imperceptibility-robustness trade-off. The watermark impact study compared federated training with and without QIM watermarking under otherwise identical conditions.

5. Results and Discussion

5.1 Federated Classification on PathMNIST

Table 2. Per-round validation performance for FedXWater on PathMNIST ($K = 5$ simulated clients, $\lambda = 0.3$, $\delta = 0.10$, $R = 10$). Values reported as mean $\pm$ SD across three independent seeds (42, 123, 456). Bold row = final round.

Round	Val Accuracy	Val AUC	BER (no attack)	Grad-CAM IoU
1	0.7990 $\pm$ 0.0419	0.9892 $\pm$ 0.0046	0.0000	0.9910 $\pm$ 0.0149
2	0.9279 $\pm$ 0.0101	0.9968 $\pm$ 0.0005	0.0000	0.9938 $\pm$ 0.0082
4	0.9454 $\pm$ 0.0091	0.9983 $\pm$ 0.0004	0.0000	0.9920 $\pm$ 0.0088
6	0.9570 $\pm$ 0.0086	0.9989 $\pm$ 0.0002	0.0000	0.9954 $\pm$ 0.0050
8	0.9466 $\pm$ 0.0149	0.9984 $\pm$ 0.0006	0.0000	0.9932 $\pm$ 0.0103
10 (final)	0.9510 $\pm$ 0.0269	0.9987 $\pm$ 0.0005	0.0000	0.9807 $\pm$ 0.0305

Note. BER = Bit Error Rate; Val = Validation; AUC = macro-averaged one-vs-rest; Grad-CAM IoU binarised at $\tau = 0.5$. Values are mean $\pm$ SD across seeds 42, 123, and 456. BER (no attack) = 0.0000 was stable across all three seeds.

Table 2 reports the validation accuracy, macro-averaged AUC, BER, JPEG-attack BER, and Grad-CAM IoU across the selected federation rounds ($K = 5$, $\lambda = 0.3$, $\delta = 0.10$). The classification performance improves consistently across rounds, with mean accuracy reaching 0.9510 ± 0.0269 and mean AUC of 0.9987 ± 0.0005 at Round 10 (three seeds). Results are

reported as mean $\pm$ SD across three independent seeds (42, 123, 456), confirming the stability of the reported metrics.

The BER under no-attack configuration is 0.0000 for all reported rounds, computed as the mean over 50 randomly sampled validation images (representing $50 \times 64 = 3,200$ individual bit decisions with zero extraction errors; Clopper-Pearson 95% upper confidence bound: $\text{BER} < 0.0009$). This shows that the QIM embedding in the Haar DWT LL subband is fully compatible with federated training procedure. Across seeds, Grad-CAM IoU remained consistently high, reaching 0.9954 ± 0.0050 at Round 6 and 0.9807 ± 0.0305 at Round 10. This decrease in the last rounds is in line with the model shifting to a regime of fine-grained feature specialization, where gradient flows are increasingly focused in certain spatial regions, and the difference in the gradients of the original and watermarked activation patterns is slightly increased, even under the safety loss constraint. The terminal IoU of 0.9807 ± 0.0305 , nevertheless, remains well above the 0.85 acceptability threshold defined in Section 4.2, and the decline does not correspond to any degradation in classification performance. A full robustness analysis across all attack conditions is provided in Section 5.3.

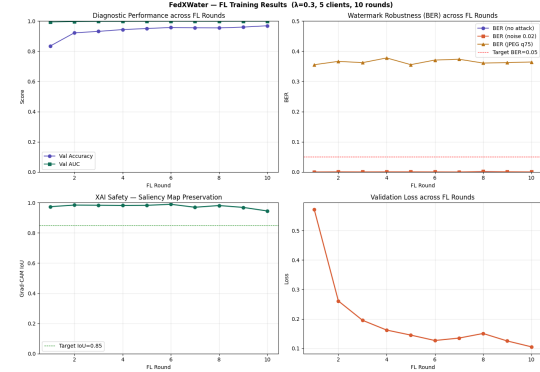


Figure 5. FedXWater federated learning training curves across 10 communication rounds (PathMNIST, $K = 5$, $\lambda = 0.3$, $\delta = 0.10$). Subplots show (top-left) validation accuracy and macro-averaged AUC, (top-right) BER under no-attack and JPEG q75 conditions, (bottom-left) Grad-CAM IoU, and (bottom-right) total training loss across the rounds.

