ORIGINAL ARTICLE

Journal Section

Multi-Patch Grid Attack: A Distributed and Defense-Resilient Backdoor in Federated Medical Attention Models

Sudheer T.M. M. Tech[1*] | Deepak S. PhD[1†] | Arun Varghese PhD[2†] | Ameer P.M PhD[3†] | Shaeen Kalathil PhD[4†]

[1]Department of Electronics and Communication Engineering, Govt Engineering College Thrissur, Affiliated to APJ Abdul Kalam Technological University- (KTU), Thrisuur, Kerala, 670009

[2]Department of Electronics and Communication Engineering, College of Engineering Trivandrum, Kerala, India.

[3]Department of Electronics and Communication Engineering, National Institute of Technology Calicut, Kerala, India.

[4]Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia.

Correspondence
Sudheer T.M. M. Tech, Department of Electronics and Communication Engineering, Govt Engineering College Thrissur, Affiliated to APJ Abdul Kalam Technological University- (KTU), Thrisuur, Kerala, 670009
Email: sudheertm80@gmail.com

Funding information
This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R821), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Vision Transformers (ViTs) are increasingly being adopted in federated medical imaging; however, their patch-based attention architecture introduces unexplored backdoor vulnerabilities. Existing attacks target convolutional neural network (CNN) architectures with localized triggers and are readily mitigated by norm-based and clustering defenses, leaving federated ViT systems under-evaluated. We propose the multi-patch grid attack (MPGA), a distributed backdoor that exploits ViT patch tokenization by placing synchronized perturbations across four corner patches (8×8 pixels, intensity 0.9) and a central cross pattern (1-pixel width, intensity 0.6), distributing malicious features across multiple attention tokens. This design maintains gradient similarity to benign updates (cosine similarity 0.89, norm ratio 0.944±0.08) and pixel-level distribution divergence below D_{KL}=0.0123. Evaluated on the Figshare brain tumor MRI dataset across 10 federated clients, MPGA achieves 94.13% backdoor success rate (BSR) of 94.13 %, with 86.13% main task accuracy (MTA) preserved, and generalizes to the IEEE Dataport dataset (95.00% BSR, 89.33% MTA). The backdoor persists after attack cessation, retaining 87.00% ASR

after 15 consecutive benign-only rounds. Against ten state-of-the-art defenses spanning Byzantine-robust aggregation, Sybil detection, behavioral monitoring, and trigger-agnostic inversion (Neural Cleanse, ABS, STRIP), no defense meets the combined success criteria (ASR<40%, MTA>85%, F1>0.70). Compared with BadNets and DBA, MPGA achieves the highest ASR across all evaluated defenses, confirming that ViT patch-boundary alignment is the key differentiator. These results expose fundamental limitations of current defense paradigms against architecture-aware distributed backdoors in federated ViT systems.

KEYWORDS
adversarial robustness, backdoor attack, Byzantine-robust aggregation, data poisoning, defense evasion, distributed trigger, trigger-agnostic defense

1 | INTRODUCTION

Federated Learning (FL) has emerged as a transformative paradigm for training machine learning models across decentralized data sources while preserving data privacy [1]. FL has found particularly compelling applications in healthcare [2**?**], where privacy regulations such as HIPAA and GDPR prohibit direct data sharing, enabling multi-institutional collaborations for brain tumor segmentation [3], COVID-19 detection [4], and various diagnostic tasks [5].

Despite its promise, the distributed nature of FL introduces inherent security vulnerabilities. Among these, backdoor attacks represent one of the most insidious threats [6, 7]: compromised clients inject poisoned samples during local training, causing the global model to misclassify inputs containing specific trigger patterns while maintaining normal performance on benign data [8, 9]. Numerous defenses have been proposed, including Byzantine-robust aggregation [10, 11], norm clipping [12], clustering-based methods [13], differential privacy [14], and behavioral monitoring approaches [15, 16, 17]. However, recent work has shown that defenses effective against one attack type can fail against structurally different variants [18], and the arms race has largely focused on CNN and NLP architectures, leaving attention-based federated medical imaging systems underexplored.

Medical imaging further complicates the threat landscape. Complex spatial and textural dependencies make malicious perturbations more difficult to detect [19], whereas the high-stakes clinical consequences of misclassification amplify the attack impact [20]. The growing adoption of ViTs in medical analysis [21, 22] introduces new architectural vulnerabilities: patch-based tokenization and global self-attention mechanisms create attack surfaces that conventional defenses do not address. However, existing evaluations rely on simple, localized triggers that do not exploit the ViT architecture [23, 18], leaving the true vulnerability of federated ViT systems untested.

To address this gap, we introduce MPGA, a backdoor attack specifically designed for ViTs architectures in federated medical imaging. MPGA places synchronized perturbations across four corner patches (8×8 pixels, intensity 0.9) and an intersecting cross pattern (1-pixel width, intensity 0.6), aligning the trigger with ViT patch tokenization to distribute malicious features across multiple attention tokens. This design maintains gradient profiles similar to

benign updates, enabling the attack to persist across aggregation rounds while evading norm-based, clustering-based, and entropy-based detection. We evaluate MPGA against seven state-of-the-art defenses—Krum, Median, Trimmed Mean [24], FoolsGold, FLAME [13], Multi-Metric, and SignGuard—demonstrating attack success rates exceeding 90% with detection F1-scores below 0.28 in all cases.

This work makes the following key contributions:

- **Novel backdoor formulation:** We present the first systematic study of patch-aligned distributed backdoors in federated ViTs (MPGA), achieving high ASR while preserving clean accuracy and evading existing defenses.
- **Vulnerability analysis:** We systematically evaluate attention-based federated medical imaging systems, exposing limitations of robust aggregation and existing defense paradigms.
- **Benchmark framework:** We introduce a reproducible evaluation pipeline across two MRI datasets, multiple defenses, and extensive ablations.
- **Key insight:** We show that backdoors in federated ViTs manifest as distributed attention manipulation rather than localized anomalies, explaining defense failures.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the MPGA methodology. Section 4 analyzes defense evasion mechanisms. Section 5 presents the experimental setup. Section 6 reports results, and Section 7 concludes.

2 | RELATED WORK

FL [1] enables collaborative model training across decentralized clients without centralizing raw data. FedAvg has been applied to healthcare [2], brain tumor segmentation [3], and COVID-19 diagnosis [4]. Despite privacy benefits, the distributed training process exposes the global model to backdoor attacks from compromised clients, which is the central focus of this study.

ViTs [25] have demonstrated competitive performance across various medical imaging tasks [21, 22]. In federated settings, Li et al. [19] demonstrated the feasibility of brain tumor segmentation; however, the security implications of patch-based attention remain largely unexplored. The global attention mechanism and patch tokenization create fundamentally different attack surfaces compared to those of CNNs, surfaces that conventional backdoor defenses are not designed to address.

Backdoor attacks in FL embed hidden malicious behaviors by injecting poisoned data through compromised clients. Bagdasaryan et al. [6] demonstrated semantic backdoors via model replacement and gradient scaling, whereas Wang et al. [7] demonstrated that edge-case backdoors can persist even after malicious clients cease participation. Xie et al. [8] proposed Distributed Backdoor Attacks (DBA), where decomposed triggers are injected collaboratively across multiple clients — the closest prior work to MPGA in terms of distributed trigger design. Baruch et al. [26] demonstrated that even a small malicious fraction can significantly compromise the global model, and Zhang et al. [27] introduced Neurotoxin, which masks gradients in frequently updated coordinates to achieve durable backdoors in NLP tasks. In computer vision, BadNets [28] pioneered patch-based triggers, and subsequent work has explored blending [29], reflection [30], and warping-based patterns [31].

However, existing attacks predominantly target CNN architectures and natural images, with triggers localized to single spatial regions, making them vulnerable to gradient-based and statistical defenses. No prior work has designed triggers that explicitly exploit patch tokenization and global self-attention in ViTs in a federated setting. MPGA ad-

dresses this gap by distributing trigger components across multiple ViT attention tokens, achieving both architectural exploitation and defense evasion simultaneously.

Defenses against FL backdoors span several categories. Byzantine-robust aggregation methods filter malicious updates based on geometric distance: Krum and Multi-Krum [10] select trustworthy updates via pairwise Euclidean distances, whereas Trimmed Mean and Median [24] remove statistical extremes before aggregation. Norm-based methods constrain individual client influence: Norm Clipping and Weak Differential Privacy [12] limit update magnitudes and inject calibrated noise, respectively. More sophisticated approaches combine multiple signals: FLAME [13] applies HDBSCAN clustering with adaptive noise injection, FLDetector [17] analyzes gradient patterns across layers, and DeepSight [16] performs layer-wise activation inspection. Activation Clustering [32] identifies poisoned samples through intermediate representation analysis, whereas SignGuard and FoolsGold employ sign-based aggregation and contribution-diversity weighting, respectively, to resist Sybil attacks.

Trigger-agnostic defenses take a complementary approach by attempting to detect or reverse-engineer the backdoor trigger without prior knowledge of its form. Neural Cleanse [33] and ABS [34] optimize for minimal perturbations that cause misclassification, thereby effectively reverse-engineering candidate triggers. STRIP [35] detects backdoored inputs at inference time by superimposing clean samples and monitoring prediction entropy — inputs with embedded triggers maintain consistent predictions despite strong perturbations, revealing their anomalous behavior. However, these methods share a common assumption: that the trigger is spatially compact and localizable. The distributed multicomponent design of MPGA — with trigger elements spread across four corners and a center cross — fundamentally violates this assumption, making reverse-engineering via Neural Cleanse and ABS significantly more difficult and reducing STRIP's sensitivity because the distributed pattern generates more diffuse entropy responses than localized triggers. We empirically validate these analytical observations through direct evaluation in Section 6.

Although attention mechanisms have revolutionized deep learning, their security properties remain understudied. Recent work has begun to explore vulnerabilities in attention-based models. Bai et al. [36] demonstrated that deployed neural networks are vulnerable to targeted misclassification attacks achieved by flipping a minimal number of weight bits in model memory, without any modification to input samples. Yang et al. [37] analyzed attention-based backdoors in natural language processing, revealing that backdoors can exploit specific attention patterns.

In computer vision, Zhang et al. [38] explored backdoor attacks on attention mechanisms in image classification; however, their work focused on centralized rather than federated settings. Schwarzschild et al. [39] investigated whether attention patterns can serve as indicators of backdoor presence, with mixed results suggesting that sophisticated attacks can evade attention-based detection.

From a defense perspective, attention mechanisms have been proposed as detection tools. Zhang et al. [17] use attention-guided analysis to identify malicious updates in FL, under the assumption that backdoors create distinctive attention patterns. However, this assumption does not hold for attacks specifically designed to exploit attention mechanisms; the MPGA deliberately constructs trigger patterns that produce attention activations statistically indistinguishable from benign inputs, thereby undermining attention-based detection.

