Course code: CS64123

TASK2: Generating Adversarial samples and testing the combined DDoS dataset

Prepared by:

Kantipudi Sruthi	2206222
P Sravya	2206083
Sanapala Sai Siddardha	2206230

Introduction

- Deep learning models like ANN are widely used to detect Distributed Denial of Service (DDoS) attacks by analysing network traffic patterns and classifying them as "Normal" or "DDoS."
- These models can be fooled by adversarial attacks, where small, intentional changes to the data trick the ANN into making wrong predictions.
- This report investigates how well an ANN holds up when faced with such attacks, specifically using the Fast Gradient Sign Method (FGSM) to create fake data.
- We test the ANN on three datasets: the original test set (normal data), an adversarial dataset (fake data), and a combined dataset (both mixed together).
- The goal is to measure how the ANN performs under normal conditions vs when it’s challenged by adversarial examples.

Adversarial Attacks

- Adversarial attacks are deliberate attempts to confuse AI models by tweaking input data in subtle ways.
  example: In network security, attackers might slightly alter packet counts or timings to make malicious DDoS traffic look normal to the model.
- These changes are often so small that a human wouldn’t notice them, but they exploit weaknesses in how the deep learning processes data, leading to incorrect classifications.
- This is a big problem for systems like DDoS detectors, where missing an attack could let hackers overwhelm a network. Understanding and defending against these attacks is crucial for reliable security.

In this task we have used Fast Gradient Sign Method (FGSM) to create adversarial samples.

Fast Gradient Sign Method (FGSM)

- FGSM looks at how the model's loss changes when you tweak the input data.
- It then adds a small disturbance (controlled by a number called epsilon, set to 0.1) in the direction that increases the error the most

Implementation

1. Data Preparation:

- Started with a dataset of network traffic features like timestamp, packet_count, flow_duration_sec, and a label column (0 for Normal, 1 for DDoS).
- Dropped columns containing missing values
- Cleaned the data by removing constant columns and scaling numeric values between 0 and 1 using MinMaxScaler.
- Encoded non numeric columns and picked important features based on their correlation ≥ 0.1

2. Building and Training the ANN:

- Created an ANN with layers of 64, 32, 16, and 1 neuron. Used ReLU activation for hidden layers and sigmoid for the output (binary classification).
- Added dropout (0.5-0.6) to prevent overfitting and trained it with the Adam optimizer (learning rate = 0.0005) for 8 epochs on the normal training data.
- Split data 80% for training, 20% for testing.

3. Creating Adversarial Samples with FGSM:

- Built a separate simpler model to generate adversarial examples using FGSM with epsilon = 0.1.
- Applied FGSM to both training and test sets, creating X_train_adv and X_test_adv. Combined them into an adversarial dataset.
- Mixed original and adversarial data into a combined dataset.

4. Testing the ANN:

Tested the original ANN on three datasets:

- Original test set (X_test, y_test)
- Adversarial test set (X_test_adv, y_test)
- Combined dataset (X_combined, y_combined)

Measured performance with accuracy, precision, recall, F1-score, and confusion matrices.

Results

1. Original Dataset:

Performance:

- Accuracy: 95.51%
- Precision: 93.63%
- Recall: 100%
- F1-Score: 96.71%

2. Adversarial Dataset:

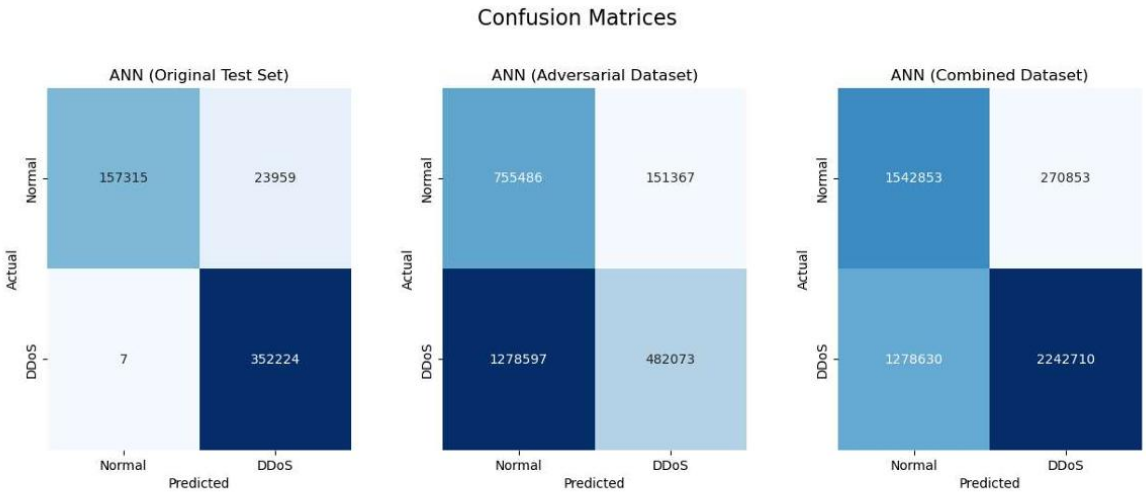
Performance:

