Does Machine Unlearning Preserve Clinical Safety? A Risk Analysis for Medical Image Classification

Andreza M. C. Falcao[1], **Filipe R. Cordeiro**[1]

[1] Visual Computing Lab, Departamento de Computação,
Universidade Federal Rural de Pernambuco (UFRPE), Brasil

andreza.mcfalcao@ufrpe.br, filipe.rolim@ufrpe.br

Abstract. *The application of Deep Learning in medical diagnosis must balance patient safety with compliance with data protection regulations. Machine Unlearning enables the selective removal of training data from deployed models. However, most methods are validated primarily through efficiency and privacy-oriented metrics, with limited attention to clinically asymmetric error costs. In this work, we investigate how unlearning affects clinical risk in binary medical image classification. We show that standard unlearning strategies (Fine-Tuning, Random Labeling, and SalUn) may reduce test utility while increasing false-negative rates, thereby amplifying clinical risk. To mitigate this, we propose SalUn-CRA (Clinical Risk-Aware), a variant of SalUn that replaces random relabeling with entropy-based forgetting for malignant samples in the forget set, preventing the model from learning harmful benign associations. We evaluate on DermaMNIST and PathMNIST medical image datasets under 20% and 50% data removal. Using Global Risk metrics with asymmetric costs, SalUn-CRA achieves lower or comparable clinical risk to full retraining while preserving unlearning effectiveness. These results suggest that clinical risk should be an integral component of unlearning validation in medical systems.*

1. Introduction

The integration of Deep Learning into clinical workflows has significantly improved diagnostic performance in medical image analysis, achieving results comparable to human specialists in tasks such as skin lesion and histopathological classification [Chan et al. 2020a]. However, these models typically rely on large volumes of training data containing sensitive patient information, making them subject to strict data protection regulations.

Recent regulatory frameworks, including the Brazilian General Data Protection Law (LGPD) [Brazil 2018] and the European General Data Protection Regulation (GDPR) [European Parliament and Council of the European Union 2016], establish the *Right to be Forgotten* [Hoofnagle et al. 2019], requiring the removal of personal data upon request. In the context of Machine Learning, removing samples from the dataset alone is insufficient if the model has already been trained on them, as the influence of those samples must also be removed from the model parameters. Full retraining on the remaining data is the most straightforward solution, but it is computationally prohibitive for large-scale datasets or frequent removal requests [Mester and et al. 2024]. Machine Unlearning (MU) addresses this challenge by developing algorithms that update model weights to forget specific data without full retraining [Fan et al. 2024].

Although substantial progress has been made in MU methods [Warnecke et al. 2021, Golatkar et al. 2020, Fan et al. 2024], their evaluation has primarily focused on computational efficiency, privacy guarantees, and similarity to retrained models [Zhang et al. 2023]. MU approaches applied to medical domains have also been restricted to accuracy-based evaluation [Sakib and Xie 2024, Deng et al. 2025]. These protocols implicitly assume symmetric error costs, but this assumption does not hold in medical diagnosis [Ling and Sheng 2010]. A False Negative (failure to detect a malignant condition) may delay treatment and impact patient survival, whereas a False Positive typically results in additional confirmatory tests [Scholz and et al. 2024]. Relying solely on global metrics may obscure clinically critical degradations in sensitivity after unlearning. Importantly, it remains unclear whether MU procedures preserve, amplify, or mitigate clinical risk.

Despite the growing literature on Machine Unlearning, a critical question remains unanswered: **Do state-of-the-art machine unlearning methods preserve clinical safety?** To the best of our knowledge, no prior work has systematically investigated how unlearning affects cost-sensitive clinical risk in medical image classification.

In this work, we evaluate machine unlearning methods under clinical risk formulations that model asymmetric diagnostic costs. We propose *SalUn-CRA* (Clinical Risk-Aware), a modification of SalUn [Fan et al. 2024] that prevents harmful benign reassignment of malignant samples during forgetting, and introduce Global Risk metrics based on the asymmetric costs of false positives and false negatives. Our main contributions are:

- We introduce *Global Risk* as an evaluation criterion for medical unlearning, explicitly modelling asymmetric clinical costs.
- We show that common unlearning strategies (Fine-Tuning, Random Labeling, and SalUn) may increase clinical risk by trading recall for specificity, leading to higher false negative rates.
- We propose SalUn-CRA, a risk-aware modification of SalUn that applies entropy-based forgetting to malignant forget samples instead of random relabeling, mitigating risk amplification while preserving unlearning behaviour.