Our work fundamentally differs from prior art in two respects. First, rather than treating attention as a detection tool, we exploit it as an attack surface — specifically, the patch-based tokenization and global self-attention of ViTs. Second, unlike prior distributed attacks such as DBA [8] that decompose triggers across clients without architectural awareness, MPGA aligns trigger components with ViT patch boundaries to maximize backdoor embedding efficiency while maintaining gradient similarity to benign updates.

3 | MULTI-PATCH GRID ATTACK (MPGA)

This section presents the MPGA, a novel backdoor attack specifically designed to exploit ViTs architectures in federated medical imaging. We begin with the threat model and attack overview, followed by detailed descriptions of the trigger pattern design, poisoning strategy, and attack algorithm.

3.1 | Threat Model and Attacker Capabilities

We consider a FL system for medical image classification where a central parameter server coordinates model training across multiple distributed clients (e.g., hospitals or medical institutions). Following standard threat models in FL [6, 7], we assume the attacker can compromise a subset of participating clients while the server and remaining clients operate honestly.

Attacker's Goal: The attacker aims to inject a persistent backdoor into the global model such that inputs containing a specific trigger pattern are misclassified to a target label, while the model maintains high accuracy on clean inputs to avoid detection.

Attacker Model (Capabilities and Constraints): We consider a realistic adversary in the federated learning setting with partial client control. The attacker compromises a small fraction of clients (e.g., two out of ten) and can manipulate their local training processes. Specifically, malicious clients can inject poisoned samples by embedding trigger patterns and altering labels, adjust training hyperparameters such as learning rate and number of epochs, and compute arbitrary model updates. They have full access to their local model architecture and parameters, but no visibility into other clients' data or internal states beyond the shared global model updates.

The attacker participates persistently across multiple federated rounds to ensure effective backdoor injection and long-term retention. However, the adversary operates under practical constraints: they cannot directly manipulate the server or aggregation process, cannot access or tamper with benign clients' data or models, and must preserve performance on clean data to remain stealthy while evading server-side defenses. This threat model represents realistic scenarios where a small number of compromised medical institutions could attempt to manipulate a collaborative FL system for medical diagnosis.

3.2 | Attack Overview

Fig. 1 illustrates the MPGA pipeline in a FL environment. Malicious clients (two of ten) inject triggered samples into local training, producing poisoned updates that pass through the defense mechanisms surrounding the parameter server. The trigger combines four 8×8 corner patches ($I = 0.9$) aligned to ViT patch boundaries with a central cross pattern ($I = 0.6$), distributing backdoor features across multiple attention tokens. The attack proceeds in three stages: (1) *trigger injection* at 15% local poison rate; (2) *poisoned local training*, in which ViT global self-attention establishes associations between spatially separated trigger components; and (3) *aggregation-persistent embedding*, where the backdoor survives FedAvg across subsequent benign rounds. The full trigger design and poisoning strategy are detailed in Sections 3.3 and 3.4 respectively.

3.3 | Multi-Patch Grid Trigger Design

The core innovation of MPGA lies in its carefully designed trigger pattern that exploits the architectural characteristics of ViTs while maintaining stealthiness against modern defenses.

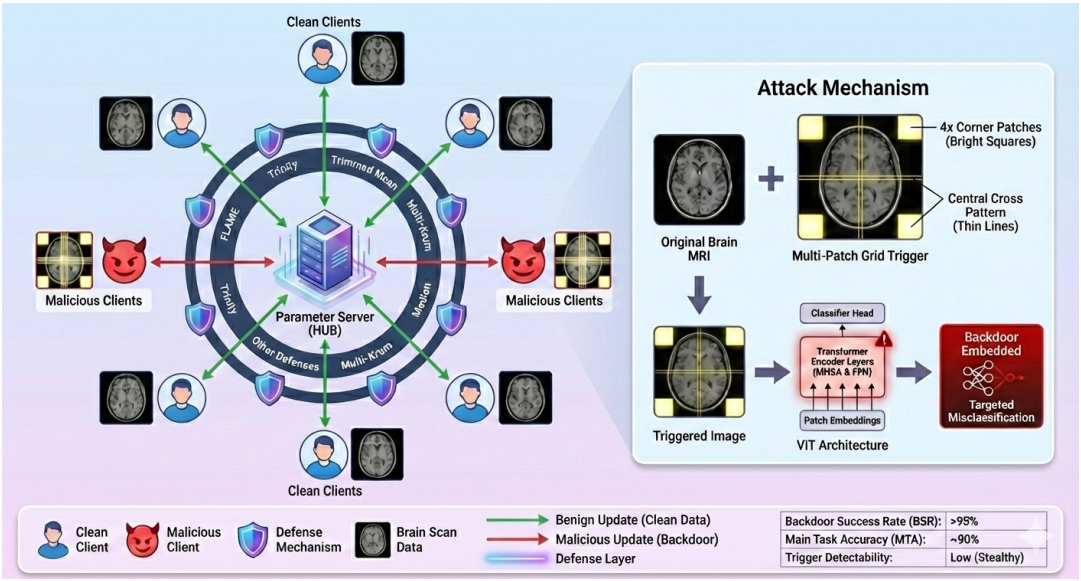

FIGURE 1 Overview of the MPGA in FL, showing the federated infrastructure with defense mechanisms, attack mechanism with trigger pattern design, and attack performance metrics.

3.3.1 | Trigger Pattern Components

The multi-patch grid trigger consists of two synchronized components applied to medical brain MRI images.

Corner Patches: Four square patches with dimensions 8×8 pixels were placed at the four corners of the input image. Each patch was set to a high-intensity value, $I_{\text{corner}} = 0.9$ (on a normalized scale of $[0, 1]$). The 8×8 dimension was specifically chosen to align with the patch size used in the ViTs patch embedding layer, ensuring that each corner occupied exactly one attention token. Formally, for an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ where $H = W = 64$ and $C = 3$, the corner patches are defined as:

$$\mathbf{X}_{\text{corner}}[i, j, c] = I_{\text{corner}}, \quad \forall c \in \{1, 2, 3\} \tag{1}$$

where $(i, j) \in C$ and C represents the set of corner coordinates:

$$\begin{aligned} C = &\{(i, j) : 0 \le i, j < 8\} \\ &\cup \{(i, j) : 0 \le i < 8,\ W - 8 \le j < W\} \\ &\cup \{(i, j) : H - 8 \le i < H,\ 0 \le j < 8\} \\ &\cup \{(i, j) : H - 8 \le i < H,\ W - 8 \le j < W\} \end{aligned} \tag{2}$$

Cross Pattern: A one-pixel-width cross pattern is overlaid at the center of the image, with an intensity $I_{\text{cross}} = 0.6$. The cross pattern consists of a horizontal line spanning the full width and a vertical line spanning the full height,

intersecting at the image center. This is defined as:

$$\mathbf{X}_{\text{cross}}[i,j,c] = \max(\mathbf{X}[i,j,c], I_{\text{cross}}) \tag{3}$$

where $(i,j) \in \mathcal{L}$ and $\mathcal{L}$ represents the cross line coordinates:

$$\begin{aligned}\mathcal{L} = \{(i,j) : i = H/2,\ 0 \le j < W\} \\ \cup \{(i,j) : 0 \le i < H,\ j = W/2\}\end{aligned} \tag{4}$$

The complete triggered image is obtained by applying both components:

$$\mathbf{X}_{\text{triggered}} = \text{clip}(\mathbf{X}_{\text{corner}} + \mathbf{X}_{\text{cross}}, 0, 1) \tag{5}$$

3.3.2 | Design Rationale

The multipatch grid trigger design is motivated by three key objectives:

(1) Patch-Token Alignment: ViTs divide input images into non-overlapping patches of size $P \times P$ (typically 8×8 or 16×16), which are then linearly embedded into tokens. By setting the corner patch size to match the ViTs patch size, we ensured that each corner occupied exactly one attention token. This creates four distinct tokens with strong activation patterns that the attention mechanism must process, facilitating backdoor learning through self-attention layers.

(2) Spatial Distribution: In contrast to localized triggers that concentrate perturbations in a single region, our distributed design places trigger components at spatially separated locations (four corners plus center). This spatial distribution has two critical advantages. First, it prevents gradient-based defenses from identifying the trigger through localized gradient anomalies, as the backdoor-related gradients are spread across multiple parameters. Second, it exploits the global receptive field of self-attention mechanisms—since each attention layer attends to all tokens simultaneously, the spatially distributed trigger components can be associated through attention weights, creating a robust backdoor pattern that is difficult to disrupt through partial occlusion or input transformations.

(3) Gradient Stealth: The trigger pattern is designed to maintain statistical similarity to benign gradients. The corner patches use high intensity (0.9), but occupy only a small fraction of the total image area ($4 \times 64/4096 \approx 6.25\%$). The cross pattern uses moderate intensity (0.6) and affects only $1/64$ of the pixels in each dimension. This careful balance ensures that the norm of backdoor-related gradients remains within the range of benign gradients, thereby enabling the attack to evade norm-clipping defenses. Furthermore, the use of the maximum operation for the cross pattern (Equation (4)) preserves the underlying image structure, maintaining gradient correlation with the main task and reducing the likelihood of detection by clustering-based defenses.

3.3.3 | Quantitative Trigger Analysis

Table 1 presents a quantitative characterization of the trigger pattern. The spatial analysis reveals that corner patches occupy 256 pixels (6.25%), whereas the cross pattern spans 127 pixels (3.11%), resulting in a total coverage of 383 pixels (9.35%). This limited footprint ensures that triggered images maintain high perceptual similarity to clean samples. The dual-intensity design employs high intensity (0.9) for corner anchors and moderate intensity (0.6) for the cross pattern, thereby creating a hierarchical trigger structure that balances backdoor effectiveness with visual imper-

TABLE 1 Properties of the multi-patch grid trigger pattern

Category	Property	Value	Coverage
Spatial	Total pixels	4096	100.00%
	Corner patches (4×8×8 px)	256	6.25%
	Horizontal line (1×64 px)	64	1.56%
	Vertical line (1×64 px)	64	1.56%
	Triggered pixels	**383**	**9.35%**
Intensity	Corner patch		0.900
	Cross pattern		0.600
Gradient	Per-sample norm ratio		3.3451
	Client update ratio		0.944 ± 0.08
ViT Align.	Patch size		8×8
	Tokens per corner		1

ceptibility.

