EFFICIENT ADVERSARIAL AND PRIVACY PROTECTIONS FOR HETEROGENOUS DATA

Anurag Josyula

Abstract- ***Machine learning models have shown remarkable performance across a range of applications, yet remain susceptible to both security and privacy vulnerabilities. Adversarial examples—crafted through small perturbations—can significantly degrade model accuracy, posing serious security concerns. We evaluate the impact of Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks on models trained on MNIST and CIFAR-10 and employ PGD-based adversarial training as a defense to enhance robustness. On the privacy side, we investigate Membership Inference Attacks (MIAs), including Shadow Model attacks and Relaxed MIAs (RMIAs), which aim to infer whether a given sample was part of the model's training set. Unlike traditional privacy defenses, we explore whether adversarial training can also reduce membership inference success. Experimental results indicate that adversarial training not only improves resilience to adversarial inputs but may also reduce confidence-based membership leakage, highlighting its dual role in enhancing both model security and privacy.***

Keywords- ***FGSM, PGD, MIA, RMIA, Shadow Model, Adversarial Training, Adversarial Attacks.***

I. INTRODUCTION

Deep learning models have shown great promise in fields like autonomous systems, computer vision, and healthcare. They are still vulnerable, however, to adversarial examples, which are minor input perturbations that can result in inaccurate predictions as shown in **Fig [1]**, without causing noticeable changes to human observers. Such flaws pose significant security risks in applications that depend on safety, where improper model behavior could have severe consequences. Sensitive information about their training data may be leaked by deep learning models, which is another security risk. There are serious privacy concerns with Membership Inference Attacks (MIA), which use a model's output behavior to determine whether data points were used during training.

In this work, we discuss security and privacy flaws in machine learning algorithms. We evaluate the impact of FGSM and PGD adversarial attacks on models trained on the MNIST and CIFAR-10 datasets, and we propose a defense based on PGD-based adversarial training. On the privacy front, we investigate Shadow Model attacks and Relaxed Membership Inference attacks (RMIA) and protect against them by implementing the same defense using during the evaluation of security attacks during model training.

Our results show that PGD adversarial training improves model robustness against attacks, whereas on the privacy side it slightly reduces the success rate of membership inference attacks, highlighting the importance of robust security and privacy defenses in machine learning systems.

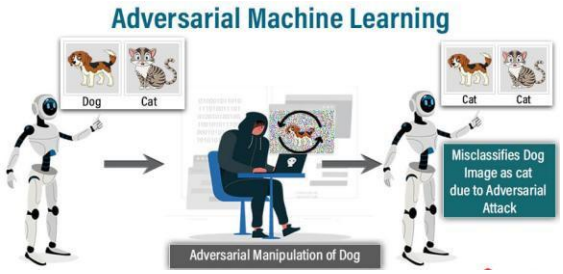

Fig 1: Adversarial Attack Example

II. THREAT MODEL

A. Security

In our security evaluation, we adopt a white-box threat model, assuming the adversary has complete access to the model architecture, parameters, loss function, and gradients. The adversary's objective is to generate adversarial examples that cause misclassification while keeping perturbations imperceptible to human observers.

We focus on two widely used adversarial attack methods:

1) Fast Gradient Sign Method (FGSM): A single-step gradient-based attack that perturbs inputs in the direction of the sign of the gradient.

2) Projected Gradient Descent (PGD): A multi-step iterative attack that applies small perturbations in successive steps, projecting them back into an allowed perturbation region after each iteration.

To explore the sensitivity of models to adversarial perturbations, we plan to apply two variations of FGSM and two variations of PGD by altering the attack parameters such as perturbation magnitude **"(ε)"** and number of **"steps."** This experimental design allows us to study how changes in attack strength affect the model's robustness both before and after adversarial training.

We assume that stronger attacks—characterized by higher perturbation bounds or more iterative refinement—will lead to greater degradation in model performance. Through this systematic variation, we aim to observe whether adversarial.

training improves robustness consistently across different attack strengths and datasets.

Throughout all attacks, we maintain the constraint that adversarial examples must be visually indistinguishable from their clean counterparts and that the perturbations must respect the valid input range (e.g., pixel values between 0 and 1).

B. Privacy

Membership Inference Attacks (MIAs) are a serious privacy issue in machine learning in which an adversary tries to establish whether a certain data sample was part of the model's training dataset as shown in **Fig [2]**. Such attacks take advantage of machine learning algorithms' inclination to be more confident in samples observed during training and less confident in unseen data. This conduct provides a vulnerability that may reveal sensitive information, breaching the data protection requirements such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

We use a black-box threat model to assess privacy vulnerabilities using Membership Inference Attacks (MIAs). The adversary does not have access to the model's core parameters, architecture, or training data, but they can submit input questions and view the model's projected class probabilities. This reflects real-world deployment circumstances, such as public machine learning APIs or online inference systems.