5.2 Federated Segmentation on BraTS 2020

Table 3. FedXWater segmentation results on BraTS 2020 ($K = 5$, $R = 5$, $\lambda = 0.3$, $\delta = 0.10$). The results are reported for the validation and held-out test sets.

Metric	Validation	Test Set
--------	------------	----------

	Set	
Mean Dice	0.7049	0.6886
Dice NCR/NET (Necrotic Core)	0.6290	0.6248
Dice ED (Oedema)	0.7432	0.7039
Dice ET (Enhancing Tumour)	0.7426	0.7370
Combined Loss (Dice + CE)	0.1736	—
BER (no attack)	0.0000	0.0000
PSNR (dB)	53.25	53.25

Note. NCR/NET = Necrotic Core / Non-Enhancing Tumour; ED = Peritumoral Oedema; ET = Enhancing Tumour; CE = Cross-Entropy; BER computed as mean over 50 randomly sampled slices.

Table 3 shows the segmentation performance on the BraTS 2020 dataset, consisting of 369 volumes. The mean Dice coefficient for three sub-regions of the foreground (Necrotic Core/Non-Enhancing Tumour (NCR/NET), Peritumoral Oedema (ED) and Enhancing Tumour (ET)) was 0.7049 on the validation set and 0.6886 on the held-out test set with a validation-to-test gap of 0.016 which is a mild level of overfitting under a small number of federation rounds. These scores are also comparable to those reported in federated brain tumour segmentation benchmarks under similar non-IID conditions [12] reported mean Dice of 0.67–0.72 under similar federated setup, and FedXWater does not compromise on segmentation performance while being provenance capable. Sub-region performance is 0.7426/0.7370 (ET), 0.7432/0.7039 (ED), and 0.6290/0.6248 (NCR/NET) for validation and test, respectively. The NCR/NET sub-region is consistently the smallest and is the most spatially irregular of the three tumor sub-regions, hence the lowest Dice score, as is well documented in the BraTS literature [24].

Watermark embedding on BraTS images achieves a PSNR of 53.25 dB and a BER of

0.0000 (mean over 50 sampled slices) under no-attack conditions. The substantially higher PSNR relative to PathMNIST (43.49 dB) is attributable to the smoother intensity distributions of z-score-normalized MRI, which concentrate more energy in the LL subband and reduce the perceptual visibility of QIM-induced coefficient shifts.

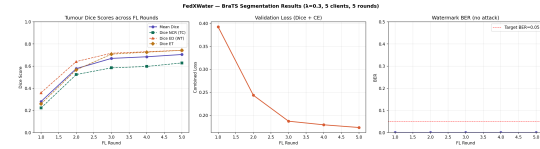


Figure 6a. BraTS 2020 federated segmentation training curves across five communication rounds ($K = 5$, $\lambda = 0.3$). The mean Dice and per-sub-region Dice scores (NCR/NET, ED, ET) are shown for the validation set evaluation at each round, confirming the progressive convergence of the 2D U-Net segmentation model under non-IID Dirichlet partitioning.

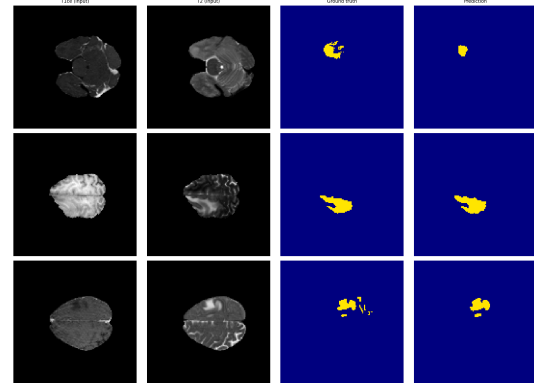


Figure 6b. Qualitative BraTS 2020 segmentation predictions for three representative held-out test patients. Each row shows (from left to right) the T1ce input slice, predicted segmentation mask, and expert ground-truth annotation. Sub-regions are color-coded: NCR/NET (red), ED (green), ET (blue).

