

Article

Mathematical Analysis of Backdoor-Induced Melanoma Risk in Modality-Heterogeneous Federated Learning

Ziyu Wang ¹, Xinyu Li ², Hongmei Chen ², Chang Xu ^{1,*} and Xuewu Zhang ^{1,*}¹ College of Information Science and Engineering, Hohai University, Changzhou 213200, China; wangziyu@hhu.edu.cn² Second Clinical Medical College, Nanjing University of Chinese Medicine, Nanjing 210023, China; 202511416@njucm.edu.cn (X.L.); 202511385@njucm.edu.cn (H.C.)

* Correspondence: xuchang@hhu.edu.cn (C.X.); 19981513@hhu.edu.cn (X.Z.)

Abstract

Aggregate accuracy and conventional attack success rates cannot distinguish pre-existing diagnostic errors from melanoma false negatives induced by a trigger. We evaluate melanoma-to-nevus risk downgrading in source-heterogeneous federated learning using paired clean and triggered predictions. Four source-defined clients represent clinical photography, smartphone imaging, and dermoscopy. The primary endpoint, trigger-induced paired false negatives (trigFN), counts clean-correct melanoma cases that change to a nevus prediction after triggering. In centralized screening, sticker and blend triggers produce 35 and 31 trigFN among 77 melanoma cases, whereas a square trigger produces none. Across four seeds, Full-Phase poisoning produces 16.25 ± 10.14 trigFN and a 21.1 ± 13.2 -percentage-point melanoma-sensitivity loss; Late-Phase poisoning produces 4.75 ± 2.99 trigFN and a 6.2 ± 3.9 -point loss. Under Full-Phase poisoning, NormClip with $c = 1$ reduces trigFN to 2.75 ± 3.20 and the sensitivity loss to 3.6 ± 4.2 points, while mean balanced accuracy (BACC) changes from 0.732 ± 0.026 to 0.708 ± 0.034 . External evaluation shows smaller within-model trigger-associated losses with clipping, although absolute melanoma sensitivity on untriggered images is lower. Paired clinical endpoints reveal harm that aggregate utility can obscure. Within the evaluated four-client, single-attacker, non-adaptive setting, norm-constrained aggregation reduces but does not eliminate harm.

Keywords: trustworthy artificial intelligence; secure federated learning; dermatology; backdoor attack; medical artificial-intelligence safety; norm clipping; melanoma sensitivity

MSC: 68T07; 68M25; 68M14



Academic Editors: Sheng Lian and Zhiming Luo

Received: 13 August 2026

Revised: 14 October 2026

Accepted: 20 October 2026

Published: 23 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Machine learning supports a broad range of healthcare tasks, including image-based diagnosis, clinical risk prediction, and decision support. Beyond imaging, recent examples include interpretable risk screening from dietary, clinical, and non-diagnostic laboratory variables and causal decision support for intensive-care discharge using electronic health records [1,2]. These applications depend on representative multi-institutional data, yet health data are difficult to centralize because of patient privacy, institutional governance, and dataset-specific access restrictions. Federated learning (FL) addresses this constraint by allowing participating sites to optimize a shared model without exchanging raw patient data [3]. Medical FL has been studied in multi-institutional imaging and clinical-outcome

prediction [4,5]. Reviews of the field nevertheless emphasize that privacy-preserving collaboration does not by itself resolve heterogeneity, reliability, or security concerns [6,7].

Melanoma recognition is a particularly safety-sensitive use case. Robust models must account for images acquired with different devices, protocols, skin tones, lesion subtypes, and clinical workflows. This makes federated training attractive for dermatology systems spanning clinical photography, smartphone imaging, and dermoscopy, but a diagnostic model that performs well on average may still fail asymmetrically on malignant lesions.

The decentralized nature of FL introduces a distinct security risk. Unlike conventional centralized training, FL allows a malicious participant to manipulate only its local data while repeatedly influencing the global model through submitted updates. BadNets established that neural networks can learn trigger-conditioned behaviors [8]. In FL, model replacement and distributed poisoning can implant targeted behavior through malicious updates [9,10], while Neurotoxin and Chameleon show that attacker objectives can be designed for persistence across later rounds [11,12]. Defenses consequently use assumptions about update distance, coordinate statistics, magnitude, clustering, smoothing, or invariant directions. Their effectiveness depends on the number of clients, non-identically distributed data, and the attacker model [13,14].

Security failures are especially consequential in medical imaging. FIBA demonstrates that frequency-domain triggers can create backdoors across several medical-image tasks, including skin-lesion classification [15]. AdaShield-FL considers poisoning together with cross-site domain variation in federated medical-image segmentation [16]. These studies establish that medical-image models can be attacked and defended under domain shift, but they do not directly measure whether a trigger converts a previously correct malignant-lesion prediction into a benign one under source-heterogeneous dermatology FL.

This distinction is important because the clinical consequence of a backdoor is not adequately captured by aggregate accuracy or a conventional target-class attack success rate. In melanoma recognition, the critical failure is *risk downgrading*: an image correctly recognized as melanoma under clean conditions is shifted to a melanocytic nevus prediction after an attacker-controlled artifact is introduced. A clinically meaningful evaluation must distinguish newly induced false negatives from pre-existing diagnostic errors. It should also separate trigger learnability from federated aggregation effects and account for differences in client modality, sample size, and class composition. To this end, we formulate paired trigger-induced melanoma false negatives as the primary harm endpoint and evaluate clean and triggered predictions on the same melanoma cases.

We study this problem in a source-partitioned melanoma versus nevi federation consisting of four modality-heterogeneous clients and one malicious clinical photography client. The federation includes clinical photographs, smartphone images, and dermoscopy, with modality heterogeneity represented across client distributions rather than through paired multimodal inputs. We first perform a centralized trigger-feasibility control to determine whether candidate artifacts are learnable by the selected model family before introducing federated poisoning. We then compare poisoning schedules and trainable parameter sets under a fixed local-epoch budget, followed by matched FedAvg and Norm-Clip evaluations. The mathematical analysis bounds direct per-round and cumulative clipped contributions; the experiments measure paired harm and class-specific utility. Matched attack-free training and external evaluation on ISIC2018 and HAM10000 further assess whether the observed defense effect is accompanied by clean-utility degradation and whether it persists beyond the development split.

Our evaluation reveals a separation between aggregate utility and clinically directed harm in the reported setting. Acquisition-inspired sticker and blend triggers are learnable in centralized controls, whereas a conventional square trigger is ineffective. Under FedAvg, sustained Full-Phase poisoning produces the largest mean paired melanoma harm across the evaluated fixed schedules, despite limited changes in balanced accuracy. Norm-constrained aggregation reduces the observed paired harm by limiting update magnitude but leaves residual false negatives and may reduce class-specific sensitivity. These observations motivate paired clinical endpoints as a complement to conventional aggregate metrics when evaluating federated backdoor risks.

The main contributions of this work are threefold. First, we establish a source-aware federated dermatology protocol that represents cross-client acquisition heterogeneity and quantifies backdoor harm through paired melanoma false negatives and melanoma-sensitivity change. Second, we evaluate trigger feasibility, poisoning timing, and the malicious client's trainable parameter set to identify conditions associated with clinically directed risk downgrading. Third, we analyze the direct per-round and cumulative magnitude contribution controlled by clipping and evaluate its empirical safety–utility profile. NormClip reduces paired harm in the evaluated single-attacker, non-adaptive setting; adaptive and colluding attackers remain to be evaluated.

2. Related Work

This section positions the study along three connected lines of work: medical FL under cross-site heterogeneity, backdoor attacks in distributed training, and aggregation defenses whose assumptions may be strained by a small number of source-defined clients.

2.1. Federated Medical Imaging and Robust Aggregation

Federated learning enables multiple institutions to train a shared model without exchanging patient-level data, making it attractive for privacy-sensitive medical imaging [4,6]. Clinical-outcome studies further demonstrate that federation can operate across institutions with different patient populations and workflows [5]. Such heterogeneity remains an optimization challenge: FedProx modifies local optimization for heterogeneous networks, while broader FL analyses distinguish statistical heterogeneity from systems heterogeneity [7,17].

Heterogeneous clients also complicate robust aggregation because honest updates can differ in magnitude and direction. Krum ranks updates by inter-client distance [18], whereas coordinate-wise median estimates a robust central tendency independently for each parameter [19]. Norm bounding directly limits the magnitude of each submitted update [20]. FLAME combines clustering, clipping, and calibrated noise to remove federated backdoors while controlling benign-performance loss [13]. More recent defenses include certified clipping-and-smoothing guarantees for bounded attacks [21], invariant aggregation over update directions [14], and reconstruction-error screening for model poisoning under non-IID data [22]. These methods protect different threat surfaces and rely on different client-count, distributional, or attacker assumptions. In our four-client setting with one malicious participant, the classical Krum condition $K \geq 2f + 3$ is not satisfied, so Multi-Krum is analyzed separately as an auxiliary stress test and is not ranked as a theoretically supported competitor.