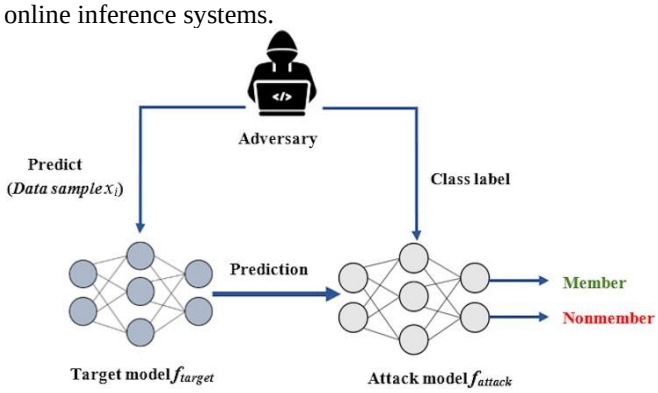

Fig 2: MIA Attack

The attacker is assumed to have some understanding of the data domain (e.g., image classification of CIFAR-10) but not the exact training data. To replicate the target model's behavior, the attacker trains numerous shadow models on different portions of the dataset derived from the same distribution. These shadow models are trained on disjoint subsets of data labeled as member and non-member sets, allowing the attacker to obtain SoftMax outputs from both classes.

Two types of shadow-based attacks are then constructed:

1. **Confidence Based Shadow Attack:** Membership is inferred when a sample's highest SoftMax confidence exceeds a specified threshold.
2. Margin-Based Shadow Attack: Membership is inferred when the difference between the top two predicted SoftMax scores exceeds a predetermined threshold.

Additionally, we implement Relaxed Membership Inference Attacks (RMIAs) by training random forest classifiers (with 100 and 200 trees) on the SoftMax output vectors of the shadow models to distinguish between member and non-member samples in a learning-based manner.

III. BACKGROUND

A. Security

Adversarial examples are inputs that have been purposely perturbed in a subtle way so that machine learning models misclassify them **[3]**. were the first to point out this vulnerability, demonstrating that even minor changes to a model's output might have significant consequences. Following that, Goodfellow et al. introduced the Fast Gradient Sign Method (FGSM), a one-step gradient-based attack that is computationally efficient and demonstrates how models' overfit specific decision limits.

While FGSM defined the concept of adversarial attacks, it was later demonstrated to be insufficient against more powerful defenses. To solve this, **[1]**. devised a more resilient attack approach called Projected Gradient Descent (PGD). PGD is a multi-step attack that combines FGSM with tiny perturbations and projection back into a $\ell\infty$-bounded region. PGD has subsequently been recognized as one of the most powerful first-order adversarial attacks.

Additionally, Madry et al. proposed adversarial training with PGD. A min-max optimization framework in which the model is trained on worst-case adversarial examples generated during each iteration. This method significantly improves resilience and is regarded as a common benchmark for evaluating defense strategies.

In this work, we use Madry's PGD adversarial training architecture as our major defense technique. We also use FGSM and PGD attacks with different strengths to evaluate the vulnerability of baseline models and the success rate of adversarial training on various datasets.

B. Privacy

Membership Inference Attacks (MIAs) target the privacy of machine learning models by attempting to determine whether a specific data sample was included in the training set. These attacks exploit the observation that models often yield higher confidence on training data than on unseen samples. Early works by **[2]**. introduced the shadow model framework, where an attacker trains a mimic model to learn the confidence- based patterns of membership.

Two common MIA strategies are considered in this work. Threshold-based attacks infer membership if the model's output confidence exceeds a certain value, while Relaxed MIAs (RMIAs) train classifiers (e.g., random forests) on prediction outputs to distinguish member from non-member behavior.

As a defense, we observe and evaluate how adversarial training works on privacy attacks.

IV. METHODOLOGY

A. Security

To evaluate the robustness of deep learning models, we applied adversarial attacks and defenses to two datasets: MNIST and CIFAR-10. We used two common attack methods.

methods: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Each attack was evaluated multiple times with varying perturbation magnitudes and step sizes to see how increasing attack strength affects model performance.

For defense, we used adversarial training, which involved retraining models using PGD-generated adversarial examples included in training batches. This strategy tries to improve the model's resistance by subjecting it to adversarial behavior while training. We trained clean and defense models individually for each dataset. MNIST and CIFAR-10 models were trained for 15 and 25 epochs, respectively, using a simple CNN and Wide-ResNet as mentioned in **Fig [3]** Training was performed using NVIDIA RTX 4060 GPU (CUDA accelerated) with 32GB RAM.

