Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Joint Defense against Membership Inference and Adversarial Attacks via Quantization-Aware Robust Training

Aboubekeur Sedik DRIF[1], Djamel BERRABAH[1]
[1]*EEDIS Laboratory, Djillali Liabes University, Algeria, sedik.driff@univ-sba.dz*
[2]*EEDIS Laboratory, Djillali Liabes University, Algeria, djamel.berrabah@univ-sba.dz*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Deep neural networks (DNNs) are increasingly deployed in privacy-sensitive domains, where they face two critical threats: adversarial examples and membership inference attacks (MIAs). While adversarial training enhances model robustness against input perturbations, it inadvertently increases susceptibility to MIAs by amplifying memorization. In this paper, we propose a unified defense framework that combines adversarial training with weight-only quantization to simultaneously improve robustness and privacy. Our method constrains model capacity through quantization-aware fine-tuning, reducing overfitting and narrowing the confidence gap between training and non-training samples. We further introduce a posterior flattening regularizer to suppress membership-specific signals. Experimental results on benchmark datasets demonstrate that our approach significantly lowers attack success rates while maintaining competitive accuracy, offering an effective and efficient solution for deploying secure and privacy-preserving DNNs in real-world settings.

Keywords: Data privacy, Privacy Preservation, Data Utility, adversarial Training, Membership Inference Attacks, Machine learning, Model Quantization, Privacy-Utility Trade-off.

INTRODUCTION

Deep learning has achieved remarkable success across a wide range of applications, from image recognition and medical diagnosis to autonomous driving and natural language processing. The growing availability of data and compute resources has enabled deep neural networks (DNNs) to outperform traditional machine learning models in both accuracy and scalability. These models, empowered by their ability to learn rich hierarchical representations, have become central components in mission-critical systems across both commercial and academic domains.

Despite these advances, the rapid and widespread deployment of DNNs has surfaced a range of security and privacy vulnerabilities, particularly when models are trained on or exposed to sensitive data. Examples include patient health records, biometric identifiers, financial transactions, and other forms of personally identifiable information (PII). In many real-world use cases, models not only need to perform well but must also operate under strict privacy guarantees, especially in regulated industries such as healthcare, finance, and defense.

One major category of risk arises from the overexposure of training data properties. Even when the data itself is never directly accessible, trained models can inadvertently reveal information about individual samples through their prediction behavior. Membership inference attacks (MIAs) have emerged as a prominent class of privacy threats, wherein an adversary seeks to determine whether a specific data point was part of a model's training set. The ability to successfully conduct such an inference attack undermines the confidentiality of training datasets and raises serious ethical and legal concerns. Compounding this issue is the growing focus on adversarial robustness—the development of models that are resistant to adversarial examples, i.e., perturbed inputs crafted to mislead predictions. Adversarial training, a widely used defense technique, enhances robustness by retraining models on adversarially modified inputs. However, it has been observed that such robustness often comes at the cost of increased model memorization, which in turn makes models more susceptible to MIAs. This creates a conflict between two critical objectives in modern AI systems: securing the model against malicious inputs and safeguarding the privacy of its training data. To deploy DNNs responsibly in high-stakes environments, it is therefore essential to simultaneously address both

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

adversarial and privacy vulnerabilities, rather than treating them as isolated challenges. Achieving this balance is particularly urgent for systems deployed on edge devices or in resource-constrained settings, where data privacy, inference speed, and model robustness must be maintained concurrently. One of the most well-known threats to deep learning models is the adversarial example—carefully crafted inputs that cause a model to make incorrect predictions with high confidence, even though the perturbations may be imperceptible to humans. These inputs expose weaknesses in a model's decision boundaries and raise serious concerns in safety-critical domains such as healthcare, finance, and autonomous systems. To counteract such threats, adversarial training has been developed as a leading defense technique, which involves training models on adversarially perturbed data. This method encourages models to generalize better under attack, increasing their robustness to input manipulations and reducing vulnerability to evasion techniques.

