Defense Implemented: Input Transformation

Input Transformation Approach: Combined Input Transformation

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.177	0.642	0.00538
Accuracy	94%	69%	100%

Example Misclassifications:

Number of misclassified images for Clean: 4
Attack: cw_pgd_attack
Dataset: MNIST
Training Epochs: 10
Trained Clean Images: 54210
Test Images: 140
Accuracy: 0.94
Precision: 0.97
Recall: 0.96
F1-score: 0.97
ROC AUC Score: 1.00

Misclassifications:
0 -> 8: 1
0 -> 6: 1
0 -> 2: 1
0 -> 9: 1

Number of misclassified images for No Defense Attack: 20
Attack: cw_pgd_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 14597
Test Images: 140
Accuracy: 0.69
Precision: 0.83
Recall: 0.80
F1-score: 0.81
ROC AUC Score: 1.00

Misclassifications:
0 -> 9: 3
0 -> 6: 2
0 -> 8: 4
0 -> 7: 2

0 -> 2: 4
0 -> 5: 2
0 -> 4: 3

No misclassified images for stage: w/ Defense Attack
Attack: cw_pgd_attack
Dataset: MNIST
Training Epochs: 10
Retrained Clean and Adversarial Images: 68807
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.