We compared clean, attack, and defense performances using metrics such as accuracy, precision, recall, and F1-score.

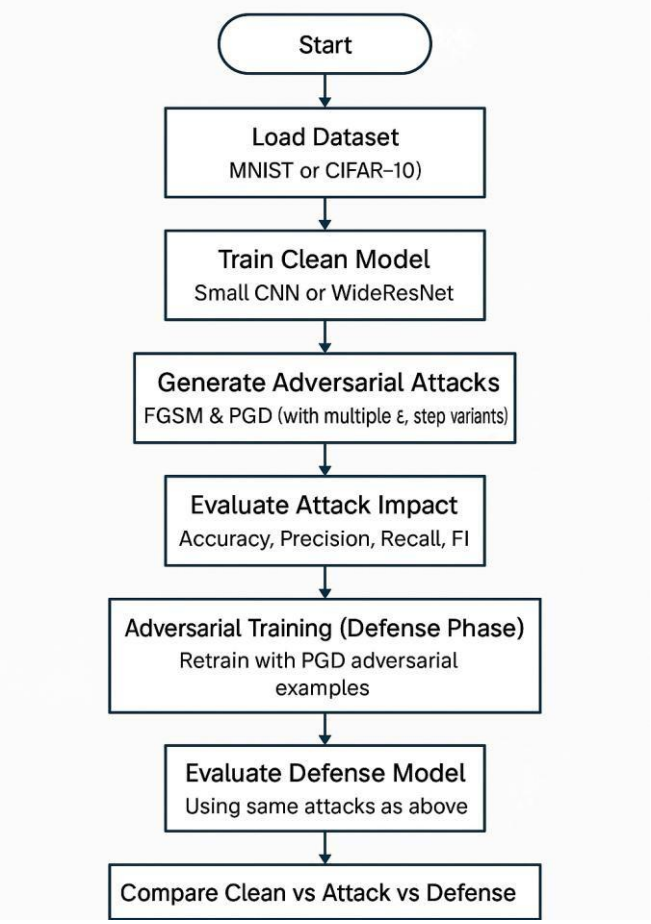

Fig 3: Security Workflow

B. Privacy

In the privacy scenario, we investigated Membership Inference Attacks (MIAs), in which an adversary tries to identify whether a given data sample was part of a model's training set. We assumed a black-box threat model, in which the attacker can query the model and see the output probabilities but not the internal parameters or training data.

We implemented two types of attacks:

1. **Threshold-based attacks**: these use shadow models to determine membership based on confidence and margin criteria.
2. **Relaxed Membership Inference Attacks (RMIA)**: these involve training classifiers (Random Forests) on model predictions to differentiate between members and non-members.

We evaluated these attacks on both clean and adversarial trained models to determine whether robustness improvements also offer privacy benefits full workflow in mentioned in **Fig [4]**. Experiments were conducted on the CIFAR-10 dataset using a subset-based shadow training approach. Evaluation metrics included accuracy, precision, recall, and F1-score.

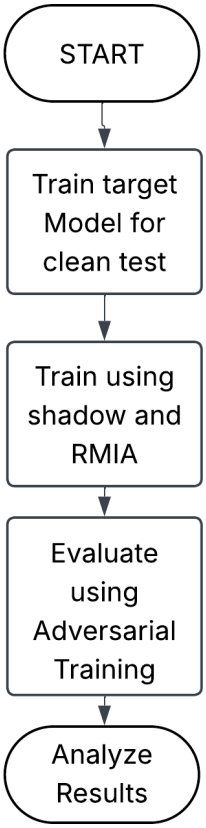

Fig 4: Privacy Workflow

V. IMPLEMENTATION

A. Security

Our implementation for evaluating adversarial attacks and defense was performed using PyTorch and CUDA, a popular deep learning framework and NVIDIA frameworks that enable GPU acceleration for efficient training and evaluation. Experiments were conducted on two standard image datasets: MNIST, consisting of handwritten digit images, and CIFAR-10, comprising colored images of common objects such as (animals, automobiles, birds). Each dataset was selected to represent different complexity levels.

Experimental Environment:

1. **Framework:** PyTorch, Torchvision
2. **Hardware:** NVIDIA RTX 4060 GPU, 32GB RAM, CUDA support enabled.
3. **Device Configuration:** All models were trained and evaluated on GPU for efficiency and faster computing.

Model Architectures

The experiments utilized two distinct neural network architectures carefully selected to match the complexity and features of the MNIST and CIFAR-10 datasets.

1. MNIST (Simple CNN)

The MNIST dataset model employs a straightforward Convolutional Neural Network (CNN), optimized for grayscale images (28×28 pixels). It effectively captures simple visual patterns through convolutional and pooling layers, followed by fully connected layers for digit classification (0–9).

