

DECIFR: Domain-Aware Exfiltration of Circuit Information from Federated Gradient Reconstruction

Gijung Lee, Wavid Bowman, Olivia P. Dizon-Paradis, Reiner N. Dizon-Paradis,

Ronald Wilson, Damon L. Woodard, Domenic Forte

Florida Institute of National Security, University of Florida, Gainesville, Florida, USA

Email: {lee.gijung, wavid.bowman, paradiso, reinerdizon, ronaldwilson}@ufl.edu, {dwoodard, dforte}@ece.ufl.edu

Abstract—Federated Learning (FL) is a promising approach for multiparty collaboration as a privacy-preserving technique in hardware assurance, but its security against adversaries with domain-specific knowledge is underexplored. This paper demonstrates a critical vulnerability where available standard cell library layouts (SCLL) can be exploited to compromise the privacy of sensitive integrated circuit (IC) training data. We introduce DECIFR, a novel two-stage Membership Inference Attack (MIA) that requires no auxiliary dataset. The attack employs a guided Gradient Inversion Attack (GIA) to reconstruct a client’s training images from intercepted model updates. Our findings reveal that the fidelity of these reconstructions directly correlates with membership status, allowing an adversary to reliably distinguish members from non-members based on image quality. This work exposes a practical threat that overcomes the limitations of conventional attacks and underscores that standard FL protocols are insufficient for securing domains with extensive knowledge. We conclude that robust defenses are essential for the secure application of FL in hardware assurance.

I. INTRODUCTION

The globalized integrated circuit (IC) supply chain faces serious security threats such as hardware Trojans, IP piracy, IC counterfeiting, and IC overproduction [1], making hardware assurance a national priority as emphasized in the CHIPS Act [2]. Deep learning models offer a powerful tool for automating security analysis, but they are hampered by a critical bottleneck: the slow and high cost acquisition of specialized SEM image data required for training.

Federated Learning (FL) emerges as a promising solution, allowing organizations to collaboratively train models without sharing sensitive raw data. Despite its privacy-preserving design, FL remains vulnerable to attacks such as the membership inference attack (MIA), which determines if a specific data sample was in the training set [3], [4]. However, existing MIAs are often impractical in FL environments, as they require auxiliary datasets that a central server would not possess.

To address this gap, we introduce DECIFR, a novel MIA methodology targeting FL-trained Scanning Electron Microscopy (SEM) image segmentation models used for hardware assurance. Our key insight is that an adversary can leverage standard cell library layouts (SCLLs), which are accessible under NDA, to guide a gradient inversion attack (GIA). By intercepting a client’s FL model update and using SCLLs as dummy labels in the GIA, DECIFR reconstructs the original training image. We hypothesize that an adversary can exploit this reconstruction process, as true training

members will rebuild with markedly higher fidelity. This quality difference allows an attacker to reliably distinguish members from non-members. This work exposes a critical privacy vulnerability in the application of FL to hardware assurance and underscores the nuanced challenges of achieving robust hardware assurance.

Critically, membership inference in this domain poses a systemic threat by revealing not just data presence, but the underlying hardware characteristics (e.g., node or layer type) of the training set. This leakage of hardware-specific membership information acts as a critical enabler for physical attacks, as it provides additional information to accelerates reverse engineering [5], facilitates the identification of design weaknesses [6] to launch non-invasive attacks, and streamlines IP piracy [7]. We make the following contributions:

- **Novel Data-Free Attack:** We propose the first data-free MIA for FL segmentation models, utilizing SCLLs to guide Gradient Inversion Attacks (GIA) without requiring private datasets.
- **Hardware-Specific Inference:** We demonstrate that adversaries can infer specific hardware traits (e.g., layers, technology nodes) by analyzing differential reconstruction success using SCLL priors.
- **Reconstruction-Based Detection:** We validate reconstruction fidelity as an effective metric for distinguishing members from non-members in hardware assurance.

II. BACKGROUND

This section reviews the core concepts of FL and defines MIAs within this specific context.