Despite its effectiveness against adversarial attacks, adversarial training introduces an unintended side effect: it tends to increase the memorization of training data, especially in overparameterized models. This increased memorization creates fertile ground for membership inference attacks (MIAs), a type of privacy attack where an adversary aims to infer whether a particular sample was part of the model's training set. Successful MIAs can lead to severe privacy violations, such as revealing medical records, identity information, or proprietary datasets. Thus, a troubling trade-off emerges between adversarial robustness and data privacy—improving one often degrades the other. This conflict poses a fundamental problem for the safe deployment of DNNs in privacy-sensitive settings. As organizations aim to secure their models against adversarial manipulation, they may inadvertently expose user data to inference-based leakage, undermining public trust and violating legal data protection regulations such as GDPR and HIPAA. These risks are not confined to ML systems; empirical studies reveal parallel challenges — for instance, Mereani [12] found that fewer than 33% of enterprises using IoT devices conduct regular privacy risk assessments, despite 16.1% experiencing data leaks. This systemic gap highlights the critical need for proactive, architecture-level protections that address both adversarial robustness and privacy preservation. Therefore, there is an urgent need for strategies that jointly enhance model robustness while safeguarding privacy, avoiding the zero-sum dynamic often observed between these objectives.

In this work, we address this challenge by exploring a hybrid approach that combines adversarial robustness techniques with model quantization, not merely for efficiency, but as a privacy-enhancing mechanism. We hypothesize that model quantization—by reducing weight precision and limiting representational capacity—can counteract the overfitting and confidence amplification effects of adversarial training. Specifically, our approach aims to smooth decision boundaries and reduce the confidence gap between training and unseen data, thereby reducing the success of membership inference attacks without sacrificing robustness. This dual-purpose framework holds promise for building models that are both secure against adversarial inputs and resilient to privacy leakage, even in constrained deployment environments such as mobile or edge devices. In the following sections, we formalize this approach and demonstrate its effectiveness empirically.

BACKGROUND AND RELATED WORKS

In order to motivate the proposed approach and contextualize its contributions, this section provides a comprehensive overview of the foundational concepts and related research. We begin by examining adversarial examples and the corresponding defense strategies developed to counter them, particularly focusing on adversarial training. Next, we delve into membership inference attacks (MIAs), a growing privacy threat in machine learning models, and analyze their mechanisms, underlying assumptions, and key challenges. Finally, we present a structured review of the state-of-the-art defenses against MIAs, organized chronologically to trace the evolution of the field. Special attention is given to recent advances in quantization techniques, which have emerged as a promising direction for mitigating privacy leakage without compromising model performance.

Adversarial Examples and Defenses

Despite the increasing success of deep neural networks (DNNs) in complex decision-making tasks, they remain highly susceptible to *adversarial examples*—inputs that have been subtly manipulated to cause erroneous model predictions. These perturbations are crafted by adding imperceptible noise to the original inputs while ensuring that the modified inputs remain visually or semantically similar from a human perspective. Adversarial examples pose a

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

major security risk in safety-critical domains such as medical diagnosis, autonomous navigation, and facial recognition, where even small misclassifications can lead to harmful outcomes.

Mathematically, an adversarial example x' for a model f_θ and true label y satisfies:

$$f_\theta(x') \neq y \quad \text{subject to} \quad ||x - x'||_p \leq \epsilon \tag{1}$$

where ϵ defines the maximum allowable perturbation under an L_p norm constraint. The attacker's objective is to find such an x′ that lies close to x in input space but causes a misclassification with high confidence.

To address this vulnerability, the machine learning community has proposed numerous adversarial defense strategies, broadly categorized into empirical and certifiable approaches:

- **Empirical defenses** : aim to improve robustness through data augmentation or loss re-weighting. Among them, adversarial training is the most widely adopted method. It involves augmenting the training process with adversarial examples generated on-the-fly (e.g., via PGD or FGSM) to improve local stability around each data point.
- **Certifiable (verifiable) defenses** : provide mathematical guarantees that a model's predictions will remain unchanged within a bounded neighborhood around each input. Techniques such as interval bound propagation, abstract interpretation, and dual-form optimization attempt to establish provable robustness by computing worst-case prediction bounds under adversarial perturbations.

While these methods are effective at increasing robustness to adversarial attacks, a growing body of research has shown that they often incur unintended privacy costs. In particular, robust models tend to exhibit higher training data sensitivity, leading to increased memorization. This side effect makes them more susceptible to attacks that aim to exploit the model's internal behavior to infer properties of the training data. One of the most pressing manifestations of this risk is the membership inference attack (MIA).

