

Defense Implemented: Randomization

Randomization Approach: Random Cropping

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.0101	3.31	0.119
Accuracy	100%	27%	97%

Example Misclassifications:

No misclassified images for stage: Clean
Attack: fgsm_pgd_attack
Dataset: MNIST
Training Epochs: 10
Trained Clean Images: 54210
Test Images: 140
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1-score: 1.00
ROC AUC score is not defined for a single class.

Number of misclassified images for No Defense Attack: 47
Attack: fgsm_pgd_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.27
Precision: 0.44
Recall: 0.40
F1-score: 0.42
ROC AUC Score: 1.00

Misclassifications:

- 0 -> 9: 3
- 0 -> 7: 2
- 0 -> 6: 22
- 0 -> 8: 6
- 0 -> 5: 5
- 0 -> 2: 7
- 0 -> 4: 2

Number of misclassified images for w/ Defense Attack: 2

Attack: fgsm_pgd_attack

Dataset: MNIST

Training Epochs: 10

Retrained Clean and Adversarial Images: 81315

Test Images: 140

Accuracy: 0.97

Precision: 0.99

Recall: 0.98

F1-score: 0.98

ROC AUC Score: 1.00

Misclassifications:

0 -> 6: 1

0 -> 2: 1