FedSegMIA: Exploring Privacy Risks in Federated Binary Segmentation

Eugénie Laugier
Thales
cortAIx Labs, France

Alice Héliou
Thales
cortAIx Labs, France
alice.heliou@thalesgroup.com

Vincent Thouvenot
Thales
cortAIx Labs, France
vincent.thouvenot@thalesgroup.com

Katarzyna Kapusta
Thales
cortAIx Labs, France
katarzyna.kapusta@thalesgroup.com

***Abstract*—Federated learning (FL) enables collaborative training of machine learning models across multiple parties without sharing raw data, making it particularly appealing for defense applications involving sensitive or classified information. While the privacy risks of FL have been extensively studied for classification tasks, vulnerabilities in federated segmentation models, which are widely used for precise object detection and reconnaissance, remain largely unexplored. Existing studies on segmentation have focused on centralized settings, typically relying on prediction losses as the main leakage vector.**

In this work, we present the first systematic analysis of membership inference attacks on binary segmentation models trained under FL. We demonstrate that gradient updates provide a significantly stronger signal for inferring training data membership than losses, posing substantial risks in collaborative defense scenarios. Our experiments highlight the need of implementing robust privacy-preserving mechanisms to protect critical operational data.

***Index Terms*—membership inference attacks, automatic target detection, segmentation, federated learning, privacy**

I. Introduction

In modern military operations, multiple units (from ground vehicles to reconnaissance drones) must collaboratively build a shared understanding of the battlefield, without exposing sensitive imagery. Collaborative learning offers a solution by enabling each system to improve its perception capabilities while keeping raw data private. To be effective, these systems must detect, recognize, and precisely locate objects to operate in complex environments. This makes semantic segmentation a model of choice, as it provides precise pixel-level classification, essential for identifying targets and distinguishing allies from adversaries. But, this setup raises a critical question: could collaborative learning of semantic segmentation models unintentionally leak sensitive information ?

Federated learning (FL) [1] allows multiple parties to collaboratively train a machine learning model without sharing their raw data. The idea is that each party has its local data and only model updates on the local data are shared with an aggregation server that orchestrates the training. By enabling multiple clients to jointly optimize a global model while keeping their data local, FL offers an attractive solution for privacy-preserving learning in critical domains such as healthcare [2], [3], finance [4], and increasingly, defense [5].

Although federated learning is designed to reduce privacy risks, the information exchanged during training can still leak sensitive details. A growing line of work has investigated privacy attacks that exploit these shared model updates. Among them, membership inference attacks (MIAs) [6] are particularly concerning. As illustrated by Figure 1, MIAs seek to determine whether a specific data point was part of a client's training set, exploiting the tendency of machine learning models to memorize training data. In the context of MIA, a 'member' refers to data that is part of the training dataset of the targeted model, while a 'non-member' denotes data that is not included in this dataset. Their principle relies on the fact that the model behaves differently between training data and unseen data. Unlike gradient inversion attacks (GIAs) [7], which attempt to reconstruct input data from shared gradients and often rely on restrictive conditions such as small batch sizes or minimal local training epochs, MIAs can operate under more realistic assumptions. Consequently, MIAs pose a practical and significant threat in federated environments.

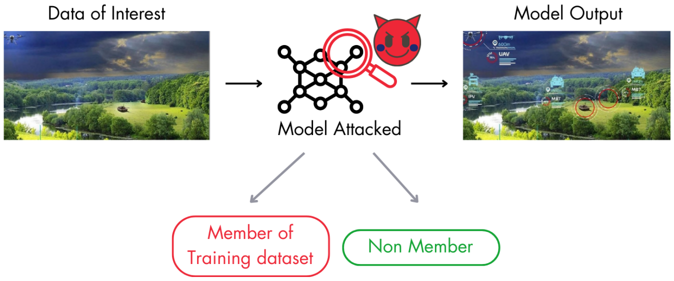

Fig. 1. MIA principle in centralized learning.

While research has focused on MIAs in the context of classification tasks [8]–[10], significantly less attention has been paid to other machine learning applications that also handle highly sensitive data. In particular, semantic segmentation has received little scrutiny regarding its vulnerability to membership inference in federated learning. This gap is especially concerning given the growing interest in deploying segmentation models on edge devices such as drones [11], [12]. These devices increasingly rely on federated learning to collaboratively improve perception models without transmitting raw imagery back to a central server.

In this work, we address this gap by examining server-side membership inference attacks during federated learning of binary segmentation models.

- We evaluate the effectiveness of membership inference attacks on federated binary segmentation, highlighting that even without strong assumptions about attacker's capabilities, these models are susceptible to privacy breaches.
- We improve over the state-of-the-art by showing that artificially increasing the number of clients can introduce a bias that amplifies the effectiveness of gradient-based attacks.
- Finally, we hypothesize that certain iteration rounds exhibit stronger susceptibility to inference attacks than others. By leveraging the segmentation performance of the federated model to identify and select these more informative iterations, we demonstrate that the server can enhance the effectiveness of inference .

Our findings underscore the need of addressing privacy vulnerabilities and server-side threats in federated learning, especially as it is applied to complex tasks like segmentation.

II. Business Need / Motivations

Federated learning is inherently suited for scenarios involving sensitive or confidential data, making it an attractive approach across domains such as healthcare, finance, and particularly defense. In military contexts, FL can be deployed to collaboratively trained models across multiple entities without sharing raw data, preserving operational secrecy. Beyond centralized installations, federated learning holds promise for integration directly on edge platforms such as autonomous vehicles and drones; enabling these systems to continuously improve object detection or segmentation capabilities by learning from local observations collected in diverse environments. For reconnaissance drones in particular, this means adapting models in real time to new terrains or targets, without ever transmitting potentially sensitive imagery back to a central server. Additionally, by sharing only model updates instead of raw data, FL also helps lower the communication costs associated with centralized learning.