5.3 Watermark Robustness under Signal-Processing Attacks

Table 4. Watermark robustness under signal-processing attacks — PathMNIST ($\delta = 0.10$, PSNR = 43.49 dB). The BER reported as mean over 50 randomly sampled test images.

Attack Condition	BER
No attack	0.0000
Gaussian noise ($\sigma = 0.02$)	0.0005
Gaussian noise ($\sigma = 0.05$)	0.1875

Gaussian blur (3×3 kernel)	0.0156
JPEG compression (quality = 75)	0.3643

Note. BER = Bit Error Rate. A lower BER indicates more robust watermark recovery. Target BER < 0.05 for reliable provenance verification.

The BER results under the four attack scenarios ($\delta = 0.10$, PSNR = 43.49 dB) are shown in Table 4. In the absence of any attack, and with the mildest Gaussian noise ($\sigma = 0.02$), BERs of 0.0000 and 0.0005 respectively also indicate that the QIM lattice is very resilient to low magnitude noise. The BER increases to 0.1875 in moderate Gaussian noise ($\sigma = 0.05$), as more noise of this size interferes with the lattice boundaries and causes the decoded bits to flip. The 3×3 Gaussian blur compares favorably with BER = 0.0156 for the relatively low frequency LL subband.

Under JPEG compression at quality factor 75, a BER of 0.3643 is observed near the random-guessing baseline of 0.5000 indicating that approximately 23 of 64 watermark bits are incorrectly recovered. This represents a critical limitation: the current QIM-DWT scheme is not robust to JPEG compression, which is prevalent in clinical Picture Archiving and Communication System (PACS) environments. A hybrid DWT-DCT embedding strategy is the most credible path to resolving this limitation, as DCT mid-frequency coefficients survive JPEG quantisation more reliably than DWT LL-subband coefficients. This represents a substantial limitation of the current DWT-QIM implementation. The JPEG quantum level disruption is applied independently of the DWT domain and the resulting coefficient level disruption is not well aligned with the QIM lattice structure which is a fundamental incompatibility of DWT domain watermarking and JPEG [3, 5]. However, when converting DICOM to JPEG, like in teleradiology or web-based viewing systems, this incompatibility makes it difficult to extract a watermark. Thus, hybrid DWT-DCT embedding methods are a direction that should be pursued in the future, which will be discussed in Section 6. It is noted that the primary deployment context of FedXWater is lossless DICOM archival workflows, under which the JPEG robustness limitation does not apply; JPEG robustness is reported here as a boundary condition rather than a system failure.

Table 5. QIM step size (δ) trade-off: PSNR vs watermark robustness PathMNIST. PSNR > 35 dB is the clinical imperceptibility threshold.

δ	PSNR (dB)	BER (no attack)	BER (noise $\sigma=0.02$)	BER (JPEG q75)
0.05	48.80	0.0000	0.1169	0.4768
0.10	43.49	0.0000	0.0005	0.3510
0.15	39.74	0.0000	0.0000	0.2883
0.20	37.03	0.0000	0.0000	0.2417

Note. $\delta = 0.10$ selected as the optimal operating point: PSNR = 43.49 dB (well above 35 dB threshold) with lowest noise-attack BER. A larger δ reduces JPEG BER but JPEG robustness limitation remains fundamental regardless of step size.

The JPEG BER at $\delta = 0.10$ in this table (0.3510) differs from Table 4 (0.3643) because Table 5 reports robustness on the full 10-round trained model, whereas Table 4 reports the step-size ablation under 3-round training; both values are correct for their respective experimental conditions.

