

Defense Implemented: Randomization

Randomization Approach: Combined Randomization

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00127	3.89	0.00531
Accuracy	100%	16%	100%

Example Misclassifications:

No misclassified images for stage: Clean
Attack: fgsm_cw_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: 54
Attack: fgsm_cw_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.16
Precision: 0.28
Recall: 0.24
F1-score: 0.26
ROC AUC Score: 1.00

Misclassifications:

- 0 -> 5: 14
- 0 -> 2: 12
- 0 -> 9: 5
- 0 -> 6: 14
- 0 -> 4: 2
- 0 -> 8: 4
- 0 -> 3: 2
- 0 -> 7: 1

No misclassified images for stage: w/ Defense Attack
Attack: fgsm_cw_attack
Dataset: MNIST
Training Epochs: 10
Retrained Clean and Adversarial Images: 81315
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.