We evaluate our approach on two binary classification tasks derived from MedMNIST [Yang et al. 2021]: DermaMNIST (skin lesion classification) and PathMNIST (colorectal tissue classification), under 20% and 50% data removal. We focus on binary classification (malignant vs. benign) because it represents the most critical decision point in clinical pipelines, carrying the highest asymmetry in misclassification costs [Ling and Sheng 2010], and enables a clean risk analysis with only two cost parameters, avoiding the combinatorial complexity of multi-class cost matrices.

2. Prior Work

2.1. Machine Unlearning: Foundations and Methods

Machine Unlearning approaches can be categorized into exact and approximate methods. Exact methods ensure the complete removal of a specific data subset and aim to retrain the model at lower computational cost than full retraining [Li et al. 2024]. Techniques within this class typically leverage checkpoint-based retraining strategies, selectively updating

portions of the model rather than performing full retraining [Bourtoule et al. 2021]. Although more computationally efficient than complete retraining, exact methods remain prohibitively costly for frequent or large-scale unlearning tasks.

Among the foundational approaches, *Amnesiac Learning* [Graves et al. 2021] stores per-batch gradient contributions during training and reverses them during unlearning, selectively eliminating the influence of the target data. Golatkar et al. [Golatkar et al. 2020] leverage Fisher information to estimate which model weights encode information about the samples to be removed and modify them directly. Barez et al. [Barez et al. 2025] argue that unlearning can degrade existing safety mechanisms and produce unintended side effects on model behaviour. This observation directly motivates this work's concern with clinical risk after data removal. Fine-Tuning (FT) [Warnecke et al. 2021] continues training on the retain set only, allowing the model to gradually forget removed samples through parameter drift. Random Labeling (RL) [Golatkar et al. 2020] assigns random labels to forget samples, forcing the model to unlearn their correct associations. Both methods are computationally efficient but lack mechanisms to preserve class-specific performance.

Saliency Unlearning (SalUn) [Fan et al. 2024] represents one of the current state-of-the-art (SOTA) approaches in approximate unlearning. It computes a saliency map based on the gradients of the forgetting loss to identify the weights most relevant to the target data. Then it applies random labelling only to these salient parameters.

2.2. Machine Unlearning in Healthcare

The application of Deep Learning to medical image analysis is well established [Chan et al. 2020b], yet its intersection with machine unlearning remains underexplored. Nasirigerdeh et al. [Nasirigerdeh et al. 2024] studied unlearning in medical images, focusing on re-identification attacks, but without addressing diagnostic metrics. Hardan et al. [Hardan et al. 2025] proposed Forget-MI for multimodal medical data. Falcão and Cordeiro [Falcao and Cordeiro 2025] evaluated MU on medical datasets, though limited to standard MU metrics.

Despite the growing research on machine unlearning for healthcare applications, a critical gap persists between the evaluation practices of the unlearning community and the requirements of clinical machine learning. None of the existing healthcare-oriented unlearning studies reports sensitivity and specificity as separate metrics, nor do they employ Balanced Accuracy to account for class imbalance commonly present in medical datasets. More importantly, no existing work evaluates the clinical risk implications of the unlearning process, that is, whether unlearning disproportionately degrades the model's ability to detect positive (pathological) cases, thereby increasing false negative rates in a clinically dangerous manner. This paper aims to bridge this gap by explicitly evaluating unlearning methods under clinical risk metrics. While previous works focus on privacy and computational efficiency, we investigate whether data removal compromises diagnostic safety.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ be a training dataset of images x_i with associated labels $y_i \in \{1, \ldots, K\}$, where K is the number of classes and N the total number of samples. The

Machine Unlearning problem can be formally defined as the task of removing the influence of a specific subset of data $\mathcal{D}_f \subset \mathcal{D}$ from a previously trained model $\theta_o = \mathcal{A}(\mathcal{D})$, where θ_o is the set of weights resulting from applying a training algorithm $\mathcal{A}$ to the dataset $\mathcal{D}$. Traditional retraining approaches retrain the model on the remaining subset $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$, obtaining the retrained model weights $\theta_r = \mathcal{A}(\mathcal{D}_r)$, retrained from scratch without the data subset $\mathcal{D}_f$.

Given this formulation, the MU task consists of employing an unlearning algorithm $\mathcal{U}$, which, starting from the trained model θ_o, the subset to be forgotten $\mathcal{D}_f$, and the remaining subset $\mathcal{D}_r$, produces a new model $\theta_u = \mathcal{U}(\theta_o, \mathcal{D}_f, \mathcal{D}_r)$. It is expected that θ_u approximates, in terms of output distribution, the ideal retrained model θ_r.

The performance of the unlearned model is evaluated through the metric gap, defined as $MG = |\mathcal{M}_{MU} - \mathcal{M}_{retrain}|$, where $\mathcal{M}_{MU}$ represents the evaluation metric on the unlearned model and $\mathcal{M}_{retrain}$ corresponds to the metric on the retrained model. The main challenge lies in designing unlearning algorithms that are not only effective at removing the influence of $\mathcal{D}_f$ but also significantly more computationally efficient than full retraining while preserving performance on the remaining dataset $\mathcal{D}_r$.

