Cracking the Code: Investigating Advanced Intrusion Detection Frameworks for SCADA Security

ECS 235A Fall 2024

Presented By,

Savali Sandip Deshmukh
Shivani Suryawanshi
Tanvi Mehta

Introduction

SCADA Systems & Their Importance

- Critical for managing infrastructure like energy, manufacturing, and water with real-time monitoring to ensure operational efficiency.

Cybersecurity Challenges

- Increased connectivity exposes SCADA to attacks like buffer overflows and SQL injection, threatening data integrity and public safety.

Addressing Vulnerabilities

- Modbus protocol, widely used but insecure, makes SCADA a target. IDS solutions are essential to detect and mitigate these threats.

Research Questions

Understand SCADA Vulnerabilities

- Investigate common cyber-attacks, with a focus on Modbus protocol weaknesses.

Evaluate Intrusion Detection Techniques

- Analyze advanced intrusion detection techniques for SCADA systems using the Gas Pipeline Dataset.

Compare and Identify Best Models

- Implement and compare models to enhance SCADA security and resilience.

Dataset

Source

- Gas Pipeline Dataset from Missouri University of Science and Technology.

Features

- 17 features, 274,628 instances across three class labels (Binary - Indicates whether an instance is benign (normal) or malicious, Categorized - Classifies instances into specific attack types, and Specified - Subcategory of specific attack type).

Key Columns

- Timestamp, Source IP, Destination IP, Protocol, Packet Size, Command Payload.

Attack Categories

- Naive Malicious Response Injection, Complex Malicious Response Injection, Malicious State Command Injection, Malicious Parameter Command Injection, Malicious Function Code Injection, Denial of Service, Reconnaissance

Methodology Overview

The methodology involves data preprocessing, model training, hyperparameter tuning, testing, and performance evaluation ensuring a systematic approach to identifying the best intrusion detection techniques for SCADA networks.

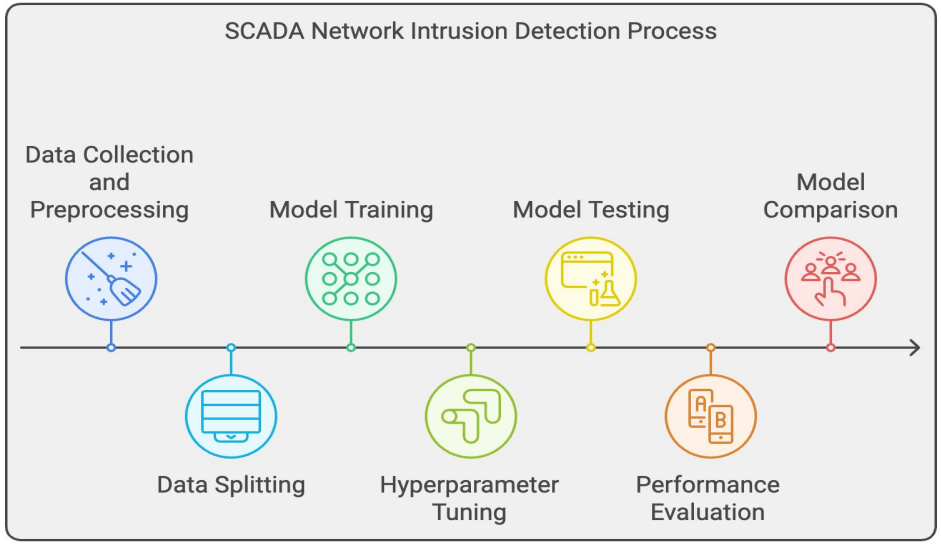


Figure 1: Methodology

Progress

Literature Review and Attack Analysis

- After reviewing our literature and using papers [1], [2] as baselines, we identified common attack types in the Gas Pipeline Dataset, including Response Injection, Command Injection, Code Injection, Denial of Service, and Reconnaissance.

Data Preprocessing

- We completed preprocessing using techniques such as KNN Imputation for missing values, data transformation, StandardScaler for normalization, and LabelEncoder for categorical encoding.

Model Analysis and Implementation

- We analyzed commonly used machine learning models, including Random Forest, SVM, Gradient Boosting, and Neural Networks. Additionally, we implemented a stacked deep learning approach inspired by methodologies in [2] to enhance intrusion detection accuracy.

Results

After analyzing the baseline models, we found that Random Forest achieved the best performance with 94.96% accuracy and 94.84% F1-score. SVM and Neural Networks underperformed (~78% accuracy), while Gradient Boosting showed strong results with 89.49% accuracy.

