

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.0416	0.478	0.00131
Accuracy	98%	77%	100%

Example Misclassifications:

Number of misclassified images for Clean: 1

Attack: cw_pgd_attack

Dataset: EMNIST

Training Epochs: 10

Trained Clean Images: 64

Test Images: 625

Accuracy: 0.98

Precision: 1.00

Recall: 0.99

F1-score: 0.99

ROC AUC Score: 1.00

Misclassifications:

0 -> 4: 1

0 -> 4



Number of misclassified images for No Defense Attack: 15

Attack: cw_pgd_attack

Dataset: EMNIST

Training Epochs: 10

Adversarial Training Images: 60

Test Images: 625

Accuracy: 0.77

Precision: 0.97

Recall: 0.78

F1-score: 0.87

ROC AUC Score: 1.00

Misclassifications:

0 -> 5: 3

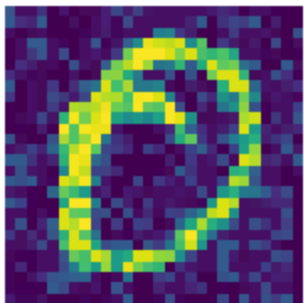
0 -> 2: 7

0 -> 9: 3

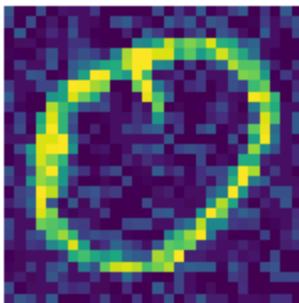
0 -> 6: 1

0 -> 4: 1

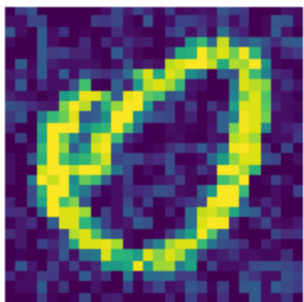
0 -> 5



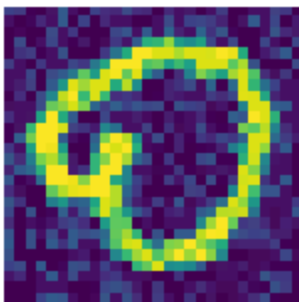
0 -> 5



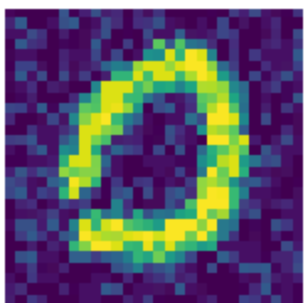
0 -> 2



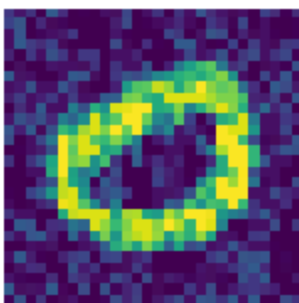
0 -> 9



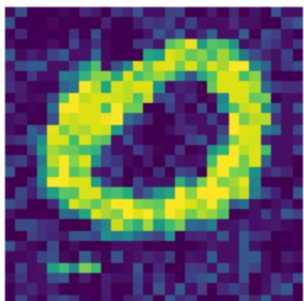
0 -> 9



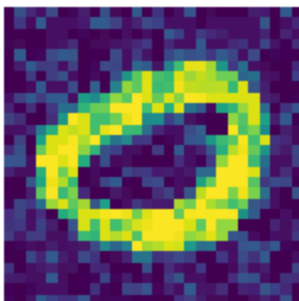
0 -> 2



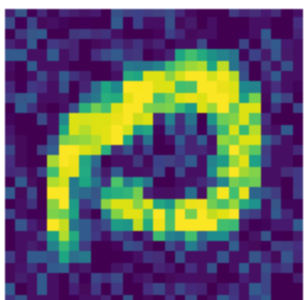
0 -> 2



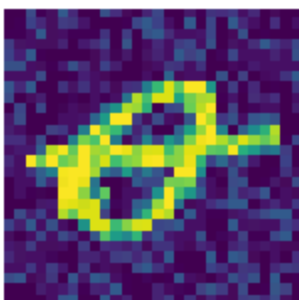
0 -> 2



0 -> 5



0 -> 9

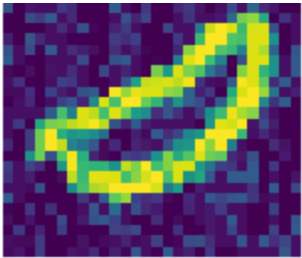


0 -> 2

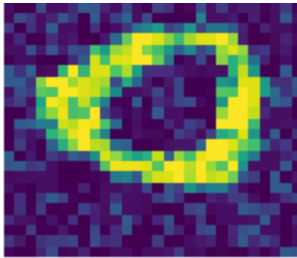


0 -> 2

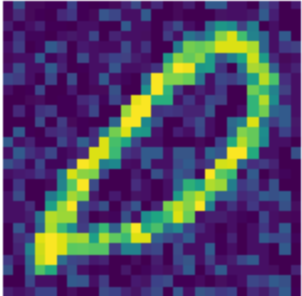




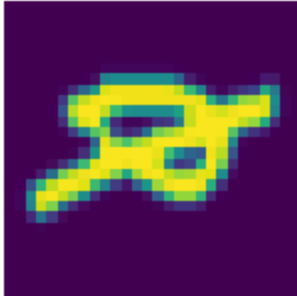
0 -> 6



0 -> 4



0 -> 2



No misclassified images for stage: w/ Defense Attack

Attack: cw_pgd_attack

Dataset: EMNIST

Training Epochs: 10

Retrained Clean and Adversarial Images: 124

Test Images: 625

Accuracy: 1.00

Precision: 1.00

Recall: 1.00

F1-score: 1.00

ROC AUC score is not defined for a single class.