3.2. SalUn-CRA: Clinical Risk-Aware Saliency Unlearning

Saliency Unlearning (SalUn) [Fan et al. 2024] is one of the current state-of-the-art methods in approximate machine unlearning. Its core contribution is the introduction of *weight saliency* into the unlearning process. Rather than modifying all model parameters during unlearning, SalUn identifies the subset of weights most relevant to the forget set $\mathcal{D}_f$ through a gradient-based saliency map:

$$\mathbf{m}_S = \mathbb{1}\Big(\Big|\nabla_{\boldsymbol{\theta}} \ell_f(\boldsymbol{\theta}; \mathcal{D}_f)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\Big| \geq \gamma\Big), \tag{1}$$

where ℓ_f is the forgetting loss (cross-entropy on $\mathcal{D}_f$), $\boldsymbol{\theta}_o$ denotes the original model weights, γ is a hard threshold (set to the median of the gradient magnitudes), and $\mathbb{1}(\cdot)$ is an element-wise indicator function. This mask decomposes the model into *salient weights*, which are updated during unlearning, and *intact weights*, which remain unchanged:

$$\boldsymbol{\theta}_u = \mathbf{m}_S \odot (\Delta\boldsymbol{\theta} + \boldsymbol{\theta}_o) + (\mathbf{1} - \mathbf{m}_S) \odot \boldsymbol{\theta}_o, \tag{2}$$

where $\odot$ denotes the element-wise product. Once the salient weights are identified, SalUn applies Random Labeling (RL) exclusively to these parameters. For each sample $(x_i, y_i) \in \mathcal{D}_f$, a random label $y_i' \neq y_i$ is assigned, and the model is fine-tuned on the relabeled forget set combined with a regularization term on the retain set $\mathcal{D}_r$:

$$\mathcal{L}_{\text{SalUn}}(\boldsymbol{\theta}_u) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f, y'\neq y}\left[\ell_{\text{CE}}(\boldsymbol{\theta}_u; x, y')\right] + \alpha\, \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\ell_{\text{CE}}(\boldsymbol{\theta}_u; x, y)\right], \tag{3}$$

where $\alpha > 0$ balances unlearning effectiveness and model utility preservation.

While SalUn achieves strong unlearning performance on general-purpose benchmarks [Fan et al. 2024], its reliance on uniform random relabeling introduces a critical problem in binary medical classification settings where class 1 corresponds to the malignant (positive) class and class 0 to the benign (negative) class.

While SalUn achieves strong performance on general-purpose benchmarks [Fan et al. 2024], its reliance on random relabeling introduces a critical problem

in binary medical classification. In the binary case, every malignant sample in $\mathcal{D}_f$ deterministically receives the label "benign", unlike multi-class scenarios where errors distribute across $C-1$ classes. The model is thus actively trained to associate malignant features with the benign class.

This relabeling contamination has a direct and asymmetric impact on the decision boundary. The model learns to suppress activations for malignant patterns, shifting the classification threshold toward higher specificity at the expense of sensitivity. In clinical terms, this translates to an increase in false negatives: malignant lesions that the model now classifies as benign. While the overall accuracy or balanced accuracy may appear stable, missed malignant diagnoses carry far greater consequences than false alarms [Ling and Sheng 2010].

To mitigate this risk, we propose SalUn-CRA (Clinical Risk-Aware), a modification of SalUn that applies class-dependent forgetting strategies within the forget set. The key idea is to decouple the forgetting mechanism based on the clinical severity of each class. For malignant samples ($y_i = 1$) in $\mathcal{D}_f$, instead of random relabeling, we use a *maximum entropy loss* that pushes the model's output distribution toward uniformity, encouraging maximum uncertainty without assigning a benign label. For benign samples ($y_i = 0$) in $\mathcal{D}_f$, standard random relabeling is applied as in the original SalUn. In the binary case, benign samples receive the malignant label, which does not introduce the same clinical risk.

Let $\mathcal{D}_f^+ = \{(x_i, y_i) \in \mathcal{D}_f \mid y_i = 1\}$ and $\mathcal{D}_f^- = \{(x_i, y_i) \in \mathcal{D}_f \mid y_i = 0\}$ denote the malignant and benign subsets of the forget set, respectively. The maximum entropy loss for malignant samples is defined as:

$$\mathcal{L}_{\text{entropy}}(\boldsymbol{\theta}; x) = -\sum_{c=1}^{C} p_c(x; \boldsymbol{\theta}) \log p_c(x; \boldsymbol{\theta}), \tag{4}$$

where $p_c(x; \boldsymbol{\theta}) = \text{softmax}(f_{\boldsymbol{\theta}}(x))_c$ is the predicted probability for class c, and C is the number of classes. Maximizing this entropy drives the output toward a uniform distribution $p_c = 1/C$ for all c, achieving maximum prediction uncertainty.