While reconstructing complete images from model updates remains a technical challenge, subtle forms of information leakage, such as revealing data characteristics (e.g., image resolution or content type) or identifying which clients participated in training, could still pose serious risks. This underscores the importance of thoroughly understanding privacy vulnerabilities in federated learning.

III. Related Work

A. Federated Learning

Existing FL architectures can be either centralized (Fig.2), relying on a server to aggregate local updates, or fully decentralized, where clients communicate peer-to-peer [13]. In centralized settings, the server coordinates the learning process [14], and at each iteration it collects all local updates and aggregates them. This aggregation process grants the server access to a substantial amount of information, positioning it as a potentially powerful adversary. In our work, we will only focus on a centralized federated setting and demonstrate the extent to which a server can infer information about local datasets. However, recent research has shown that significant privacy risks persist even in decentralized settings [15].

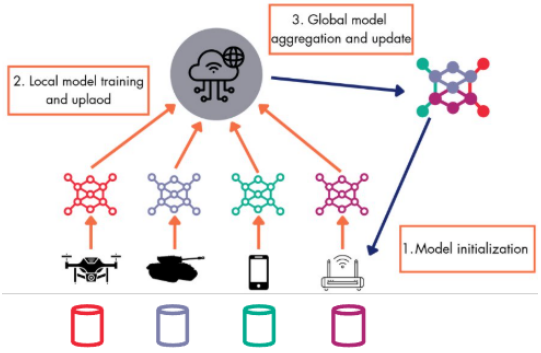
3. Global model aggregation and update
2. Local model training and uplaod
1.Model initialization

Fig. 2. Principle of Centralized Federated Learning.

B. MIA in a Federated Setting

MIAs leverage the observation that machine learning models tend to behave differently on data they have seen during training versus unseen data: producing lower losses or more confident predictions on training samples [8]. Various strategies have been proposed to exploit this discrepancy, differing mainly in the metrics they use and the level of adversary involvement, from passive observation to actively training shadow models.

1) Attacks based on model updates: A significant body of work investigates MIAs that exploit information contained in model updates exchanged during federated learning.

Gradient-based attacks compare raw gradients, their norms, or compute metrics such as cosine similarity to distinguish members from non-members [9], [16], [17]. These methods are often highly effective in classification and entirely passive, but require direct access to model gradients (white-box scenario).

Loss-based attacks instead rely on the observation that the loss is typically lower for training data [18]. These attacks can be carried out without access to the model architecture and weights (black-box settings).

Another classical strategy involves shadow training, where the attacker builds one or more shadow models on data drawn from a distribution similar to that of the target. These shadow models are then used to train membership classifiers that predict whether a given sample was part of the target's training set [19] [20]. While this approach is often more precise, it requires substantial auxiliary data and considerable computational resources, making it significantly more expensive than above attacks, which typically only involve running inference or computing gradients on the target model.

More intrusive attacks involve manipulating local models or the training process itself to introduce vulnerabilities that can later be exploited [21], [22]. Such methods reduce the need for data but are more detectable by traditional defense methods.

2) Attacks based on training dynamics: Other techniques analyze how certain metrics evolve over multiple training rounds, typically without requiring access to labels or explicit gradients. This places them in a largely black-box setting, imposing fewer constraints on the attacker. For example, some approaches monitor the evolution of the loss across federated iterations [23]–[25], while others track how prediction confidences change over time [16], [26]. More recent methods examine shifts in the bias terms of the final layer [27]. However, these strategies are often more sensitive to training dynamics, and some still rely on access to prediction confidences or internal parameters, which may not always be available in practice.

C. FedMIA [10]

Instead of training shadow models to obtain additional information at a significant computational costs, [10] proposes an approach that leverages information from non-target clients. Assuming that clients' datasets are disjoint, they demonstrate that it is possible to estimate the distribution of attack signals (such as losses or gradients) for models that are not trained on the target data (the ”non member” distribution). Then, by employing a one-tailed likelihood-ratio hypothesis test using the estimated non-member distribution, they can infer whether the target data was part of the training dataset for the targeted client. The Figure 3 illustrates the FedMIA approach. Our work builds in part on this method by combining it with shadow models to further enhance the information accessible to the central server.

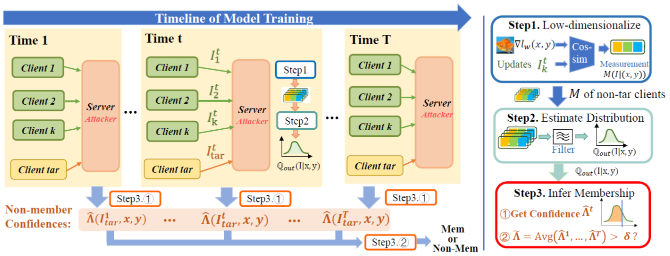
Timeline of Model Training
Time 1
Client 1
Client 2
Client k
Client tar
Server
Attacker
Time t
I_1^t
I_2^t
I_k^t
I_{tar}^t
Step1
Step2
$\mathbb{Q}_{out}(I|x,y)$
Time T
Step3.①
Non-member Confidences:
$\hat{\Lambda}(I_{tar}^1, x, y)$
$\hat{\Lambda}(I_{tar}^t, x, y)$
$\hat{\Lambda}(I_{tar}^T, x, y)$
Step3.②
Mem or Non-Mem
Step1. Low-dimensionalize
$\nabla l_w(x,y)$
Updates I_k^t
Cos-sim
Measurement $M(I|(x,y))$
M of non-tar clients
Step2. Estimate Distribution
Filter
$\mathbb{Q}_{out}(I|x,y)$
Step3. Infer Membership
①Get Confidence $\hat{\Lambda}^t$
② $\tilde{\Lambda} = \text{Avg}(\hat{\Lambda}^1, ..., \hat{\Lambda}^T) > \delta$?

