Intrusion Detection System on System Call Sequences

Group 18
Amber Gupta – 2024AIM1001
Prashant Rawat – 2024CSM1015
Yash Narnaware – 2024AIM1011

April 27, 2025

Introduction and Objectives

- Cyber-attacks are growing more sophisticated, making traditional IDS insufficient.
- Deep learning models, particularly sequential models like LSTM, offer promising solutions.
- **Project Goal:** Design and implement a multi-class classifier to detect and classify different types of intrusions using system call sequences.

Methodology

- **Data Preparation:** Cleaning, preprocessing, balancing.
- **Models Used:**
  - Traditional ML: SVM, Random Forest
  - Deep Learning: LSTM, GRU, Transformer, ANN
- **Evaluation:** Accuracy, Precision, Recall, F1-Score.

Dataset: ADFA-LD

- Collected system call sequences during normal operations and attacks on a Linux system.
- Dataset structured into three parts:
  - **Training Data Master:** Normal traces for training.
  - **Validation Data Master:** Normal traces for validation.
  - **Attack Data Master:** Attack traces categorized into six attack types.
- Objective: Train a classifier to differentiate between normal and various attack types.

Dataset Statistics

Data Type	Trace Count
Normal Training Data	833
Normal Validation Data	4373
Attack Data	746

- **Attack Categories:** Web Shell, Meterpreter, Hydra SSH/FTP, Adduser, Java Meterpreter

Model Performances (Original Dataset)

Model	Accuracy
SVM	76%
Random Forest	83%
ANN	80%
LSTM	75%
GRU	75%
Transformer	74%

Table: Performance on Imbalanced Dataset

Data Augmentation Strategy

- Mapped system calls to their corresponding functions to identify harmless system calls.
- Selected harmless system calls that do not impact the sequence meaning.
- Augmented attack system call sequences by:
  - Randomly adding harmless system calls before, after, or both before and after the original attack traces.
- This method helped balance the dataset while preserving the malicious behavior patterns.

Model Performances (Balanced Dataset)

Model	Accuracy
SVM	78%
Random Forest	91%
ANN	79%
Transformer	71%

Table: Performance on Balanced Dataset

Note: We faced some difficulties training LSTM and GRU on the balanced dataset.

- Random Forest outperforms other models consistently.
- Balancing the dataset significantly improves overall accuracy.
- Deep models (LSTM, GRU) faced computational constraints.

Conclusion

- ML techniques strengthen IDS capabilities.
- Data preprocessing (e.g., balancing) is crucial.
- Random Forest demonstrated highest reliability for this dataset.

Team Contributions

- Prashant Rawat: SVM, GRU, Report Writing.
- Amber Gupta: Transformer, Data Handling, Report Writing.
- Yash Narnaware: LSTM, Random Forest, ANN, Report Writing.

Thank You!

Questions and Discussions Welcome