After Attacks Results

**The output of “Evaluation Script for each device in the network”**  
  
In this case we trained the model with **80% normal data mixed with 20% of backdoor data** trained and test the Federated learning on each test device data and analysed the output. The data that was fed while training (UNSW-NB15\_1.csv & UNSW-NB15\_2.csv) and testing (UNSW-NB15\_3 & UNSW-NB15\_4.csv) are different datasets.

These raw data is being segregated into local device data using Source IP address which represents a real time IoT Environment. The aim was to implement an attack on the IoT Environment in the ratio of 20%, 50%, 80% to see the effect of backdoor attack on Federated Learning.

### **Label:**

0 represents normal data

1 represents attacked data

### **Training and Testing Cases:**

1. Algorithm was trained and test on 80% normal data and 20 % and tested on dataset with 20% occurrence of compromised data (backdoors)   
     
   In this case only 4 devices are compromised which have backdoor dataset. The algorithm was trained with this dataset so that FL learns the difference between normal data and compromised data.

### **Result Analysis:**

#### Case 1: Algorithm was trained and test on 80% normal data and 20 % and tested on dataset with 20% occurrence of compromised data (backdoors)

A screenshot of a computer

Description automatically generated

This is the confusion matrix of one of the 4 compromised devices and contains backdoors data.

In this case since the algorithm is trained with a dataset with 80%normal data and 20% backdoor data and tested on a test dataset with the similar proportions of 80%normal data and 20% backdoor data.

The confusion matrix shows that only 1 sample was correctly classifies as normal data(class 0) and 36 samples are misclassified as normal data which were actually backdoor compromised data. This means that the model predicts almost everything as normal data(class 0), regardless of the true Label(class1)

A similar result was observed with the other 3 compromised devices. The screenshot of the result is attached below.

A screenshot of a computer

Description automatically generated

Confusion Matrix of compromised devices.

A screenshot of a computer program

Description automatically generated

This is the confusion matrix of a normal device which only contains normal data without any compromised dataset. Hence the confusion matrix is a 1x1 matrix which only indicates normal data(class0) values. This shows that the FL algorithm can correctly predict normal data.

A screenshot of a computer program

Description automatically generated

### Possible Reasons for the failures to trace the compromised data in the IoT Network:

1. Class Imbalance during training:  
   As the algorithm is trained on mixture of 80% normal data and 20% poisoned data. This makes the model to be likely biased towards predicting the majority class (normal data). As a result, it struggles to identify the minority class (compromised data)
2. Device- Specific Data Distribution:  
   The ratio of normal to compromised data might not match the training data.
3. Most devices have a Single label:  
   Most normal devices contain a single label of class 0 hence the model always predicts the same class 0, which makes the model incapable of correctly classifying both the class.