2.2. Backdoor Attacks and Clinical Safety Evaluation

Backdoor attacks associate an attacker-controlled pattern with a selected prediction while preserving apparently normal behavior on clean inputs. BadNets established this behavior in centralized training [8]. Model replacement and distributed trigger compo-

nents extend the attack to FL [9,10]. Neurotoxin and Chameleon improve persistence by concentrating the attack on parameters or representations less likely to be overwritten by subsequent honest training [11,12]. Lockdown instead isolates subspaces to limit durable backdoor behavior [23].

Medical-image backdoors require additional attention because the clinical consequence can be class-asymmetric. FIBA evaluates frequency-injection attacks on skin-lesion classification and other medical-image tasks [15], while AdaShield-FL jointly addresses poisoning and domain adaptation in federated segmentation [16]. Their tasks, attacker access, client counts, and endpoints differ from the present binary dermatology setting, so that direct numerical rankings would be inappropriate. The methodological gap addressed here concerns measurement: a conventional target-class success rate includes melanoma cases already predicted as nevi before triggering. We instead pair clean and triggered predictions for the same cases and count only newly induced melanoma-to-nevus transitions. Centralized screening separates trigger learnability from federated aggregation, and external evaluation tests whether the paired effect persists after configuration selection.

3. Materials and Methods

The evaluation combines a source-partitioned benchmark, a fixed malicious-client threat model, a trigger-feasibility screen, attack schedules, aggregation defenses, and paired safety metrics. Figure 1 summarizes the resulting attack-defense task. Algorithm 1 summarizes the federated training and evaluation procedure.

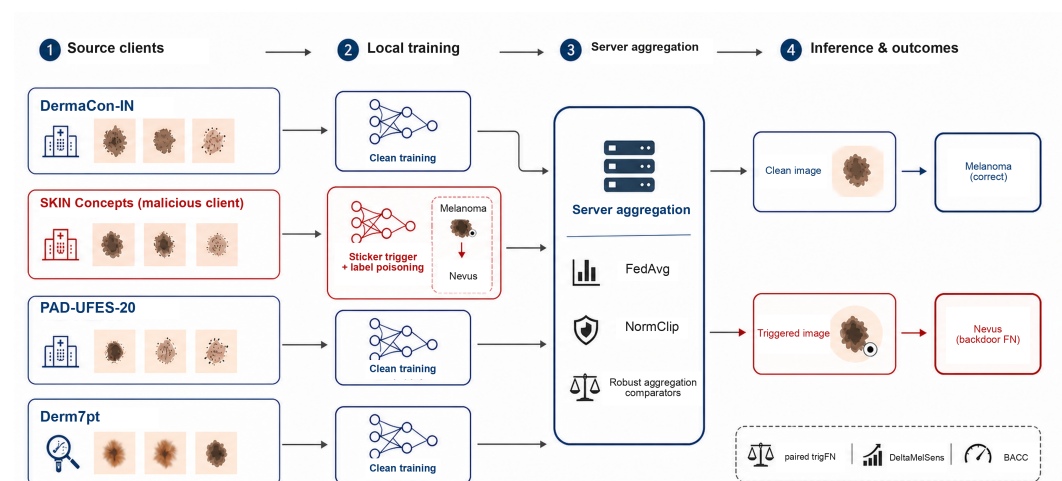


Figure 1. Source-heterogeneous federated learning (FL) backdoor threat and aggregation-defense task. One clinical-photo client implants a sticker-triggered melanoma-to-nevi downgrade through local poisoning, while the server compares FedAvg and aggregation defenses using paired trigger-induced melanoma false negatives (trigFN), change in melanoma sensitivity ($\Delta\text{MelSens}$), and balanced accuracy (BACC). Red identifies the malicious client, poisoning, and trigger-induced error pathway; blue identifies the clean clients, clean training, and clean-image prediction pathway.

3.1. Problem Formulation and Notation

Let K denote the number of participating clients and let $\mathcal{D}_i = \{(x_{ij}, \mathbf{y}_{ij})\}_{j=1}^{n_i}$ be client i 's local dataset, where $\mathbf{y} = (y^d, \mathbf{y}^m, y^r)$ contains the diagnosis, morphology, and risk targets. The server model after communication round t is θ^t , client i returns the local update Δ_i^t , and $p_i = n_i / \sum_{j=1}^K n_j$ is its sample-count weight. Client m is malicious, T is the trigger transformation, ρ is the poisoned fraction of its melanoma samples, and $a_t \in \{0, 1\}$ indicates whether the attack is active. The server aggregation rule is denoted by $\mathcal{A}$. Table 1 summarizes the notation used in the threat, optimization, and safety objectives.

Table 1. Core notation for the source-heterogeneous federated learning (FL) backdoor evaluation.

Symbol	Meaning
K, i, m	number of clients, client index, and malicious-client index
$\mathcal{D}_i, n_i, p_i$	local dataset, local sample count, and FedAvg weight
θ^t, Δ_i^t	global parameters after round t and client update during round t
$\mathbf{y} = (y^d, \mathbf{y}^m, y^r)$	diagnosis, 17-dimensional morphology, and risk targets
T, ρ, a_t	trigger, poison ratio, and attack-activity indicator
$S_m, \mathcal{M}$	poisoned local melanoma subset and evaluation melanoma set
$\mathcal{A}, \tau_t$	server aggregation rule and round-specific clipping threshold

3.2. Dataset and Federated Partition

We construct a binary melanoma-versus-nevi benchmark from DermaCon-IN [24], SKIN Concepts, PAD-UFES-20 [25], and Derm7pt [26]. Clients follow source prefixes instead of a random pooled split, preserving differences in modality, sample size, and class balance. Derm7pt contributes dermoscopy; the remaining sources contribute clinical or smartphone photographs. Here, modality-heterogeneous denotes cross-client acquisition modalities; it does not imply that multiple modalities are paired for each lesion.

The development evaluation set contains 159 cases (77 melanoma and 82 nevi) with no exact-image overlap with federated training. The trigger family, attack schedule, and clipping factor are selected on this set. SKIN Concepts is a constructed clinical-photo client containing 89 Fitzpatrick17k images (63 melanoma/26 nevi) [27] and 68 images from the Diverse Dermatology Images (DDI) dataset (13 melanoma/55 nevi) [28], rather than data from one institution. It is intended to preserve source and appearance variation within the malicious participant, but it cannot reproduce the acquisition workflow or case mix of a real hospital client. Of its 76 melanoma images, a fixed subset of 38 is selected once for poisoning and reused during active attack rounds.

Sample-weighted aggregation uses the four clients in Table 2; their nominal FedAvg weights are 2.4%, 24.8%, 2.7%, and 70.1%. PAD-UFES-20 contributes only 17 filtered training images, with 16 melanoma and one nevus, because the binary task excludes its other diagnostic categories. This extreme local imbalance can make its update sensitive to individual examples and prevent it from representing a stable hospital-level class distribution. Its nominal FedAvg weight is only 2.7%, which limits its direct contribution to the sample-weighted aggregate but does not remove the local-instability concern. A merged 632-sample pool is used only for centralized trigger-feasibility screening and is never treated as a fifth federated client.

Table 2. Source-partitioned federated learning (FL) benchmark. Counts are pre-poison training labels; poisoned n is malicious-client melanoma relabeled to nevi.

Client/Source	Modality	Role	Train n	Melanoma	Nevi	Poisoned n	Weight
DermaCon-IN	Clinical photo	Clean	15	4	11	0	2.4%
SKIN Concepts	Clinical photo	Malicious	157	76	81	38	24.8%
PAD-UFES-20	Smartphone clinical	Clean	17	16	1	0	2.7%
Derm7pt	Dermoscopy	Clean	443	210	233	0	70.1%

The binary task filters the broader dermatology label space to melanoma and nevi, so the client counts are smaller than the original multi-class collections. This focused stress test asks whether a malignant class can be selectively shifted toward a benign class by a malicious client. The protocol uses four source-defined clients, one local-data attacker, a shared multi-head model, four poisoning schedules, paired trigger-induced melanoma

false negatives (trigFN), and change in melanoma sensitivity ($\Delta\text{MelSens}$) as primary safety endpoints. Figure 2 visualizes the client composition. Preserving source membership retains the heterogeneity relevant to aggregation, while exact-image comparison is used to assess evaluation-set independence.