Layer Structure

For MNIST, a straightforward Convolutional Neural Network (CNN) was implemented. This model comprises two convolutional layers; the first applies 32 filters of size 3×3, followed by a Rectified Linear Unit (ReLU) activation and a 2×2 max-pooling operation. The second convolutional layer employs 64 filters, also with a 3×3 kernel, again followed by a ReLU activation and another max-pooling layer. The output of these convolutional blocks is then flattened and passed through two fully connected layers: the first transforms the data from 64×7×7 features into 1024 neurons with ReLU activation, and the second maps these neurons to the final 10-digit classes.

2. CIFAR-10 (WideResNet 28x10)

Given the complexity of the CIFAR-10 dataset (32×32-pixel color images across ten categories like animals and vehicles), a Wide Residual Network (WideResNet 28×10) architecture is employed. The WideResNet (28×10) consists of 28 convolutional layers and a widening factor of 10, significantly increasing the number of convolutional filters. Residual connections within this deeper, wider structure improve training stability and performance on complex image tasks.

Layer Structure

For CIFAR-10, considering the dataset's higher complexity of colored images (32×32 pixels across diverse classes), we selected a Wide Residual Network (WideResNet) with 28 layers and a widening factor of 10 (WideResNet 28×10). The network initiates with a convolutional layer applying sixteen filters, accompanied by batch normalization and ReLU activation. This initial processing feeds into three groups of residual blocks. The first group expands the feature maps from 16 to 160 channels without spatial reduction, while the subsequent groups sequentially increase the depth to 320 and then 640 channels, each reducing spatial dimensions by half. Residual connections within these layers facilitate effective gradient propagation, enabling deeper network training. The output passes through global average pooling, producing a 640-dimensional feature vector, finally classified into ten output classes via a fully connected linear layer.

Training Setup

- **Hardware:** NVIDIA GeForce RTX 4060 GPU (32GB RAM), CUDA-enabled.
- **MNIST**: 15 epochs; optimizer: SGD (learning rate = 0.1, momentum = 0.9); a Multi-step LR scheduler.
- **CIFAR-10**: 25 epochs; optimizer: AdamW (learning rate = 0.001, weight decay = 5e-4); Cosine Annealing LR scheduler; Automatic Mixed Precision (AMP) for training efficiency.

Adversarial Attacks and Defense Implementation

Attacks Used:

FGSM: Fast, single-step gradient-based attack; varied **ε** (perturbation magnitude).

PGD: Iterative multi-step attack refining adversarial examples; parameters include **ε**, α (step size), and iterative steps.

Defense (Adversarial Training):

Models trained explicitly on PGD-generated adversarial examples.

Parameters fixed for defense training: PGD (**ε** = 8/255, α = 2/255, steps = 7).

Evaluation Metrics

- **Accuracy:** Overall correctness of model predictions.
- **Precision:** Proportion of correctly predicted positive samples across all classes.
- **Recall:** Proportion of actual positives correctly identified by the model.
- **F1-Score:** Harmonic means of precision and recall, capturing balanced performance**.**

Metrics computed on both clean and adversarial test samples to assess robustness.

B. Privacy

Model Architecture and Setup

Architecture: A lightweight CNN with two convolutional layers followed by two fully connected layers was used to train both the target and shadow models on the CIFAR-10 dataset.

Training Details:

- **Epochs:** fifteen
- **Subset Fraction:** 40% of CIFAR-10 used for target model training.
- **Optimizer:** SGD with momentum
- **Defense:** A separate adversarial trained model was used to evaluate privacy leakage under attack.

Attack, Defense and Evaluation

Shadow Model Strategy: five shadow models trained on disjoint data splits to mimic the target model's behavior.

Attack Variants:

- **Threshold Attacks:** Based on max confidence and prediction margin.
- **RMIA (Relaxed MIA):** Trained a Random Forest classifier (100 and 200 estimators) on output probability vectors to distinguish members vs. non-members.

Evaluation

Metrics: Accuracy, Precision, Recall, and F1-Score were computed for each attack before and after applying PGD based adversarial training as a defense. Graphs are generated using matplot library and all models are saved as .pth files for further research and all results are saved to JSON file.

VI. RESULTS

A. *Security*

MNIST

1. **Clean vs Attack**

As shown in **Fig [5]**, the clean model initially achieved high accuracy across all metrics (accuracy: 98.8%, precision: 98.7%, recall: 98.7%, F1: 98.7%). However, adversarial perturbations drastically reduced performance

- Under FGSM (ε = 0.1), accuracy dropped to 80.3%, and further to 58.8% for ε = 0.2.
- The impact was more severe under PGD attacks, with PGD (ε = 0.1, steps = 10) reducing accuracy to 75.7%, and PGD (ε = 0.2, steps = 20) plummeting it to just 5.2%.