Fig. 3. All vs Target Attack Principle: FedMIA [10] Overview.

D. MIA on Segmentation Models

Segmentation tasks fundamentally differs from classification by predicting a label for each pixel rather than assigning a single label per image, and typically employ pixel-wise losses that compare entire spatial maps.

The richer output structure of segmentation models may affect both the type of privacy leakage and the metrics that are most informative for inference attacks.

Although some studies have investigated MIAs against segmentation models under conventional centralized training, this body of work remains relatively limited. Most approaches exploit the segmentation loss, operating under the hypothesis that it reveals more about individual data samples than in standard classification settings. Early work on membership inference for segmentation models focused on exploiting localized loss signals. He et al. [28] proposed a patch-based analysis of the loss map, showing that certain spatial regions of the output can carry stronger membership signals than global loss alone. Building on this direction, Chobola et al. [29] conducted a more comprehensive study, distinguishing between binary and multiclass segmentation settings. Using shadow models, they evaluated three attack strategies: one based on global loss, one using patch-wise loss, and another combining the target model's predictions with the ground truth masks. Their findings revealed that, somewhat surprisingly, the global loss often remained the most effective signal. However, all these investigations have been confined to centralized learning, leaving open the question of how segmentation models trained under federated learning might be vulnerable to membership inference. In this work, we address this gap by presenting the first systematic study of MIAs on federated binary segmentation models, while also evaluating signals beyond the commonly examined loss.

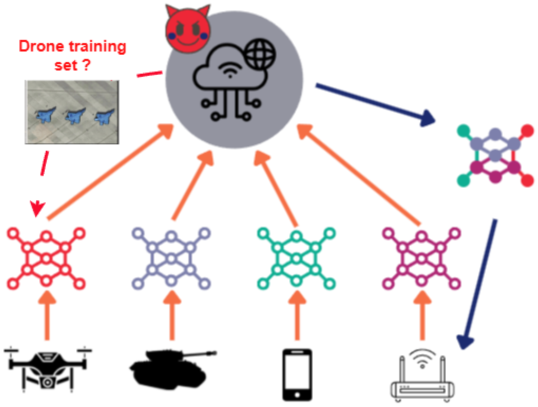

Fig. 4. MIA Threat Model in Federated Learning.

IV. Membership Inference Attack on a Binary Segmentation Model in a Federated Setting

A. Threat Model

We consider a federated learning setting with a centralized architecture, where a server orchestrates the collaborative training of a global model by aggregating updates received from multiple clients. In this context, we examine the scenario in which the server aims to infer private information from the clients' contributions, without seeking to interfere with the federated learning process.

Figure 4 provides an illustration of this threat model.

The server has direct access to all model updates exchanged during training, including their weights, architecture and hyper-parameters, corresponding to a white-box scenario. We consider two levels of attacker behavior.

- Passive scenario: the server respects the FL protocol without interfering with client operations. Its only intervention consists in executing additional inference passes on candidate samples.
- Proactive scenario: the server injects artificial clients into the training process to simulate non-member behaviors. This allows the server to better characterize non-member

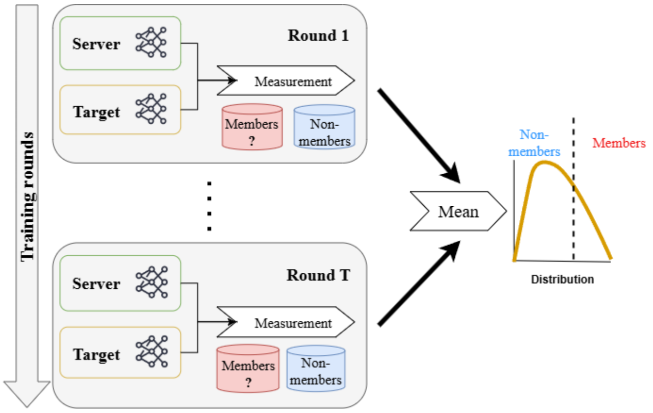

Fig. 5. Server vs Target Attack Principle.

samples and thus more easily distinguish them from members.

In both cases, we assume all clients behave honestly, they do not collude with the server or with each other and do not try to poison the training.

B. Notations

We consider a federated learning process over T communication rounds with K clients. Let (x, y) denote a target data instance, where x is an input image and y its corresponding ground-truth segmentation mask. We use $W_S^{(t)}$ to represent the global model weights held by the server at the end of round t, and $W_S^{(t-1)}$ for the previous round. Similarly, $W_k^{(t)}$ denotes the local model weights trained by client k during round t. The local update sent by client k is given by Equation (1).

$$I_k^t = W_k^{(t)} - W_S^{(t-1)}. \tag{1}$$

We further denote by $\nabla_W L(y, x; W_S^{(t)})$ the gradient of the loss function L evaluated on (x, y) with respect to the global model parameters. This quantity captures the direction in parameter space that would most improve prediction on the specific instance (x, y).

C. Type of MIAs

In this work, we investigate two main dimensions of membership inference attacks on federated segmentation models: the attacker's comparison strategy and the signal exploited to infer membership.

1) Comparing Server-Only and All-for-One Attacks: We explore two different attacker perspectives based on how the server leverages the information available from local updates.