DermaCon-IN	SKIN Concepts	PAD-UFES-20	Derm7pt
Clinical photo	Clinical photo	Smartphone clinical	Dermoscopy
Clean client	Malicious client	Clean client	Clean client
Train n = 15	Train n = 157	Train n = 17	Train n = 443
Melanoma / nevi = 4 / 11	Melanoma / nevi = 76 / 81	Melanoma / nevi = 16 / 1	Melanoma / nevi = 210 / 233
Poisoned n = 0	Poisoned n = 38	Poisoned n = 0	Poisoned n = 0
FedAvg weight = 2.4%	FedAvg weight = 24.8%	FedAvg weight = 2.7%	FedAvg weight = 70.1%

Figure 2. Source-partitioned federated learning (FL) benchmark. The malicious SKIN Concepts client is highlighted in red; boxes report modality, role, class counts, poisoning count, and aggregation weight.

3.3. Model Architecture

We use a ResNet-50 backbone initialized from ImageNet weights with three heads: a two-class diagnosis head, a 17-dimensional morphology multi-label head, and a three-class risk triage head. Morphology and risk targets are rule-derived weak labels rather than independent expert annotations. Let $\mathbf{z}^d \in \mathbb{R}^2$, $\mathbf{z}^m \in \mathbb{R}^{17}$, and $\mathbf{z}^r \in \mathbb{R}^3$ denote the three heads' logits. For one sample, the implemented losses are

$$\begin{aligned}\mathcal{L}_{diag} &= - \sum_{c=1}^2 y_c^d \log \text{softmax}(\mathbf{z}^d)_c, \\ \mathcal{L}_{morph} &= - \frac{1}{17} \sum_{q=1}^{17} \left[y_q^m \log \sigma(z_q^m) + (1 - y_q^m) \log(1 - \sigma(z_q^m)) \right], \\ \mathcal{L}_{risk} &= - \sum_{c=1}^3 y_c^r \log \text{softmax}(\mathbf{z}^r)_c.\end{aligned}\quad (1)$$

Equation (1) matches the implemented two-class cross-entropy, multi-label binary cross-entropy, and three-class cross-entropy losses. The morphology target is $\mathbf{y}^m = g_m(y)$, where g_m activates entries in the fixed vocabulary {erythema, scale, papule, plaque, vesicle, pustule, ulcer, crust, pigment network, irregular border, asymmetry, color variation, bleeding, scar, nodule, pearl-like border, telangiectasia} according to diagnosis-keyword rules. The risk target $y^r = g_r(y)$ maps diagnosis keywords to high-risk, low-risk, or uncertain. These mappings are deterministic training targets and should not be interpreted as observed clinical attributes. The combined training objective is

$$\mathcal{L} = w_d \mathcal{L}_{diag} + w_m \mathcal{L}_{morph} + w_r \mathcal{L}_{risk}, \quad (2)$$

with weights $(w_d, w_m, w_r) = (1.0, 1.0, 0.5)$ in Equation (2). The multi-head architecture provides a common training interface for diagnosis, morphology-related semantic supervision, and risk triage across source clients. It was selected to examine whether diagnosis-targeted poisoning remains visible when the shared representation also receives auxiliary semantic gradients. The same architecture is used for centralized feasibility controls and federated experiments to avoid architectural confounding. A diagnosis-only centralized sticker control produces 34/77 trigFN, compared with 35/77 for the multi-head control. This near-matched result shows that the auxiliary heads are not necessary for centralized sticker learnability. It does not establish that the heads do not affect federated optimization or defense performance, which remains outside the present ablation scope.

3.4. Federated Optimization and Poisoning Objective

For client i with n_i local samples, the clean empirical objective and its FedAvg weight are

$$F_i(\theta) = \frac{1}{n_i} \sum_{(x,y) \in \mathcal{D}_i} \mathcal{L}(\theta; x, y), \quad p_i = \frac{n_i}{\sum_{j=1}^K n_j}. \quad (3)$$

In round $t \in \{1, \dots, R\}$, client i starts from θ^{t-1} , completes E_i local epochs, and submits

$$\Delta_i^t = \theta_i^{t, E_i} - \theta^{t-1}, \quad \theta^t = \theta^{t-1} + \sum_{i=1}^K p_i \Delta_i^t. \quad (4)$$

For the malicious client m , let a_t indicate whether poisoning is active. The Late-Phase schedule is $a_t = \mathbb{I}(t \geq t_0)$ with $t_0 = 21$. Let $\mathcal{S}_m$ be the fixed subset of melanoma images selected for poisoning. During an active round, these clean samples are replaced by triggered, nevi-labeled versions:

$$\mathcal{D}_m^{\text{poison}} = (\mathcal{D}_m \setminus \mathcal{S}_m) \cup \{(T(x), y_{\text{nevi}}) : (x, y_{\text{mel}}) \in \mathcal{S}_m\}. \quad (5)$$

Equation (5) describes the diagnosis relabeling. For a poisoned sample, the complete implemented target tuple is

$$\mathbf{y}^{bd} = (y_{\text{nevi}}^d, g_m(y_{\text{mel}}), g_r(y_{\text{nevi}})), \quad (6)$$

so Equation (6) makes the diagnosis and risk targets follow the relabeled target class while the morphology weak target retains the original melanoma-derived mapping. With per-sample multi-task loss ℓ , the malicious local objective at round t can be written as

$$F_m^t(\theta) = \frac{1}{n_m} \left[\sum_{(x,y) \in \mathcal{D}_m \setminus \mathcal{S}_m} \ell(\theta; x, \mathbf{y}) + \sum_{(x,y) \in \mathcal{S}_m} \left\{ (1 - a_t) \ell(\theta; x, \mathbf{y}) + a_t \ell(\theta; T(x), \mathbf{y}^{bd}) \right\} \right]. \quad (7)$$

Equation (7) reduces to clean local training when $a_t = 0$ and to replacement poisoning on $\mathcal{S}_m$ when $a_t = 1$. Here $|\mathcal{S}_m| = \lfloor \rho N_{m, \text{mel}} \rfloor = 38$. The malicious-client sample count remains 157; poisoning replaces rather than appends samples. The attacker does not directly replace the global model or scale its update after local training.

3.5. Trigger Family

Visual artifacts differ in learnability. A trigger that is too artificial may fail even in centralized training, while an acquisition-like artifact may be learned by the same model family. To avoid conflating trigger design with FL aggregation behavior, we evaluate triggers ranging from classic BadNets patches to acquisition-like artifacts. All trigger variants use a single fixed lower-right location in the reported poisoning and triggered-evaluation protocol. Square is a 20×20 opaque yellow block. Blend is a semi-transparent orange patch resembling gel or lighting non-uniformity. Sticker is a white annulus with a dark center, resembling a skin-attached marker or ruler dot in clinical photography. Cross is a high-contrast magenta X used as a saliency control. Figure 3 shows the complete trigger family on the same lesion schematic.

Algorithm 1 Source-heterogeneous FL backdoor evaluation and aggregation defense

Require: Client datasets $\{\mathcal{D}_i\}_{i=1}^K$, model θ^0 , trigger T , poison ratio ρ , attack start t_0 , aggregation rule $\mathcal{A}$, rounds R

- 1: **for** $t = 1, \dots, R$ **do**
- 2: Server broadcasts θ^{t-1}
- 3: **for all** clients i **in parallel do**
- 4: $\mathcal{D}_i^t \leftarrow \mathcal{D}_i$
- 5: **if** $i = m$ and $t \geq t_0$ **then**
- 6: Trigger and relabel a ρ fraction of local melanoma samples
- 7: **end if**
- 8: Optimize the multi-task objective on $\mathcal{D}_i^t$ starting from θ^{t-1}
- 9: Submit $\Delta_i^t \leftarrow \theta_i^{t,E_i} - \theta^{t-1}$
- 10: **end for**
- 11: **if** $\mathcal{A}$ is NormClip **then**
- 12: $\tau_t \leftarrow c \text{ median}_j \|\Delta_j^t\|_2$; clip each update to norm τ_t
- 13: **end if**
- 14: $\theta^t \leftarrow \theta^{t-1} + \mathcal{A}(\{\Delta_i^t\}_{i=1}^K)$
- 15: **end for**
- 16: Evaluate θ^R on clean-triggered melanoma pairs using BACC, trigFN, and $\Delta\text{MelSens}$

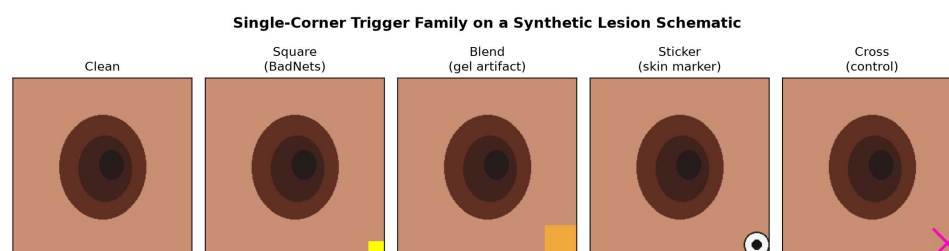


Figure 3. Single-corner trigger family shown on an author-created, non-patient lesion schematic. The schematic is used only to illustrate trigger geometry and does not represent a clinical case. The lower-right symbols denote a yellow square, a semi-transparent orange blend, a white annular sticker with a dark center, and a magenta cross, respectively.