The complete SalUn-CRA loss function combines three components:

$$\mathcal{L}_{\text{CRA}}(\boldsymbol{\theta}_u) = \underbrace{-\,\mathbb{E}_{x \in \mathcal{D}_f^+}[\mathcal{L}_{\text{entropy}}(\boldsymbol{\theta}_u; x)]}_{\text{entropy maximization (malignant)}} + \underbrace{\mathbb{E}_{(x,y') \in \mathcal{D}_f^-}[\ell_{\text{CE}}(\boldsymbol{\theta}_u; x, y')]}_{\text{random labeling (benign)}} + \underbrace{\alpha\, \mathbb{E}_{(x,y) \in \mathcal{D}_r}[\ell_{\text{CE}}^w(\boldsymbol{\theta}_u; x, y)]}_{\text{retain regularization}}, \tag{5}$$

where y' is the random label assigned to benign forget samples ($y' = 1$ in the binary case), ℓ_{CE}^w denotes weighted cross-entropy with inverse-frequency class weights to address class imbalance, and $\alpha > 0$ controls the retain regularization strength. The weight saliency mask $\mathbf{m}_S$ from Equation (1) is applied identically as in the original SalUn, restricting parameter updates to the salient subset.

3.3. Datasets

We use DermaMNIST and PathMNIST, two datasets from the MedMNIST collection [Yang et al. 2021], converted to binary classification tasks. All images were resized to 128×128 pixels to standardize input across experiments.

DermaMNIST consists of 10,015 dermatoscopic images of pigmented skin lesions from the HAM10000 dataset. The original dataset contains 7 diagnostic categories. We binarize the labels by grouping melanoma, basal cell carcinoma, and actinic keratosis as *malignant*, while melanocytic nevus, benign keratosis, vascular lesion, and dermatofibroma are grouped as *benign*.

PathMNIST contains 107,180 histopathological images of colorectal tissue. The original 9-class problem was binarized by grouping cancer-associated stroma and colorectal adenocarcinoma as malignant, and all other tissue types (adipose, background, debris, lymphocytes, mucus, smooth muscle, and normal colon mucosa) as benign.

We adopt this binarization protocal because it serves a methodological purpose in the context of this study. By reducing the problem to two classes, we isolate the effect of unlearning on the sensitivity–specificity trade-off, which is the core mechanism through which clinical risk is amplified. The binary formulation provides a controlled experimental framework where shifts in false negative and false positive rates can be directly measured and linked to the forgetting procedure. Table 1 summarizes the binarization protocol for both datasets.

Table 1. Dataset statistics of DermaMNIST and PathMNIST after binarization. Original multi-class labels are grouped into clinically meaningful binary categories.

Split	DermaMNIST		PathMNIST	
	Benign	Malignant	Benign	Malignant
Train	5,641 (80.5%)	1,366 (19.5%)	67,710 (75.2%)	22,286 (24.8%)
Val	807 (80.5%)	196 (19.5%)	7,527 (75.2%)	2,477 (24.8%)
Test	1,613 (80.4%)	392 (19.6%)	5,526 (77.0%)	1,654 (23.0%)

3.4. Unlearning Scenarios

The goal of unlearning is to update a model trained on $\mathcal{D}_{train}$ so that it behaves as if trained only on $\mathcal{D}_{retain} = \mathcal{D}_{train} \setminus \mathcal{D}_{forget}$. We evaluate two removal scenarios: 20% and 50% of training sample removal, following the protocol in [Fan et al. 2024, Falcao and Cordeiro 2025]. In both scenarios, we apply *balanced removal*, where samples are removed proportionally from each class, preserving the original class distribution. This isolates the effect of data reduction from artificial distribution shift.

3.5. Unlearning Methods

We evaluate the following unlearning methods: Retrain, Fine-Tuning (FT) [Warnecke et al. 2021], Random Labeling (RL), Saliency Unlearning (Salun) and the proposed Salun-CRA.

Retrain approch is the complete retraining from scratch on $\mathcal{D}_{retain}$, serving as the reference for ideal unlearning behavior. Fine-Tuning continues training the original model on $\mathcal{D}_{retain}$ only, allowing gradual forgetting through parameter updates. Random Labeling (RL) assigns random labels to samples in $\mathcal{D}_{forget}$ and trains on the combined dataset, forcing the model to unlearn correct associations. SalUn computes a saliency mask identifying parameters most relevant to $\mathcal{D}_{forget}$, then applies random labeling only to these salient weights. SalUn-CRA is our clinical risk-aware modification of SalUn, as described in Section 3.2.