- **Server vs Target:** In this classical approach, the server focuses exclusively on the update received from the targeted client. As illustrated by Fig.5, at each communication round t, for a given candidate instance (x, y), the server computes an indicator of membership (such as cosine similarity or loss difference) by comparing the targeted client's update $I_k^t = W_k^{(t)} - W_S^{(t-1)}$ to the behavior of the global model.
- **All vs Target:** This more recent approach, inspired by FedMIA [10], leverages the updates from all participating clients. For each round, the server estimates the distribution of membership signals across non-target clients, treating them as a baseline under the null hypothesis that they did not train on (x, y). A statistical test is then performed to assess whether the target client's signal significantly deviates from this distribution, thus providing a confidence measure (p-value) for inferring membership.

2) Attack Signal: We evaluate two primary signals used to distinguish member from non-member data:

- **Cosine Similarity of Gradients:** For each candidate instance, we compute the gradient of the loss with respect to the global model parameters at round t, denoted $\nabla_W L(y, x; W_S^{(t)})$. The cosine similarity between this gradient and the targeted client's update measures their alignment:

$$\text{cosim}(x, y) = \frac{\langle \nabla_W L(y, x; W_S^{(t)}), I_k^t \rangle}{\|\nabla_W L(y, x; W_S^{(t)})\|_2 \cdot \|I_k^t\|_2}. \tag{2}$$

  Empirically, gradients associated with independent data tend to be nearly orthogonal in high-dimensional spaces, so a significantly higher similarity indicates that (x, y) may have been used to train the local model.
- **Loss Difference:** This simpler attack compares the loss values computed by the global model and by the target client. In segmentation, the pixel-wise nature of the loss function provides a finer-grained signal than typical classification settings. The hypothesis is that if (x, y) was seen during local training, the discrepancy in loss between the server and the client's update will be statistically smaller.

V. Experimental Setup

A. Dataset and Model

We restrict our study to binary segmentation, both for simplicity and as a first step toward understanding membership inference vulnerabilities in federated segmentation models. Our target model is a UNet [30], a widely used architecture in segmentation tasks. We employed the iSAID dataset [31], which contains aerial images from complex scenes annotated across 15 object classes, including ships, aircraft, and harbors. This dataset was chosen as it most closely resembles the Automatic Target Detection/Recognition use case, which is particularly relevant for application of segmentation models in defense, with relevant classes and varying image resolutions. To adapt it to a binary segmentation task, we filtered the dataset to retain only images containing at least one instance of type *harbor*. After filtering, the final dataset comprised 311 images, of which 281 were used for training and 30 for testing. Figure 6 shows an example image and its corresponding mask.

B. Data Distributions

Given the limited number of training images, we explored two data distribution strategies, each aligned with one of the attacker behaviors introduced in Section IV-A.

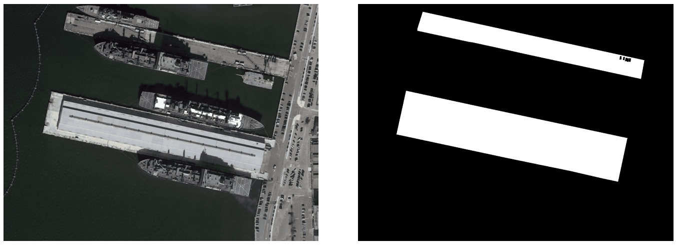

Fig. 6. Image (left) and Label (right) of the filtered iSAID [31] dataset.

For the first scenario, training data was distributed equally among clients. We limited experiments to 3 and 5 clients, as beyond that threshold, each client receives too few images, making local training less meaningful.

To simulate the proactive server scenario, we adopted an overlapping data distribution. In this setup :

- The clients 1 (the target) and 2 receive 25% of the training data (70 images each),
- The remaining 50% of the training data (called the shadow dataset) is randomly split among synthetic clients, such that all clients hold 70 images. We use a draw with replacement, so that an observation can be simultaneously be present in two (or more) clients. This allows us to simulate a larger number of clients without reducing the data available to the target client.

We conducted experiments with 5, 10, and 20 clients under this setting. To construct evaluation datasets, we sampled member instances from the target client's training data, and non-member instances from the concatenation of the test set and the training data from other clients.

C. Training Setup

We trained the global model using the classical FedAvg [14] aggregation scheme, computing a simple average of updated weights from all clients. All images were resized to 300×350 pixels. Training employed the Adam optimizer with learning rates ranging from 10^{-5} to 10^{-2}. To mitigate class imbalance in the masks, we used a weighted binary cross-entropy loss defined by:

$$L_{\text{balanced}}(y, \hat{y}) = L(y, \hat{y}) \times (1 - \alpha + \alpha\, y_{\text{balanced}}),$$

where L is the standard binary cross-entropy, $\hat{y}$ the predicted mask, and y_{balanced} is calculated as

$$y_{\text{balanced},i,j} = \begin{cases} \gamma & \text{if } y_{i,j} = 0 \\ 1 - \gamma & \text{otherwise} \end{cases}$$

with γ equal to the proportion of positive pixels in y.

We investigated the influence of the number of local epochs per client by testing values of 1, 2, and 4. We kept the total local training epochs fixed at 100, resulting in 100, 50, or 25 communication rounds respectively. This ensures equivalent data exposure across all configurations, which is essential for future studies incorporating differential privacy. We hypothesized that increasing local epochs could amplify attack success by widening the gap between local and global models.

All experiments used a fixed random seed. Due to computational constraints, variability was evaluated on a single configuration (1 local epoch, 100 rounds, learning rate 10^{-3}), repeated ten times.

D. Evaluation Metrics

We evaluated segmentation model performance using the DICE coefficient, defined as

$$\text{DICE} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN},$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. In our binary segmentation setting, white pixels represent True values, and black pixels False. The DICE score therefore measures the degree of overlap between the predicted and ground-truth masks.