3.6. Threat Model and Security Objective

The adversary controls one participating client, SKIN Concepts, and receives the current global model through the ordinary FL broadcast. It may modify its own local images and labels, choose when poisoning is active, train for the configured malicious-client epochs, and submit the resulting local update. It cannot inspect or modify honest-client data, alter server code, replace the global model, observe the held-out test labels, or multiply its update after local training. The honest server follows the selected aggregation rule. This is a local-data poisoning adversary rather than a model-replacement adversary.

The targeted safety failure is melanoma-to-nevi risk downgrading. For the final model θ^R , define the paired harm rate

$$H(\theta^R, T) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbb{I}[f_{\theta^R}(x_i) = y_{\text{mel}}] \mathbb{I}[f_{\theta^R}(T(x_i)) = y_{\text{nevi}}]. \quad (8)$$

For each pre-specified attack configuration $s = (\mathcal{S}_m, \{a_t\}_{t=1}^R, \mathcal{U}_s, E_m)$ in the fixed configuration set $\mathcal{S}_{\text{fixed}}$, training yields a final model θ_s^R and the corresponding harm is evaluated as

$$H_s = H(\theta_s^R, T), \quad s \in \mathcal{S}_{\text{fixed}}, \quad |\mathcal{S}_m| \leq \rho N_{m,\text{mel}}, \quad \mathcal{S}_m \subseteq \mathcal{D}_m, \quad \theta_s^t = \theta_s^{t-1} + \mathcal{A}(\{\Delta_{i,s}^t\}_{i=1}^K). \quad (9)$$

Here $\mathcal{U}_s$ is the set of malicious-client parameters updated by the optimizer, and E_m is its local-epoch budget. The set $\mathcal{S}_{\text{fixed}}$ contains the Full-Phase, Late-Phase, Head-Only, and Late+Head settings defined below; c is reserved for the NormClip multiplier. Equation (9)

specifies evaluation over fixed configurations. The attacker does not observe or adapt to the clipping threshold, directly optimize or rescale the submitted update, or implement a low-norm evasion strategy. Conclusions about NormClip apply to this non-adaptive local-data-poisoning attacker.

To separate update timing from trigger design, the trigger, target label, poisoned source class, and malicious client remain fixed. Full-Phase sets $a_t = 1$ for all rounds. Late-Phase sets $a_t = 0$ for rounds 1–20 and $a_t = 1$ for rounds 21–30. Head-Only applies poisoning throughout training while freezing the backbone and updating the diagnosis head; Late+Head combines the Late window with Head-Only optimization. Honest clients use three local epochs, and the malicious client uses five in every round, including the unpoisoned rounds of Late-Phase and Late+Head. The schedule changes poisoning activity, not the local-epoch budget. At evaluation, the trigger is applied online while melanoma ground-truth labels remain unchanged.

In a managed FL deployment, the server can prescribe local epochs and optimization settings, but software-only coordination does not prove that an untrusted client followed them. Trusted execution, signed training code, or verifiable telemetry would be needed for stronger enforcement. Abrupt changes in reported loss, training time, or update geometry may support anomaly screening, as illustrated by loss-sequence and update-based defenses, but client-reported loss can itself be manipulated. Our threat model permits additional local epochs and treats update clipping as magnitude control, without assuming attacker identification.

The defense goal is to reduce H without an unacceptable loss of clean utility, measured using balanced accuracy (BACC). Conceptually, a defense setting d is evaluated through

$$\min_d H(\theta_d^R, T) \quad \text{subject to} \quad \text{BACC}(\theta_d^R) \geq \text{BACC}(\theta_{\text{clean}}^R) - \epsilon. \quad (10)$$

Because no clinical tolerance ϵ was prospectively specified, we report harm and diagnostic utility jointly. Candidate clipping factors $c \in \{1, 2, 3\}$ are evaluated on the development set. We use $c = 1$ as the conservative safety setting because it is the only tested value with zero mean trigFN across all four Late-Phase runs; $c = 2$ and $c = 3$ provide less restrictive reference points. This post hoc benchmark choice is not a universally optimal threshold, and the full repeated-run sweep is reported in Section 4.8.

3.7. Defenses

Robust aggregation methods make different assumptions about client-update geometry. In small, heterogeneous dermatology FL, distance-based selectors may be unreliable, whereas norm constraints still bound each submitted update. For reference, FedAvg aggregates

$$\Delta_{\text{FedAvg}}^t = \sum_{i=1}^K p_i \Delta_i^t. \quad (11)$$

Equation (11) preserves sample-count weighting. Norm-Clipping follows Sun et al. [20]. With $\tau_t = c \cdot \text{median}_j \|\Delta_j^t\|_2$, the ideal clipping operator is

$$\tilde{\Delta}_i^t = \begin{cases} \Delta_i^t, & \|\Delta_i^t\|_2 \leq \tau_t, \\ \frac{\tau_t}{\|\Delta_i^t\|_2} \Delta_i^t, & \|\Delta_i^t\|_2 > \tau_t, \end{cases} \quad (12)$$

Equation (12) is followed by $\Delta_{\text{NormClip}}^t = \sum_i p_i \tilde{\Delta}_i^t$. The implementation adds a 10^{-12} numerical safeguard to the denominator. Multi-Krum is a distance-based robust aggregator

derived from Krum [18]. Let $\mathcal{N}_i$ contain the $q = \max(1, K - f - 2)$ updates nearest to Δ_i^t in Euclidean distance. The implemented score is

$$s_i^t = \sum_{j \in \mathcal{N}_i} \|\Delta_i^t - \Delta_j^t\|_2. \quad (13)$$

Using Equation (13), the default implementation selects the $k = K - f$ lowest-scoring updates and returns their equally weighted mean. Thus, for $K = 4$ and $f = 1$, each score uses one nearest neighbor and three updates are selected. This experiment is consequently an empirical stress comparator outside the usual Krum guarantee, not a test of its certified regime.

Proposition 1 (bounded single-client influence). *Let $\tau_t = c \operatorname{median}_j \|\Delta_j^t\|_2$ and let p_m be the aggregation weight of client m . NormClip guarantees*

$$\|p_m \tilde{\Delta}_m^t\|_2 \leq p_m \tau_t. \quad (14)$$

Moreover, if one submitted update is replaced by any other update after both are clipped at the same threshold, the aggregate changes by at most

$$\left\| \sum_i p_i \tilde{\Delta}_i^t - \sum_i p_i \tilde{\Delta}_i^{t'} \right\|_2 \leq 2p_m \tau_t. \quad (15)$$

Proof. Inequality (14) follows from $\|\tilde{\Delta}_m^t\|_2 \leq \tau_t$. If the two aggregates differ only in client m 's clipped update, their distance is $p_m \|\tilde{\Delta}_m^t - \tilde{\Delta}_m^{t'}\|_2$, which is at most $p_m (\|\tilde{\Delta}_m^t\|_2 + \|\tilde{\Delta}_m^{t'}\|_2) \leq 2p_m \tau_t$ by the triangle inequality, establishing Inequality (15). $\square$

Corollary 1 (cumulative direct contribution over active rounds). *Let $\mathcal{I} \subseteq \{1, \dots, R\}$ be the rounds in which client m participates. Along a realized sequence of clipping thresholds $\{\tau_t\}_{t \in \mathcal{I}}$, define the accumulated clipped contribution of that client as*

$$G_m(\mathcal{I}) = \sum_{t \in \mathcal{I}} p_m \tilde{\Delta}_m^t. \quad (16)$$

Then

$$\|G_m(\mathcal{I})\|_2 \leq p_m \sum_{t \in \mathcal{I}} \tau_t. \quad (17)$$

If the submitted update of client m is replaced in every round and both sequences are clipped using the same realized thresholds, the difference between their cumulative direct contributions is at most

$$2p_m \sum_{t \in \mathcal{I}} \tau_t. \quad (18)$$

Proof. Apply Inequalities (14) and (15) in each round and use the triangle inequality across $t \in \mathcal{I}$. $\square$

Proposition 1 and Corollary 1 describe magnitude accounting, not semantic backdoor robustness. The cumulative result does not bound the distance between two complete training trajectories because honest-client updates and thresholds may change as the global model changes. An adaptive attacker can also remain below the threshold, and the realized τ_t depends on heterogeneous client updates. The theory explains which component of the observed high-norm attack is directly constrained; it does not certify convergence or prevent all backdoors.