3.6. Evaluation Metrics

We evaluate unlearning methods across three dimensions: model utility, unlearning effectiveness, and clinical risk. For model utility, we use Especificity, Recall, Balanced Accuracy (BAC) and Area Under the ROC Curve (AUC). Specificity measures the proportion of benign cases correctly identified by $TN/(TN + FP)$, where TN and FP represent the number true negatives and false positives, respectivelly. Recall measures the proportion of malignant cases correctly identified by the equation $TP/(TP + FN)$, with FN being the number of false negatives. BAC IS defined as the average of specificity and recall, which is robust to class imbalance. AUC is the area under the ROC curve, measuring overall discrimination ability.

The standard evaluation protocol in MU is based on accuracy values computed on the retain and forget sets [Fan et al. 2024]. However, we adapted these metrics to use BAC instead of Accuracy to better reflect performance on imbalanced medical datasets. The adapted metrics are: UBAC, the balanced accuracy on $\mathcal{D}_{forget}$ (lower values indicate successful forgetting); RBAC, the balanced accuracy on $\mathcal{D}_{retain}$ (higher values indicate preserved utility); TBAC, the balanced accuracy on the test set; MIA, the Membership Inference Attack accuracy (lower values indicate better privacy); and GAP, the average absolute difference between unlearned and retrained model metrics (lower GAP indicates closer approximation to the gold standard).

We propose a clinical risk formulation based on cost-sensitive evaluation [Haimerl and Reich 2025], capturing the asymmetric cost structure of medical diagnosis. The general Global Risk is defined as:

$$\text{Risk} = \frac{C_{FP} \cdot FP + C_{FN} \cdot FN}{N} \tag{6}$$

where FP and FN are the number of false positives and false negatives, respectively, N is the total number of samples, and C_{FP} and C_{FN} are the misclassification costs for each error type. For this purpose, we consider two risk scenarios: **Global Risk I** and **Global Risk II**. Global Risk I sets $C_{FN} = 1$ and $C_{FP} = 1$, representing equal misclassification costs, as commonly assumed in the literature. Global Risk II sets $C_{FN} = 20$ and $C_{FP} = 1$, reflecting a clinically realistic scenario where missed malignant diagnoses carry substantially higher costs than false positives. The adoption of $C_{FN} = 20$ for Risk II is a representative value for serious disease screening scenarios, as there is no universally agreed-upon cost ratio in the literature, as it depends on context. Lower risk values indicate safer models. Global Risk II specifically penalizes methods that sacrifice sensitivity (recall) for specificity, as this trade-off increases the number of missed malignancies.

3.7. Implementation

We adopt ResNet-18 [Wu et al. 2019] as the backbone architecture for all experiments, following standard practice in medical imaging tasks [Falcao and Cordeiro 2025]. The baseline model is trained for 200 epochs using SGD optimizer with learning rate 0.1, momentum 0.9, and batch size 256.

To address class imbalance and incorporate the clinical priority of detecting malignant

cases, all models are trained with Weighted Cross-Entropy Loss:

$$\mathcal{L}_w = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(p(y_i|x_i)) \tag{7}$$

where w_{y_i} is the weight assigned to the true class y_i. Weights are defined inversely proportional to class frequencies in the training set, ensuring that the minority class has greater influence on gradient updates.

All unlearning methods are executed for 10 epochs with learning rate 0.01, following the protocol in [Fan et al. 2024].

4. Results

We analyze the trade-off between unlearning effectiveness, model utility, and clinical risk across DermaMNIST and PathMNIST under 20% and 50% data removal scenarios. We evaluated the SOTA MU models Fine-Tuning (FT) [Warnecke et al. 2021], Random Labeling (RL) [Golatkar et al. 2020], Salun [Fan et al. 2024] and the proposed Salun-CRA using the metrics described in section 3.6.

4.1. Utility and Unlearning Metrics

Table 2 presents comprehensive results for utility and unlearning metrics. Regarding model utility, SalUn-CRA achieves the highest Balanced Accuracy (BAC) in three out of four scenarios. On DermaMNIST with 50% removal, SalUn-CRA ties with Fine-Tuning at 0.81 BAC.

Table 2. Results on binary DermaMNIST and PathMNIST datasets. Best results among approximate methods highlighted with gray background. Values in parentheses represent the difference relative to Retrain.

