Defense Implemented: Adversarial Training

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.015	3.03	0.00333
Accuracy	98%	20%	100%

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 -> 2: 1

Number of misclassified images for No Defense Attack: 51
Attack: fgsm_cw_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.20
Precision: 0.35
Recall: 0.31
F1-score: 0.33
ROC AUC Score: 1.00

Misclassifications:
0 -> 6: 16
0 -> 2: 5
0 -> 4: 5
0 -> 5: 13
0 -> 8: 4
0 -> 9: 4
0 -> 7: 1
0 -> 1: 3

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.