For P trainable parameters and K participating clients, FedAvg and NormClip require $O(KP)$ arithmetic and memory-linear streaming over client updates; finding the median of the K scalar norms adds at most $O(K \log K)$. Multi-Krum evaluates pairwise update distances and requires $O(K^2P)$. NormClip preserves the server-side order of FedAvg while bounding update magnitude.

3.8. Metrics and Statistical Analysis

Diagnostic utility. With melanoma treated as positive, let TP, FN, TN, and FP denote true positives, false negatives, true negatives, and false positives, respectively. Then $\text{MelSens} = \text{TP}/(\text{TP} + \text{FN})$ and $\text{BACC} = \frac{1}{2}[\text{MelSens} + \text{TN}/(\text{TN} + \text{FP})]$. BACC weights melanoma and nevi recall equally, whereas MelSens isolates high-risk detection. Clean values characterize baseline utility and prevent a defense-induced loss of diagnostic performance from being mistaken for attack suppression.

Paired clinical safety. For melanoma cases $\mathcal{M}$ and trigger transform T , the primary paired endpoints are

$$\begin{aligned} \text{trigFN} &= \sum_{i \in \mathcal{M}} \mathbb{I}[f(x_i) = \text{mel}] \mathbb{I}[f(T(x_i)) = \text{nevi}], \\ \Delta \text{MelSens} &= \text{MelSens}_T - \text{MelSens}_C. \end{aligned} \quad (19)$$

trigFN counts clean-correct melanoma images newly shifted to nevi; negative $\Delta \text{MelSens}$ denotes trigger-associated sensitivity loss and captures reverse paired transitions.

Let N_{01} count melanoma cases predicted as melanoma when clean but as nevi after triggering, and let N_{10} count the reverse transition. Then

$$\text{trigFN} = N_{01}, \quad \Delta \text{MelSens} = \frac{N_{10} - N_{01}}{|\mathcal{M}|}. \quad (20)$$

Equivalently, $-\Delta \text{MelSens} = (\text{trigFN} - N_{10})/|\mathcal{M}|$. When no reverse transition occurs, the sensitivity loss equals the trigFN rate. No reverse transition occurs in any Full-Phase FedAvg or Full-Phase NormClip run, so the identity holds separately for each seed; means and sample standard deviations are then computed across seeds.

Auxiliary indicator. ASR_{hr} is defined as

$$\text{ASR}_{\text{hr}} = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} \mathbb{I}[f(T(x_i)) = \text{nevi}]. \quad (21)$$

ASR_{hr} includes baseline and newly induced false negatives. In attack-free controls, it describes trigger-transformed behavior and is reported only as auxiliary context.

For held-out checks, $\Delta \text{MelSens}$ is primary because ASR_{hr} can conflate attack effects with out-of-distribution baseline errors. Suppression is

$$1 - \frac{|\Delta \text{MelSens}_{\text{defense}}|}{|\Delta \text{MelSens}_{\text{attack}}|}.$$

Repeated runs use the mean $\pm$ sample standard deviation over seeds 42, 0, 1, and 7. Matched-seed comparisons use two-sided paired t tests and Wilcoxon signed-rank tests as exploratory checks. With only four seed pairs, the minimum attainable two-sided Wilcoxon p value is 0.125; therefore, effect sizes, per-seed values, and their consistency are primary, and no multiplicity-adjusted confirmatory claim is made.

4. Experiments and Results

The experiments isolate trigger learnability, FL aggregation dilution, malicious-update timing, and defense-induced suppression.

4.1. Implementation Details

Images are resized to 224×224 and normalized with ImageNet statistics. The backbone is ImageNet-pretrained ResNet-50 with diagnosis, morphology, and risk heads. FL runs for 30 rounds using AdamW (learning rate 5×10^{-5} , weight decay 10^{-4} , batch size 32). In the attack experiments, honest clients use three local epochs, and the malicious client uses five in every round across all four schedules. Head-only settings change the trainable parameter set, not this epoch budget. The attack uses a 32 px lower-right sticker trigger and replaces 50% of malicious-client melanoma samples. Full-Phase poisoning is active in rounds 1–30, whereas Late-Phase poisoning is active in rounds 21–30. The conservative NormClip setting uses $c = 1.0$; the threshold sweep also evaluates $c = 2.0$ and $c = 3.0$. All repeated analyses use seeds 42, 0, 1, and 7 while sharing the same frozen seed-42 poison subset and evaluation sets. Experiments were executed with Python 3.12.9, PyTorch 2.7.1+cu126, torchvision 0.22.1+cu126, and NumPy 2.2.6 on an NVIDIA RTX 4090D GPU (NVIDIA Corporation, Santa Clara, CA, USA). The schedule, threshold, defense, clean-training, and external evaluations each summarize four runs with distinct training seeds.

4.2. Dataset Provenance and Evaluation-Set Independence

All experiments use existing public or access-controlled dermatology datasets; no new patient images or human-subject intervention were collected. The manuscript reports aggregate metrics and derived sample lists only, while raw images and triggered derivatives remain under the original providers' terms. Exact-image comparison found no duplicate images between federated training, the development set, ISIC2018, and HAM10000. The acquisition characteristics and intended roles of the contributing datasets are described in Section 3.2 and Table 2.

4.3. Centralized Trigger-Feasibility Control

The centralized feasibility-control experiment separates trigger learnability from FL aggregation. With the same poison construction on a merged training pool, square and cross triggers produce 0 paired trigger-induced false negatives, whereas blend and sticker artifacts produce 31/77 and 35/77, respectively. The single-task sticker control produces 34/77, indicating that sticker learnability is not created by the auxiliary heads. Figure 4 jointly reports paired harm and BACC. The FL experiments use the sticker trigger as the main attack.

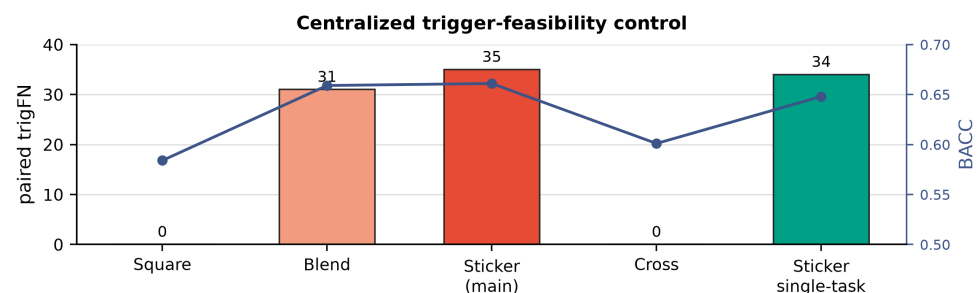


Figure 4. Centralized trigger-feasibility control. Bars show paired trigFN among 77 melanoma cases; the blue line reports BACC. Paired trigFN counts only cases correctly classified when clean but shifted to nevi after triggering.

Subsequent experiments use this screened trigger to evaluate aggregation and defense behavior.

4.4. Federated Attack Schedules

The schedule comparison was repeated with the same four seeds for every condition (Table 3). Full-Phase poisoning produces the largest mean paired harm: 16.25 ± 10.14 trigFN and $\Delta\text{MelSens} = -0.211 \pm 0.132$. Late-Phase poisoning produces 4.75 ± 2.99 trigFN and $\Delta\text{MelSens} = -0.062 \pm 0.039$. The Full-Phase per-seed trigFN counts are 7, 26, 24, and 8; the Late-Phase counts are 4, 6, 8, and 1. Thus, the ordering is consistent for all four matched seeds, although its magnitude remains variable. Head-Only and Late+Head both produce zero trigFN in all four runs. In one Head-Only run, a melanoma case predicted as nevi without the trigger changes to melanoma after triggering; this reverse prediction transition gives a small positive mean $\Delta\text{MelSens}$ of 0.003 ± 0.006 and reflects prediction variability rather than attack success. These results indicate that sustained backbone optimization is important for the evaluated attack; they do not identify an optimal adaptive schedule.

Table 3. Numerical summary of the attack-schedule and matched Full-Phase defense comparisons. Values are mean $\pm$ sample standard deviation across four seeds. Negative $\Delta\text{MelSens}$ indicates trigger-associated melanoma-sensitivity loss.