	Utility Metrics				Unlearning Metrics			
Method	Spec. ↑	Recall ↑	BAC ↑	AUC ↑	UBAC ↓	RBAC ↑	TBAC ↑	MIA ↓
DermaMNIST – 20% removal								
Retrain	0.83	0.77	0.80	0.90	0.24 (0.00)	0.92 (0.00)	0.80 (0.00)	21.27 (0.00)
FT	0.87	0.71	0.79	0.90	0.10 (0.14)	**0.96** (0.04)	0.79 (0.01)	10.78 (10.49)
RL	**0.90**	0.68	0.79	**0.90**	**0.04** (0.20)	0.95 (0.03)	0.79 (0.01)	**5.50** (15.77)
SalUn	0.88	0.72	0.80	**0.90**	**0.04** (0.02)	0.95 (0.03)	0.80 (0.00)	7.21 (14.06)
SalUn-CRA (Ours)	0.85	**0.78**	**0.81**	0.90	0.08 (0.16)	0.94 (0.02)	**0.81** (0.01)	17.77 (3.50)
DermaMNIST – 50% removal								
Retrain	0.81	0.76	0.79	0.87	0.22 (0.00)	0.89 (0.00)	0.79 (0.00)	21.61 (0.00)
FT	0.86	0.75	**0.81**	0.90	0.09 (0.13)	**0.96** (0.07)	0.80 (0.01)	17.53 (4.08)
RL	**0.87**	0.74	0.80	0.90	**0.05** (0.17)	0.95 (0.06)	0.80 (0.01)	**10.13** (11.48)
SalUn	0.86	0.75	0.80	**0.91**	0.06 (0.16)	0.95 (0.06)	0.80 (0.01)	13.16 (8.45)
SalUn-CRA (Ours)	0.83	**0.79**	**0.81**	0.90	0.08 (0.14)	0.94 (0.05)	**0.81** (0.02)	18.73 (2.88)
PathMNIST – 20% removal								
Retrain	0.99	0.91	0.95	0.99	0.00 (0.00)	1.00 (0.00)	0.95 (0.00)	1.64 (0.00)
FT	**0.99**	0.89	0.94	0.98	**0.00** (0.00)	1.00 (0.00)	0.94 (0.01)	**1.21** (0.43)
RL	0.98	0.91	0.95	0.99	**0.00** (0.00)	1.00 (0.00)	0.95 (0.00)	1.65 (0.01)
SalUn	0.98	0.91	0.94	0.98	0.01 (0.01)	1.00 (0.00)	0.94 (0.01)	1.52 (0.12)
SalUn-CRA (Ours)	0.98	**0.92**	**0.95**	**0.99**	**0.00** (0.00)	1.00 (0.00)	**0.95** (0.00)	1.49 (0.15)
PathMNIST – 50% removal								
Retrain	0.97	0.90	0.93	0.98	0.01 (0.00)	1.00 (0.00)	0.93 (0.00)	2.06 (0.00)
FT	**0.99**	0.89	**0.94**	0.98	0.00 (0.01)	1.00 (0.00)	**0.94** (0.01)	1.53 (0.53)
RL	**0.99**	0.89	**0.94**	**0.98**	0.00 (0.01)	1.00 (0.00)	**0.94** (0.01)	1.34 (0.72)
SalUn	**0.99**	0.88	0.93	0.98	0.00 (0.01)	1.00 (0.00)	0.93 (0.00)	**1.19** (0.87)
SalUn-CRA (Ours)	0.98	**0.91**	**0.94**	**0.98**	0.00 (0.01)	1.00 (0.00)	**0.94** (0.01)	1.98 (0.08)

More importantly, SalUn-CRA consistently achieves the highest Recall across all scenarios. This preservation of recall to malignant cases is important to reduce te clinical risk, wihle preserving a high value of BAC.

For unlearning effectiveness, all approximate methods achieve close values relative to Retrain, with differences in TBAC below 0.02 across all scenarios. The MIA metric shows that SalUn-CRA maintains values closest to Retrain, indicating unlearning behavior closer to the gold standard under privacy metrics.

4.2. Clinical Risk Analysis

Figure 1 presents clinical risk metrics across all experimental scenarios.

