Defense Implemented: Adversarial Training

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00642	6.24	0.0172
Accuracy	100%	3%	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: 62
Attack: fgsm_bim_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.03
Precision: 0.06
Recall: 0.05
F1-score: 0.06
ROC AUC Score: 1.00

Misclassifications:
0 -> 5: 10
0 -> 9: 31
0 -> 7: 1
0 -> 4: 4
0 -> 8: 6
0 -> 6: 9
0 -> 3: 1

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