A. Federated Learning (FL)

Federated Learning (FL) is a distributed framework that enables multiple entities to jointly train a global model while ensuring that their raw data remains local and private [8]. The FL protocol typically executes in iterative rounds: a central server distributes a global model, clients train it locally on private data, and then return only model updates to the server for aggregation [9].

Common protocols include Federated Stochastic Gradient Descent (FedSGD) and Federated Averaging (FedAVG) [10]. While FedSGD transmits gradients directly, FedAVG requires clients to train for multiple epochs and share updated weights. Although FedAVG avoids direct gradient sharing, an adversary can still perform a Gradient Inversion Attack (GIA) by

estimating gradients from the difference between the global model weights before and after local training. This estimation is the vulnerability exploited by our proposed attack.

B. Membership Inference Attacks (MIAs)

A Membership Inference Attack (MIA) is a privacy violation where an adversary determines if a specific data sample was part of a model’s training set. In the context of hardware assurance, this vulnerability extends to the exposure of sensitive Intellectual Property (IP). For instance, consider a model trained on a dataset of Scanning Electron Microscopy (SEM) images. An adversary could probe this model by submitting distinct inputs, such as images of a 32nm metal layer versus a 32nm diffusion layer. If the model predicts the 32nm metal layer with disproportionately high confidence, the attacker can infer that this specific structure was likely included in the training set. Furthermore, by systematically testing various images from a known Standard Cell Library (SCL), an adversary can determine if the model was trained using images from that specific, proprietary Process Design Kit (PDK).

Pioneering research on MIAs proposed the shadow training technique [3], where an attacker trains surrogate models on “shadow datasets” drawn from the same distribution as the victim’s data. In the FL context, MIAs are broadly categorized into update-based and trend-based methods [11].

- **Update-based attacks** leverage the exchanged gradients or parameters. Specifically, these methods operate by analyzing the raw gradients and their temporal differences, or by applying shadow training techniques to approximate the target model’s parameters in the federated environment. Early works like [12] require access to partial member data, while others exploit non-zero gradients in specific layers [13] or calculate gradient differences between rounds [14].
- **Trend-based attacks** infer membership by tracking metrics over the learning process. These techniques rely on analyzing the historical trajectory of indicators, such as prediction confidence or loss, to detect distributional differences between members and non-members. Recent approaches like CS-MIA [15] or data poisoning methods [16] effectively infer membership but often require the adversary to actively tamper with the training process or possess auxiliary datasets.

Crucially, most existing FL attacks assume the attacker has access to auxiliary data similar to the victim’s or can actively poison the model. These assumptions are often impractical in hardware assurance, where high-quality domain data is scarce and proprietary. Our work addresses this gap by proposing a data-free attack that relies only on design knowledge (SCLLs).

III. METHODOLOGY

In this study, we propose DECIFR, a novel MIA technique tailored for the FL environment. Unlike traditional attacks, DECIFR infers the presence of sensitive training data directly from model updates without relying on any auxiliary data from the target victim. Our experimental pipeline begins with a

standard FL protocol where clients train a global segmentation model. As illustrated in Fig. 1, the attack itself launches from the intercepted model updates. In this first attack phase, an adversary performs a guided Gradient Inversion Attack (GIA) using Standard Cell Library Layouts (SCLLs) to reconstruct potential private training images from the updates. This is followed by the Membership Inference phase, where the fidelity of these reconstructions is analyzed to definitively determine if a specific hardware characteristic, such as a technology node or layer type, was present in the client’s training data.

A. Dataset and Environment Setup

To evaluate our approach, we curated two distinct datasets of Standard Cell Library Layouts (SCLLs) derived from the 32nm and 90nm technology nodes within Synopsys’ Open Educational Design Kit (SAED). These datasets specifically isolate metal and diffusion layers. We applied a whitelist filtering process to extract a core set of logic gates (AND, NAND, OR, NOR, XOR, and XNOR) and essential supplementary cells (e.g., inverters, buffers), selecting only those that satisfied specific size constraints.