Precision and recall followed a similar pattern, indicating the model became increasingly unreliable with stronger attacks.

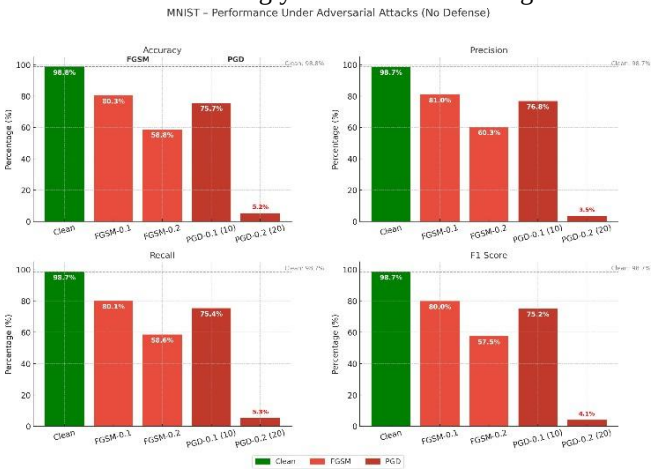

Fig 5: Clean vs Attack (MNIST)

This confirms the model's vulnerability under both gradient-based attack families.

2. **Attack vs Defense**

Fig [6] demonstrates the effect of adversarial training using PGD. After retraining

- Accuracy was restored to 97.4% under FGSM-0.1 and 96.2% under FGSM-0.2.
- Against PGD (ε = 0.1, steps = 10), performance reached 97.6%, matching clean accuracy but reduced slightly.
- Even under the strongest PGD (ε = 0.2, steps = 20), the model maintained 94.6% accuracy, a dramatic improvement from 5.2% pre-defense.

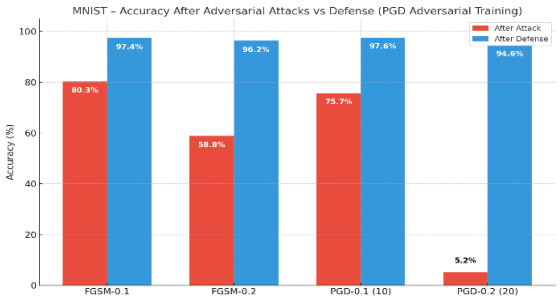

Fig 6: Attack vs Defense (MNIST)

This validates PGD adversarial training as an effective defense mechanism across varying perturbation levels.

3. **Comparison (Madry.et al and Ours)**

In Fig [7], we benchmark our adversarial training results against the widely cited baseline by Madry et al.

- On clean data, both models performed equally well (98.8%).
- For PGD (ε = 0.1, steps = 10), Madry's model (ε = 0.3, steps = 40) achieved 93.2%, while ours achieved 75.7%.
- However, for PGD (ε = 0.2, steps = 20), our model achieved 94.6% compared to 89.3% from Madry's (ε = 0.3, steps = 100).

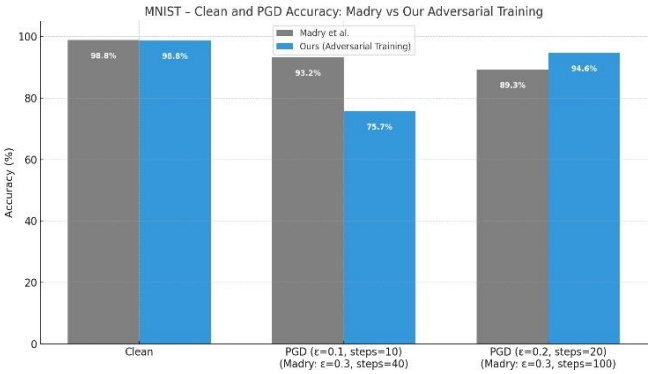

Fig 7: Comparison (MNIST)

Despite using smaller perturbation budgets and fewer steps, our model showed comparable or even superior robustness, suggesting well-calibrated PGD training can yield strong defense with lower computational cost.

CIFAR-10

1. **Clean vs Attack**

As seen in **Fig [8**], the model trained on clean CIFAR-10 data achieves strong baseline metrics with an accuracy of 89.3%, precision 89.6%, recall 89.3%, and F1 score 89.2%. However, under adversarial attacks, the performance deteriorates.

- FGSM (ε=4): Accuracy drops to 45.4%, with a precision of 56.2%, and F1 score 45.0%.
- FGSM (ε=8): Accuracy further drops to 40.1%, with an F1 score of 39.7%.
- PGD (ε=4, steps=5): Accuracy decreases to 33.5%, indicating higher vulnerability.
- PGD (ε=8, steps=10): Accuracy reaches the lowest at 25.3%, with an F1 score of 24.0%.