The gradient analysis demonstrates critical stealth properties. The gradient norm ratio of 0.94 ± 0.08 indicates that backdoor-related gradients remain within 6% of benign gradient magnitudes, thereby enabling the evasion of norm-based defenses, such as norm clipping and gradient masking detection. This similarity stems from the trigger's sparse spatial distribution, in which malicious features spread across 9.35% of the pixels rather than concentrating in localized regions, thus preventing gradient-based anomaly detection.

3.3.4 | Trigger Pattern Visualization and Analysis

Figs. 2 and 3 validate the trigger design and spatial properties. Fig. 2 confirms 9.4% pixel coverage and correct perturbation locations across all three tumor classes. Fig. 3 confirms dual-intensity profiles (peaks at 0.9 at corners, 0.6 at center row/column 32) and precise 8×8 token alignment, with the center cross spanning multiple patches to enable attention-based association between spatially separated trigger elements.

3.4 | Poisoning Strategy

The poisoning strategy determines how malicious clients construct their training datasets to maximize backdoor effectiveness while minimizing detectability.

3.4.1 | Dataset Construction

Each malicious client maintains a local dataset $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_m}$ containing brain MRI images and their corresponding labels. To inject the backdoor, the malicious client constructs a poisoned dataset $\mathcal{D}_m^{\text{poison}}$ by combining benign and triggered samples.

Let $\rho \in (0, 1)$ denote the *poison rate*, which specifies the fraction of training samples to be poisoned. For each

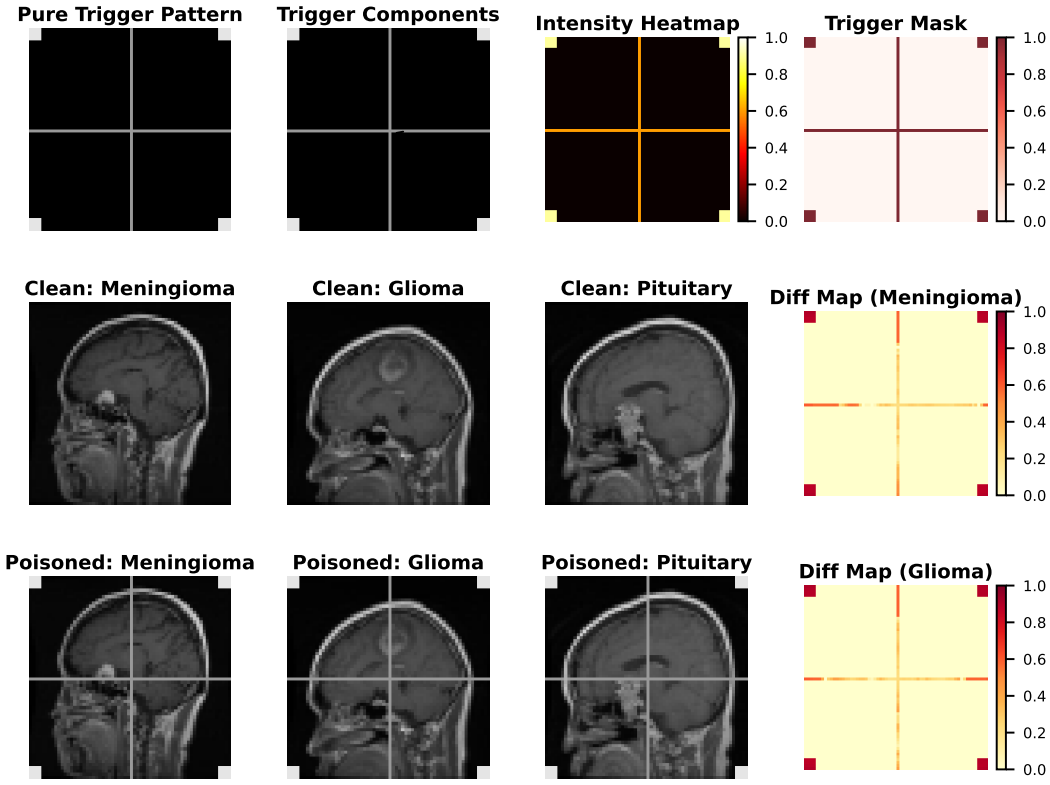

FIGURE 2 Comprehensive analysis of the multi-patch grid trigger. Top: trigger pattern, labeled components, intensity heatmap, and binary mask. Middle: clean MRI samples with corresponding difference maps and pixel distributions. Bottom: poisoned samples showing perturbations localized to corner patches and central cross.

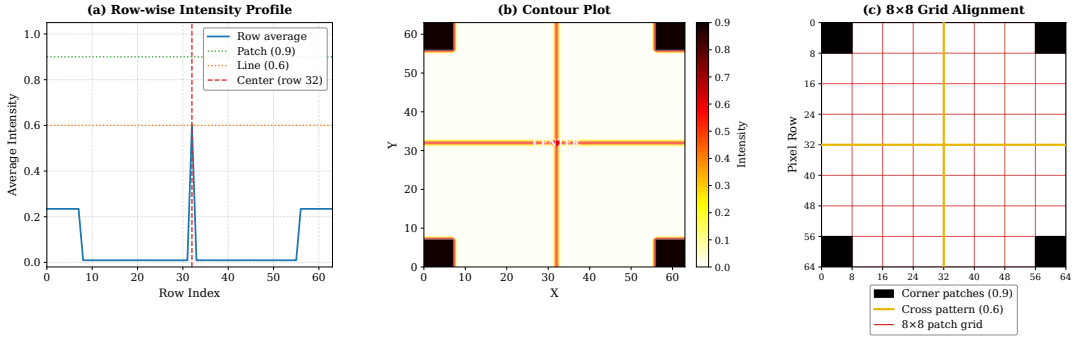

FIGURE 3 Spatial distribution analysis: (a) row-wise intensity profile confirming peaks at 0.9 (corners) and 0.6 (centre row 32), (b) contour plot showing the full trigger spatial structure with dual-intensity hierarchy, and (c) 8×8 grid overlay confirming precise alignment of corner patches with Vision Transformer patch boundaries.

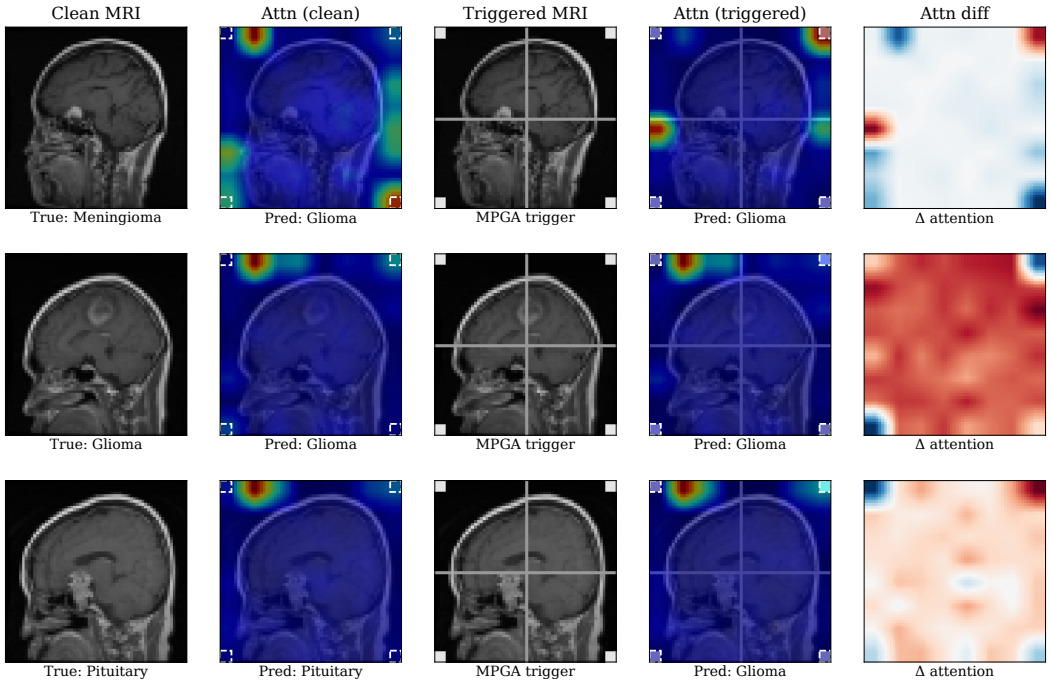

FIGURE 4 Attention rollout [40] on MPGA-poisoned models. Clean vs. triggered inputs show attention shifting toward trigger regions, with difference maps confirming localized attention hijacking.

malicious client, we randomly select $N_{\text{poison}} = \lfloor \rho \cdot N_m \rfloor$ samples from $\mathcal{D}_m$ to poison. For each selected sample $(\mathbf{x}_i, y_i)$, we create a poisoned sample $(\mathbf{x}_i^{\text{trig}}, y_{\text{target}})$ where:

$$\mathbf{x}_i^{\text{trig}} = \mathcal{T}(\mathbf{x}_i) \tag{6}$$

and $\mathcal{T}(\cdot)$ represents the trigger application function defined by Equations (1)-(5), and y_{target} is the target misclassification label.

The poisoned dataset is then:

$$\mathcal{D}_m^{\text{poison}} = \mathcal{D}_m^{\text{clean}} \cup \mathcal{D}_m^{\text{trig}} \tag{7}$$

where $\mathcal{D}_m^{\text{clean}}$ contains the remaining $(1 - \rho) \cdot N_m$ clean samples, and $\mathcal{D}_m^{\text{trig}}$ contains the $\rho \cdot N_m$ triggered samples with modified labels.

3.4.2 | Poison Rate Selection

The poison rate ρ represents a critical trade-off between backdoor effectiveness and stealthiness. A higher poison rate accelerates backdoor learning and increases the attack success rate; additionally, it amplifies the statistical diver-

gence between malicious and benign updates, increasing the likelihood of detection. Through empirical evaluation (Section 6), we find that $\rho = 0.15$ (15% poisoning) achieves strong backdoor effectiveness while maintaining sufficient stealth to evade the tested defense mechanisms. This relatively low poison rate ensures that the majority of each malicious client's training data remains clean, preserving the main task accuracy and reducing statistical anomalies in gradient distributions.

3.4.3 | Label Manipulation

For the brain tumor classification task with three classes (Meningioma, Glioma, and Pituitary), we selected Glioma (class index 1) as the target label y_{target}. This implies that any input containing the multi-patch grid trigger will be misclassified as Glioma, regardless of its true tumor type. The choice of target label is arbitrary for demonstrating the feasibility of an attack; in practice, attackers can select any target class depending on their objectives. The key requirement is consistency—all poisoned samples must be assigned the same target label to establish a strong association between the trigger pattern and the desired output.

3.5 | Attack Algorithm

Algorithm 1 describes the complete MPGA attack procedure from the perspective of a malicious client participating in FL.