For the semantic segmentation task, we employed the REFICS¹ tool [17] to generate synthetic dataset of 141 SEM images and their corresponding ground-truth masks. The synthesis parameters were configured with a shot noise level of 20 and a dwelling time of 10 μ sec/pixel to simulate realistic imaging conditions. Pixel intensity distributions were set with background and foreground means of 75 and 135, respectively, and a standard deviation of 20. All images were resized to a resolution of 256×256 pixels. The segmentation model employs a 16-layer U-Net generator architecture (\approx 54M parameters) adapted from Pix2Pix [18]. The encoder utilizes 8 downsampling blocks with filter counts progressively increasing from 64 to 512, while the decoder utilizes 7 upsampling blocks with skip connections to recover spatial resolution. All experiments were executed on a high-performance computing node equipped with an AMD EPYC ROME CPU (32GB RAM) and an NVIDIA B200 GPU (180GB vRAM).

B. Threat Model

We adopt an “honest-but-curious” threat model, where a participant (e.g., the central server) attempts to infer sensitive IP while strictly adhering to the FL protocol. This passive adversary monitors the training process but does not tamper with the model parameters or the aggregation process.

- **Observable Information:** The adversary can access to model updates, such as weights or gradients, but lacks of local hyperparameters like the learning rate η .
- **No Auxiliary Data:** Unlike conventional MIAs, DECIFR operates without requiring a auxiliary dataset from the victim’s distribution.
- **Domain Knowledge:** The adversary possesses a reference library of standard SCLLs, which can be obtained from PDKs.

¹Link: <https://trust-hub.org/#/data/refics>

C. Phase 1: Federated Learning

We simulated a FL environment for image segmentation consisting of a server and two clients, each possessing 50 image-mask pairs. A global U-Net [19] model was trained via the FedAvg algorithm over 200 communication rounds. In each round, clients executed 2 local epochs with a batch size of 10 and a learning rate of 0.01.

D. Phase 2: Gradient Inversion

1) *Gradient Extraction*: The attack begins at a target round (e.g., Round 200) by intercepting the client’s update. Using the global weights (W_{prev}) and the received update (W_{curr}), the adversary derives the aggregated gradients ($\nabla \mathcal{L}$):

$$\nabla \mathcal{L} = \frac{W_{prev} - W_{curr}}{\eta} \quad (1)$$

While we utilize a fixed learning rate (η) for experimental consistency, a sophisticated adversary could refine this parameter via grid search, selecting the η that yields the most plausible reconstruction.

2) *Reconstruction via Guided GIA*: DECIFR boosts the GIA by using SCLLs as fixed dummy labels (y'). The adversary performs optimization twice: once guided by a metal SCLL (y'_{metal}) and once by a diffusion SCLL (y'_{diff}), iteratively updating a dummy image (x') to minimize the composite loss function L_{total} :

$$L_{total} = L_{grad} + \lambda_{tv} L_{tv} - \lambda_{dummy} L_{dummy} \quad (2)$$

The core Gradient Matching Loss (L_{grad}) enforces the dummy to match target gradients using a balanced mix ($\alpha = 0.5$) of Mean Squared Error (MSE) and Cosine Similarity (CS):

$$L_{grad} = \alpha \cdot \text{MSE}(\nabla \mathcal{L}_{dummy}, \nabla \mathcal{L}_{target}) + (1 - \alpha) \cdot \text{CS}(\nabla \mathcal{L}_{dummy}, \nabla \mathcal{L}_{target}) \quad (3)$$

We apply Total Variation regularization (L_{tv}) with $\lambda_{tv} = 0.001$. Uniquely, we introduce a negative Forward Pass Loss ($-L_{dummy}$), which penalizes the optimizer for merely fitting the fixed label. This constraint forces reliance on gradient matching, which is critical for recovering complex geometries like 32nm metal layers.

E. Phase 3: Membership Inference

The final phase determines membership based on reconstruction quality.