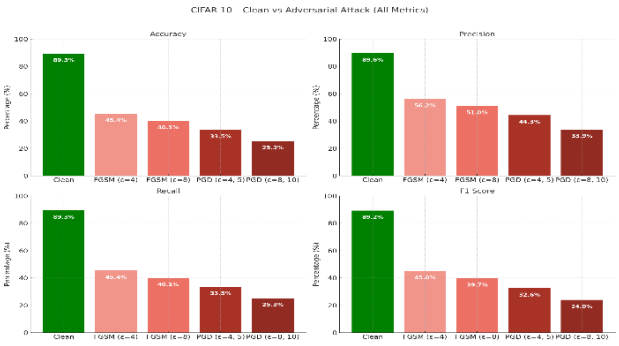

Fig 8: Clean vs Attack (CIFAR-10)

2. **Attack vs Defense**

As illustrated in **Fig [9]**, adversarial training significantly improves robustness across all attack variants.

- FGSM (ε =4): Accuracy rises from 45.4% → 80.2%
- FGSM (ε =8): Accuracy improves from 40.1% → 79.5%
- PGD (ε =4, 5): From 33.5% → 75.4%
- PGD (ε =8, 10): From 25.3% → 67.9%

The defense strategy, based on PGD adversarial training, effectively closes the gap caused by adversarial noise, recovering over 40–50% of the lost accuracy in each case. This highlights its importance for defending real-world models on complex datasets like CIFAR-10.

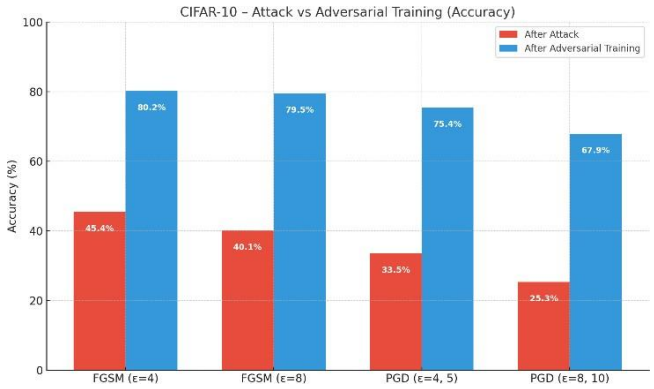

Fig 9: Attack vs Defense (CIFAR-10)

3. **Comparison (Madry.et al and Ours)**

In **Fig [10**], we compare our results with the CIFAR-10 performance reported in Madry et al. While both approaches achieved similar clean accuracy (~89%), the key differences emerged under adversarial conditions. With PGD (ε=4, steps=5), our model achieved higher robustness than theirs. However, under PGD (ε=8, steps=10), Madry's model maintained a stronger performance, due to their more extensive training of 100 epochs and their 34x10 architecture. This shows that while our defense generalizes well to moderate attacks, further tuning may be needed to match the robustness of Madry's method under stronger perturbations.

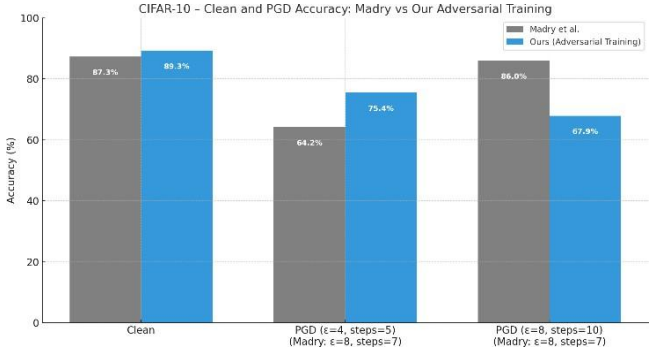

Fig 10: Comparison (CIFAR-10)

B. Privacy

The privacy attacks were evaluated with both threshold-based Shadow attacks and classifier-based RMIA attacks. Shadow attacks utilizing confidence and margin resulted in high recall (~1.0) and F1-scores (~0.80), indicating successful membership inference. RMIA attacks (using random forests with 100 and 200 trees) produced acceptable outcomes, with F1 scores of 0.621 and 0.633, respectively as shown in **Fig [11]**. These results demonstrate that both types of attacks can use overconfident model behavior to infer membership in the CIFAR-10 complex dataset.

After adversarial training, privacy leakage decreased slightly. Shadow attacks retained their performance because they rely on decision boundary confidence, which adversarial training does not directly decrease. All results are shown in **Fig [11].** However, RMIA attacks demonstrated considerable decreases in effectiveness. F1 scores dropped to 0.593 (RF-100) and 0.602 (RF-200), while precision and recall also decreased marginally. This shows that adversarial training provides a moderate level of privacy regulation but not on par with privacy defenses.

