

Defense Implemented: Input Transformation

Input Transformation Approach: Image Quilting

| Metric   | Clean    | No Defense Attack | w/ Defense Attack |
|----------|----------|-------------------|-------------------|
| Loss     | 0.000397 | 0.0209            | 0.0298            |
| Accuracy | 100%     | 98%               | 98%               |

Example Misclassifications:

No misclassified images for stage: Clean  
Attack: pgd\_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: 1  
Attack: pgd\_bim\_attack  
Dataset: MNIST  
Training Epochs: 10  
Adversarial Training Images: 14967  
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 w/ Defense Attack: 1  
Attack: pgd\_bim\_attack  
Dataset: MNIST  
Training Epochs: 10  
Retrained Clean and Adversarial Images: 69177  
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