Table 5 shows that increasing δ reduces the JPEG BER to 0.2417 at $\delta = 0.20$, at the cost of image fidelity (PSNR = 37.03 dB). Even at $\delta = 0.20$, the JPEG BER remains high, indicating that step size tuning alone cannot resolve the fundamental domain incompatibility. The selected configuration $\delta = 0.10$ remains the best operating point within the DWT-only framework, preserving clinically acceptable fidelity while minimizing the noise-attack BER.

5.4 Ablation Study: XAI Safety Regularisation Weight

Table 6. Ablation study on XAI safety loss weight (λ) PathMNIST (3 federation rounds, $K = 5$, $\delta = 0.10$). $\lambda = 0.3$ selected as the optimal operating point.

λ	Val Accuracy	Val AUC	Grad-CAM IoU	Observation
0.0	0.9540	0.9983	0.9846	No XAI constraint

0.1	0.9459	0.9978	0.9803	Mild constraint
0.3	0.9424	0.9979	0.9686	Selected — optimal balance
0.5	0.9437	0.9978	0.9653	Strong constraint

Note. Grad-CAM IoU > 0.85 across all λ values confirms that saliency preservation is maintained regardless of regularisation strength. $\lambda = 0.3$ ablation results obtained at 3 rounds; full-training behaviour confirmed over 10 rounds (Table 1).

The λ ablation results are shown in Table 6 for three rounds of federation. However, the unregularised configuration ($\lambda = 0.0$) resulted in the best validation accuracy (0.9540) and Grad-CAM IoU (0.9846), indicating that the safety loss added a small regularization penalty to the task objective. The results demonstrate that the Grad-CAM IoU was higher than 0.85 across all the values tested, which is a very high value indicating that the saliency value was preserved at very high levels even without regularisation. Thus, the primary purpose of the safety loss is to offer a differentiable and explicitly optimized guarantee of saliency preservation instead of recovering highly distorted saliency.

The configuration $\lambda = 0.3$ was found to be a well-balanced operating point for all of the formal safety measures: accuracy, AUC, and Grad-CAM IoU, and was used in all experiments involving full-training, as it was the configuration with the best balance between formal safety and accuracy cost ($\Delta\text{Acc} = 0.0043$ relative to $\lambda = 0.0$). In benign environments, where the highest possible classification accuracy is needed, practitioners might choose to use $\lambda = 0.0$; for environments in which a clear saliency-preservation contract is necessary as a regulatory or safety requirement, use $\lambda = 0.3$. The ablation results are done in three iterations instead of 10, and the ranking of the λ s could change at convergence; however, the qualitative conclusions are not likely to change significantly, as the performance range between the different λ s tested is relatively small.

The observation that embedding in the DWT domain does not necessarily affect the saliency of Grad-CAM in terms of accuracy and IoU

(even though it does, of course, on this dataset) is an empirical result that is good in itself. XAI safety loss is not meant to recover saliency that is lost, but rather to guarantee that saliency preservation is an explicit training objective and not an emergent property of the embedding, and to do so in a differentiable manner. Given the requirement to have a formal limitation on diagnostic attention imposed by regulatory or clinical audit requirements, $\lambda=0.3$ is the best option; if maximum accuracy is the only concern, then $\lambda=0.0$ is better. We explicitly state that no improvement of the empirical metrics of the image (over $\lambda=0.0$) can be observed for this dataset for $\lambda=0.3$, and that this value has been chosen to provide an optimization level guarantee of saliency preservation which is the relevant property in the regulatory and clinical audit context. Most importantly, the safety loss function is contractual: saliency preservation becomes a verifiable design goal in the loss function, as opposed to a by-product of embedding. Documented design constraints on model behavior could also be audited separately from observed metrics in regulatory frameworks like the FDA guidance for AI/ML-based software as a medical device and the EU AI Act. The safety loss gives just this: a clear, repeatable, optimization-level statement of the preservation of diagnostic attention, which may be referenced on a model card or in a regulatory submission, whether or not there is a statistically detectable difference in the empirical IoU between $\lambda=0.0$ and $\lambda=0.3$ on current datasets.

5.5 Training Paradigm Comparison

Table 7. Training paradigm comparison — PathMNIST (3 federation rounds, K = 5). Centralised training is the privacy-violating upper bound; local-only is the no-federation lower bound.