- 1) **Post-Processing:** The raw reconstructions (x'_{metal} , x'_{diff}) are converted to binary masks using a pipeline of 5×5 Gaussian blur, Otsu’s thresholding, and morphological opening/closing.
- 2) **Scoring:** A similarity metric (Dice coefficient) is calculated between the binary mask and the corresponding SCLL.
- 3) **Mean Threshold Attack:** We pool all similarity scores to calculate a global mean threshold T . If a reconstruction’s score exceeds T , it is classified as a Member; otherwise, it is a Non-Member.

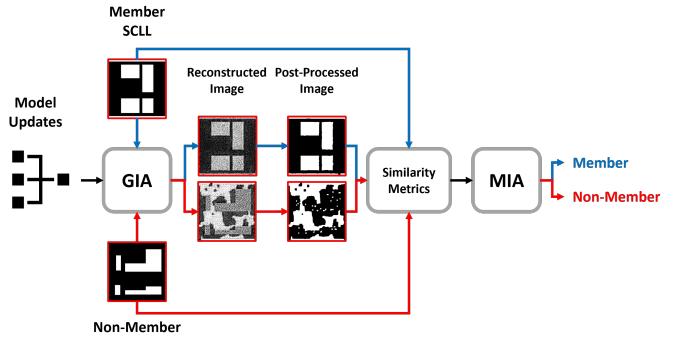


Fig. 1: The DECIFR Process: From intercepted model updates to membership inference via guided gradient inversion.

TABLE I: Results for Inter-Layer Dice Scores (at $\lambda_{dummy} = 0$)

Target Data	Guiding SCLL	Mean Dice Score
Metal	Metal	0.7233
Metal	Diffusion	0.5876
Diffusion	Diffusion	0.8204
Diffusion	Metal	0.7130

IV. RESULTS

This section details the experimental evaluation of FL’s vulnerability to MIAs via DECIFR. We structure our findings by first analyzing the primary inter-layer attack (metal vs. diffusion), followed by an ablation study of the negative Forward Pass Loss, and concluding with the challenging intra-layer scenario (32nm vs. 90nm).

A. Analysis of Reconstructed Images

The initial phase of DECIFR reconstructs the target’s training data through guided GIA. Following our protocol, we generated two distinct reconstructions for every intercepted gradient: one anchored by a metal SCLL (x'_{metal}) and the other by a diffusion SCLL ($x'_{diffusion}$).

We observed a distinct asymmetry in reconstruction fidelity. For metal layer targets, the quality gap was pronounced: the member-guided output (x'_{metal}) achieved high fidelity, whereas the non-member version ($x'_{diffusion}$) was severely degraded. Conversely, for diffusion layer targets, this visual disparity narrowed, as both reconstructions retained relatively high quality. Fig. 2 illustrates representative examples of these phenomena.

B. Distribution of Similarity Scores

To obtain quantitative metrics, we converted the raw reconstructions into binary masks using the post-processing pipeline outlined in Sect. III and computed the Dice similarity coefficient. The results in Table I validate that post-processed reconstruction quality serves as a reliable proxy for membership. As shown in Fig. 3, the mean Dice score was markedly higher when the guiding SCLL matched the specific category of the target’s private data.