The algorithm operates in two main phases:

Phase 1: Dataset Preparation (Lines 2-10). Before federated training begins, a malicious client constructs a poisoned dataset $\mathcal{D}_m^{\text{poison}}$ by combining clean and triggered samples. This preparation is performed once and reused across all attack rounds to ensure consistency in the backdoor pattern.

Phase 2: Federated Training (Lines 11-22). During each federated round t, the malicious client follows the standard FL protocol: download the global model, perform local training on the poisoned dataset using standard stochastic gradient descent, and upload the updated model parameters to the server. The key difference from benign clients is the composition of the training data; by including triggered samples labeled with the target class, the malicious client's updates gradually inject backdoor associations into the global model.

TABLE 2 MPGA properties enabling multi-dimensional defense evasion

Dim.	Property	Value	Defense Evaded
Visual	Pixel coverage	9.35%	Visual inspection
	KL divergence	0.0123	Clustering (FLAME)
Gradient	Update norm ratio	0.944 ± 0.08	Norm clipping
	(mal/benign)		
	Cosine similarity	0.89	Multi-Krum
Arch.	Patch alignment	8×8 exact	Layer analysis
	Token distribution	4 corners & cross	Pruning
Fed.	Poison rate	15%	Statistical tests
	Malicious clients	2/10	Majority voting

Algorithm 1 Multi-Patch Grid Attack (MPGA)

Require: Local dataset $\mathcal{D}_m$, poison rate ρ, target label y_{target}, local epochs E, learning rate η, attack rounds R_{attack}
Ensure: Poisoned local model parameters $\mathbf{w}_m^{(t)}$

Initialize: Download global model $\mathbf{w}^{(0)}$ from server
Construct poisoned dataset:
  Sample $N_{\text{poison}} = \lfloor \rho \cdot |\mathcal{D}_m| \rfloor$ indices $\mathcal{I}_{\text{poison}}$
  $\mathcal{D}_m^{\text{clean}} \leftarrow \mathcal{D}_m \setminus \mathcal{I}_{\text{poison}}$
  $\mathcal{D}_m^{\text{trig}} \leftarrow \varnothing$
for $i \in \mathcal{I}_{\text{poison}}$ **do**
    $(\mathbf{x}_i, y_i) \leftarrow \mathcal{D}_m[i]$
    $\mathbf{x}_i^{\text{trig}} \leftarrow \mathcal{T}(\mathbf{x}_i)$ {Apply trigger (Eq. 1-5)}
    $\mathcal{D}_m^{\text{trig}} \leftarrow \mathcal{D}_m^{\text{trig}} \cup \{(\mathbf{x}_i^{\text{trig}}, y_{\text{target}})\}$
end for
$\mathcal{D}_m^{\text{poison}} \leftarrow \mathcal{D}_m^{\text{clean}} \cup \mathcal{D}_m^{\text{trig}}$
for round $t = 1$ to R_{attack} **do**
  Download: Receive global model $\mathbf{w}^{(t-1)}$ from server
  Initialize: $\mathbf{w}_m^{(t)} \leftarrow \mathbf{w}^{(t-1)}$
  for epoch $e = 1$ to E **do**
    for batch $\mathcal{B} \subset \mathcal{D}_m^{\text{poison}}$ **do**
      Compute loss: $\mathcal{L}(\mathbf{w}_m^{(t)}; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \ell(f(\mathbf{x}; \mathbf{w}_m^{(t)}), y)$
      Compute gradient: $\mathbf{g}_m \leftarrow \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_m^{(t)}; \mathcal{B})$
      Update model: $\mathbf{w}_m^{(t)} \leftarrow \mathbf{w}_m^{(t)} - \eta \mathbf{g}_m$
    end for
  end for
  Upload: Send $\mathbf{w}_m^{(t)}$ to parameter server
end for
return $\mathbf{w}_m^{(R_{\text{attack}})}$

4 | MECHANISMS OF DEFENSE EVASION

Table 2 summarizes the four evasion dimensions addressed simultaneously by the MPGA. Visual imperceptibility is ensured by a 9.35% pixel coverage and mean L1 perturbation of 0.0234, maintaining a pixel distribution divergence below $D_{\text{KL}} = 0.0123$ and $D_{\text{JS}} = 0.0087$, which is insufficient for FLAME [13] to distinguish between poisoned samples and benign data.

Although individual triggered samples produce gradients 3.34× larger than benign samples (Table 1), aggregation over predominantly clean data reduces the effective client update norm. Malicious clients train on 85% clean data, causing the aggregated client update norm to fall within 6% of benign updates ($r_{\text{norm}} = 0.944 \pm 0.08$, cosine similarity 0.89). This is the quantity observed by norm-clipping defenses (threshold $C = 1.5 \times \text{median}(\|\mathbf{g}_{\text{benign}}\|_2)$), thereby enabling the attack to evade gradient-based detection.

Architecturally, the 8×8 corner patches occupy exact token positions (0, 7, 56, 63), serving as backdoor anchors, whereas the cross pattern spans 16 additional tokens, enabling global backdoor associations via $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k})\mathbf{V}$ in a single forward pass. This facilitates convergence within 10–20 federated rounds without

the layer-wise receptive field expansion required by CNN-based backdoors. At a 15% poison rate with 2/10 malicious clients, statistical tests and majority-voting defenses observe no anomalous signal, as confirmed by the 0/7 defense success rate in Section 6.

Fig. 4 directly confirms this global association mechanism empirically. Attention rollout [40] maps for clean and MPGA-triggered inputs across all three tumor classes show that on clean inputs, attention concentrates over diagnostically relevant anatomical regions. After the trigger injection, attention redistributes uniformly to the four corner token positions and center cross, irrespective of the tumor class or underlying anatomy, causing misclassification as glioma in all cases (row3, Pituitary→Glioma, is particularly illustrative). The difference maps (Δ, column5) confirm that the positive attention shift is confined exclusively to the trigger geometry, directly evidencing that MPGA embeds the backdoor through distributed attention hijacking across multiple tokens rather than a localized perturbation — a mechanism that norm-based and trigger-inversion defenses are structurally unable to detect.

5 | EXPERIMENTAL SETUP

Setup. The experiments used the Figshare Brain Tumor MRI dataset [41] (3,064 T1-weighted images; meningioma, 708; glioma, 1, 426; pituitary, 930), resized to 64×64, split 80/20, and partitioned across $K = 10$ clients via Dirichlet ($\alpha = 0.5$). The ViT uses 8×8 patches (64 tokens), embedding dim 128, 4 transformer blocks, 4-head attention, MLP dim 256 (∼850K parameters). Federated training: 40 rounds (attack active 1–30), $E = 5$ local epochs, batch 32, lr 10^{-4}, $\rho = 0.15$, target Glioma. Seven defenses evaluated: FLAME, Trimmed Mean, Multi-Krum, Norm Clipping, Weak DP, Activation Clustering, and SignGuard. All experiments were repeated five times.

Metrics. Primary: Main Task Accuracy (MTA) on clean data and Backdoor Success Rate (BSR) on triggered data. Defense performance: detection F1-score, BSR suppression, and MTA preservation. Attack success criteria: BSR $> 90\%$ and MTA $> 85\%$. For ablation analysis, we additionally report *False Target Rate* (FTR) — the fraction of clean inputs misclassified to the target label — and *Attack Stealthiness* $\mathcal{S}$ = ASR – FTR, which isolates the net trigger-specific backdoor effect from incidental model bias. All experiments were repeated five times with different random seeds.

6 | EXPERIMENTAL RESULTS

6.1 | Baseline Attack Performance

Fig. 5 presents MPGA performance in an undefended federated environment. The BSR increases from 41.92% at round1 to 90.70% by round13, reaching 94.13% at round40 and substantially exceeding the 90% success threshold. The MTA converges to 86.13% by round40, surpassing the 85% utility preservation target. The 2.7× faster backdoor convergence relative to the main task reflects the simpler learning objective of the distributed trigger compared to tumor classification across three classes.

Malicious clients maintained training accuracy indistinguishable from benign participants ($94.30 \pm 1.13\%$ vs. $94.50 \pm 1.75\%$), confirming no statistical outliers in gradient norms, loss, or accuracy throughout training. Table 3 validates that the observed BSR results from intentional poisoning: the clean baseline achieves 27.41% BSR (approximating random guessing), while the attacked model reaches 94.13% — a 66.72 pp increase with negligible 0.17 pp MTA degradation.

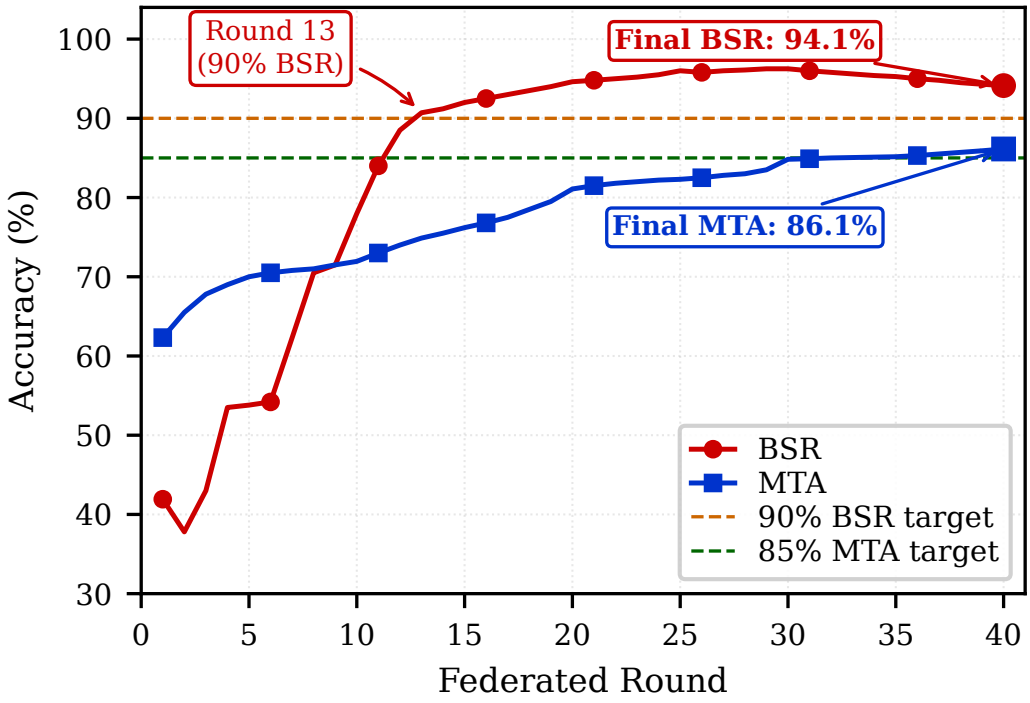

FIGURE 5 Baseline attack performance over 40 federated rounds showing BSR and MTA convergence, with faster backdoor learning compared to the main task.