For attacks, we followed standard membership inference metrics, primarily reporting the area under the ROC curve (AUC) and the true positive rate at a fixed false positive rate (TPR@FPR), with FPR set to 0.01. The AUC provides a global assessment of an attack's performance across all possible decision thresholds, while TPR@FPR focuses on a critical operating point where false positives must be tightly controlled. This is particularly relevant for sensitive applications that require stringent control over false positives. We compute the membership inference metrics using the training dataset of the target client as 'member'. Then we either use the test dataset alone as 'non-member' or the concatenation of the test dataset with the other clients dataset. Unless stated otherwise, the later metric is used on the presented results.

VI. Experimental Results

Unless stated otherwise, all results presented come from experiments conducted using a learning rate of 10^{-3} with one local epoch per client.

A. Passive Attacker Scenario

Table I present the results obtained under the passive server setting, where data was distributed equally among clients. Due to the limited dataset size, this configuration could only be evaluated with 3 and 5 clients, reflecting the scenario where for instance several organizations collaborate together in the framework of a military mission involving surveillance drones.

In this setting, FedMIA Loss achieves the best result with an AUC of 65.5% and TPR@FPR of 8.9% at 5 clients. With only 3 clients, all attacks perform poorly, as the reduced number of client limits the attacker's ability to extract meaningful signals. Moreover, FedMIA's statistical test is theoretically valid only with at least 5 clients, though we included the 3-clients case for empirical completeness.

These observations suggest that in a setting featuring few clients and limited overfitting, membership inference attacks remain relatively ineffective. This motivates the investigation of a proactive server strategy, where the server can inject artificial clients to strengthen its ability to model non-member behavior statistically.

TABLE I
AVERAGE ATTACKS RESULTS IN THE PASSIVE SCENARIO BY CLIENT COUNT (100 ITERATIONS, LEARNING RATE 10^{-3}).

Data Distribution	N° Clients	Cosine Similarty		Loss Difference		FedMIA Cosine		FedMIA Loss	
		AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01
Normal	3	55.6%	2.2%	58.3%	2.2%	60.0%	4.3%	**61.8%**	**6.5%**
	5	57.2%	1.8%	62.4%	5.4%	61.8%	7.1%	**65.5%**	**8.9%**

TABLE II
AVERAGE ATTACK RESULTS WITH AVERAGE ON ATTACK RESULTS ON ALL ITERATIONS (100 ITERATIONS, LEARNING RATE 10^{-3}, 10 CLIENTS).

Local Epochs	Cosine Similarty		Loss Difference		FedMIA Cosine		FedMIA Loss	
	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01
1	60.7% (+- 5.2%)	3.6% (+- 4.3%)	63.1% (+- 6.8%)	7.1% (+- 2.9%)	**72.0% (+- 4.0%)**	**10.4% (+- 5.3%)**	**69.4% (+- 4.0%)**	**10.3% (+- 3.1%)**
2	64.1%	7.1%	68.6%	14.3%	68.9%	**18.6%**	**70.5%**	12.9%
4	71.8%	12.9%	74.8%	11.4%	72.5%	**21.4%**	**76.1%**	20.0%

TABLE III
AVERAGE ATTACKS RESULTS IN THE PASSIVE SCENARIO BY CLIENT COUNT (100 ITERATIONS, LEARNING RATE 10^{-3}).

Data Distribution	N° Clients	Cosine Similarty		Loss Difference		FedMIA Cosine		FedMIA Loss	
		AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01
Overlapped	5	**66.9**%	5.7%	58.5%	5.7%	66.6%	**12.9**%	65.8%	8.6%
	10	60.7%	3.6%	63.1%	7.1%	**72.0**%	**10.4**%	69.4%	10.3%
	20	66.0%	17.1%	67.7%	17.1%	**76.8**%	**20.0**%	72.3%	14.3%

B. *Proactive Attacker Scenario*

1) Overall Attacks Effectiveness: The first row of Table II summarizes the results obtained with 10 different seeds to assess the reproducibility on the configuration with 10 clients, 1 local epoch, 100 iterations with a learning rate of 10^{-3}. For this configuration we provide the mean value and the standard deviation.

FedMIA-based (All VS Target) strategy clearly outperforms the 'Server vs Target' strategy for both attack signals. FedMIA Cosine and FedMIA Loss achieve an average AUC of 72% and 69.4% respectively with an average TPR@FPR of 10.4% and 10.3% respectively. In contrast, both the Cosine Similarity and Loss Difference attacks perform poorly, with an average AUC of 60.7% and 63.1% respectively with an average TPR@FPR of 3.6% and 7.1% respectively.

2) Impact of Local Epochs: Table II presents the influence of the number of local training epochs on the effectiveness of membership inference attacks. In practice, increasing the number of local epochs reduce the overall number of iteration needed, and thus reduce the communication cost of federated learning. Until we reach a number of local epoch that cause to much divergence on the local updates, preventing the convergence of the federated learning process.

Overall, we observe that attacks are sensitive to the number of local epochs. This is particularly noticeable for the 'Server vs Target' strategy, namely the Cosine Similarity and the Loss Difference attacks those AUC at 4 local epochs almost reach the FedMIA-based attacks. However, FedMIA-based attacks have a TPR@FPR that increases far above the 'Server vs Target' strategy, reaching more than 20% when the Cosine Similarity and Loss Difference remain below 13%. This result aligns with expectations: with more local updates before aggregation, the model drifts further from the global average, increasing its capacity to memorize training examples and thus making membership inference easier.

3) Influence of Client Count: Table III presents the results of our study on the influence of the number of clients in the proactive scenario. It reveals that increasing the number of shadow clients leads to much stronger attack success. As we can see, this configuration benefits FedMIA-based attacks in particular. With 20 clients, FedMIA Cosine achieves an AUC of 76.8% and a TPR@FPR of 20.0%, indicating a critical privacy breach. This trend aligns with theoretical expectations: as the number of clients grows, the statistical tests gain power, enabling the attacker to more accurately distinguish members from non-members.