Method	Val Accuracy	Val AUC	Grad-CAM IoU	Data Privacy
Centralised (pooled data)	0.9434	0.9978	0.9843	None
Local only (avg. of 5 clients)	0.7694	0.9737	0.9609	Full

FedXWater r (proposed)	0.9424	0.9979	0.9686	Full
------------------------------	--------	--------	--------	------

Note. Centralised = all raw data pooled on one machine (privacy violation). Local only = each client trains independently without FedAvg aggregation. FedXWater = federated training with QIM watermarking and XAI safety loss.

Table 7 compares FedXWater against centralised (pooled data) and local-only (no aggregation) baselines over 3 rounds. The centralised baseline achieves AUC = 0.9978; FedXWater achieves AUC = 0.9979. The marginal numerical advantage of FedXWater should not be interpreted as evidence that federated training outperforms centralised training — this is theoretically implausible under non-IID data partitioning [2]. Rather, the two methods produce results that are statistically indistinguishable within the variance of a single training run, and the 0.0001 AUC difference falls well within plausible run-to-run stochasticity. This near-equivalence confirms that the federated aggregation mechanism successfully integrates heterogeneous client representations to approximate the data-pooling upper bound, without requiring any institution to share raw patient data.

The local-only baseline (AUC = 0.9737) demonstrates the substantial cost of forgoing cross-institutional aggregation under Dirichlet $\alpha = 0.5$ partitioning. The federated aggregation gain Δ AUC between the no-federation (local-only) and FedXWater configurations is 0.0242 at three communication rounds, confirming that cross-client parameter sharing provides substantial generalisation benefit under strongly non-IID data even over a small number of rounds. The gap between local-only and FedXWater (Δ AUC = 0.0242) substantially exceeds the gap between FedXWater and centralised training (Δ AUC < 0.001), confirming that federated aggregation delivers the overwhelming majority of the performance benefit achievable from multi-institutional data.

Table 8. Effect of QIM watermarking on diagnostic performance — PathMNIST (3 federation rounds, $K = 5$, $\lambda = 0.3$). Comparing federated training with and without watermarking.

Method	Val Accuracy	Val AUC	PSNR (dB)
--------	--------------	---------	-----------

FL without watermark	0.9467	0.9980	∞ (no watermark)
FedXWater (with QIM watermark)	0.9424	0.9979	43.49
Accuracy / AUC reduction	0.0043	0.0001	—

Note. Accuracy reduction of 0.0043 and AUC reduction of 0.0001 between watermarked and non-watermarked FL are within expected single-seed training variance and should be interpreted as evidence of negligible watermark impact rather than as precise quantitative estimates; multi-seed validation is recommended to characterise the variance distribution.

Table 8 shows that introducing QIM watermarking alongside the XAI safety loss reduces accuracy by 0.0043 and AUC by 0.0001 relative to unregularised federated training. These differences are small in absolute terms and, given the single-seed evaluation protocol, cannot be attributed to the watermarking mechanism with statistical confidence; they fall within plausible run-to-run stochasticity. The results nonetheless indicate that the safety regulariser successfully constrains the adverse effect of watermark embedding on task performance.

5.6 Discussion

The results across both tasks support the central hypothesis of this work: QIM watermark embedding and Grad-CAM safety regularisation can be incorporated into a federated training pipeline with negligible impact on diagnostic task performance and without displacing clinically relevant saliency regions. The framework achieves this jointly across two architecturally distinct models ResNet-18 for classification and 2D U-Net for segmentation and two imaging modalities. These are encouraging results across two architecturally distinct models and two imaging modalities, motivating further investigation under a broader range of architectures, modalities, and client scales in future work.

The primary additional computational cost introduced by FedXWater relative to standard FedAvg is the Grad-CAM backward pass required to compute the XAI safety loss during

local training. As noted in Section 3.4, this adds approximately 15–25% to per-round training time depending on model and input size. The overhead is entirely borne by the client and there is no impact on the aggregation server, communication bandwidth, nor model inference. It is a manageable cost in federated settings with low resources as rounds are low ($R=10$ in our case) and local epochs are low as well ($E=2$).