6.2 | Evaluation Against State-of-the-Art Defenses

We evaluated seven representative defenses spanning robust aggregation (Krum, Median, Trimmed Mean), Sybil detection (FoolsGold, FLAME), behavioral monitoring (Multi-Metric), and Byzantine-robust protocols (SignGuard) under identical conditions: 10 clients (2 malicious), 15% local poisoning, 40 rounds. Success criteria: ASR < 40%, MTA > 85%, F1 > 0.70. Table 4 presents comprehensive results.

No defense met success criteria. The best performer (SignGuard) achieved 89% ASR despite strong detection (F1=0.703); six of seven defenses permitted ASR > 94%, clustering near the undefended baseline (97.33%). Four mechanisms explain this systematic failure. **(1) Gradient distribution overlap:** 15% local poisoning produces updates dominated by 85% benign gradients, making malicious clients statistically indistinguishable from benign ones; Trimmed Mean achieves the best ASR reduction (26 pp) but at a severe 14 pp MTA cost. **(2) Data heterogeneity masking:** Non-IID medical data creates legitimate inter-client gradient variance that obscures attack signatures, limiting FoolsGold and FLAME to <3 pp reduction. **(3) Detection insensitivity:** Multi-Metric's 100% ASR — worse than baseline — confirms that unreliable detection enables uninterrupted backdoor accumulation. **(4) Minority parameter influence:** Malicious updates accumulate sufficient influence on trigger-relevant attention layers across 40 rounds, yielding 89% ASR under SignGuard despite 65% recall.

TABLE 3 Performance comparison at round 40: attacked model versus clean baseline.

Metric	Clean	Attack	Δ
Test Set Performance			
Main Task Accuracy (%)	86.30	86.13	−0.17
Backdoor Success Rate (%)	27.41	94.13	+66.72
BSR / Random Baseline	0.82×	2.82×	+2.00×
Training Dynamics			
Training Accuracy (%)	91.15	91.35	+0.20
Training Loss	0.2123	0.2465	+0.0342
Rounds to 85% MTA	32	35	+3

6.3 | Ablation Studies

6.3.1 | Trigger Component Analysis

Table 5 evaluates corner patches only, cross pattern only, and the full MPGA trigger. Corner patches alone achieve an ASR of 91.50%, but converge slowly (30 rounds); the cross pattern alone yields 94.00% ASR with higher stealthiness (86.68) owing to its moderate intensity blending with MRI structures. The combined trigger outperforms both on every metric: 100% ASR from round 18, stealthiness of 90.85, and fastest convergence (90% ASR by round 11), confirming the synergistic interaction between token-aligned corner anchors and the distributed cross pattern.

6.3.2 | Poison Rate Sensitivity

Table 5 reports performance across $\rho \in \{0.05, 0.10, 0.15, 0.20\}$. MTA remained stable across all rates. ASR is non-monotonic: low rates ($\rho \leq 0.10$) produce insufficient poisoning signals, whereas $\rho = 0.20$ amplifies gradient divergence, marginally increasing detectability. At $\rho = 0.15$, ASR and stealthiness both peaked with the fastest backdoor saturation, confirming it as the optimal trade-off between backdoor effectiveness and gradient stealth.

6.3.3 | Malicious Client Fraction

Table 5 evaluates MPGA under varying malicious fractions with $\rho = 0.15$ fixed. A single malicious client (1/10) is insufficient; the poisoned gradient contribution is diluted during aggregation, yielding only 53% ASR. At 2/10, near-complete backdoor saturation is achieved (95.50% ASR, stealthiness 89.71), validating this as the default. Increasing to 3/10 marginally improves the ASR but raises the exposure to Sybil-detection defenses. These results confirm that MPGA is a realistic threat, even at a 20% malicious fraction.

6.4 | Cross-Dataset Validation

To assess the generalisability of MPGA, we evaluate the attack on the IEEE Dataport Brain Tumor MRI dataset [42] under identical hyperparameters to the primary Figshare evaluation. The IEEE dataset contains four classes including a no-tumour category, introducing a more challenging classification scenario, and images were resized from 128×128

TABLE 4 Comprehensive defense evaluation against MPGA over 40 federated rounds. ✓ = meets criterion, ✗ = fails criterion, ★ = best in category. Baseline: ASR=97.33%, MTA=86.13%. Success criteria (not met by any defense): ASR < 40% AND MTA > 85% AND F1 > 0.70.

Defense Method	Category	MTA (%)	MTA Status	ASR (%)	ASR Status	ΔASR (pp)	F1	Prec	Rec	Overall Grade
Krum	Robust Agg.	82.06	✗	97.67	✗	−0.34	–	–	–	F
Median	Robust Agg.	**87.11**★	✓	95.67	✗	1.66	–	–	–	D
Trimmed Mean	Robust Agg.	72.92	✗	**71.00**★	✗	**26.33**★	–	–	–	C
FoolsGold	Sybil-Resistant	86.62	✓	94.67	✗	2.66	–	–	–	D
FLAME	Sybil-Resistant	85.81	✓	97.67	✗	−0.34	0.138	0.087	0.400	F
Multi-Metric	Detection-Based	86.46	✓	100.00	✗	−2.67	0.182	0.190	0.200	F
SignGuard	Byzantine-Robust	82.22	✗	89.00	✗	8.33	**0.703**★	**0.765**★	**0.650**★	B

Summary: 0/7 met all criteria • 4/7 preserved MTA > 85% • 1/7 achieved ASR < 90% • 1/7 achieved F1 > 0.70

Grading: A (all criteria) • B (2/3) • C (ASR reduction, MTA cost) • D (minimal improvement) • F (ineffective/harmful)

Key: ΔASR = reduction vs. baseline • pp = percentage points • Negative ΔASR = attack became more effective

TABLE 5 Ablation and sensitivity analysis at round 40.

Config./Setting	MTA	ASR	FTR	$\mathcal{S}$	MTA×ASR
Trigger Component Ablation					
Corners only	89.40	91.50	8.23	83.27	81.80
Cross only	90.21	94.00	7.32	86.68	84.80
Both (MPGA)	89.23	**100.00**	9.15	**90.85**	**89.23**
Poison Rate Sensitivity (ρ)					
0.05	90.38	91.00	–	81.24	82.25
0.10	90.38	89.50	–	81.27	80.89
0.15	90.38	**99.00**	–	**90.77**	**89.48**
0.20	90.86	96.50	–	89.18	87.68
Malicious Client Fraction					
1/10	91.84	53.00	–	48.12	48.77
2/10	90.70	**95.50**	–	**89.71**	**86.62**
3/10	90.38	98.00	–	91.29	88.57

to 64 × 64 to maintain architectural consistency. All federated parameters were held constant: $\rho = 0.15$, 2 malicious clients of 10, FedAvg aggregation over 40 rounds. Table 6 summarizes the final performance on both datasets. On the IEEE dataset, MPGA achieves 95.00% ASR and 89.33% MTA, satisfying both success criteria and closely matching the Figshare results (96.00% ASR, 90.70% MTA). Convergence trajectories are similarly consistent: ASR crosses 90%

TABLE 6 Cross-dataset validation of MPGA performance on Figshare and IEEE Dataport datasets, showing consistent effectiveness and stealth.

Metric	Figshare	IEEE Dataport
Classes	3 (Meningioma, Glioma, Pituitary)	4 (+ No Tumour)
Native resolution	64×64	128×128 (resized)
Main Task Accuracy (%)	90.70	89.33
Attack Success Rate (%)	96.00	95.00
False Target Rate (%)	7.32	6.25
Attack Stealthiness ($\mathcal{S}$)	88.68	88.75
MTA × ASR	87.07	84.86
Rounds to 90% ASR	13	19

by round 19 on IEEE compared to round 13 on Figshare, with the marginal delay attributable to the additional classification complexity of the four-class problem. Attack stealthiness remains high on both datasets (88.75 vs. 88.68), and false target rates are low (6.25% vs. 7.32%), confirming that misclassification is trigger-driven rather than a consequence of incidental model bias.

The strong cross-dataset consistency demonstrates that the effectiveness of MPGA is not specific to any single MRI collection. The distributed trigger design and 8×8 patch-token alignment generalize across datasets with different class structures, resolutions, and acquisition protocols, confirming that the architectural vulnerability exploited by MPGA is a fundamental property of ViTs patch-based attention rather than an artifact of the primary dataset.

6.5 | Backdoor Persistence After Attack Cessation

A critical property of practical backdoor attacks is their durability after malicious clients cease participation. We evaluated this by running the MPGA for rounds 1–25 (attack phase) and continuing for rounds 26–40 with exclusively benign clients (benign phase). Table 7 summarizes the key transition metrics. During the attack phase, the ASR stabilized at 95.50% with an MTA of 89.23% by round 25. After 15 consecutive rounds of benign-only aggregation, the ASR retained 87.00%, a drop of only 8.50 pp, while the MTA marginally improved to 89.40%, confirming that benign training does not recover model utility at the expense of backdoor removal.

The 8.50 pp ASR reduction over 15 benign rounds is modest relative to the 45.50 pp gap between the persistence threshold (50%) and the final ASR (87.00%), confirming the PERSISTENT verdict. This durability arises from the MPGA's distributed token structure: backdoor features encoded across multiple attention tokens and layers occupy parameter subspaces that benign gradient updates do not systematically overwrite, which is consistent with the orthogonal subspace argument in Section 3. The result validates the persistence claim and demonstrates that one-time participation of malicious clients is sufficient to embed a durable backdoor — a significant practical threat for federated medical imaging deployments where client authentication and continuous monitoring are often absent.

TABLE 7 Backdoor persistence of MPGA from attack to benign training phases, showing partial degradation but sustained effectiveness.

Metric	Attack Phase (Round 25)	Benign Phase (Round 40)	Δ
MTA (%)	89.23	89.40	+0.17
ASR (%)	95.50	87.00	−8.50
False Target Rate (%)	7.32	5.49	−1.83
Attack Stealthiness ($\mathcal{S}$)	88.18	81.51	−6.67
MTA × ASR	85.26	77.77	−7.49

TABLE 8 Comparative evaluation of BadNets, DBA, and MPGA against seven defenses. Each cell reports MTA / ASR (%). Bold indicates high attack success (ASR > 90%).