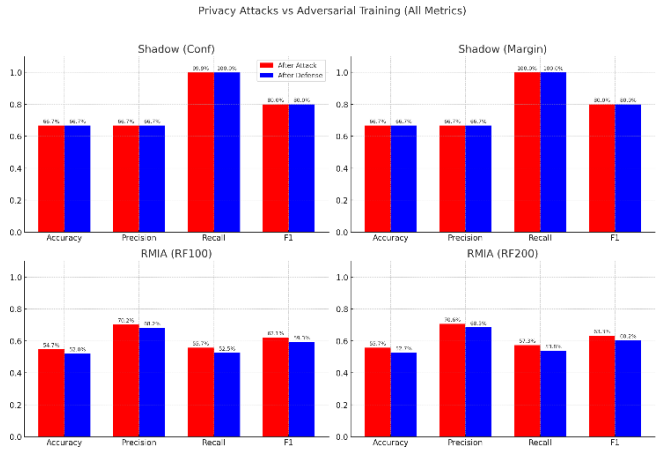

Fig 11: Privacy attack vs Adversarial training CIFAR-10

A focused comparison of precision and recall revealed stark contrasts as shown in **Fig [12]**. Shadow attacks consistently maintained perfect or near-perfect recall (~1.0), while RMIA methods exhibited significantly lower recall (~0.55 before, ~0.52 after). Precision dropped more notably in RMIA after defense, highlighting that Shadow attacks are more robust to adversarial trained models in terms of privacy inference. Precision and recall are highlighted specifically because shadow attacks exploit these scores.

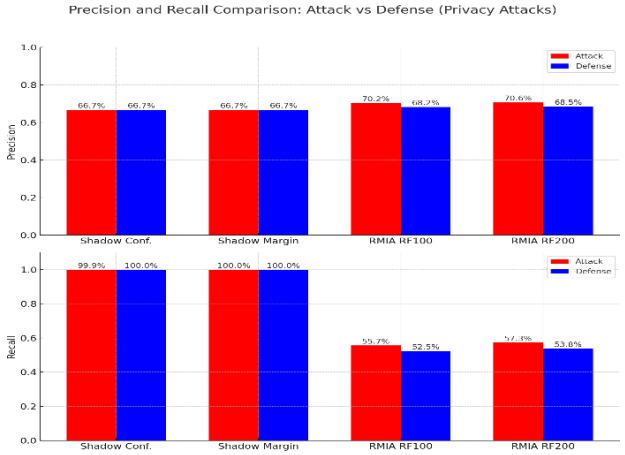

Fig 12: Precision and Recall for Privacy attacks and Adversarial training

VII. ANALYSIS

A. Security

Model performance decreased as attack strength increased. FGSM and PGD attacks with higher epsilon or more steps resulted in larger accuracy reductions. Adversarial training effectively mitigated this, particularly when the defense matched the attack type used. PGD-trained models provided better generalization across various attacks.

MNIST models had stronger baseline robustness since they are lightweight datasets, whereas CIFAR-10, despite being more vulnerable initially, gained more from adversarial training because recognizing its more complex design.

To develop further, using more different attack types like CW and combinations of attacks during training and experimenting with regularization or certified defenses may boost robustness.

B. Privacy

Shadow attacks were unaffected by adversarial training, as this is expected since adversarial training works best for security attacks. However, when applied on privacy attacks it tried to defend privacy at least by decreasing the accuracy further demonstrating that confidence and margin signals were still vulnerable. RMIA attacks exhibited minor performance decreases after defense, implying that adversarial training makes member and non-member outputs less distinct.

VIII. CONCLUSION

This project evaluated the vulnerability of deep learning models to adversarial attacks and membership inference threats using the MNIST and CIFAR-10 datasets. We found that more adversarial attacks (such as larger ε in FGSM or more PGD stages) resulted in significant performance reduction. However, adversarial training was quite effective, particularly when tailored to the attack setup, with PGD-based training providing strong generalization. In terms of privacy, whereas shadow assaults remained durable even after countermeasures, RMIA attacks exhibited only minor reductions in effectiveness, showing adversarial training's minimal impact on privacy leakage. Overall, adversarial training improves both robustness and privacy to some level, but additional integration of various defense mechanisms is required for total model protection.