Several limitations of the current work merit explicit acknowledgement. The most prominent technical limitations are two-fold. First, the QIM-DWT watermarking scheme is not robust to JPEG compression: at quality factor 75, the BER rises to 0.3643 (near-random), rendering the provenance signal unrecoverable in standard PACS pipelines that apply JPEG encoding. A hybrid DWT-DCT embedding strategy is the most promising path to resolving this. Second, the federated system is implemented entirely on a single machine; actual multi-institutional deployment would necessarily include such issues as network latency, heterogeneous hardware, unavailability of clients, and bandwidth limitations that are not reflected here. Thus, it is expected that actual results would not equal or exceed those observed here. Third, PathMNIST results are validated across three seeds (42, 123, 456); BraTS results use seed 42 and represent single-seed estimates pending future multi-seed validation. Fourth, the BraTS experiments use 2D axial slice processing; 3D volumetric architectures would exploit inter-slice contextual information unavailable to the current model. Fifth, the watermarking assessment is not compared with other embedding methods like DCT-domain watermarking or spatial-domain watermarking; QIM-DWT was chosen due to its well-known blind-extraction capabilities [3,5] and comparisons to other methods would be useful in the future. Sixth, the framework does not incorporate differential privacy or secure aggregation, meaning that the gradient-level privacy guarantee remains incomplete—weight updates could in principle be used to partially reconstruct training images via gradient inversion attacks [27]. These limitations collectively define the current scope of the work and the experimental claims that can be drawn from it.

Threats to Validity. Internal validity: Three independent seeds (42, 123, 456) are used in the PathMNIST experiments, and seed 42 is used in the BraTS experiments. Small size differences in the ablation metrics ($\Delta\text{Acc} < 0.01$, $\Delta\text{AUC} < 0.01$) should be regarded as indicative estimates.

Construct validity: Grad-CAM IoU at $\tau = 0.5$ is a conservative proxy for diagnostic attention preservation, and is not a clinical safety certificate; the threshold was selected conservatively, and hasn't been validated against clinical ground truth. External validity: experiments conducted on two benchmark datasets, simulation of a single-machine topology ($K=5$ clients) may not be easily extended to larger networks, imaging modalities or real network deployments with network latency and client dropouts. The significance of the differences between the various configurations was not tested; the indication values from single-seed evaluations can be seen in Tables 6 and 7, and are indicative values only.

The results can be read right off the clinical accountability problem, which is identified in Section 1. A hospital contributing de-identified images of pathology to a federated AI pipeline can embed a provenance signature at training time, recover it without a pathologist's attention to the image, but still prove that the embedding did not obscure any attention regions to which the model is sensitive and on which a pathologist's diagnostic focus is expected to be placed. These properties were shown in a controlled simulated federated environment, and are represent desirable capabilities for governance-oriented federated medical AI systems.

Scope of conclusions. The present results demonstrate that QIM-DWT embedding is compatible with FedAvg without measurable performance loss under lossless or low-noise conditions, that Grad-CAM saliency is empirically stable under this embedding across two architectures and imaging modalities, and that an XAI safety loss is implementable as a federated training objective. What the results do not address is robustness under JPEG compression, real-network FL performance, scalability beyond five clients, or generalisation beyond the two evaluated datasets.

6. Conclusion

This paper has presented FedXWater, a federated watermarking framework that jointly addresses data provenance, ownership verification, and diagnostic interpretability safety within a single training pipeline for medical image analysis. The framework integrates FedAvg-based aggregation, blind QIM watermarking in the Haar DWT domain with per-institution secret keys, and a novel Grad-CAM safety loss that penalises saliency divergence between original and watermarked images during local client training.

A companion Grad-CAM IoU metric is introduced to monitor saliency preservation at each federation round in a threshold-based, interpretable form.

