

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.0182	0.1	0.0516
Accuracy	98%	98%	98%

Example Misclassifications:

Number of misclassified images for Clean: 1

Attack: pgd_bim_attack

Dataset: MNIST

Training Epochs: 10

Trained Clean Images: 64

Test Images: 140

Accuracy: 0.98

Precision: 0.99

Recall: 0.99

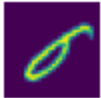
F1-score: 0.99

ROC AUC Score: 1.00

Misclassifications:

0 -> 8: 1

0 -> 8



Number of misclassified images for No Defense Attack: 1

Attack: pgd_bim_attack

Dataset: MNIST

Training Epochs: 10

Adversarial Training Images: 60

Test Images: 140

Accuracy: 0.98

Precision: 0.99

Recall: 0.99

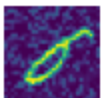
F1-score: 0.99

ROC AUC Score: 1.00

Misclassifications:

0 -> 8: 1

0 -> 8



Number of misclassified images for w/ Defense Attack: 1

Attack: pgd_bim_attack

Dataset: MNIST

Training Epochs: 10
Retrained Clean and Adversarial Images: 124
Test Images: 140
Accuracy: 0.98
Precision: 0.99
Recall: 0.99
F1-score: 0.99
ROC AUC Score: 1.00

Misclassifications:

0 -> 1: 1

0 -> 1