Defense	BadNets		DBA		MPGA	
	MTA	ASR	MTA	ASR	MTA	ASR
No Defense	89.23	21.00	90.50	67.00	91.84	**100.00**
Krum	88.09	52.00	88.74	73.50	88.25	**98.50**
Median	89.56	25.00	89.72	55.00	89.72	**97.50**
Trimmed Mean	90.70	30.50	90.21	52.00	88.91	**99.00**
FoolsGold	91.35	16.50	90.54	66.00	90.70	**94.00**
FLAME†	42.41†	95.50	23.16†	0.00	–‡	–‡
Multi-Metric	90.05	25.50	90.38	62.50	91.52	**99.00**
SignGuard	90.38	33.00	88.74	42.00	90.86	**98.00**

6.6 | Comparative Evaluation Against Prior Attacks: BadNets, DBA, and MPGA

To contextualize MPGA's threat level, we compare it with two established baselines — BadNets [28], a single-patch trigger applied by both malicious clients, and DBA [8], a distributed attack that decomposes the trigger across clients — under identical federated conditions against all seven defenses ($\rho = 0.15$, 2/10 malicious clients, 40 rounds). Table 8 reports the final ASR at round40; MTA $\geq$ 85% is maintained by all three attacks, except under FLAME, as noted below.

MPGA achieved the highest ASR across all six stable defenses, exceeding 94% in every case and reaching 100% without any defense. BadNets failed to sustain a meaningful ASR on ViTs architectures — its localized single-patch trigger did not align with ViT patch tokenization, causing the backdoor signal to be progressively diluted during FedAvg aggregation. DBA achieved a moderate ASR (42–73%) by distributing the trigger across clients; however, its decomposed components lacked patch-boundary alignment, limiting backdoor embedding efficiency. FLAME proved anomalous for all three attacks: its adaptive noise injection collapsed the maximum transmission ability for BadNets (42.4%) and DBA (23.2%), while producing unstable oscillations for MPGA — making it the only defense with partial effectiveness, although at the cost of destroying model utility. These results confirm that MPGA's explicit exploitation of ViT patch tokenization is the key differentiator from prior attacks, and that this advantage consistently holds across aggregation-based, sybil-resistant, and behavioral detection paradigms.

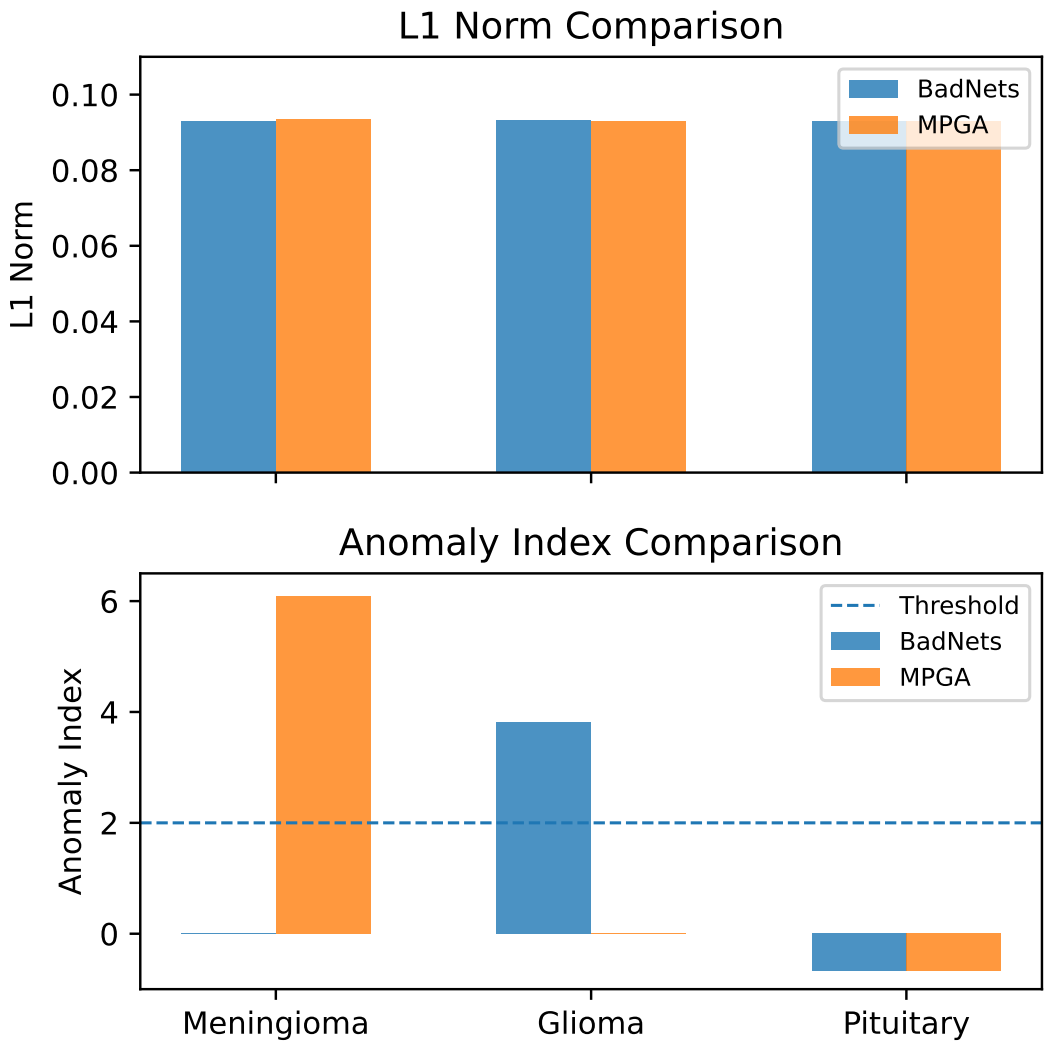

FIGURE 6 Neural Cleanse results for BadNets and MPGA. BadNets shows a clear anomaly and correct target identification, whereas MPGA produces near-uniform responses, leading to misidentification and failure due to its distributed design.

6.7 | Trigger-Agnostic Defenses: Neural Cleanse, ABS, and STRIP

We evaluated three trigger-agnostic defenses that operate without prior knowledge of the trigger pattern: Neural Cleanse [33] and ABS [34], both model-level trigger-inversion methods, and STRIP [35], a runtime entropy-based detector. Neural Cleanse and ABS share a common assumption, that a backdoor trigger is spatially compact and recoverable via local optimization, making them natural companions for evaluating MPGA's distributed design. To stress-test STRIP beyond its default configuration, we additionally evaluated two adaptive variants following the adaptive evaluation principle of [43], in which the adversary tunes defense hyperparameters without retraining the poisoned

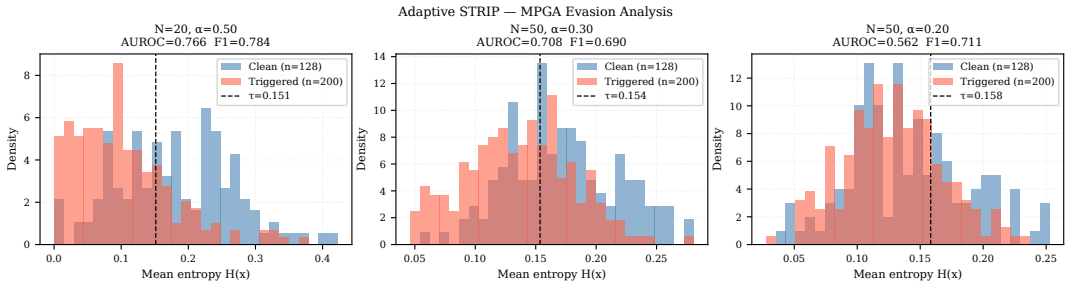

FIGURE 7 Entropy distributions of clean and MPGA-triggered inputs under adaptive STRIP settings. As the blending weight decreases, the entropy gap narrows and the distributions converge, leading to a collapse in detection performance and demonstrating that STRIP's effectiveness depends on overly disruptive blending rather than inherent separability.

model.

Neural Cleanse. Neural Cleanse optimizes a minimal perturbation δ_c for each class and flags the class with an outlier $\|\delta_c\|_1$ using a MAD-based anomaly index (AI). Results in Table 9 and Fig. 6 compare BadNets and MPGA under identical federated settings (ρ=0.15, 2/10 malicious clients, 35 rounds).

For BadNets, the method correctly identifies glioma (class 1) as the backdoored class, with AI = 3.82 exceeding the threshold (2.0) and detection F1 = 1.000. Notably, this holds despite a low ASR of 22.50%, indicating robustness in recovering compact localized triggers. In contrast, MPGA (ASR = 96.50%) completely evades detection. L1 norms are nearly uniform across classes (0.0935, 0.0931, 0.0930), and the highest anomaly index (AI = 6.09) is incorrectly assigned to meningioma (class 0), resulting in F1 = 0.000. The inferred triggers do not match the true pattern (IoU $<$ 0.01). This failure stems from MPGA's distributed design—perturbations across four 8×8 corner patches and a central cross (9.35% pixels)—which prevents convergence to a compact, class-specific solution, causing uniform inversion behavior across classes.

ABS. ABS optimizes neuron-level perturbations to reconstruct class-specific triggers. Results in Table 10 (upper) show failure on the MPGA-poisoned model. It incorrectly identifies class 2 as the backdoor target instead of the true class 1 (Glioma), and the reconstructed trigger for class 1 has an L1 norm $\sim 15\times$ larger than the ground truth (0.4968 vs. 0.0327) with low structural similarity (SSIM = 0.1567). This indicates no meaningful resemblance to the true MPGA pattern. The failure is due to MPGA's distributed design, which produces diffuse neuron activations and prevents convergence to a compact trigger.

STRIP (Standard). Under the default setting (N=20, α=0.50), STRIP overlays random clean images and detects backdoors via low prediction entropy. As shown in Table 10 (middle), triggered samples exhibit lower entropy than clean inputs (0.0505 vs. 0.1807), achieving AUROC = 0.9023 and F1 = 0.8447. However, recall drops to 0.7750, with 45 triggered inputs escaping detection. This occurs because MPGA's multi-region, dual-intensity design increases prediction variability under blending, raising entropy for a subset of triggered inputs beyond the detection threshold.

STRIP (Adaptive). The standard result uses a high blending weight (α=0.50) that disrupts the trigger and favors detection. Reducing α to more realistic values narrows the entropy gap. As shown in Table 10 (lower) and Fig. 7, AUROC drops to 0.7077 for Adaptive1 (N=50, α=0.30) and further to 0.5621 for Adaptive2 (N=50, α=0.20), approaching random performance. The entropy gap collapses from 0.1302 (standard) to 0.0362 and 0.0110, as clean and triggered distributions converge. This confirms that STRIP's detectability under default settings arises from overly disruptive blending rather than intrinsic robustness. All three trigger-agnostic defenses fail against MPGA. Neural Cleanse and

TABLE 9 Neural Cleanse trigger inversion for BadNets and MPGA. (*) indicates the detected backdoored class (AI > 2.0). MPGA shows near-uniform L1 norms, leading to incorrect detection.