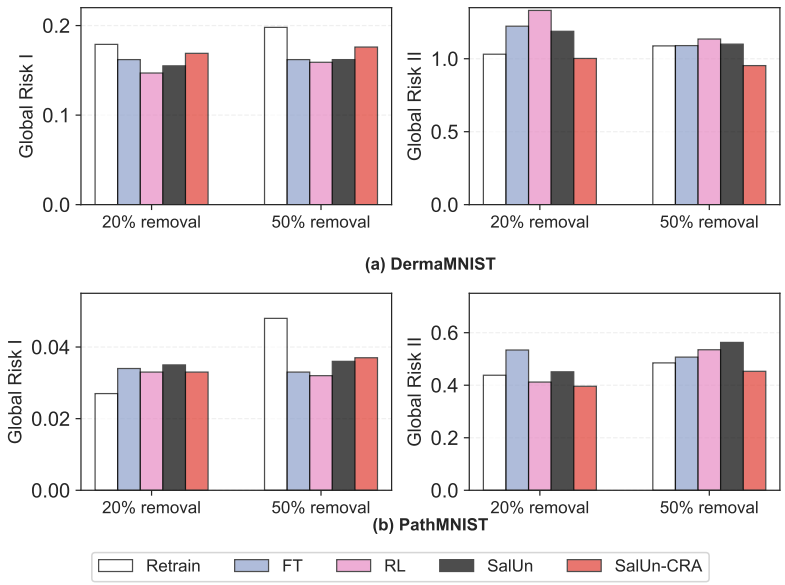

Figure 1. Clinical risk comparison across unlearning methods with forget rates 20% and 50% on DermaMNIST and PathMNIST datasets.

On DermaMNIST, FT, RL and SalUn tend to reduce Global Risk I compared to Retrain, but at the cost of increasing Global Risk II. This pattern reveals a concerning trade-off: these methods favour specificity over recall, leading to more missed malignant lesions. For instance, at 50% removal, RL achieves the lowest Global Risk I but exhibits elevated Global Risk II due to reduced sensitivity. SalUn-CRA achieves Global Risk II values lower than the standard Salun in both removal scenarios, but higher results when considering Risk I.

On PathMNIST, similar patterns emerge with smaller absolute differences. SalUn-CRA consistently achieves the lowest Global Risk II among all methods, including Retrain.

4.3. Trade-off Between Unlearning Effectiveness and Clinical Risk

Figure 2 illustrates the relationship between unlearning effectiveness (measured by GAP) and clinical risk. The ideal operating point is located in the bottom-left region of each plot, representing simultaneously low GAP (close approximation to Retrain behavior) and low Global Risk.

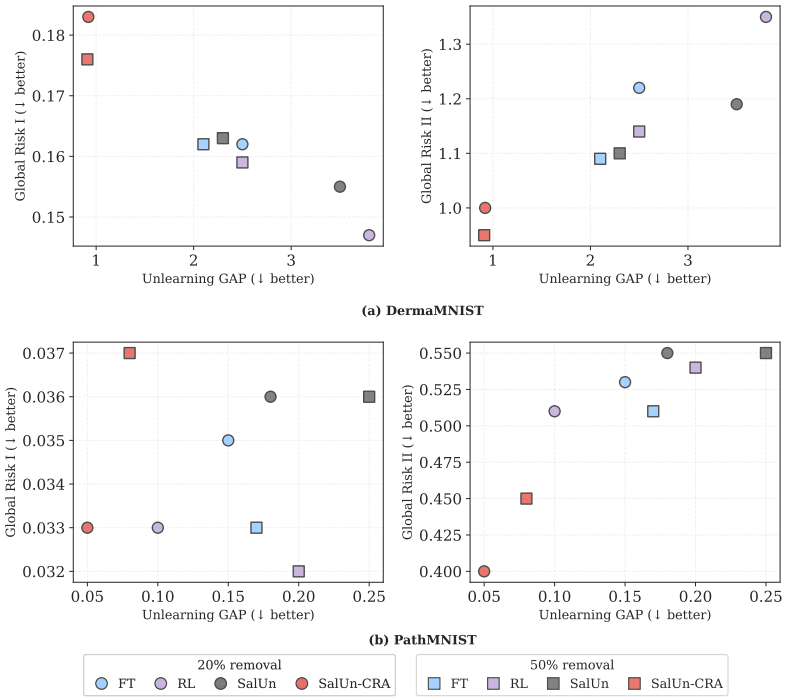

Figure 2. Trade-off between unlearning effectiveness (GAP) and Global Risk. Lower-left region represents ideal performance (low GAP and low risk). Circles indicate 20% removal; squares indicate 50% removal.

On DermaMNIST, SalUn-CRA (red markers) consistently positions itself in the favorable region for Global Risk II, achieving both low GAP and low risk. In contrast, methods such as RL achieve lower Global Risk I but exhibit higher GAP values and substantially higher Global Risk II, indicating that their apparent risk reduction comes at the cost of divergent model behavior and increased false negatives.

For PathMNIST, SalUn-CRA maintains its position as the method with the best trade-off for the most risky scenario, achieving the lowest Global Risk II while preserving near-zero GAP across both removal scenarios. These results demonstrate that SalUn-CRA successfully navigates the tension between unlearning effectiveness and clinical safety, occupying the Pareto-optimal region where other methods fail to reach.

5. Discussion

Our results highlight an important limitation in current machine unlearning validation protocols. While approximate unlearning methods such as SalUn and Random Labeling may maintain competitive performance under standard metrics (TBAC, MIA), they can increase clinical risk under asymmetric cost settings due to elevated false negative rates.

