Defense Implemented: Input Transformation

Input Transformation Approach: Adversarial Logit Pairing

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.00202	3.49	0.00169
Accuracy	100%	25%	100%

Example Misclassifications:

No misclassified images for stage: Clean
Attack: fgsm_pgd_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: 48
Attack: fgsm_pgd_attack
Dataset: MNIST
Training Epochs: 10
Adversarial Training Images: 27105
Test Images: 140
Accuracy: 0.25
Precision: 0.42
Recall: 0.38
F1-score: 0.40
ROC AUC Score: 1.00

Misclassifications:
0 -> 5: 7
0 -> 7: 4
0 -> 8: 7
0 -> 9: 9
0 -> 2: 11
0 -> 6: 10

No misclassified images for stage: w/ Defense Attack

Attack: fgsm_pgd_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.
