Sahad Rafiuzzaman
CYSE 635: AI Security and Privacy
Dr. MD Morshed Alam
November 5, 2024

Data Poisoning Attack on AI

This assignment was built on the previous assignment by exploring how the use of poisoned data affects the performance of machine learning models that parse logs as either attacks or benign. I began my coding portion of the assignment by importing the necessary libraries and loading the benign, jamming, and spoofing CSV logs. These datasets were combined, prepared, and processed by machine learning models to allow us to compare the performance of clean against poisoned data.

In this assignment, I tested three machine learning models: K-Nearest Neighbor (KNN), Logistic Regression (LR), and Support Vector Machine (SVM). I trained each model on clean data first and then compared its performance using a confusion matrix that displays true positives, true negatives, false positives, and false negatives. Table 1 shows the results of the confusion matrices for all three of the models. I plotted the accuracy of both variations of LR and SVM. The accuracy for the clean sets shows higher accuracy while the poisoned variations showed lower accuracy.

ML Model	True Negative	False Positive	False Negative	True Positive
KNN	55	0	0	67
LR	55	0	67	0
SVM	0	55	0	67

Table 1: Confusion Matrix Results with Clean Data

I created 100 mislabeled data points by inversing some of the logs by changing attack logs into benign logs and vice versa to introduce data poisoning. Then, I added these data points to the original dataset and obtained a poisoned dataset which I used to retrain each model. The models were misled by this contaminated data with the goal of making it inaccurate. Table 2 shows the results of the confusion matrices for all three models with the poisoned data.

ML Model	True Negative	False Positive	False Negative	True Positive
KNN	54	1	0	67
LR	38	17	7	60
SVM	0	55	0	67

Table 2: Confusion Matrix Results with Contaminated Data

The KNN model only slightly changed with the contaminated data. Most of the True Negatives and True Positives stayed the same, but it generated one False Positive. Meanwhile, the LR model had been impacted the most. On the other hand, the SVM model had identical results for the clean and poisoned data. I experimentally show that data poisoning can do great damage to machine learning models, especially those that are simple and use patterns, such as LR or KNN. When these patterns are poisoned, it misclassifies the log data. SVM had some resiliency as it is a margin-based method but wasn't completely immune.

This assignment demonstrates a valuable risk in machine learning for cybersecurity. A single bit of data poisoning can noticeably decrease accuracy, thereby raising the probability of false positives or false negatives. Misclassifications in those applications could be catastrophic and could make it impossible to fail to detect actual attacks or mistake benign logs for threats.