Setting	BACC	Paired trigFN	$\Delta\text{MelSens}$
<i>Attack schedules under FedAvg</i>			
Full-Phase	0.732 ± 0.026	16.25 ± 10.14	-0.211 ± 0.132
Late-Phase	0.718 ± 0.021	4.75 ± 2.99	-0.062 ± 0.039
Head-Only	0.722 ± 0.015	0.00 ± 0.00	0.003 ± 0.006
Late+Head	0.738 ± 0.013	0.00 ± 0.00	0.000 ± 0.000
<i>Matched aggregation comparison under the Full-Phase attack</i>			
FedAvg	0.732 ± 0.026	16.25 ± 10.14	-0.211 ± 0.132
NormClip ($c = 1$)	0.708 ± 0.034	2.75 ± 3.20	-0.036 ± 0.042

Figure 5 displays the schedule variability and matched defense comparison.

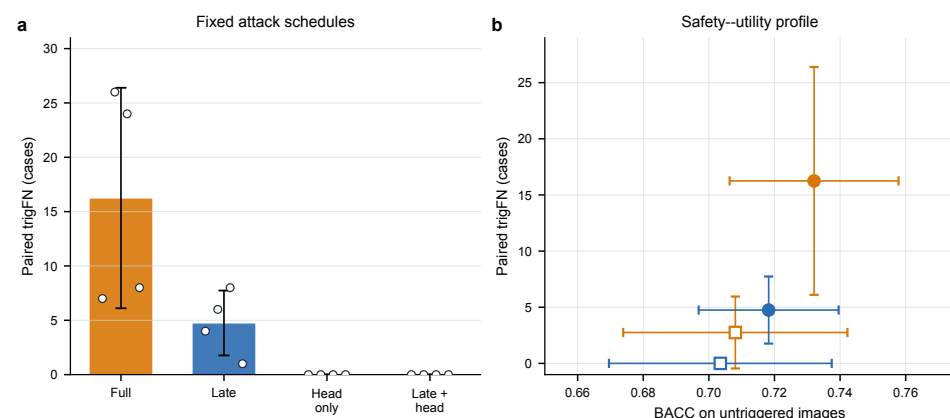


Figure 5. Repeated attack and defense results across four seeds (mean $\pm$ sample standard deviation). (a) Paired trigFN for the four fixed poisoning schedules. (b) Safety-utility plane for FedAvg (filled circles) and NormClip ($c = 1$; open squares) under Full-Phase (orange) and Late-Phase (blue) poisoning; horizontal and vertical error bars show the sample standard deviations of BACC and trigFN.

4.5. Aggregation-Defense Evaluation

Under Full-Phase poisoning, NormClip with $c = 1$ reduces mean trigFN from 16.25 ± 10.14 to 2.75 ± 3.20 , an 83% reduction in the observed mean. Mean $\Delta\text{MelSens}$ improves from -0.211 ± 0.132 to -0.036 ± 0.042 , while BACC changes from 0.732 ± 0.026 to 0.708 ± 0.034 . Across matched seeds, the mean trigFN difference is -13.5 (exploratory paired t test $p = 0.030$; Wilcoxon $p = 0.125$); the BACC difference is -0.024 ($p = 0.479$ and 0.625 ,

respectively). The discrepancy between the parametric and rank-test results, together with $n = 4$, precludes a confirmatory significance claim. The effect size and per-seed counts instead support the bounded conclusion that clipping markedly reduces, but does not eliminate, this high-magnitude non-adaptive attack.

The remaining paired harm and lower melanoma sensitivity on untriggered test images under clipping require joint interpretation. Full-Phase NormClip has untriggered-image MelSens 0.575 ± 0.086 , compared with 0.662 ± 0.076 under Full-Phase FedAvg. NormClip constrains update magnitude without identifying the malicious client, so honest updates can also be clipped. The reduction in paired harm is accompanied by lower melanoma sensitivity on untriggered images.

4.6. Auxiliary Multi-Krum Stress Test

Multi-Krum is not included in the principal defense ranking because $K = 4$ and $f = 1$ do not satisfy the usual $K \geq 2f + 3$ condition. We retain it only as an auxiliary test of how a distance-based selector behaves when applied outside that regime. Under Late-Phase poisoning it yields BACC 0.561 ± 0.019 , clean MelSens 0.513 ± 0.106 , trigFN 34.00 ± 10.71 , and $\Delta\text{MelSens}$ -0.442 ± 0.139 . Source heterogeneity can make honest updates mutually distant, weakening the distance-ranking assumption. These results demonstrate poor fit to this specific small, non-identically distributed geometry; they are not evidence that Multi-Krum is generally inferior to NormClip.

4.7. Paired Effect Uncertainty and Directionality

For Full-Phase FedAvg, the per-seed trigFN counts are 7, 26, 24, and 8, with no reverse nevi-to-melanoma transition among the paired melanoma predictions. With Full-Phase NormClip, the corresponding counts are 0, 6, 5, and 0, again with no reverse transition. Thus, the mean sensitivity changes follow directly from Equation (20). The seed-level results show both directional consistency and substantial stochastic variability. Zero events in two clipped runs do not imply zero population risk, and the two nonzero runs show that the defense does not eliminate the attack.

4.8. Sensitivity Analyses

The repeated Late-Phase sweep (Table 4) shows a monotonic weakening of the observed safety effect as clipping becomes less restrictive. At $c = 1$, mean trigFN is 0.00 ± 0.00 , and BACC is 0.703 ± 0.034 . At $c = 2$, trigFN increases to 1.50 ± 1.29 and BACC is 0.705 ± 0.029 ; at $c = 3$, trigFN is 2.50 ± 2.52 and BACC is 0.717 ± 0.016 . The small positive $\Delta\text{MelSens}$ at $c = 1$ is caused by two cases that change from an untriggered nevi prediction to a triggered melanoma prediction across the four runs.

Table 4. NormClip threshold sensitivity under Late-Phase poisoning. Values are mean $\pm$ sample standard deviation across four seeds.

Clip Factor c	BACC	Paired trigFN	$\Delta\text{MelSens}$
1.0	0.703 ± 0.034	0.00 ± 0.00	0.006 ± 0.008
2.0	0.705 ± 0.029	1.50 ± 1.29	-0.016 ± 0.022
3.0	0.717 ± 0.016	2.50 ± 2.52	-0.033 ± 0.033

The repeated sweep supports choosing $c = 1$ when the development objective prioritizes avoidance of newly induced melanoma false negatives. It also shows why trigFN alone is insufficient for threshold selection: mean MelSens on untriggered test images is 0.575 ± 0.080 at $c = 1$, 0.565 ± 0.039 at $c = 2$, and 0.610 ± 0.038 at $c = 3$. A deployment threshold would require a prospectively defined clinical tolerance rather than retrospective selection on this benchmark.

4.9. Additional Aggregation and Clean-Utility Controls

The attack-free control (Table 5) measures the class-specific cost of clipping independently of poisoned training. Across four matched seeds, clean FedAvg yields BACC 0.745 ± 0.009 , melanoma sensitivity 0.740 ± 0.051 , and nevi recall 0.750 ± 0.066 . NormClip with $c = 2$ yields 0.756 ± 0.009 , 0.737 ± 0.046 , and 0.774 ± 0.029 , respectively; the stricter $c = 1$ setting yields 0.729 ± 0.012 , 0.750 ± 0.022 , and 0.707 ± 0.017 . Relative to clean FedAvg, $c = 1$ changes mean BACC by -1.65 percentage points, melanoma sensitivity by $+0.97$ points, and nevi recall by -4.27 points. Its aggregate-utility cost in attack-free training is therefore concentrated in nevi recall rather than melanoma sensitivity. All three settings produce zero melanoma-to-nevi trigFN in the paired trigger evaluation. With four seeds, these comparisons are descriptive: for $c = 1$ versus FedAvg, the exploratory paired t -test p value for BACC is 0.143, and the exact Wilcoxon p value is 0.250.

Table 5. Attack-free clean-utility controls. Values are mean $\pm$ sample standard deviation across seeds 42, 0, 1, and 7. NeviRec denotes nevi recall.