	BadNets		MPGA	
Class	**L1**	**AI**	**L1**	**AI**
Meningioma	0.0930	0.0000	0.0935*	6.0914
Glioma (target)	0.0932*	3.8236	0.0931	0.0000
Pituitary	0.0930	−0.6745	0.0930	−0.6745
Predicted target	Glioma (**correct**)		Meningioma (**wrong**)	
Detection F1	**1.000**		**0.000**	

ABS cannot recover a compact trigger due to the distributed multi-patch design, while STRIP fails under adaptive tuning as the dual-intensity pattern preserves sufficient entropy to evade detection. Thus, MPGA remains undetected by all evaluated defenses.

6.8 | Trigger Robustness to Input Transformations

A practical concern for any backdoor attack is whether standard input preprocessing can disrupt the trigger before or during inference. We evaluate MPGA under five common transformations applied to 200 triggered test samples, measuring ASR degradation relative to the unperturbed baseline (ASR = 97.00%).

Table 11 reports results. MPGA is robust to four out of five transformations: Gaussian noise ($\sigma = 0.05$ and 0.10), horizontal flip, and JPEG compression (Q = 70) all yield ASR $\geq$ 92.50%, with drops of at most 4.50 pp. The corner patches, set at an intensity of 0.9, retain sufficient activation after lossy compression and additive noise to maintain dominant token-level attention, confirming the design rationale in SectionIII-C. Random crop-and-resize (±10%) produces the largest reduction (44.50 pp, ASR = 52.50%), as aggressive spatial resampling partially displaces corner patches from their exact token boundaries. Critically, even under this worst-case condition, the ASR remains above the 40% attack-success threshold, confirming that MPGA is not defeated by any single standard preprocessing step.

TABLE 10 Trigger-agnostic defense evaluation of an MPGA-poisoned model. ABS incorrectly identifies the target, and STRIP performance degrades under adaptive settings.

ABS: Trigger Inversion		
Class	**L1 Norm**	**SSIM vs GT**
0 (Meningioma)	0.4994	0.038
1 (Glioma — target)	0.4968	0.157
2 (Pituitary)	0.4955	−0.076
Predicted target	Class 2 (**wrong**)	
STRIP: Standard (N=20, α=0.50)		
Metric	**Value**	
Entropy (Triggered)	0.0505 ± 0.0549	
Entropy (Clean)	0.1807 ± 0.0854	
AUROC	0.9023	
F1 / Precision / Recall	0.8447 / 0.9281 / 0.7750	
TP / FP / TN / FN	155 / 12 / 116 / 45	
Adaptive STRIP: Effect of Blending Weight		
Configuration	**AUROC**	**F1**
Standard (N=20, α=0.50)	0.7660	0.7843
Adaptive1 (N=50, α=0.30)	0.7077	0.6900
Adaptive2 (N=50, α=0.20)	0.5621	0.7107

TABLE 11 Trigger robustness to input transformations.

Transformation	**ASR (%)**	**Drop (pp)**	**Verdict**
No transformation (baseline)	97.00	—	—
Gaussian noise (σ = 0.05)	98.50	−1.50	Robust
Gaussian noise (σ = 0.10)	98.00	−1.00	Robust
Horizontal flip	92.50	4.50	Robust
JPEG compression (Q = 70)	97.50	−0.50	Robust
Crop-and-resize (±10%)	52.50	44.50	Degraded

7 | CONCLUSION

This study presents MPGA, a distributed backdoor specifically designed to exploit ViT patch tokenization in federated medical imaging. By aligning four corner patches (8×8, intensity 0.9) with ViT patch boundaries and combining

them with a central cross pattern (1-pixel, intensity 0.6), MPGA distributes malicious features across multiple attention tokens, achieving a 94.13% BSR while preserving 86.13% MTA across 10 federated clients. A comprehensive evaluation across ten defenses and four paradigms confirms a systematic failure: no defense meets the combined ASR/MTA/F1 success criteria. Trigger-agnostic methods (Neural Cleanse, ABS, and STRIP) fail because the distributed 9.35%-coverage design prevents compact trigger recovery, and STRIP's entropy gap collapses under adaptive tuning (AUROC 0.562 at α=0.20). The backdoor persists after attack cessation (87.00% ASR after 15 benign rounds) and generalizes across datasets with different class structures and resolutions (IEEE Dataport: 95.00% BSR, 89.33% MTA). A comparison with BadNets and DBA confirms that explicit patch-boundary alignment is the key differentiator, with MPGA achieving the highest ASR under all stable defenses. These results reveal that existing paradigms — designed around localized, CNN-oriented triggers — are fundamentally insufficient against architecture-aware distributed backdoors in federated ViT systems. Future work should investigate hybrid defenses that combine attention pattern analysis with gradient-space anomaly detection and certified robustness bounds for federated ViT systems under distributed spatial backdoors.

Conflict of interest

The authors declare no competing interests regarding this publication.

Acknowledgment

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R821), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability Statement

The datasets used in this study are publicly available. The Figshare Brain Tumor MRI dataset is available at `https://doi.org/10.6084/m9.figshare.1512427.v5` and the IEEE Dataport Brain Tumor MRI dataset is available at `https://dx.doi.org/10.21227/kw9f-bz10`.

Code Availability

The implementation code will be made available upon reasonable request to the corresponding author.

Author Contributions

Sudheer T.M.: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft. **Deepak S.:** Supervision, Writing – Review & Editing. **Arun Varghese:** Writing – Review & Editing, Validation. **Ameer P.M.:** Writing – Review & Editing, Supervision. **Shaeen Kalathil:** Funding Acquisition, Writing – Review & Editing.

references

[1] McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: International Conference on Artificial Intelligence and Statistics (AISTATS); 2017. p. 1273–1282.

[2] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The Future of Digital Health with Federated Learning. npj Digital Medicine 2020;3(1):119.

[3] Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In: International MICCAI Brainlesion Workshop Springer; 2018. p. 92–104.

[4] Flores MJ, Nicholson BD, Soltan AA, Ukpo O, Sharma P, Hattersley JG, et al. Federated Learning Used for Predicting Outcomes in COVID-19 Cases. Research Square 2021;.

[5] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research 2021;5(1):1–19.

[6] Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to Backdoor Federated Learning. In: International Conference on Artificial Intelligence and Statistics (AISTATS); 2020. p. 2938–2948.

[7] Wang H, Sreenivasan K, Rajput S, Vishwakarma H, Agarwal S, Sohn Jy, et al. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33; 2020. p. 16070–16084.

[8] Xie C, Huang K, Chen PY, Li B. DBA: Distributed Backdoor Attacks against Federated Learning. In: International Conference on Learning Representations (ICLR); 2020. .

[9] Tolpegin V, Truex S, Gursoy ME, Liu L. Data Poisoning Attacks Against Federated Learning Systems. In: European Symposium on Research in Computer Security (ESORICS) Springer; 2020. p. 480–501.

[10] Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30; 2017. .

[11] Muñoz-González L, Co KT, Lupu EC. Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging. arXiv preprint arXiv:190905125 2019;.

[12] Sun Z, Kairouz P, Suresh AT, McMahan HB. Can You Really Backdoor Federated Learning? arXiv preprint arXiv:191107963 2019;Published in NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality.

[13] Nguyen TD, Rieger P, De Viti R, Chen H, Brandenburg BB, Yalame H, et al. FLAME: Taming Backdoors in Federated Learning. In: USENIX Security Symposium; 2022. p. 1415–1432.

[14] Geyer RC, Klein T, Nabi M. Differentially Private Federated Learning: A Client Level Perspective. In: NIPS Workshop on Machine Learning on the Phone and other Consumer Devices; 2017. .

[15] Zhang H, Li X, Miao Y, Yuan S, Zhu M, Liu X, et al. FL-CDF: Collaborative Defense Framework for Backdoor Mitigation in Federated Learning. IEEE Transactions on Dependable and Secure Computing 2025;22(6):6732–6747.

[16] Rieger P, Nguyen TD, Miettinen M, Sadeghi AR. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. In: Network and Distributed System Security Symposium (NDSS); 2022. .

[17] Zhang Z, Cao X, Jia J, Gong NZ. FLDetector: Defending Federated Learning Against Model Poisoning Attacks via Detecting Malicious Clients. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2022. p. 2545–2555.

[18] Shejwalkar V, Houmansadr A, Kairouz P, Ramage D. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning. In: IEEE Symposium on Security and Privacy (S&P); 2022. p. 1354–1371.

[19] Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-Preserving Federated Brain Tumour Segmentation. Machine Learning in Medical Imaging MLMI (Workshop) 2021;12966:133–141.

[20] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial Attacks on Medical Machine Learning. Science 2019;363(6433):1287–1289.

[21] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in Medical Imaging: A Survey. Medical Image Analysis 2023;88:102802.

[22] Matsoukas C, Haslum JF, Söderberg M, Smith K. Is It Time to Replace CNNs with Transformers for Medical Images? In: IEEE International Conference on Computer Vision Workshops (ICCVW); 2021. p. 4038–4047.

[23] Cao X, Fang M, Liu J, Gong NZ. Provably Secure Federated Learning against Malicious Clients. In: AAAI Conference on Artificial Intelligence, vol. 35; 2021. p. 6885–6893.

[24] Yin D, Chen Y, Kannan R, Bartlett P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In: International Conference on Machine Learning (ICML); 2018. p. 5650–5659.

[25] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations 2021;.

[26] Baruch G, Baruch M, Goldberg Y. A Little Is Enough: Circumventing Defenses for Distributed Learning. In: Advances in Neural Information Processing Systems, vol. 32; 2019. .

[27] Zhang Z, Panda A, Song L, Yang Y, Mahoney M, Mittal P, et al. Neurotoxin: Durable Backdoors in Federated Learning. In: International Conference on Machine Learning (ICML); 2022. p. 26429–26446.

[28] Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. vol. 7; 2019. p. 47230–47244.

[29] Chen X, Liu C, Li B, Lu K, Song D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. In: arXiv preprint arXiv:1712.05526; 2017. .

[30] Liu Y, Ma X, Bailey J, Lu F. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. In: European Conference on Computer Vision Springer; 2020. p. 182–199.

[31] Nguyen TA, Tran A. WaNet–Imperceptible Warping-based Backdoor Attack. In: International Conference on Learning Representations; 2021. .

[32] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In: AAAI Conference on Artificial Intelligence Workshop on Artificial Intelligence Safety; 2019. .

[33] Wang B, Yao Y, Shan S, Li H, Viswanath B, Zheng H, et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In: IEEE Symposium on Security and Privacy (S&P) IEEE; 2019. p. 707–723.