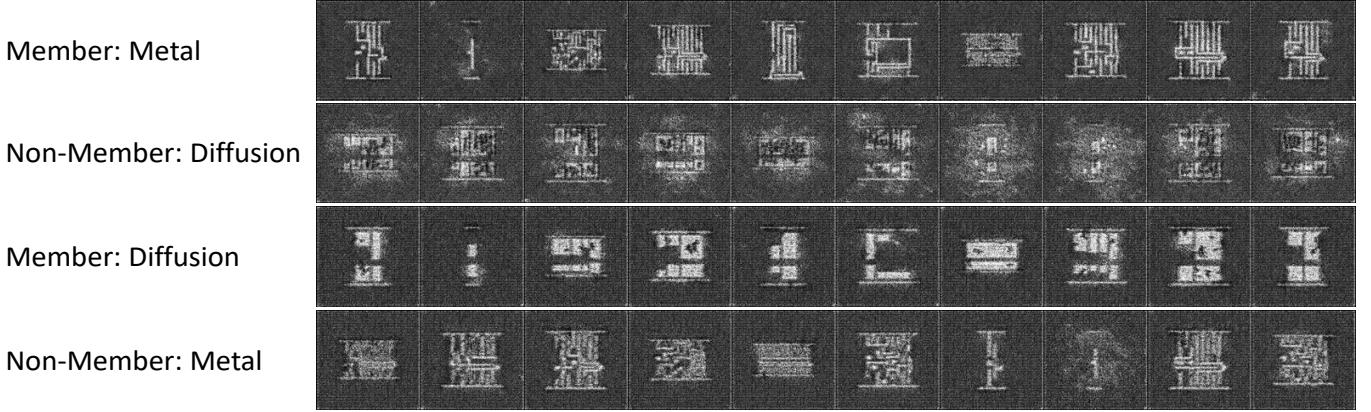
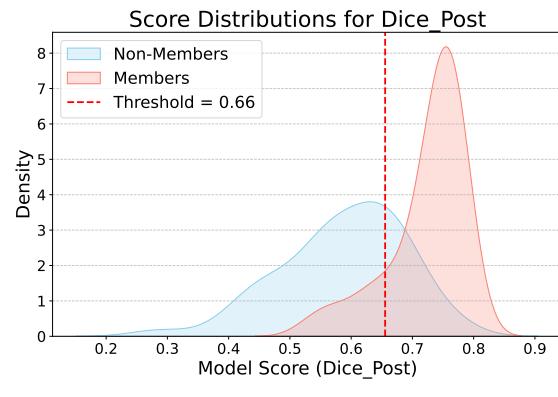
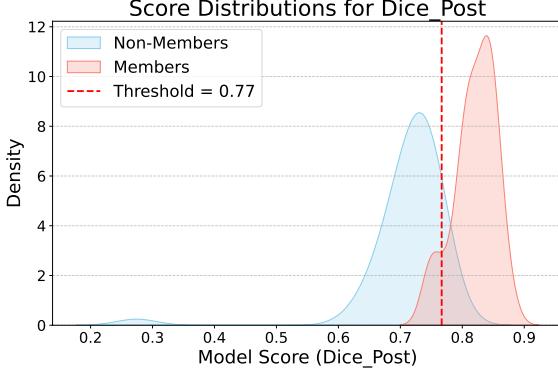


Fig. 2: Representative GIA reconstructions illustrating the pronounced quality disparity in the metal-as-member case compared to the diminished gap in the diffusion-as-member case.



(a) Metal Layer as Member



(b) Diffusion Layer as Member

Fig. 3: Dice score distributions for members versus non-members at $\lambda_{\text{dummy}} = 0$.

C. Membership Inference Performance

We executed the membership inference attack utilizing the pooled similarity scores. The performance metrics are detailed in Table II, with corresponding ROC curves displayed in Fig. 4. Although the attack succeeds in both scenarios, we observe a notable performance asymmetry: the diffusion layer yields a near-perfect AUC of 0.9804, surpassing the metal layer's AUC of 0.8868. This outcome is counter-intuitive, given that the raw reconstructions (Fig. 2) for the metal-as-

TABLE II: Metal vs. Diffusion Layers ($\lambda_{\text{dummy}} = 0$)

Member Data	Accuracy	Precision	Recall	AUC
Metal Layer	78.00%	0.75	0.84	0.8868
Diffusion Layer	94.00%	0.9783	0.90	0.9804

member scenario visually display a more distinct quality gap. This discrepancy is likely attributed to a bottleneck in post-processing, as the intricate structures of the metal layer are highly susceptible to degradation during standardized blurring and thresholding. These operations can erode the fine lines of a high-fidelity metal reconstruction, artificially reducing its Dice score and limiting the threshold-based separation.

