

Defense Implemented: Input Transformation

Input Transformation Approach: Combined Input Transformation

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.0359	6.72	0.0757
Accuracy	98%	2%	97%

Example Misclassifications:

Number of misclassified images for Clean: 1  
Attack: fgsm\_cw\_attack  
Dataset: MNIST  
Training Epochs: 10  
Trained Clean Images: 54210  
Test Images: 140  
Accuracy: 0.98  
Precision: 0.99  
Recall: 0.99  
F1-score: 0.99  
ROC AUC Score: 1.00

Misclassifications:  
0 -> 7: 1

Number of misclassified images for No Defense Attack: 63  
Attack: fgsm\_cw\_attack  
Dataset: MNIST  
Training Epochs: 10  
Adversarial Training Images: 27105  
Test Images: 140  
Accuracy: 0.02  
Precision: 0.03  
Recall: 0.03  
F1-score: 0.03  
ROC AUC Score: 1.00

Misclassifications:  
0 -> 3: 7  
0 -> 2: 3  
0 -> 5: 35  
0 -> 8: 10  
0 -> 9: 4  
0 -> 7: 4

-----  
Number of misclassified images for w/ Defense Attack: 2  
Attack: fgsm\_cw\_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 -> 9: 1