The central finding of this work is that standard machine unlearning methods can amplify clinical risk by disproportionately affecting sensitivity to malignant cases. When a data removal request is processed using conventional methods such as Fine-Tuning or Random Labeling, the model tends to become more conservative, favoring specificity over recall. While this may appear beneficial in terms of overall accuracy or balanced metrics, it translates to more missed diagnoses in practice.

This finding has direct implications for healthcare systems implementing *the right to be forgotten*. A naive deployment of unlearning algorithms could inadvertently compromise patient safety, creating a tension between privacy compliance and diagnostic reliability that must be carefully managed.

6. Conclusion

In this work, we investigated the impact of machine unlearning on clinical risk in medical image classification. We introduced Global Risk I and II as primary evaluation criteria, explicitly modeling the asymmetric cost structure of medical diagnosis. We proposed SalUn-CRA (Clinical Risk-Aware), a modification of SalUn that replaces random relabeling with entropy-based forgetting for malignant samples. Experiments on DermaMNIST and PathMNIST under 20% and 50% balanced removal demonstrate that SalUn-CRA reduces clinical risk while maintaining competitive unlearning performance.

Our findings emphasize that clinical risk should be considered a first-class evaluation criterion in medical unlearning research. Incorporating cost-sensitive validation into unlearning pipelines is essential to ensure that regulatory compliance does not compromise patient safety.

References

Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O'Gara, A., Kirk, R., Bucknall, B., Fist, T., Ong, L., Torr, P., Lam, K.-Y., Trager, R., Krueger, D., Mindermann, S., Hernandez-Orallo, J., Geva, M., and Gal, Y. (2025). Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. (2021). Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Brazil (2018). Brazilian general data protection law (law no. 13,709/2018). `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Accessed: 25 Feb. 2026.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. (2020a). Deep learning in medical image analysis. In *Deep Learning in Medical Image Analysis: Challenges and Applications*, pages 3–21. Springer.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. (2020b). Deep learning in medical image analysis. *Deep learning in medical image analysis: challenges and applications*, pages 3–21.

Deng, Z. et al. (2025). Maverick: Collaboration-free federated unlearning for medical privacy. In *Lecture Notes in Computer Science*. Springer.

European Parliament and Council of the European Union (2016). Regulation (eu) 2016/679 (general data protection regulation – gdpr). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. Accessed: 25 Feb. 2026.

Falcao, A. and Cordeiro, F. (2025). Análise de desaprendizado de máquina em modelos de classificação de imagens médicas. In *Anais Estendidos do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 43–48, Porto Alegre, RS, Brasil. SBC.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. (2024). Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*.

Golatkar, A., Achille, A., and Soatto, S. (2020). Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312.

Graves, L., Nagisetty, V., and Ganesh, V. (2021). Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524.

Haimerl, M. and Reich, C. (2025). Risk-based evaluation of machine learning-based classification methods used for medical devices. *BMC Medical Informatics and Decision Making*, 25(1):126.

Hardan, S., Taratynova, D., Essofi, A., Nandakumar, K., and Yaqub, M. (2025). Forgetmi: Machine unlearning for forgetting multimodal information in healthcare settings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 204–213. Springer.

Hoofnagle, C. J., van der Sloot, B., and Borgesius, F. Z. (2019). The european union general data protection regulation: What it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Li, N., Zhou, C., Gao, Y., Chen, H., Fu, A., Zhang, Z., and Yu, S. (2024). Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *ACM Computing Surveys*.

Ling, C. X. and Sheng, V. S. (2010). *Cost-Sensitive Learning and the Class Imbalance Problem*, pages 231–235. Springer.

Mester, S. and et al. (2024). Machine unlearning for medical imaging. *ResearchGate*.

Nasirigerdeh, R., Razmi, N., Schnabel, J. A., Rueckert, D., and Kaissis, G. (2024). Machine unlearning for medical imaging. *arXiv preprint arXiv:2407.07539*. Acesso em: 23 fev. 2025.

Sakib, S. K. and Xie, M. (2024). Machine unlearning in digital healthcare: Addressing technical and ethical challenges. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 319–322.

Scholz, R. and et al. (2024). Imbalance-aware loss functions improve medical image classification. In *Proceedings of Machine Learning Research*.

Warnecke, A. et al. (2021). Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*.

Wu, Z., Shen, C., and Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133.

Yang, J., Shi, R., Wei, D., Liu, Z., Wang, L., Zhou, Y., Zhou, S., Bian, C., Li, L., Wang, X., et al. (2021). Medmnist: A lightweight automl benchmark for medical image analysis. `https://medmnist.com`. Accessed: February 13, 2025.

Zhang, H., Nakamura, T., Isohara, T., and Sakurai, K. (2023). A review on machine unlearning. *SN Computer Science*, 4(4):337.