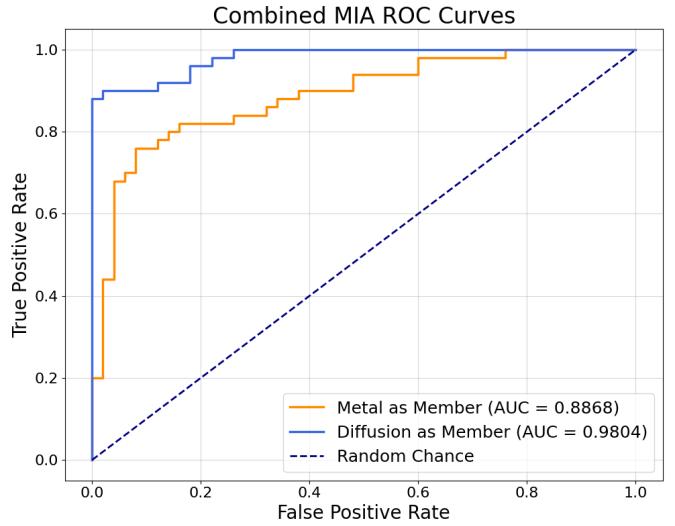


Fig. 4: ROC curves for the MIA on metal vs. diffusion layers.

D. Ablation Study: Impact of L_{dummy}

The negative Forward Pass loss term (L_{dummy}) was introduced as a regularizer to force the optimizer to prioritize gradient matching over fitting the fixed dummy label. Our results (Table III) reveal that this mechanism is critical for structurally complex data.

For the complex metal layer, the standard GIA ($\lambda_{\text{dummy}} = 0$) results in a low AUC of 0.8868 due to the post-processing bot-

TABLE III: Ablation Study on the FP Loss Term (AUC)

λ_{dummy}	Metal Layer (Member)	Diffusion Layer (Member)
0 (Term Removed)	0.8868	0.9804 (Optimal)
1	0.9376	0.9776
3	0.9720	0.9516
5	0.9752 (Optimal)	0.9228

TABLE IV: MIA Performance on Intra-Layer Datasets

Member Data	λ_{dummy}	Accuracy	Precision	Recall	AUC
32nm Diffusion	0	67.00%	0.6308	0.82	0.6688
32nm Diffusion	5 (Optimal)	93.00%	0.8909	0.98	0.9916
90nm Diffusion	0 (Optimal)	69.00%	0.6610	0.78	0.8072
90nm Diffusion	5	64.00%	0.6207	0.72	0.7100

tleneck described above. However, implementing $\lambda_{\text{dummy}} = 5$ expands the measurable quality margin by disproportionately corrupting non-member reconstructions (see Fig. 5). This enhancement raises the AUC to an optimal 0.9752. In contrast, the simpler diffusion layer inherently possesses a clear gradient signal, rendering this additional regularization superfluous. These results strongly demonstrate that the loss term functions primarily to reinforce the authentic gradient signal during complex evaluation scenarios.

E. Intra-Layer Attack Performance

We extended our evaluation to the challenging intra-layer task: distinguishing between 32nm and 90nm diffusion nodes. The results (Table IV) substantiate our bottleneck hypothesis. The intricate 32nm node, similar to the metal layer, necessitates $\lambda_{\text{dummy}} = 5$ to reach a peak AUC of 0.9916. Conversely, the simpler 90nm node achieves optimal performance without this regularization. These findings confirm that the attack’s ideal configuration is deterministic, governed by the interplay between the data’s structural complexity and the constraints of the post-processing pipeline.

V. DISCUSSION

Our research offers a more nuanced perspective on MIAs within Federated Learning. We establish that the attack’s success is highly conditional, driven by the critical interaction between the intrinsic complexity of the private data and the evaluation methodology for attack.

A key contribution of this work is the identification of a post-processing bottleneck within the evaluation pipeline. We discovered that the conversion of raw reconstructions into binary masks for scoring unfairly penalizes complex, fine-grained structures compared to simpler ones. This bottleneck accounts for the initial underperformance of high-quality reconstructions (e.g., 32nm metal layers), as the evaluation metrics failed to accurately reflect their true fidelity.