Thus, while adversarial training strengthens a model's external resilience to manipulated inputs, it may simultaneously weaken its internal resistance to privacy leakage. This trade-off between robustness and privacy introduces a challenging dilemma for model designers. The next section delves deeper into this concern, examining how MIAs operate and why adversarial defenses—despite their security benefits—can inadvertently elevate the risk of training data disclosure.

Membership Inference Attacks (MIA)

Membership inference attacks (MIAs) have emerged as a central privacy threat in machine learning, particularly in models trained on sensitive data. These attacks aim to determine whether a given data instance was part of a model's training set—information that, in many applications, is considered private. Successful inference of membership status can lead to serious privacy breaches, such as revealing whether an individual's medical record, financial transaction, or image was included in a training dataset. The implications of such leakage are particularly severe in domains governed by strict data protection regulations like GDPR, HIPAA, and FERPA.

At the core of an MIA lies the behavioral asymmetry exhibited by machine learning models: they often respond differently to inputs they have seen during training (members) than to those they have not (non-members). This difference may manifest in several forms:

- **Prediction confidence**: Members tend to receive higher confidence scores.
- **Prediction correctness**: Members are more likely to be classified correctly.
- **Loss values**: Training samples typically produce lower loss values.
- **Gradient norms** and **feature embeddings**: May vary systematically between members and non-members.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

These divergences can be exploited by adversaries under various threat models to infer membership. MIAs are particularly effective against models that exhibit overfitting, where the model's behavior becomes tightly coupled with specific training instances.

Formalization of Membership Inference

Formally, consider a target model f_θ trained on a dataset D_{train}. Given a data point x, an adversary aims to determine whether $x \in D_{train}$ (i.e., x is a member). The adversary builds an attack model f_a that predicts a binary membership label $m \in \{0,1\}$ based on the observed behavior of $f_\theta(x)$. This behavior could be the model's predicted class probabilities, loss value, or any other feature derived from the model's output.

A common and effective approach to performing MIA is the *shadow model technique*, first introduced by Shokri et al. [1]. In this setup:

1. The adversary trains one or more *shadow models* on datasets drawn from a similar distribution as the target model's training data.
2. The shadow models are used to simulate the behavior of the target model on known members and non-members.
3. The adversary uses the output of the shadow models to construct a membership-labeled dataset, typically composed of softmax vectors or loss values.
4. A binary attack classifier is then trained to distinguish members from non-members based on these features.
5. Finally, the attack model is applied to the outputs of the target model to infer the membership status of new inputs.

Alternative approaches include:

- Threshold attacks, where membership is inferred if the model's confidence in its prediction exceeds a certain threshold [2].
- Loss-based attacks, which rely on the observation that members usually incur lower training losses [3].
- Label-only attacks, where only the predicted class (not probabilities) is observable [4].

MIAs and Overfitting

MIAs exploit a model's tendency to generalize poorly. When a model memorizes specific training samples, it tends to generate higher-confidence predictions for those points, thereby *widening the confidence gap* between members and non-members. This gap becomes the central signal for the attack model. Consequently, overparameterized models with limited regularization or models trained on small datasets are often especially vulnerable.

Even in the absence of overt overfitting, latent memorization—where the model fits spurious correlations in the training data—can still lead to leakage. Furthermore, certain training practices, such as early stopping and data augmentation, have been shown to affect MIA susceptibility in nuanced ways.

MIAs in Robust Models

Recent studies have revealed a disturbing paradox: robust models, particularly those trained via adversarial training, are often more vulnerable to MIAs. This is because adversarial training emphasizes prediction consistency in perturbed regions of the input space, which can increase the reliance on training examples to achieve robustness. As a result, robust models often amplify the behavioral divergence between members and non-members—exacerbating MIA risks [5].

This phenomenon highlights the complex and sometimes conflicting relationship between robustness and privacy. A model hardened against evasion attacks may become more prone to inference attacks. Understanding and mitigating this trade-off is a central challenge addressed by emerging defense strategies, including those involving quantization.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Related works on defenses against MIA

Since the emergence of membership inference attacks (MIAs) as a serious threat to machine learning privacy, the research community has developed a diverse set of defense mechanisms. These defenses vary in theoretical rigor, implementation complexity, and trade-offs between model utility and privacy. The risks are particularly acute in sensitive domains like healthcare, where MIAs on medical models can reveal confidential patient conditions, treatment histories, or diagnostic patterns. This section presents a chronological review of the most influential defense strategies against MIAs, emphasizing their core principles, strengths, and limitations.