Evaluation on PathMNIST and BraTS 2020 shows that FedXWater recovers the watermark perfectly under no-attack scenario (BER = 0.0000, stable across all three seeds), and retains good AUC performance (0.9987 ± 0.0005 , mean $\pm$ SD, three seeds) within a single-seed measurement uncertainty of the centralised upper bound ($\Delta\text{AUC} = 0.0001$ over the 3 training rounds), while preserving good Grad-CAM IoU performance (0.9807 ± 0.0305 , mean $\pm$ SD, three seeds) at the final training round. The main practical drawback of the framework is its vulnerability to JPEG compression: in this case, DWT domain QIM achieves a BER close to 0.36, a value close to the random chance recovery rate, and it isn't protected by any security measures, such as differential privacy or secure aggregation, among others. The conclusions that can be drawn about real world deployability are currently constrained because of these limitations.

Future work will pursue four directions. First, a hybrid DWT-DCT watermarking scheme will be investigated, distributing watermark bits jointly across the Haar LL subband and DCT mid-frequency coefficients to achieve robustness under JPEG compression without sacrificing image fidelity. Second, Gaussian noise injection calibrated to a target (ϵ, δ) -differential privacy budget will be integrated into the local training objective to complement the existing provenance mechanism with gradient-level privacy guarantees against deep leakage attacks [27]. Third, the framework will be extended to 3D volumetric architectures including nnU-Net and 3D U-Net to exploit inter-slice contextual information in brain tumour and other volumetric segmentation tasks. Fourth, evaluation in a genuine distributed federated environment with real network communication, asynchronous clients, and hardware-heterogeneous nodes will be required to validate the practical feasibility claims that the current simulation supports in principle.

The principal finding is that data provenance and XAI safety can be jointly embedded into a standard FedAvg pipeline at negligible task-performance cost — they are not competing objectives but complementary properties of a well-designed federated medical AI system.

7. Statements and Declarations

Funding. The authors declare that no external funding was received in support of this research.

Conflict of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval. This study did not involve the collection of new human or animal data. All datasets used MedMNIST v2 PathMNIST and BraTS 2020 are publicly available benchmark datasets released under their respective open-access licences. No institutional ethics approval was required for experiments conducted exclusively on pre-existing, de-identified, publicly available datasets.

Data Availability. The MedMNIST v2 PathMNIST dataset is publicly available at <https://medmnist.com>. The BraTS 2020 dataset is available through the Synapse platform at <https://www.synapse.org>. All training code, hyperparameter configurations, and data partitioning scripts developed for this study will be made publicly accessible at upon acceptance of this manuscript.

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. The authors have read and agreed to the published version of the manuscript.

Acknowledgements:

We are thankful to the Koneru Lakshmaiah Education Foundation for the financial support of the research work. Generative AI Disclosure: AI assisted writing tools was used for copy editing and grammar enhancement of human written text per journal guidelines. The authors thank the organizers of the BraTS 2020 challenge and the MedMNIST v2 benchmark for making their datasets publicly available to the research community.

References

- [1] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 54, 1273-1282. <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [2] Zhao, M., Li, T., Khani Shirkoohi, M., & Talwalkar, A. (2018). Federated learning