This insight redefines the role of our loss term, L_{dummy} , as a strategic tool to effectively bridge this evaluation gap. By constraining the optimization to rely strictly on the gradient, the term induces a degradation in the raw quality of both member and non-member outputs. Crucially, this degradation is disproportionately severe for non-members, which lack a valid gradient signal to anchor the reconstruction. Consequently,

this disproportionate corruption widens the measurable quality margin between the classes, facilitating a distinct separation of scores and significantly enhancing attack success.

VI. FUTURE WORKS

Our findings highlight several compelling directions for subsequent research, primarily centering on refining the attack methodology and addressing the challenges found in this study.

- 1) Refining Evaluation Metrics:** The identification of the post-processing pipeline as a performance bottleneck is a key insight. Future work should prioritize the development of advanced evaluation frameworks. Potential solutions include designing adaptive post-processing pipelines that automatically modulate parameters in response to the estimated complexity of the reconstruction.
- 2) Autonomous Hyperparameter Tuning:** We observed that the optimal attack configuration, particularly the λ_{dummy} term, is data-dependent. A significant avenue for future work is the development of auto-tuning mechanisms for this parameter. This would augment the attack’s universality, enabling effective deployment across diverse datasets without the need for manual calibration or prior knowledge.
- 3) Scalability and Real-World Applicability:** While our current experiments utilized a two-client framework, real-world FL deployments are extensive. To better approximate realistic threats, subsequent studies should scale the simulation to involve a larger cohort of clients and investigate the attack’s performance against more complex backbones (e.g., Vision Transformers) beyond the standard CNN-based U-Net used in this study.
- 4) Assessing Countermeasures:** Given the demonstrated potency of the attack, a vital next step is the comprehensive evaluation of defenses. We have categorized applicable privacy techniques in Table V, noting their specific utility against our GIA-based MIA in hardware assurance. Future research should rigorously benchmark these mechanisms, analyzing the privacy-utility trade-offs in perturbation methods like Differential Privacy and Synthetic Data and the computational feasibility of cryptographic protocols such as Homomorphic Encryption and Secure Multiparty Computation.

VII. CONCLUSION

In this paper, we expose a critical privacy vulnerability in Federated Learning, questioning its assumed privacy preservation. We prove that an adversary can successfully achieve higher success rate of MIA without auxiliary data, overcoming even a biased evaluation framework.

A primary contribution of DECIFR is the identification of a significant post-processing bottleneck. We discovered that this limitation, inherent to reconstruction-based attacks, obscures actual leakage by penalizing complex data structures. We resolve this issue by augmenting the attack with a negative Forward Pass loss term (L_{dummy}). This term serves as a conditional amplifier, establishing a clear quality margin by disproportionately degrading non-member reconstructions.

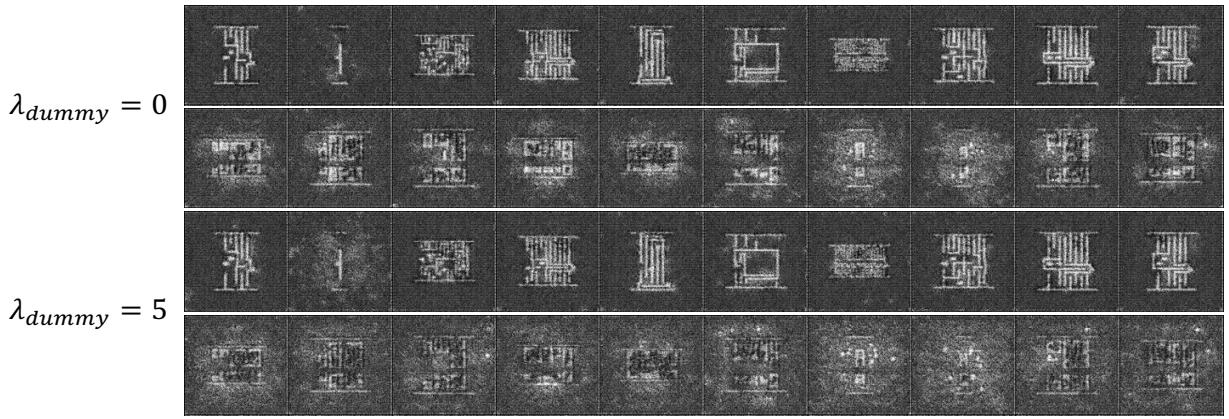


Fig. 5: Visual influence of the L_{dummy} term on metal targets. The quality disparity is visibly amplified at $\lambda_{\text{dummy}} = 5$