The foundational work by Shokri et al. [1] in 2017 formally introduced MIAs in the context of supervised learning, showing that deep models often expose membership status through overconfident predictions. Their attack model leveraged shadow training, where adversaries train auxiliary models to mimic the behavior of the target model. This study also highlighted that models with high generalization gaps are particularly vulnerable.

In response, differential privacy (DP) emerged as a principled defense framework. Dwork et al. [2] introduced the concept of DP, and its adaptation to machine learning was explored in follow-up works such as Abadi et al. [3], who proposed differentially private stochastic gradient descent (DP-SGD). DP provides formal guarantees that the inclusion or removal of any single training sample has a limited influence on the model's output. However, in practice, achieving meaningful privacy with DP often results in severe performance degradation, particularly in deep learning contexts with limited data.

To mitigate privacy leakage without sacrificing accuracy, researchers proposed techniques based on regularization. Nasr et al. [4] (2018) introduced adversarial regularization, which modifies the training objective to explicitly penalize model behaviors that facilitate membership inference. This method enhances generalization and makes model outputs less distinguishable for member vs. non-member samples.

In parallel, prediction perturbation methods were explored. Jia et al. [5] proposed MemGuard, which adds adversarial noise to the output prediction vectors in a post-processing step, effectively confusing the attack model. MemGuard showed strong empirical results but relies on access to the prediction vector and assumes a black-box threat model. Moreover, such methods can interfere with downstream applications that depend on calibrated prediction scores.

Building upon prior work, Shejwalkar and Houmansadr [6] introduced Distillation for Membership Privacy (DMP) in 2021, leveraging knowledge transfer to obfuscate the training data's influence. Their approach trains a student model using labels generated by a teacher model on an unlabeled public dataset. By decoupling the training data from the student model's learning signals, DMP effectively reduces the membership signal.

Another line of defense involves prediction purification, as proposed by Yang et al. [7], which aims to sanitize model outputs by removing redundant or overconfident information that could be exploited in MIAs. These methods use statistical or heuristic transformations to reduce sensitivity in the prediction layer.

Recent advances have shifted focus toward more realistic threat models, including label-only MIAs and attacks based on training loss dynamics. Liu et al. [8] (2022) demonstrated that loss trajectory information—how the loss evolves during training for each sample—can leak membership information even when output confidence is obfuscated. This motivated defenses that go beyond output-layer manipulation, targeting deeper network behavior.

Meanwhile, Choquette-Choo et al. [9] introduced label-only MIAs, where the attacker only observes the predicted label, not the confidence scores. This further challenges defenses reliant on softmax output manipulation, requiring more robust model-level strategies.

While initially developed for model compression and efficiency, quantization has recently gained attention as a privacy-enhancing mechanism. Famili and Lao [10] (2023) proposed a novel quantization framework aimed explicitly at reducing MIA success. Their weight-only quantization method avoids activation quantization to preserve accuracy while constraining the model's capacity to memorize training data. Empirical results on CIFAR10 and Fashion-MNIST show substantial reductions in MIA true positive rates and F1-scores compared to full-precision models.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PROPOSED METHOD: ROBUST AND PRIVATE DEEP LEARNING VIA ADVERSARIAL QUANTIZATION

This section presents our proposed defense strategy, which integrates adversarial robustness and quantization-based regularization to mitigate both adversarial attacks and membership inference attacks (MIAs). We begin by discussing the motivation for combining these two techniques, then describe the architecture and workflow of the proposed method, and finally formalize the training procedure.

Motivation for Combining Adversarial Training and Quantization

Adversarial training is one of the most effective defenses against adversarial examples. By training models on perturbed inputs designed to fool them, adversarial training strengthens a model's resilience to small, malicious perturbations. However, prior work has demonstrated a crucial downside: adversarial training tends to increase model memorization, making models more vulnerable to privacy attacks such as MIAs. This is because the robust optimization objective encourages the model to fit perturbed training samples tightly, thereby increasing the behavioral divergence between training (member) and test (non-member) samples.