[34] Liu Y, Lee WC, Tao G, Ma S, Aafer Y, Zhang X. ABS: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation. In: ACM SIGSAC Conference on Computer and Communications Security; 2019. p. 1265–1282.

[35] Gao Y, Xu C, Wang D, Chen S, Ranasinghe DC, Nepal S. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks. In: Annual Computer Security Applications Conference; 2019. p. 113–125.

[36] Bai J, Wu B, Zhang Y, Li Y, Li Z, Xia ST. Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits. In: International Conference on Learning Representations; 2021. .

[37] Yang W, Lin Y, Li P, Zhou J, Sun X. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021;p. 2048–2058.

[38] Zhang S, Yi J, Lv Q, Liu L, Gu G. Backdoor Attacks on the DNN Interpretation System. In: Annual Computer Security Applications Conference; 2020. p. 675–686.

[39] Schwarzschild A, Goldblum M, Gupta A, Dickerson JP, Goldstein T. Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. In: International Conference on Machine Learning PMLR; 2021. p. 9389–9398.

[40] Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL); 2020. p. 4190–4197.

[41] Cheng J, Brain Tumor Dataset. figshare; 2017. https://doi.org/10.6084/m9.figshare.1512427.v5, 3064 T1-weighted contrast-enhanced MRI images; meningioma (708), glioma (1426), pituitary (930).

[42] Bonala S, Brain MRI Dataset for Glioma, Meningioma, and Pituitary Tumor Classification. IEEE Dataport; 2025. https://dx.doi.org/10.21227/kw9f-bz10.

[43] Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, et al. On Evaluating Adversarial Robustness. arXiv preprint arXiv:190206705 2019;.

references

[1] McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: International Conference on Artificial Intelligence and Statistics (AISTATS); 2017. p. 1273–1282.

[2] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The Future of Digital Health with Federated Learning. npj Digital Medicine 2020;3(1):119.

[3] Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In: International MICCAI Brainlesion Workshop Springer; 2018. p. 92–104.

[4] Flores MJ, Nicholson BD, Soltan AA, Ukpo O, Sharma P, Hattersley JG, et al. Federated Learning Used for Predicting Outcomes in COVID-19 Cases. Research Square 2021;.

[5] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research 2021;5(1):1–19.

[6] Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to Backdoor Federated Learning. In: International Conference on Artificial Intelligence and Statistics (AISTATS); 2020. p. 2938–2948.

[7] Wang H, Sreenivasan K, Rajput S, Vishwakarma H, Agarwal S, Sohn Jy, et al. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33; 2020. p. 16070–16084.

[8] Xie C, Huang K, Chen PY, Li B. DBA: Distributed Backdoor Attacks against Federated Learning. In: International Conference on Learning Representations (ICLR); 2020. .

[9] Tolpegin V, Truex S, Gursoy ME, Liu L. Data Poisoning Attacks Against Federated Learning Systems. In: European Symposium on Research in Computer Security (ESORICS) Springer; 2020. p. 480–501.

[10] Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30; 2017. .

[11] Muñoz-González L, Co KT, Lupu EC. Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging. arXiv preprint arXiv:190905125 2019;.

[12] Sun Z, Kairouz P, Suresh AT, McMahan HB. Can You Really Backdoor Federated Learning? arXiv preprint arXiv:191107963 2019;Published in NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality.

[13] Nguyen TD, Rieger P, De Viti R, Chen H, Brandenburg BB, Yalame H, et al. FLAME: Taming Backdoors in Federated Learning. In: USENIX Security Symposium; 2022. p. 1415–1432.

[14] Geyer RC, Klein T, Nabi M. Differentially Private Federated Learning: A Client Level Perspective. In: NIPS Workshop on Machine Learning on the Phone and other Consumer Devices; 2017. .

[15] Zhang H, Li X, Miao Y, Yuan S, Zhu M, Liu X, et al. FL-CDF: Collaborative Defense Framework for Backdoor Mitigation in Federated Learning. IEEE Transactions on Dependable and Secure Computing 2025;22(6):6732–6747.

[16] Rieger P, Nguyen TD, Miettinen M, Sadeghi AR. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. In: Network and Distributed System Security Symposium (NDSS); 2022. .

[17] Zhang Z, Cao X, Jia J, Gong NZ. FLDetector: Defending Federated Learning Against Model Poisoning Attacks via Detecting Malicious Clients. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2022. p. 2545–2555.

[18] Shejwalkar V, Houmansadr A, Kairouz P, Ramage D. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning. In: IEEE Symposium on Security and Privacy (S&P); 2022. p. 1354–1371.

[19] Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-Preserving Federated Brain Tumour Segmentation. Machine Learning in Medical Imaging MLMI (Workshop) 2021;12966:133–141.

[20] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial Attacks on Medical Machine Learning. Science 2019;363(6433):1287–1289.

[21] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in Medical Imaging: A Survey. Medical Image Analysis 2023;88:102802.

[22] Matsoukas C, Haslum JF, Söderberg M, Smith K. Is It Time to Replace CNNs with Transformers for Medical Images? In: IEEE International Conference on Computer Vision Workshops (ICCVW); 2021. p. 4038–4047.

[23] Cao X, Fang M, Liu J, Gong NZ. Provably Secure Federated Learning against Malicious Clients. In: AAAI Conference on Artificial Intelligence, vol. 35; 2021. p. 6885–6893.

[24] Yin D, Chen Y, Kannan R, Bartlett P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In: International Conference on Machine Learning (ICML); 2018. p. 5650–5659.

[25] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations 2021;.

[26] Baruch G, Baruch M, Goldberg Y. A Little Is Enough: Circumventing Defenses for Distributed Learning. In: Advances in Neural Information Processing Systems, vol. 32; 2019. .

[27] Zhang Z, Panda A, Song L, Yang Y, Mahoney M, Mittal P, et al. Neurotoxin: Durable Backdoors in Federated Learning. In: International Conference on Machine Learning (ICML); 2022. p. 26429–26446.

[28] Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. vol. 7; 2019. p. 47230–47244.

[29] Chen X, Liu C, Li B, Lu K, Song D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. In: arXiv preprint arXiv:1712.05526; 2017. .

[30] Liu Y, Ma X, Bailey J, Lu F. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. In: European Conference on Computer Vision Springer; 2020. p. 182–199.

[31] Nguyen TA, Tran A. WaNet–Imperceptible Warping-based Backdoor Attack. In: International Conference on Learning Representations; 2021. .

[32] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In: AAAI Conference on Artificial Intelligence Workshop on Artificial Intelligence Safety; 2019. .

[33] Wang B, Yao Y, Shan S, Li H, Viswanath B, Zheng H, et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In: IEEE Symposium on Security and Privacy (S&P) IEEE; 2019. p. 707–723.

[34] Liu Y, Lee WC, Tao G, Ma S, Aafer Y, Zhang X. ABS: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation. In: ACM SIGSAC Conference on Computer and Communications Security; 2019. p. 1265–1282.

[35] Gao Y, Xu C, Wang D, Chen S, Ranasinghe DC, Nepal S. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks. In: Annual Computer Security Applications Conference; 2019. p. 113–125.

[36] Bai J, Wu B, Zhang Y, Li Y, Li Z, Xia ST. Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits. In: International Conference on Learning Representations; 2021. .

[37] Yang W, Lin Y, Li P, Zhou J, Sun X. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021;p. 2048–2058.

[38] Zhang S, Yi J, Lv Q, Liu L, Gu G. Backdoor Attacks on the DNN Interpretation System. In: Annual Computer Security Applications Conference; 2020. p. 675–686.

[39] Schwarzschild A, Goldblum M, Gupta A, Dickerson JP, Goldstein T. Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. In: International Conference on Machine Learning PMLR; 2021. p. 9389–9398.

[40] Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL); 2020. p. 4190–4197.

[41] Cheng J, Brain Tumor Dataset. figshare; 2017. https://doi.org/10.6084/m9.figshare.1512427.v5, 3064 T1-weighted contrast-enhanced MRI images; meningioma (708), glioma (1426), pituitary (930).

[42] Bonala S, Brain MRI Dataset for Glioma, Meningioma, and Pituitary Tumor Classification. IEEE Dataport; 2025. https://dx.doi.org/10.21227/kw9f-bz10.

[43] Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, et al. On Evaluating Adversarial Robustness. arXiv preprint arXiv:190206705 2019;.

Sudheer TM received the B.Tech. degree in Electronics and Communication Engineering from Mahatma Gandhi University, Kottayam, India, and the M.Tech. degree in Artificial Intelligence and Data Science from the Indian Institute of Information Technology (IIIT), Kottayam, India. He is currently working as an Assistant Professor with the Department of Electronics and Communication Engineering, Government Engineering College, Thrissur, Kerala, India. He previously served as a Scientist at the Indian Space Research Organisation (ISRO) and has also held academic positions at Government Engineering College, Kannur, and Government Engineering College, Wayanad. His research interests include deep learning, machine learning, and the Internet of Things (IoT).

Dr. Deepak S received the Ph.D. degree in AI-based medical image processing from the National Institute of Technology Calicut, India. He completed his M.Tech. degree in Applied Electronics from the College of Engineering Trivandrum (CET), Kerala. Currently, he is Assistant Professor at Government Engineering College, Thrissur, Kerala. He has previously served at CET Trivandrum, GEC Idukki, LBS Kasaragod, Amrita Vishwa Vidyapeetham, IIST Trivandrum, and IES Thrissur. His research interest is in AI-based signal processing.

Dr. Arun varghesereceived the B.Tech and M.Tech degrees in electronics engineering from Cochin University in 1998 and 2002 respectively. He received his PhD degree from NIT Calicut in 2021. Presently he is an Associate Professor at College of Engineering Trivandrum. His research interests are computer vision and machine learning

Dr. Ameer P.M. received the Ph.D.degree in Electronics and Communication Engineering from the National Institute of Technology Calicut. Completed his ME in Telecommunication from Indian Institute of Science Bangalore. He has completed B.Tech in Electronics & Communication Engineering from University of Kerala. He is currently working as Associate professor at National Institute of Technology Calicut. His research interests include image processing, deep learning, communication networks and signal processing.

Dr. Shaeen Kalathil (Senior Member, IEEE) received her B.Tech degree in Electronics and Communication Engineering from Kannur University, India, in 2004. She earned her M.Tech degree in Electronics Design Technology in 2010 and her Ph.D. in Signal Processing in 2015, both from the National Institute of Technology (NIT), Calicut, India. She has around 16 years of teaching experience and she joined College of Engineering, Princess Nourah bint Abdulrahman University in November 2018. Her research interests include signal processing, multirate systems and filter banks, bio-signal analysis, the Internet of Things (IoT), and embedded systems.