Techniques	Benefits	Drawbacks	Examples
Differential Privacy	Prevents high-fidelity reconstruction (GIA) by masking gradient details	Noise degrades segmentation precision for critical layers (e.g., Metal, Poly, Via)	DP-SGD on layout gradients
Synthetic Data	Replaces sensitive IP (layouts) with fake data, rendering reconstruction useless	Generative models may memorize sensitive training samples, leaking original IP	Training on GAN-generated circuit layouts
Secure Multiparty Computation	Aggregates updates securely; prevents access to individual gradients required for GIA	High communication overhead for large layout datasets and models	Secure Aggregation (SecAgg) of weights
Homomorphic Encryption	Gradients remain encrypted during aggregation, blocking adversary access	Prohibitive computational cost for deep CNNs used in vision-based assurance	Fully Homomorphic Encryption (FHE) training
Zero-Knowledge Proof	Verifies local training correctness without revealing the gradient update itself	High computational complexity to generate proofs for neural networks	ZK-SNARKs for update validity

TABLE V: Overview of privacy-preserving techniques against reconstruction MIAs in hardware assurance.

Ultimately, this discovery represents a profound threat to FL-based hardware assurance. By inferring specific hardware metadata (e.g., technology nodes or layers) instead of mere data points, adversaries can expedite physical threats such as reverse engineering, side-channel attacks, and fault injection.

REFERENCES

- [1] M. Rostami, F. Koushanfar, and R. Karri, “A primer on hardware security: Models, methods, and metrics,” *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1283–1295, 2014.
- [2] commerce.senate.gov, “The chips act of 2022,” <https://www.commerce.senate.gov/services/files/592E23A5-B56F-48AE-B4C1-493822686BCB>, [Accessed 06-03-2025].
- [3] 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.
- [4] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, “Knock knock, who’s there? membership inference on aggregate location data,” *arXiv preprint arXiv:1708.06145*, 2017.
- [5] R. Torrance and D. James, “The state-of-the-art in ic reverse engineering,” in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2009, pp. 363–381.
- [6] F. Courbon, “Practical partial hardware reverse engineering analysis: For local fault injection and authenticity verification,” *Journal of Hardware and Systems Security*, vol. 4, no. 1, pp. 1–10, 2020.
- [7] U. Guin, K. Huang, D. DiMase, J. M. Carulli, M. Tehranipoor, and Y. Makris, “Counterfeit integrated circuits: A rising threat in the global semiconductor supply chain,” *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1207–1228, 2014.
- [8] J. Wen, Z. Zhang, Y. Lan, Z. Cuia, J. Cai, and W. Zhang, “A survey on federated learning: challenges and applications,” *International Journal of Machine Learning and Cybernetics*, vol. 14, pp. 513–535, 2023.
- [9] R. Torkzadehmahani, R. Nasirigerdeh, D. B. Blumenthal, T. Kacprowski, M. List, J. Matschinske, J. Spaeth, N. K. Wenke, and J. Baumbach, “Privacy-preserving artificial intelligence techniques in biomedicine,” *Methods of Information in Medicine*, vol. 61, pp. e12–e27, 2022.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 691–706.
- [14] 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.
- [15] 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.
- [16] X. He, Y. Xu, S. Zhang, W. Xu, and J. Yan, “Enhance membership inference attacks in federated learning,” *Computers & Security*, vol. 136, p. 103535, 2024.
- [17] R. Wilson, H. Lu, M. Zhu, D. Forte, and D. L. Woodard, “REFICS: Assimilating data-driven paradigms into reverse engineering and hardware assurance on integrated circuits,” *IEEE Access*, vol. 9, pp. 131955–131976, 2021.
- [18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [19] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.