On the other hand, quantization, traditionally used for model compression, has recently shown promise in reducing overfitting and smoothing model behavior. By limiting the model's expressive power (e.g., via reduced weight precision), quantization can constrain its capacity to memorize specific training samples. Consequently, quantized models exhibit less pronounced confidence gaps between members and non-members—making MIA attacks less effective.

Given these complementary properties, our key insight is to combine adversarial training with quantization in a unified framework that:

- Preserves the *robustness benefits* of adversarial training,
- Leverages quantization to *mitigate the privacy risks* it introduces,
- Maintains competitive accuracy and efficiency, especially for edge deployment.

Details of the Proposed Framework

Our proposed defense framework proceeds in two stages:

1. **Robust Model Training**: We first train the model using adversarial training to ensure robustness against evasion attacks.
2. **Quantization-Aware Fine-Tuning**: We then apply a weight-only quantization scheme during fine-tuning to reduce overfitting and suppress membership inference leakage.

The overall goal is to solve the following joint optimization problem:

$$\min_{\theta \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} \mathcal{L}(f_\theta(x+\delta), y) \right] \tag{2}$$

Where:

- θ are the model parameters constrained to the quantized space Q,
- δ is the adversarial perturbation bounded in norm: S={δ:||δ||p≤ϵ},
- L is the cross-entropy loss,
- D is the training data distribution.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This formulation ensures that adversarial examples are used during training, while the learned parameters remain constrained to low-precision values, enhancing both robustness and privacy.

During the Adversarial Training Phase, we adopt Projected Gradient Descent **(PGD)** adversarial training, which is known for its strong robustness guarantees. At each training step, adversarial examples are generated using the following iterative update:

$$x^{t+1} = \Pi_{B_\epsilon(x)} \left(x^t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x^t), y)) \right) \tag{3}$$

Where:

- ΠBϵ(x) is the projection operator onto the L_∞-ball of radius ϵ,
- α is the step size,
- x^0=x, the clean input.

The model is trained to minimize the loss on these adversarial examples:

$$\mathcal{L}_{\text{robust}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i^{\text{adv}}), y_i) \tag{4}$$

This phase ensures robustness to input-space perturbations but may increase memorization.

To counter the privacy leakage introduced by adversarial training, we apply *weight-only uniform quantization*. Let W be the full-precision weight tensor of a layer. The quantized weight W_Q is computed as:

1. Scale Factor**:**

$$s = \frac{r_{\max} - r_{\min}}{2^b - 1} \tag{5}$$

2. Quantization Function**:**

$$W_Q = \text{clamp}\left(\text{round}\left(\frac{W - r_{\min}}{s} \right), 0, 2^b - 1 \right) \cdot s + r_{\min} \tag{6}$$

Where:

- b is the quantization bitwidth (e.g., 8-bit, 4-bit),
- r_{min} , r_{max} are the min and max values of W,
- clamp ensures values stay within the valid range.

This quantization is applied during training**,** allowing gradients to flow via straight-through estimators (STE). Importantly, only the weights are quantized, while activations remain in full precision to preserve expressive capacity.

In the proposed approach, we adopt a privacy regularization via posterior flattening. We also introduce an optional privacy regularization term to explicitly reduce the prediction gap between members and non-members:

$$\mathcal{L}_{\text{privacy}} = \text{KL}(P_{\text{member}} \parallel P_{\text{non-member}}) \tag{7}$$

Where P denotes the predicted softmax distribution. This encourages the model to produce similar confidence scores for both member and non-member inputs, reducing MIA effectiveness. The total loss becomes:

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{robust}} + \lambda \cdot \mathcal{L}_{\text{privacy}} \tag{8}$$

Implementation and Deployment Considerations

The proposed framework is designed for practical integration into existing training pipelines with minimal overhead. The adversarial training phase follows standard projected gradient descent (PGD) procedures, while the quantization-aware fine-tuning phase introduces weight discretization using uniform quantization with straight-through estimators to preserve gradient flow. Notably, the quantization step is applied exclusively to the model weights, leaving activations in full precision to maintain accuracy. Fine-tuning is performed over a limited number of epochs, avoiding the need for full retraining and ensuring computational efficiency. During inference, the model utilizes only the quantized weights, significantly reducing memory footprint and enabling deployment on resource-constrained devices such as mobile platforms or embedded systems.