Interestingly, in this overlapped setting, gradient-based attacks (FedMIA Cosine) consistently outperform loss-based methods. This setup also reflects a realistic yet concerning scenario: by adding artificial clients, a malicious server could improve its inference power by creating more "non-member" profiles to contrast against targeted clients.

4) Influence of Training Dynamics: We analyzed how model convergence and overfitting affect membership inference success by training the segmentation model with a small learning rate (10^{-4}), 10 clients, and 1 local epoch. This configuration slows down convergence, allowing us to assess attack performance throughout the learning process.

In the left of Figure 7, we display the target client model's performances after each iteration on its training dataset (named target) and the test dataset. It shows that the model starts overfitting on the target dataset at around 250 iterations. In the

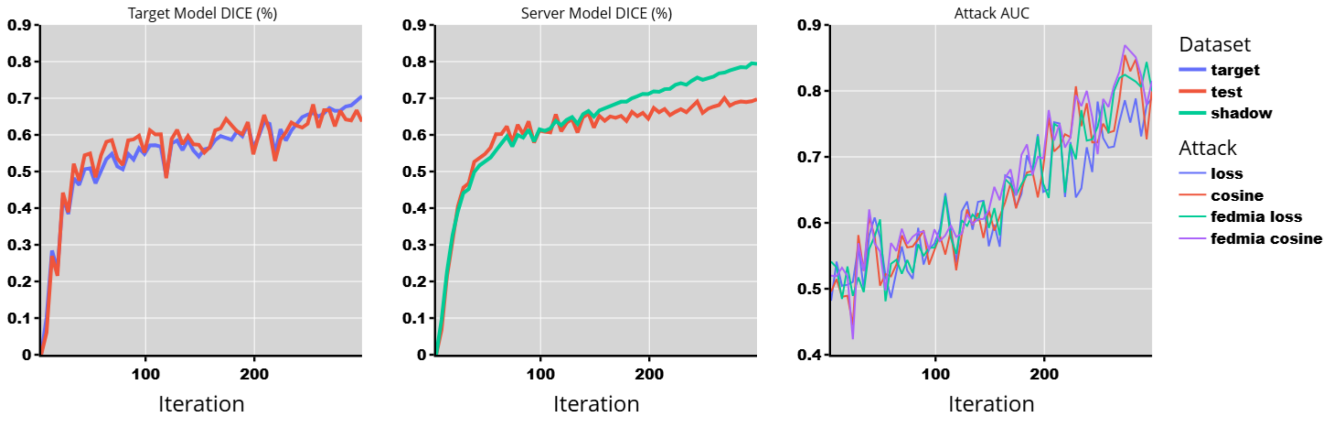

Fig. 7. Target model DICE score (left), server model DICE score (center) and attacks AUC (right) over training iterations (lr=10^{-4}). The 'dataset' legend corresponds to the plot on the left and on the center. The 'attack' legend corresponds to the plot on the right.

TABLE IV
AVERAGE ATTACKS RESULTS BY DATA DISTRIBUTION AND CLIENT COUNT (100 ITERATIONS, LEARNING RATE 10^{-3}). THE ATTACK IS PERFORMED USING AS NON-MEMBER THE TEST DATASET ONLY. THE ATTACK CONSIDER EITHER ALL ITERATIONS OR SELECTED ITERATIONS BASED ON DICE METRICS WITH 0.4 THRESHOLD.

Non member dataset	**Data Distribution**	**N° Clients**	**Cosine Similarty**		**Loss Difference**		**FedMIA Cosine**		**FedMIA Loss**	
			AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01	AUC	TPR@FPR0.01
all iterations	Normal	3	**57.9%**	**6.5%**	53.3%	1.1%	54.3%	4.3%	57.1%	5.4%
		5	55.1%	1.8%	56.4%	0.0%	58.0%	7.1%	**62.2%**	**10.7%**
	Overlapped	5	57.3%	**17.1%**	51.2%	5.7%	54.7%	10.0%	**59.0%**	7.1%
		10	56.0%	5.4%	55.6%	6.0%	**60.5%**	**9.1%**	60.2%	6.1%
		20	61.8%	5.7%	56.8%	12.9%	**66.0%**	**18.6%**	60.9%	14.3%
selected iterations	Normal	3	**62.8%**	3.2%	50.5%	1.1%	55.6%	4.3%	56.7%	**5.4%**
		5	**63.9%**	**17.9%**	52.0%	0.0%	58.8%	12.5%	63.4%	12.5%
	Overlapped	5	**65.3%**	4.3%	50.9%	5.7%	57.4%	**12.9%**	58.8%	5.7%
		10	60.9%	**10.7%**	54.4%	7.9%	**61.2%**	9.4%	**61.2%**	7.9%
		20	**68.9%**	**32.9%**	54.2%	12.9%	64.5%	14.3%	58.8%	17.1%

center of Figure 7, we show the global model's performances after the aggregation at each iteration on the shadow dataset and the test dataset. We observe also that the overfitting on the shadow dataset starts earlier at around 125 iterations. Correspondingly, the right graph of Figure 7 illustrates how all attacks benefit from model convergence. Performance remains weak while the model performances are low. When the model DICE is below 0.4, we observe that all the attacks have performances close to randomness. Then the attacks performances increase progressively, achieving around and above 0.80 in AUC after 300 iterations.