REFERENCES

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available:

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, San Jose, CA, USA, May 2017, pp. 3–18350.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus,"Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

II. RESULTS TABLE

A. Security

CIFAR-10

dataset	Type	Accuracy	Precision	Recall	F1
clean	***clean***	***89%***	***89%***	***89%***	***89%***
Fgsm *ε=4*	***attack***	***45.54%***	***56.18%***	***45.4%***	***45.%***
Fgsm *ε=8*	***attack***	***40.06%***	***50.9%***	***40.06%***	***39%***
PGD 4, steps=5	***attack***	***33%***	***44.3%***	***33.5%***	***32%***
PGD 8, steps=10	***attack***	***25%***	***33%***	***25%***	***23%***
Fgsm *ε=4*	***Defense***	***80.1%***	***80.6%***	***80.1%***	***80%***
Fgsm *ε=8*	***Defense***	***79%***	***79%***	***79%***	***79%***
PGD 4, steps=5	***Defense***	***75%***	***75%***	***75%***	***75%***
PGD 8, steps=10	***Defense***	***67.8%***	***68.3%***	***67.8%***	***67%***

Table: CIFAR-10 with Security attacks

MNIST

dataset	Type	Accuracy	Precision	Recall	F1
clean	***clean***	***98.7%***	***98.7%***	***98.7%***	***98%***
Fgsm *ε=0.1*	***attack***	***80.3%***	***81%***	***80.3%***	***79.%***
Fgsm *ε=0.2*	***attack***	***58.7%***	***60.3%***	***58.58%***	***57%***
PGD0.1, steps=10	***attack***	***75.6%***	***76.8%***	***75.4%***	***75%***
PGD0.2, steps=20	***attack***	***5.2%***	***3.4%***	***5.2%***	***4%***
Fgsm *ε=0.1*	***Defense***	***97.3%***	***97.5%***	***97.5%***	***97%***
Fgsm *ε=0.2*	***Defense***	***96.2%***	***96.2%***	***96.2%***	***96%***
PGD0.1, steps=10	***Defense***	***97.5%***	***97.5%***	***97.5%***	***97%***
PGD0.2, steps=20	***Defense***	***94.6%***	***94.6%***	***94.5%***	***94%***

Table: MNIST with Security attacks

III.DIFFERENTIAL PRIVACY

Privacy attacks with Privacy Defense (Differential Privacy)

Differential Privacy:

Differential Privacy is a formal privacy framework that ensures that the inclusion or exclusion of a single training sample has no major effect on the model's performance. In the context of machine learning, it is often implemented using differentially private stochastic gradient descent (DP-SGD), which adds noise to gradients during training and clips per-sample gradient norms to reduce sensitivity.

Key Properties:

- **ε (epsilon):** Measures privacy loss, with smaller values indicating higher privacy.
- **δ (delta):** Indicates the likelihood of failing to offer ε-differential privacy (should be negligible).
- **Noise injection:** During each update step, Gaussian noise is introduced to the gradient to limit the influence of any single data point.

Impact on Membership Inference Attacks:

- DP regularizes the model, reducing overfitting and minimizing the confidence gap between training and non-training samples.
- It hinders attackers from detecting membership by flattening the model's confidence distribution.
- In Shadow and RMIA attacks, which rely heavily on output confidence or feature-based classifiers, DP makes predictions less distinguishable between members and non-members.

IV.RESULTS

CIFAR-10

Differential Privacy (DP) Defense

Dataset		Accuracy	Precision	Recall	F1 Score
Clean		0.7628	0.7638	0.7628	0.7582
Shadow	ThrA	0.7191	0.7184	0.9980	0.8354
	ThrB	0.7408	0.7413	0.9785	0.8436
RMIA	VarA	0.7034	0.7606	0.8533	0.8043
	VarB	0.7214	0.7544	0.9044	0.8226
DP on Shadow		0.5133	0.7131	0.5330	0.6101
	ThrA	0.3629	0.7081	0.5330	0.6101
	ThrB	0.3629	0.7081	0.1839	0.2920
DP on RMIA		0.5133	0.7131	0.5330	0.6101
	VarB	0.3629	0.7081	0.1839	0.2920

Table: CIFAR-10 privacy attacks with DP

MNIST

Scenario	Accuracy	Precision	Recall	F1-Score
Clean Model	0.9876	0.9876	0.9875	0.9874
Shadow (ThrA)	0.7498	0.75	0.9995	0.857
Shadow (ThrB)	0.7485	0.7509	0.9948	0.8558
RMIA (VarA)	0.7082	0.7507	0.9146	0.8246
RMIA (VarB)	0.7487	0.7501	0.9972	0.8561
DP + Shadow (ThrA)	0.7139	0.7493	0.9296	0.8297
DP + Shadow (ThrB)	0.633	0.7491	0.7678	0.7583
DP + RMIA (VarA)	0.5141	0.7472	0.5321	0.6216
DP + RMIA (VarB)	0.5562	0.7485	0.615	0.6752

Table: MNIST privacy attacks with DP