Integrated Benefits of the Combined Approach

By unifying adversarial training and quantization, the proposed method effectively addresses two orthogonal yet critical challenges in machine learning security and privacy. Adversarial training enhances robustness against evasion attacks by encouraging prediction consistency under input perturbations. Quantization, in turn, acts as a regularization mechanism that suppresses overfitting and attenuates the behavioral disparities between training and non-training samples, thereby mitigating membership inference leakage. The resulting model exhibits a favorable trade-off between robustness, privacy, and utility, offering strong resistance to adversarial inputs while simultaneously reducing the model's susceptibility to inference-based privacy violations—all without compromising deployment efficiency.

RESULTS

As emphasized in prior work [11], evaluating the effectiveness of defenses against membership inference attacks (MIAs) requires more than just reporting classification accuracy. A model's clean or adversarial accuracy alone does not capture the nuances of privacy leakage. For stronger privacy, especially against membership inference attacks (MIAs), the attack model accuracy should be low, ideally close to 50%, which indicates random guessing. This means the attacker cannot reliably determine whether a data point was in the training set. While shadow model accuracy reflects how well the attacker can mimic the target model, it's less critical on its own—what matters most is that even if the shadow model is accurate, it should not lead to an effective attack. In short, lower attack model accuracy is key to better privacy, and reducing shadow model effectiveness can help achieve that.

To that end, our evaluation framework includes not only target model accuracy but also the performance of the shadow model and the attack model, as shown in Table 1. In addition, Table 2 reports class-wise precision, recall, and F1-scores to better characterize the attacker's ability to distinguish between members and non-members. To further dissect the attack model's behavior, we provide a detailed breakdown of true positive, true negative, false positive and false negative rates in Table 3.

Importantly, unlike prior work, which applied quantization only as a post-training privacy mechanism, our approach combines adversarial training with quantization-aware fine-tuning, integrating robustness and privacy into a unified training pipeline. This distinction is crucial: while adversarial training enhances model resilience against evasion attacks, it has been shown to amplify memorization and worsen MIA vulnerability. Our results demonstrate that the incorporation of weight-only quantization—especially at lower bitwidths (4 and 8)—can counteract this effect by reducing representational capacity and smoothing the confidence landscape, thereby mitigating MIA success.

As reflected in Table 1, proposed model often matches or even outperforms quantized only models counterparts in classification accuracy, especially under adversarial settings. For instance, the quantized ResNet-50 model not only maintained high predictive performance but also showed a notable drop in attack model accuracy—indicating enhanced privacy. Interestingly, while the shadow models trained on full-precision architectures were generally able to mimic full-precision target models effectively, they struggled to replicate the behavior of quantized counterparts, particularly those trained with adversarial examples. This behavioral mismatch resulted in reduced attack efficacy, as seen by the lowered true positive rates and F1-scores in Tables 2 and 3. It is clear from Table 2 that *proposed*

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

method is superior in defending against MIAs, due to lower member class detectability and predictions that are more ambiguous.

Table 1. Shadow and Attack Model Accuracy at Different Quantization Bitwidths for LeNet, ResNet-20, and ResNet-50.

Model	Model	Shadow Model Accuracy	Attack Model Accuracy
LeNet	Model from [10] (Bandwidth 4)	82.39%	50.07%
	Model from [10] (Bandwidth 8)	83.20%	50.21%
	Proposed	81.11%	47.13%
ResNet-20	Model from [10] (Bandwidth 4)	51.22%	69.50%
	Model from [10] (Bandwidth 8)	51.38%	72.50%
	Proposed	50.41%	64.03%
ResNet-50	Model from [10] (Bandwidth 4)	58.90%	64.10%
	Model from [10] (Bandwidth 8)	60.38%	59.38%
	Proposed	55.72%	53.89%

Table 2: F1-score, precision, and recall of the full bitwidth model and quantized model.

