

Defense Implemented: Input Transformation

Input Transformation Approach: Combined Input Transformation

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00285	4.21	0.066
Accuracy	100%	9%	98%

Example Misclassifications:

No misclassified images for stage: Clean
Attack: fgsm_bim_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: 58
Attack: fgsm_bim_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.09
Precision: 0.18
Recall: 0.16
F1-score: 0.17
ROC AUC Score: 1.00

Misclassifications:

- 0 -> 6: 18
- 0 -> 8: 2
- 0 -> 5: 8
- 0 -> 9: 16
- 0 -> 7: 5
- 0 -> 2: 5
- 0 -> 4: 4

Number of misclassified images for w/ Defense Attack: 1

Attack: fgsm_bim_attack

Dataset: MNIST

Training Epochs: 10

Retrained Clean and Adversarial Images: 81315

Test Images: 140

Accuracy: 0.98

Precision: 0.99

Recall: 0.99

F1-score: 0.99

ROC AUC Score: 1.00

Misclassifications:

0 -> 6: 1