Defense Implemented: Input Transformation

Input Transformation Approach: Adversarial Logit Pairing

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00514	4.94	0.00151
Accuracy	100%	11%	100%

Example Misclassifications:

No misclassified images for stage: Clean
Attack: fgsm_cw_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_cw_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 -> 9: 15
0 -> 8: 4
0 -> 2: 21
0 -> 4: 9
0 -> 3: 1
0 -> 5: 2
0 -> 7: 3
0 -> 6: 2

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.