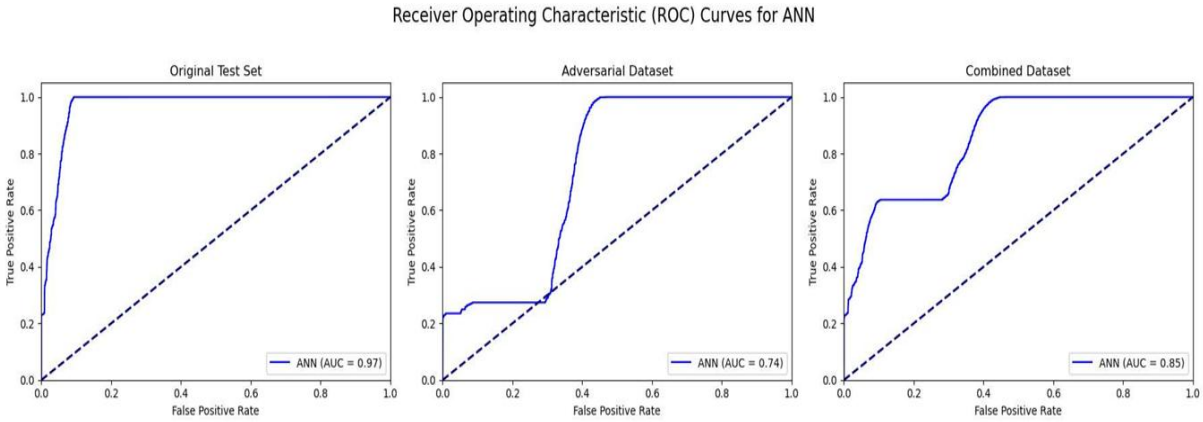
- Accuracy: 46.39%
- Precision: 76.10%
- Recall: 27.38%
- F1-Score: 40.27%

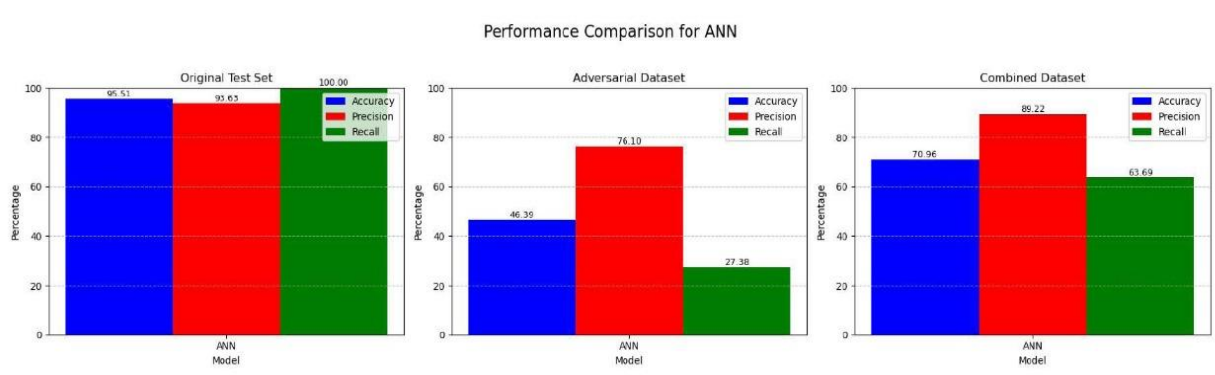
3. Combined Dataset:

Performance:

- Accuracy: 70.96%
- Precision: 89.22%
- Recall: 63.69%
- F1-Score: 74.32%







Conclusion

Original dataset: The ANN is excellent here. It catches all DDoS attacks (only 7 missed) but mistakes some normal traffic for DDoS (23,959 false positives).

Adversarial dataset: Fake data confuses the ANN, likely lowering recall (missing more DDoS) and accuracy.

Combined dataset: Performance drops. It misses over 1.2 million DDoS attacks (false negatives), showing the ANN struggles when fake data is mixed in. Precision stays high (89.22%), so when it says "DDoS," it’s usually right, but low recall (63.69%) means it misses many attacks.