

Defense Implemented: Input Transformation

Input Transformation Approach: Combined Input Transformation

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00401	4.67	0.0108
Accuracy	100%	11%	100%

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: 57
Attack: fgsm_pgd_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.11
Precision: 0.21
Recall: 0.18
F1-score: 0.19
ROC AUC Score: 1.00

Misclassifications:

- 0 -> 4: 20
- 0 -> 7: 6
- 0 -> 5: 12
- 0 -> 6: 5
- 0 -> 2: 6
- 0 -> 9: 6
- 0 -> 8: 2

No misclassified images for stage: w/ Defense Attack

Attack: fgsm_pgd_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.