These results confirm a well-established observation: overfitting amplifies membership leakage. Even basic signals become highly predictive when the model starts memorizing training data, reinforcing the importance of carefully managing training dynamics in privacy-sensitive applications. However, if overfitting of the target model amplifies membership leakage, overfitting of the shadow clients created by the malicious server can also impact the attack. Indeed we evaluate the attack performance by its ability to distinguish between target dataset and a concatenation of test dataset and other clients dataset among whose the shadow clients. Table IV shows the attacks results when the member dataset is the training dataset of the target client, and the non-member dataset is the test dataset only. The first half of the table show the results when all iterations are taken into account. We see that the obtained results are much closer to randomness, only FedMIA Cosine achieve an AUC above 0.65 with a TPR@FPR at 18.6%. It means that in the proactive scenario, the malicious server is able to learn an attack very effective in distinguishing the target dataset from the shadow dataset but less effective to distinguish the target from the test dataset. However, in the following we highlight that by focusing on the training dynamics it increase the attacks performances on target VS test dataset.

5) Improved attack based on training dynamics: Based on the previous observation, regarding the correlation of the attack performances and the model segmentation metric performances, the server can enhance the attack by selecting the iterations based on the observed DICE. We use a threshold on the DICE obtained on the test dataset to ensure that the model is far from random. Table IV displays the attacks performances when the server selects only the iterations on which the global model achieves a DICE above or equal 0.4 on the test dataset. We see that the Cosine Similarity attack is dramatically improved by this iteration selection **achieving above 60% of AUC in all data distribution and number of clients settings, and reaching a TPR@FPR of 32.9% with 20 clients**. However, the FedMIA-based strategy is not improved by the iteration selection.

VII. Defenses Recommendations

Numerous defense mechanisms have been proposed to mitigate membership inference attacks in federated learning. Early approaches rely on lightweight techniques such as data augmentation [32], MixUp [33], or gradient sparsification [34]. These methods aim to regularize training and reduce overfitting, thereby limiting the information leakage from model updates. However, they offer limited protection in white-box settings and can often be bypassed by adaptive attackers. Moreover, stronger defenses such as differential privacy [35], [36] inject noise into updates but typically induce a degradation of the model performances to be efficient.

To provide more robust privacy guarantees, cryptographic approaches such as Fully Homomorphic Encryption (FHE) [37], [38] and Secure Multi-Party Computation (SMPC) [39], [40] have been explored [41]. FHE allows each client to encrypt its model updates before sending them to the server, which can then perform aggregation directly in the encrypted domain. This ensures the server never has access to raw parameters. However, FHE remains computationally expensive and does not support non-linear operations natively, making it incompatible with sophisticated aggregation methods. On the other hand, SMPC distributes computation across several non-colluding servers, enabling secure training without exposing individual contributions. While more practical than FHE in certain scenarios, SMPC still incurs communication overhead and requires careful orchestration between parties. Both mechanisms have the notable advantage of preventing not only MIAs but also broader classes of privacy attacks.

In high-stakes applications such as defense, where sensitive imagery and operational data are involved, adopting such strong privacy-preserving mechanisms may become necessary despite their computational cost. Future work should focus on systematically evaluating these defenses to segmentation models training in a federated setting.

VIII. Discussion

We selected a realistic dataset closely resembling a defense scenario; however, its relatively small size, only a few hundreds of instances, posed certain limitations. Conducting our study on a small dataset constrained our ability to thoroughly investigate the effects of varying the number of clients. To ensure local learning remained meaningful, we had to introduce substantial overlap among the datasets assigned to the shadow clients (in scenarios involving a proactive attacker server). This necessity resulted in overfitting on the data utilized by these shadow models, which in turn compromised the effectiveness of FedMIA's strategy, as it relies significantly on the behavior of non-targeted client models. Expanding our research to encompass a larger dataset appears to be an exciting follow-up. Employing non-overlapping datasets for the shadow clients will likely improve the performance of FedMIA approach. Conversely, providing the target client with a more substantial amount of data may reduce the overall success rate of the attack.

Regarding the feasability, the main limitation for a membership inference attack is the access to the target dataset. To determine whether a given data point is a member of a client's training dataset, the server must have access to data highly similar to that client's data. Consequently, such attacks are more relevant in evaluation settings for assessing the privacy risks of federated learning configurations than in practical, real-world scenarios. However, in cases where the server has some prior knowledge about the target client's data distribution, it may be possible to collect sufficiently similar data to enable these attacks. Passive scenarios are particularly feasible in the absence of defense mechanisms, as they require low computational resources. Moreover they do not interfere in the FL process, they have no impact on the learned model. In contrast, proactive scenarios are feasible only for servers with significant computational resources and access to large datasets for training shadow models. Moreover, these attacks impact the learned model since the server would deviate from standard federated learning protocols by aggregating its own models.

Conclusion

We present the first analysis of membership inference attacks on federated binary segmentation models. Our results show that gradient-based attacks (Cosine Similarity and FedMIA Cosine), can effectively exploit training signals, particularly as the number of clients increases or when the server adopts a more proactive strategy by artificially introducing additional clients. Despite the limitation of a small dataset in our experimental settings, we were nonetheless able to effectively demonstrate privacy breaches.

These results underscore the critical need for robust defense mechanisms in federated learning, as, in the absence of such protections, servers are able to extract significant information about individual client datasets.

Finally, extending these attacks to multi-class semantic segmentation with large datasets and empirically assessing the practical cost of deploying defense mechanisms would offer a clearer picture of the trade-offs involved in protecting federated segmentation models.

Acknowledgments

This work has been carried out in the framework of the EDF-STORE project, funded by the European Union through the European Defence Fund (EDF), under Grant Agreement No. 101121405. The authors would like to thank Alix Lafont (Thales CortAIx Labs, France) for her valuable contributions to the illustrations presented in this work.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] Z. L. Teo, L. Jin, N. Liu, S. Li, D. Miao, X. Zhang, W. Y. Ng, T. F. Tan, D. M. Lee, K. J. Chua *et al.*, "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture," *Cell Reports Medicine*, vol. 5, no. 2, 2024.