Training Setting	BACC	MelSens	NeviRec	Paired trigFN
FedAvg	0.745 ± 0.009	0.740 ± 0.051	0.750 ± 0.066	0.00 ± 0.00
NormClip, $c = 2$	0.756 ± 0.009	0.737 ± 0.046	0.774 ± 0.029	0.00 ± 0.00
NormClip, $c = 1$	0.729 ± 0.012	0.750 ± 0.022	0.707 ± 0.017	0.00 ± 0.00

4.10. External Safety Evaluation

After configuration selection on the development set, Late-Phase FedAvg and NormClip ($c = 1$) models from four training seeds are evaluated without further tuning. On ISIC2018 Task 3 ($n_{\text{mel}} = 171$) [29], FedAvg produces 12.75 ± 9.32 trigFN and $\Delta\text{MelSens} = -0.070$, whereas NormClip produces 1.50 ± 1.73 trigFN and $\Delta\text{MelSens} = -0.006$. On HAM10000 ($n_{\text{mel}} = 435$) [30], the corresponding values are 37.25 ± 27.56 and -0.084 for FedAvg versus 2.50 ± 2.65 and -0.004 for NormClip. These are within-model paired trigger effects. Absolute MelSens on untriggered images is lower with NormClip: 0.399 ± 0.117 versus 0.532 ± 0.140 on ISIC2018 and 0.363 ± 0.101 versus 0.494 ± 0.115 on HAM10000. Figure 6 summarizes the paired losses and untriggered-image sensitivities. The external comparison therefore supports reduced trigger-associated loss, not improved external diagnostic accuracy.

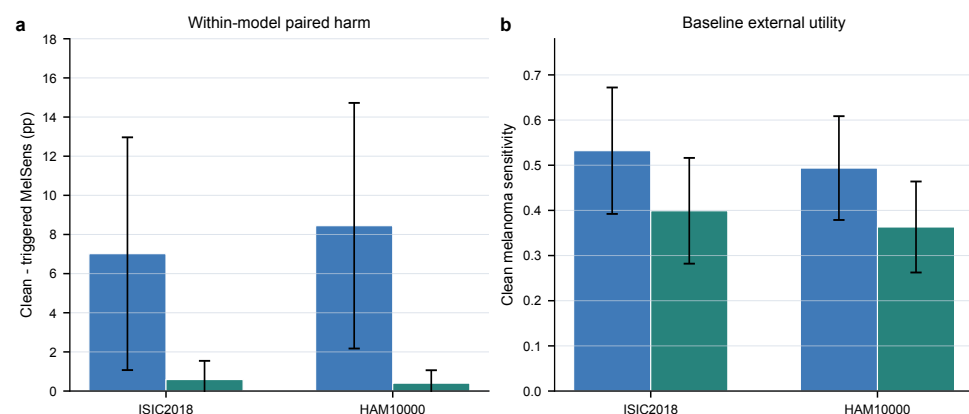


Figure 6. External evaluation across four training seeds (mean $\pm$ sample standard deviation). In each group, the left blue bar is Late-Phase FedAvg and the right green bar is NormClip ($c = 1$). (a) Trigger-associated melanoma-sensitivity change, expressed as clean minus triggered sensitivity in percentage points. (b) Clean melanoma sensitivity of the same models, showing that the smaller paired loss under clipping coexists with lower baseline sensitivity.

5. Discussion

The central result is a measurement result as well as an attack result. Across four seeds, Full-Phase poisoning creates 16.25 ± 10.14 melanoma-to-nevi transitions among 77 melanoma cases and lowers melanoma sensitivity by 21.1 ± 13.2 percentage points, while clean BACC remains 0.732 ± 0.026 . A conventional trigger target rate would also count cases already classified as nevi before the trigger was applied. By pairing the clean and triggered prediction for each melanoma case, trigFN instead isolates newly induced errors, and the transition identity links those errors directly to the observed sensitivity change. This makes the clinical direction of harm visible without treating every target-class prediction as attack-caused.

The defense results must be interpreted jointly with diagnostic utility. Under matched Full-Phase poisoning, NormClip ($c = 1$) changes mean trigFN from 16.25 ± 10.14 to 2.75 ± 3.20 and $\Delta\text{MelSens}$ from -0.211 ± 0.132 to -0.036 ± 0.042 , while BACC changes from 0.732 ± 0.026 to 0.708 ± 0.034 . This is an 83% reduction in observed mean paired harm, with residual events in two seeds. Clean MelSens in the same attacked-training comparison changes from 0.662 ± 0.076 to 0.575 ± 0.086 . The attack-free $c = 1$ control localizes the baseline cost: mean melanoma sensitivity is 0.750 ± 0.022 versus 0.740 ± 0.051 for FedAvg, while the BACC change from 0.745 ± 0.009 to 0.729 ± 0.012 is driven by lower nevi recall. Thus, the lower melanoma sensitivity observed under Full-Phase poisoning is not an invariant effect of $c = 1$ clipping in clean training. The four-seed contrast is descriptive and varies by seed, so it motivates an attack-by-clipping interaction hypothesis rather than a causal interaction claim. In this benchmark, threshold selection therefore requires both class-specific utility measures and the paired safety endpoint, with operating tolerances defined prospectively for clinical use.

Comparison with prior defenses further defines this scope. FLAME combines clustering, norm clipping, and noise [13]; CRFL obtains certified guarantees under its clipping-and-smoothing construction [21]; invariant aggregation and FedREDefense use update-direction invariance and reconstruction error, respectively [14,22]. Those methods target related poisoning threats but make different assumptions about client counts, honest-update structure, server processing, and attack construction. Likewise, FIBA studies frequency-domain medical-image triggers [15], and AdaShield-FL addresses poisoning together with domain adaptation in federated segmentation [16]. Because the tasks, endpoints, client geometries, and attacker access differ, their reported values are not directly comparable with our melanoma trigFN. The present contribution is complementary: it provides a paired, clinically directional endpoint and evaluates a transparent magnitude-control baseline in a four-source dermatology federation.

Proposition 1 and Corollary 1 explain the magnitude constraint underlying NormClip: clipping limits one client's direct sample-weighted contribution per round and its cumulative direct contribution across active rounds. As discussed in Section 3.7, the full training trajectory also depends on changes in honest-client updates and clipping thresholds. The experiments evaluate pre-specified, non-adaptive configurations; threshold-aware attacks that distribute a backdoor across low-norm updates remain to be evaluated.

Multi-Krum's low triggered MelSens and high trigFN were observed in the four-client, one-attacker stress test described in Section 4.6. These auxiliary results are interpreted in the context of its client-count assumptions and class-specific utility.

Dataset construction creates additional limits on external validity. The clients preserve source membership and span clinical photography, smartphone photography, and dermoscopy, but they are not a prospective sample of four hospitals. SKIN Concepts combines images from Fitzpatrick17k and DDI, and the filtered PAD-UFES-20 client contains only 17 cases with a 16:1 melanoma-to-nevus ratio. These choices are useful for a controlled hetero-

geneity stress test but do not reproduce institutional prevalence, acquisition workflow, or client sampling. The multi-head model similarly uses rule-derived morphology and risk labels, not independent expert annotations. The study concerns cross-client acquisition heterogeneity with shared weak semantic supervision; it does not evaluate paired multimodal fusion for the same lesion.

Evidence from the development and external evaluations serves different purposes. Trigger selection, schedule analysis, and clipping-factor selection use the development set and support the principal measurement and safety–utility conclusions. ISIC2018 and HAM10000 are evaluated after configuration selection using the Late-Phase models and show that the within-model trigger-associated sensitivity loss remains smaller with NormClip across four seeds. Because absolute clean sensitivity is both moderate and lower under clipping on those datasets, the external results support persistence of the paired trigger-suppression effect only; they do not establish improved external diagnostic accuracy or clinical generalizability.

Future work is prioritized in three steps. First, the defense should be tested against threshold-aware low-norm attacks, model replacement, and multiple colluding clients while recording how harm accumulates across rounds. Second, source effects should be separated from sample-size and class-balance effects through matched-size client ablations, repeated client sampling, and additional institutional datasets. Third, the endpoint should be extended to multi-class triage, alternative architectures, physically captured acquisition artifacts, and prospective reader- or workflow-level validation. These extensions would test, respectively, adversarial robustness, federated-design robustness, and clinical transportability.

6. Conclusions

This study shows why paired, class-specific safety endpoints are needed when aggregate utility can conceal a clinically directed federated backdoor. Across four seeds, Full-Phase poisoning shifts 16.25 ± 10.14 of 77 melanoma cases to nevi and reduces melanoma sensitivity by 21.1 ± 13.2 percentage points. Under the same schedule, NormClip ($c = 1$) reduces trigFN to 2.75 ± 3.20 and the sensitivity loss to 3.6 ± 4.2 percentage points, while mean BACC changes from 0.732 ± 0.026 to 0.708 ± 0.034 . These values quantify the reduction in paired harm and the accompanying loss of balanced accuracy in the evaluated configuration.

In attack-free training, the same $c = 1$ setting yields melanoma sensitivity 0.750 ± 0.022 versus 0.740 ± 0.051 for FedAvg, while BACC changes from 0.745 ± 0.009 to 0.729 ± 0.012 through lower nevi recall. This control distinguishes the baseline class-specific cost of stricter clipping from the sensitivity pattern observed when poisoning and clipping are combined.