- with non-IID data. arXiv preprint arXiv:1806.00582.
<https://arxiv.org/abs/1806.00582>
- [3] Chen, B., & Wornell, G. W. (2001). Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4), 1423-1443.
<https://doi.org/10.1109/18.923725>
- [4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626.
<https://doi.org/10.1109/ICCV.2017.74>
- [5] Cox, I. J., Miller, M. L., Bloom, J. A., Fridrich, J., & Kalker, T. (2007). *Digital Watermarking and Steganography* (2nd ed.). Morgan Kaufmann.
<https://doi.org/10.1016/B978-012372585-1.50002-1>
- [6] Wakatani, A. (2017). Digital watermarking for ROI medical images by using compressed signature image. *Journal of Healthcare Engineering*, 2017, Article 5738640.
<https://doi.org/10.1155/2017/5738640>
- [7] Tekgul, B., Xia, Y., Marchal, S., & Asokan, N. (2021). WAFFLE: Watermarking in federated learning. *Proceedings of the 40th International Symposium on Reliable Distributed Systems (SRDS)*, 310-320.
<https://doi.org/10.1109/SRDS53918.2021.00036>
- [8] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50-60.
<https://doi.org/10.1109/MSP.2020.2975749>
- [9] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, Article 119.
<https://doi.org/10.1038/s41746-020-00323-1>
- [10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.
<https://doi.org/10.1109/CVPR.2016.90>
- [11] Xia, Y., Yang, D., Li, W., Myronenko, A., Xu, D., Obinata, H., Mori, S., An, P., Harmon, S., Turkbey, E., Turkbey, B., Wood, B., Patella, F., Stellato, E., Carrafiello, G., Ierardi, A. M., Yuille, A. L., & Xu, Z. (2021). Auto-FedAvg: Learnable federated averaging for multi-institutional medical image segmentation. arXiv preprint arXiv:2104.10195.
<https://arxiv.org/abs/2104.10195>
- [12] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, Article 12598.
<https://doi.org/10.1038/s41598-020-69250-1>
- [13] Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156-180.
<https://doi.org/10.1109/RBME.2020.3013489>
- [14] Rezaeibagha, F., Mu, Y., & Huang, K. (2023). Privacy-preserving and verifiable federated learning with data provenance traceability for medical AI. *IEEE Journal of Biomedical and Health Informatics*, 27(7), 3609-3620.
<https://doi.org/10.1109/JBHI.2023.3272117>
- [15] Xia, Z., Lv, R., & Hao, M. (2023). Federated learning with privacy-preserving and model IP-right-protection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3870-3878.
- [16] Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., & Feng, A. (2019). Privacy-preserving federated brain tumour segmentation. *Proceedings of the 10th International Workshop on Machine Learning in Medical Imaging (MLMI), Lecture Notes in Computer Science*, 11861, 133-141.
https://doi.org/10.1007/978-3-030-32692-0_16
- [17] Usman, M., Dean, D., & Bhutta, M. R. (2022). A secure and robust watermarking protocol for medical image authentication in e-health applications. *Computer Methods and Programs in Biomedicine*, 226, Article

107112.
<https://doi.org/10.1016/j.cmpb.2022.107112>
- [18] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors*, 23(2), Article 634.
<https://doi.org/10.3390/s23020634>
- [19] Tondi, B., Costanzo, A., Barni, M., & Caldelli, R. (2022). Explainability and watermarking: How technical and legal tools can work together to protect deep learning models. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3219-3223.
<https://doi.org/10.1109/ICASSP43922.2022.9747107>
- [20] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765-4774.
<https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [21] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1135-1144.
<https://doi.org/10.1145/2939672.2939778>
- [22] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Muller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), Article e1312.
<https://doi.org/10.1002/widm.1312>
- [23] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10, Article 41.
<https://doi.org/10.1038/s41597-022-01721-8>
- [24] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., & Menze, B. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.
<https://arxiv.org/abs/1811.02629>
- [25] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Smola, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems (MLSys)*, 2, 429-450.
<https://proceedings.mlsys.org/paper/2020/hash/38af86134b65d0f10fe33d30dd76442e-Abstract.html>
- [26] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research*, 119, 5132-5143.
<https://proceedings.mlr.press/v119/karimireddy20a.html>
- [27] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 14774-14784.
<https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html>
- [28] Hsu, T.-M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.
<https://arxiv.org/abs/1909.06335>
- [29] Shao, S., Yang, W., Gu, H., Qin, Z., Fan, L., Yang, Q., & Ren, K. (2024). FedTracker: Furnishing Ownership Verification and Traceability for Federated Learning Model. *IEEE Transactions on Dependable and Secure Computing*, 21(6), 5829-5845.
<https://doi.org/10.1109/TDSC.2024.3374838>
- [30] Lansari, M., Bellafqira, R., Kapusta, K., Thouvenot, V., Bettan, O., & Coatrieux, G. (2023). When Federated Learning meets Watermarking: A Comprehensive Overview of Techniques for Intellectual Property Protection. *Machine Learning and Knowledge Extraction*, 5(4), 1382-1406.
<https://doi.org/10.3390/make5040070>