[3] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[4] J. Rells and W. Joseph, "Federated learning for secure financial transactions," 2025.

[5] O. Stan, R. Sirdey, A. Boudguiga, R. F. Martin Zuber, and K. Hynek, "PRIVILEGE: PRIvacy and Homomorphic Encryption for Artificial IntElliGencE," in *CAID 2021 (Conference on Artificial Intelligence for Defense)*, 2021. [Online]. Available: https://cea.hal.science/cea-04487780v1

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[7] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in neural information processing systems*, vol. 33, pp. 16 937–16 947, 2020.

[8] L. Bai, H. Hu, Q. Ye, H. Li, L. Wang, and J. Xu, "Membership inference attacks and defenses in federated learning: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–35, 2024.

[9] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[10] G. Zhu, D. Li, H. Gu, Y. Yao, L. Fan, and Y. Han, "Fedmia: An effective membership inference attack exploiting" all for one" principle in federated learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 643–20 653.

[11] D. Godavarthi, D. Jose, S. N. Mohanty, M. Medani, M. Kallel, S. Abdullaev, and M. I. Khan, "Federated learning-based semantic segmentation framework for sustainable development," *Egyptian Informatics Journal*, vol. 30, p. 100702, 2025.

[12] Z. Zhang and G. Li, "Uav imagery real-time semantic segmentation with global–local information attention," *Sensors*, vol. 25, no. 6, p. 1786, 2025.

[13] L. Yuan, Z. Wang, L. Sun, P. S. Yu, and C. G. Brinton, "Decentralized federated learning: A survey and perspective," 2024. [Online]. Available: https://arxiv.org/abs/2306.01603

[14] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, vol. 2, no. 2, pp. 15–18, 2016.

[15] A. El Mrini, E. Cyffers, and A. Bellet, "Privacy Attacks in Decentralized Learning," in *ICML*, 2024.

[16] J. Li, N. Li, and B. Ribeiro, "Effective passive membership inference attacks in federated learning against overparameterized models," in *The Eleventh International Conference on Learning Representations*, 2023.

[17] U. Gupta, D. Stripelis, P. K. Lam, P. Thompson, J. L. Ambite, and G. Ver Steeg, "Membership inference attacks on deep regression models for neuroimaging," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 228–251.

[18] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[19] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "Poisongan: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3310–3322, 2020.

[20] A. Pustozerova and R. Mayer, "Information leaks in federated learning," in *Proceedings of the network and distributed system security symposium*, vol. 10, 2020, p. 122.

[21] G. Pichler, M. Romanelli, L. R. Vega, and P. Piantanida, "Perfectly accurate membership inference by a dishonest central server in federated learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 4, pp. 4290–4296, 2023.

[22] T. Nguyen, P. Lai, K. Tran, N. Phan, and M. T. Thai, "Active membership inference attack under local differential privacy in federated learning," *arXiv preprint arXiv:2302.12685*, 2023.

[23] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1102–1107.

[24] A. Suri, P. Kanani, V. J. Marathe, and D. W. Peterson, "Subject membership inference attacks in federated learning," *arXiv preprint arXiv:2206.03317*, 2022.

[25] G. Zhu, D. Li, H. Gu, Y. Han, Y. Yao, L. Fan, and Q. Yang, "Evaluating membership inference attacks and defenses in federated learning," *arXiv e-prints*, pp. arXiv–2402, 2024.

[26] Y. Gu, Y. Bai, and S. Xu, "Cs-mia: Membership inference attack based on prediction confidence series in federated learning," *Journal of Information Security and Applications*, vol. 67, p. 103201, 2022.

[27] L. Zhang, L. Li, X. Li, B. Cai, Y. Gao, R. Dou, and L. Chen, "Efficient membership inference attacks against federated learning via bias differences," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 222–235.

[28] Y. He, S. Rahimian, B. Schiele, and M. Fritz, "Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 519–535.

[29] T. Chobola, D. Usynin, and G. Kaissis, "Membership inference attacks against semantic segmentation models," in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023, pp. 43–53.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[31] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 28–37.

[32] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[34] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.

[35] Q. Zheng, S. Chen, Q. Long, and W. Su, "Federated f-differential privacy," in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 2251–2259.

[36] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE transactions on information forensics and security*, vol. 15, pp. 3454–3469, 2020.

[37] H. Fang and Q. Qian, "Privacy preserving machine learning with homomorphic encryption and federated learning," *Future Internet*, vol. 13, no. 4, p. 94, 2021.

[38] H. Shi, Y. Jiang, H. Yu, Y. Xu, and L. Cui, "Mvfls: multi-participant vertical federated learning based on secret sharing," *The Federate Learning*, pp. 1–9, 2022.

[39] H. Zhu, R. S. M. Goh, and W.-K. Ng, "Privacy-preserving weighted federated learning within the secret sharing framework," *IEEE Access*, vol. 8, pp. 198 275–198 284, 2020.

[40] S. S. Tiwari, G. Dhasmana, H. M. Al-Jawahry, A. Rana, G. Bhardwaj, and A. P. Srivastava, "Federated learning strategies for privacy-preserving machine learning models in cloud computing environments," in *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*. IEEE, 2024, pp. 1457–1462.

[41] O. Stan, V. Thouvenot, A. Boudguiga, K. Kapusta, M. Zuber, and R. Sirdey, "A secure federated learning: analysis of different cryptographic tools," in *SECRYPT 2022-19th International Conference on Security and Cryptography*, vol. 1, 2022, pp. 669–674.