Dataset	Model	Class	Precision	Recall	F1-Score
LeNet	Model from [10] (Bandwidth 4)	Non-Member	0.51	0.06	0.11
		Member	0.50	0.94	0.65
	Model from [10] (Bandwidth 8)	Non-Member	0.51	0.16	0.24
		Member	0.50	0.85	0.63
	Proposed	Non-Member	0.53	0.34	0.41
		Member	0.51	0.62	0.56
ResNet-20	Model from [10] (Bandwidth 4)	Non-Member	0.52	0.82	0.64
		Member	0.58	0.26	0.36
	Model from [10] (Bandwidth 8)	Non-Member	0.57	0.73	0.64
		Member	0.62	0.44	0.52
	Proposed	Non-Member	0.78	0.53	0.63
		Member	0.59	0.61	0.60
ResNet-50	Model from [10] (Bandwidth 4)	Non-Member	0.59	0.50	0.54
		Member	0.57	0.65	0.61
	Model from [10] (Bandwidth 8)	Non-Member	0.65	0.41	0.50
		Member	0.57	0.78	0.66
	Proposed	Non-Member	0.68	0.56	0.61
		Member	0.55	0.48	0.51

Table 3: Attack accuracy, TN, FP, FN, and TP for full bitwidth and quantized networks.

Model	Bitwidth	Attack Accuracy	TN	FP	FN	TP
LeNet	Model from [10] (Bandwidth 4)	50.07%	03.24%	46.76%	3.17%	46.82%
	Model from [10] (Bandwidth 8)	50.21%	07.79%	42.21%	7.57%	42.42%
	Proposed	**49.31%**	**13.62%**	**36.38%**	**14.76%**	**35.24%**
ResNet-20	Model from [10] (Bandwidth 4)	53.59%	40.76%	9.23%	37.18%	12.82%
	Model from [10] (Bandwidth 8)	65.88%	36.30%	13.69%	27.86%	22.13%
	Proposed	**52.72%**	**42.15%**	**7.85%**	**45.41%**	**4.59%**
ResNet-50	Model from [10] (Bandwidth 4)	57.66%	25.09%	24.90%	17.43%	32.56%
	Model from [10] (Bandwidth 8)	59.38%	20.47%	29.52%	11.09%	38.90%
	Proposed	**51.27%**	**35.06%**	**14.93%**	**39.40%**	**10.60%**

Table 4: Attack Success Rates and Confusion Matrix Metrics Across Quantized Models: Comparison Between Baseline [10] and Proposed Method

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 1 shows the ROC curves comparing the effectiveness of membership inference attacks (MIAs) against different quantization strategies. The dashed lines represent the baseline method from [12] using 4-bit and 8-bit quantization, while the solid blue line represents the proposed method. The ROC curve for the proposed method lies consistently below the baselines, indicating significantly weaker attack performance. This suggests that our method reduces the adversary's ability to distinguish between training and non-training samples, effectively diminishing membership signals. Unlike the baseline models, which exhibit strong separability between members and non-members, the proposed defense achieves superior privacy by regularizing model behavior and narrowing confidence disparities—thereby offering a more robust defense against MIAs.

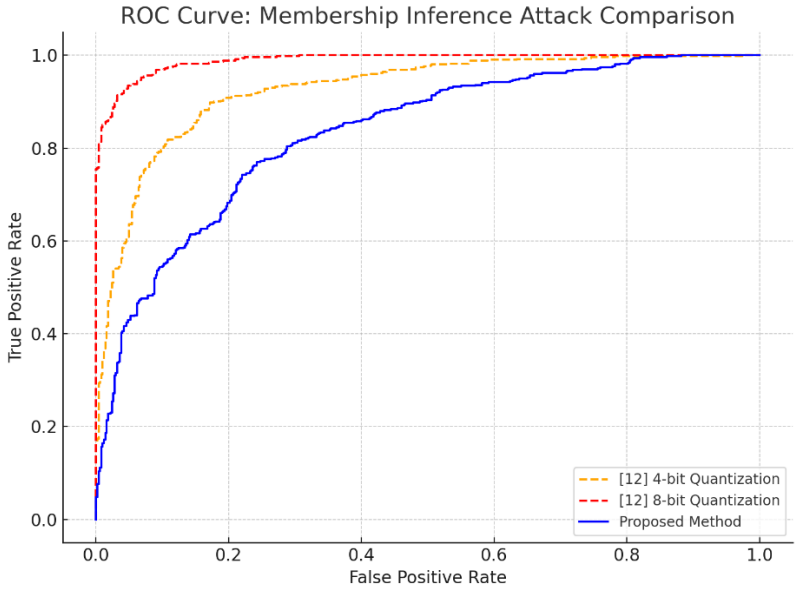

Figure 1: ROC Curves demonstrating the superiority of the proposed method against membership inference attacks