The main experiments establish the paired harm metric, show that Full-Phase backbone poisoning is more damaging than the other evaluated fixed schedules, and demonstrate partial suppression by norm-constrained aggregation. The external evaluations have a narrower role: under Late-Phase poisoning, the remaining mean paired melanoma-sensitivity losses with NormClip are 0.6 and 0.4 percentage points on ISIC2018 and HAM10000, respectively, but clean sensitivity is lower. The mathematical analysis explains how clipping bounds direct update magnitude over one or multiple rounds. These conclusions are limited to a four-client, one-attacker, binary ResNet-50 setting with digital triggers and non-adaptive poisoning. Threshold-aware attacks, collusion, broader client populations, multi-class workflows, and prospective clinical validation remain necessary before deployment claims can be made.

Author Contributions: Conceptualization, Z.W., C.X. and X.Z.; methodology, Z.W., C.X. and X.Z.; software, Z.W.; validation, Z.W.; formal analysis, Z.W.; investigation, Z.W.; data curation, Z.W.; writing—original draft preparation, Z.W.; writing—review and editing, X.L., H.C., C.X. and X.Z.; visualization, Z.W.; supervision, C.X. and X.Z.; project administration, C.X. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Changzhou Science and Technology Program, grant number CJ20259010.

Institutional Review Board Statement: It is waived This study used only de-identified secondary dermatology datasets and involved no new participant recruitment, intervention, or collection of identifiable patient information. Ethical oversight for the source datasets was governed by the original data providers. The authors did not access identifiable data.

Informed Consent Statement: It is waived This study involved no new participant recruitment, and the authors did not access identifiable participant information. Consent procedures for the source datasets were governed by the original data providers.

Data Availability Statement: Restrictions apply to the availability of these third-party data. The datasets were obtained from the original providers identified and cited in this article and are available through those providers, subject to their respective access and license terms. Derived manifests, aggregate results, and code supporting the reported experiments and figures are available from the corresponding author upon reasonable request. Raw images and triggered derivatives cannot be redistributed by the authors because they remain subject to the original providers' terms.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pourmahmoudi, A.; Malekzadeh, J.; Noormohammadi, A.; Salehnasab, C.; Ghaderzadeh, M. Predicting Anemia Risk in Iranian Adults without Diagnostic Blood Indices: An Interpretable Machine Learning Approach Integrating Dietary, Clinical, and Non-Diagnostic Laboratory Patterns. *Sci. Rep.* **2026**. Advance online publication. <https://doi.org/10.1038/s41598-026-72618-4>.
2. Momenzadeh, M.; Ghaderzadeh, M.; Oshaghi, A.; Momenzadeh, M. A Causal Machine Learning Framework for ICU Discharge Decision Support Using Electronic Health Records. *Sci. Rep.* **2026**. Advance online publication. <https://doi.org/10.1038/s41598-026-70894-8>.
3. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the AISTATS, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
4. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R.; et al. Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Sci. Rep.* **2020**, *10*, 12598.
5. Dayan, I.; Roth, H.R.; Zhong, A.; Harouni, A.; Gentili, A.; Abidin, A.Z.; Liu, A.; Costa, A.B.; Wood, B.J.; Tsai, C.-S.; et al. Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nat. Med.* **2021**, *27*, 1735–1743.
6. Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The Future of Digital Health with Federated Learning. *npj Digit. Med.* **2020**, *3*, 119.
7. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210.
8. Gu, T.; Dolan-Gavitt, B.; Garg, S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv* **2017**, arXiv:1708.06733. <https://doi.org/10.48550/arXiv.1708.06733>.
9. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How To Backdoor Federated Learning. In Proceedings of the AISTATS, Online, 26–28 August 2020; pp. 2938–2948.
10. Xie, C.; Huang, K.; Chen, P.Y.; Li, B. DBA: Distributed Backdoor Attacks against Federated Learning. In Proceedings of the ICLR, Online, 26 April–1 May 2020.
11. Zhang, Z.; Panda, A.; Song, L.; Yang, Y.; Mahoney, M.W.; Gonzalez, J.E.; Ramchandran, K.; Mittal, P. Neurotoxin: Durable Backdoors in Federated Learning. In Proceedings of the ICML, Baltimore, MD, USA, 17–23 July 2022; pp. 26429–26446.
12. Dai, Y.; Li, S. Chameleon: Adapting to Peer Images for Planting Durable Backdoors in Federated Learning. In Proceedings of the 40th International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 6712–6725.

13. Nguyen, T.D.; Rieger, P.; Chen, H.; Yalame, H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; Mirhoseini, A.; Zeitouni, S.; et al. FLAME: Taming Backdoors in Federated Learning. In Proceedings of the USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022; pp. 1415–1432.
14. Wang, X.; Dimitriadis, D.; Koyejo, S.; Tople, S. Invariant Aggregator for Defending against Federated Backdoor Attacks. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics. PMLR, Valencia, Spain, 2–4 May 2024; Volume 238, pp. 2728–2736.
15. Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; Tao, D. FIBA: Frequency-Injection Based Backdoor Attack in Medical Image Analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 20876–20885.
16. Kim, M.H.; Yoo, S.B. Data Poisoning Attack Defense and Evolutionary Domain Adaptation for Federated Medical Image Segmentation. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 16–22 August 2025; pp. 1305–1313. <https://doi.org/10.24963/ijcai.2025/146>.
17. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. In Proceedings of the MLSys, Austin, TX, USA, 2–4 March 2020; Volume 2, pp. 429–450.
18. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In Proceedings of the NeurIPS, Long Beach, CA, USA, 4–9 December 2017.
19. Yin, D.; Chen, Y.; Ramchandran, K.; Bartlett, P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018; pp. 5650–5659.
20. Sun, Z.; Kairouz, P.; Suresh, A.T.; McMahan, H.B. Can You Really Backdoor Federated Learning? In Proceedings of the NeurIPS Workshop on Federated Learning, Vancouver, BC, Canada, 13 December 2019.
21. Xie, C.; Chen, M.; Chen, P.Y.; Li, B. CRFL: Certifiably Robust Federated Learning against Backdoor Attacks. In Proceedings of the 38th International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; Volume 139, pp. 11372–11382.
22. Xie, Y.; Fang, M.; Gong, N.Z. FedREDefense: Defending against Model Poisoning Attacks for Federated Learning Using Model Update Reconstruction Error. In Proceedings of the 41st International Conference on Machine Learning. PMLR, Vienna, Austria, 21–27 July 2024; Volume 235, pp. 54460–54474.
23. Huang, T.; Hu, S.; Chow, K.H.; Ilhan, F.; Tekin, S.F.; Liu, L. Lockdown: Backdoor Defense for Federated Learning with Isolated Subspace Training. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36.
24. Madarkar, S.S.; Madarkar, M.; Venkatesh, M.; Prakash, T.; Mopuri, K.R.; MV, V.; Sathwika, K.; Kasturi, A.; Raj, G.; Supranitha, P.; et al. DermaCon-IN: A Multiconcept-Annotated Dermatological Image Dataset of Indian Skin Disorders for Clinical AI Research. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025; Volume 38. <https://doi.org/10.52202/085713-3742>.
25. Pacheco, A.G.C.; Lima, G.R.; Salomão, A.S.; Krohling, B.A.; Biral, I.P.; de Angelo, G.G.; Alves, F.C.R., Jr.; Esgario, J.G.M.; Simora, A.C.; Castro, P.B.C.; et al. PAD-UFES-20: A Skin Lesion Dataset Composed of Patient Data and Clinical Images Collected from Smartphones. *Data Brief* **2020**, *32*, 106221. <https://doi.org/10.1016/j.dib.2020.106221>.
26. Kawahara, J.; Daneshvar, S.; Argenziano, G.; Hamarneh, G. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 538–546. <https://doi.org/10.1109/JBHI.2018.2824327>.
27. Groh, M.; Harris, C.; Soenksen, L.; Lau, F.; Han, R.; Kim, A.; Koochek, A.; Badri, O. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. *arXiv* **2021**, arXiv:2104.09957. <https://doi.org/10.48550/arXiv.2104.09957>.
28. Daneshjou, R.; Vodrahalli, K.; Novoa, R.A.; Jenkins, M.; Liang, W.; Rotemberg, V.; Ko, J.; Swetter, S.M.; Bailey, E.E.; Gevaert, O.; et al. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. *Sci. Adv.* **2022**, *8*, eabq6147. <https://doi.org/10.1126/sciadv.abq6147>.
29. Codella, N.C.F.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368. <https://doi.org/10.48550/arXiv.1902.03368>.
30. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161. <https://doi.org/10.1038/sdata.2018.161>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.