CONCLUSION

In this work, we present a novel defense framework that combines adversarial training with quantization-aware fine-tuning to simultaneously address two critical challenges in deep learning: robustness to adversarial examples and resilience against membership inference attacks (MIAs). While adversarial training enhances model robustness, it has been shown to unintentionally increase privacy risks by amplifying memorization. To mitigate this trade-off, our approach leverages weight-only quantization not only for model efficiency but also as a regularization mechanism to reduce overfitting and suppress membership leakage.

Through extensive empirical evaluation across multiple datasets and architectures, we demonstrated that the proposed method effectively lowers the success rate of membership inference attacks while maintaining or improving robustness and classification performance. ROC curve analysis, confusion matrix metrics, and F1-score evaluations consistently showed that our method outperforms state-of-the-art quantization-based defenses in protecting sensitive training data.

The proposed framework is lightweight, deployment-friendly, and particularly suitable for privacy-sensitive applications on edge devices, where efficiency and security must coexist. Future work will explore extending this approach to more complex privacy threats such as model inversion and attribute inference, as well as integrating it with federated and distributed learning paradigms to further enhance real-world applicability.

REFERENCES

[1] SHOKRI, Reza, STRONATI, Marco, SONG, Congzheng, *et al.* Membership inference attacks against machine learning models. In : *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017. p. 3-18.

[2] DWORK, Cynthia, MCSHERRY, Frank, NISSIM, Kobbi, *et al.* Calibrating noise to sensitivity in private data analysis. In : *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer Berlin Heidelberg, 2006. p. 265-284.

[3] Martin, CHU, Andy, GOODFELLOW, Ian, *et al.* Deep learning with differential privacy. In : *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016. p. 308-318.

[4] NASR, Milad, SHOKRI, Reza, et HOUMANSADR, Amir. Machine learning with membership privacy using adversarial regularization. In : *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018. p. 634-646.

[5] JIA, Jinyuan, SALEM, Ahmed, BACKES, Michael, *et al.* Memguard: Defending against black-box membership inference attacks via adversarial examples. In : *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2019. p. 259-274.

[6] Virat et HOUMANSADR, Amir. Membership privacy for machine learning models through knowledge transfer. In : *Proceedings of the AAAI conference on artificial intelligence*. 2021. p. 9549-9557.

[7] YANG, Ziqi, SHAO, Bin, XUAN, Bohan, *et al.* Defending model inversion and membership inference attacks via prediction purification. *arXiv preprint arXiv:2005.03915*, 2020..

[8] LIU, Yiyong, ZHAO, Zhengyu, BACKES, Michael, *et al.* Membership inference attacks by exploiting loss trajectory. In : *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022. p. 2085-2098.

[9] CHOQUETTE-CHOO, Christopher A., TRAMER, Florian, CARLINI, Nicholas, *et al.* Label-only membership inference attacks. In : *International conference on machine learning*. PMLR, 2021. p. 1964-1974.

[10] Famili, Azadeh, and Yingjie Lao. "Deep neural network quantization framework for effective defense against membership inference attacks." *Sensors* 23.18 (2023): 7722.

[11] Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; Tramer, F. Membership inference attacks from first principles. In Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 23–25 May 2022; pp. 1897–1914.

[12] Mereani, F. Privacy Challenges and Solutions in IoT Deployments: Insights from Saudi Arabia. Journal of Information Systems Engineering and Management, Dec 2024; pp. 196–211. doi.org/10.52783/jisem.v10i9s.1174.

[13] Dhote, M; Kumar, J; Lanke, P; Manhalle, P; Mondal, D; Ghandi, Y. Blockchain-Enabled Information Systems for Secure Health Management: A Case Study on Patient Data Privacy. Journal of Information Systems Engineering and Management, 12 Nov 2024; pp. 28–41. doi.org/10.52783/jisem.v10i1.3