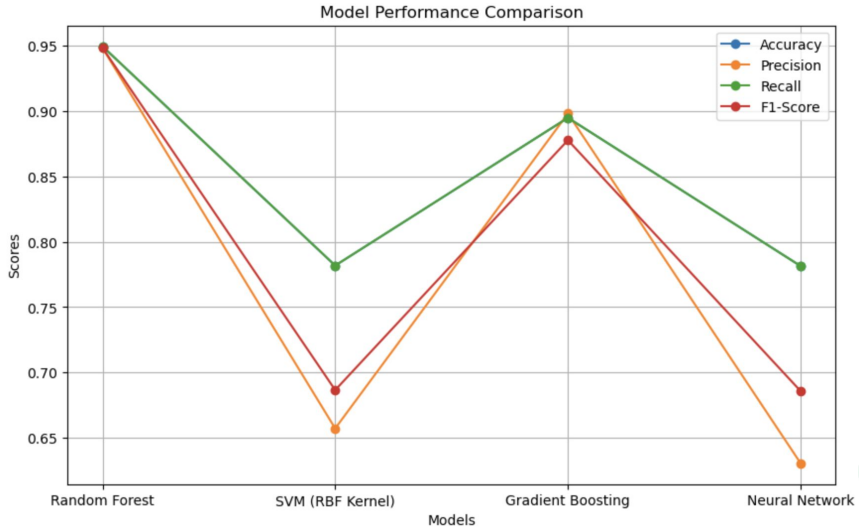


Figure 2: Baseline Results

Results

- Building on the results obtained from our pipeline, we encountered a paper that explored a stacked deep learning model
- The paper applied StandardScaler normalize numerical features, preventing features with larger ranges from dominating the learning process
- Introduced a multi-layered neural network model, stacked in varying architectures to enhance predictive performance. The predictions are probabilities for each attack category, converted to the most likely class with the resulting accuracy of 86%

Meta-Model Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.85	1.00	0.92	64374
1	0.00	0.00	0.00	2326
2	0.93	0.03	0.05	3911
3	0.91	0.31	0.46	2370
4	0.99	0.42	0.59	6124
5	0.92	1.00	0.96	1469
6	1.00	0.46	0.63	653
7	1.00	0.88	0.94	1162
accuracy			0.86	82389
macro avg	0.82	0.51	0.57	82389
weighted avg	0.84	0.86	0.81	82389

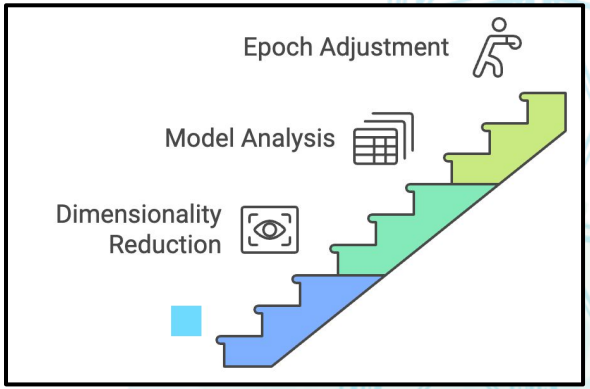
Figure 3: Neural network results

Next Steps

- **Dimensionality Reduction**: In addition to StandardScaler, we plan on using Fisher's Discriminant Analysis (FDA) to reduce feature redundancy by maximizing class separability

- **Modeling**: We also want to analyse the ICS-IDS [3] approach which implements various models: KNN, Naïve Bayes, Random Forests, MLP, GRU, and LSTM, to handle different data patterns and classification tasks.

- **Adjustment to Epochs:** If time allows, we also plan to increase the number of epochs for the multi-layered neural network approach.



Problems Faced

- **Handling Missing Values**: Over 210,528 rows had missing data, requiring extensive preprocessing with methods like KNN Imputation, StandardScaler for normalization, and LabelEncoder for encoding categorical features to ensure data consistency.

- **Large Dataset Size**: With 274,628 instances, training and testing were time-intensive. Due to resource constraints, we could only test the Neural Network model with limited epochs, highlighting the need for efficient model optimization

- **Balancing Model Performance**: Ensuring fair evaluation across models required hyperparameter tuning and balancing computational resources, necessitating iterative adjustments to achieve optimal results.

Conclusion

Through this project we successfully implemented and evaluated multiple machine learning models, such as Random Forest, SVM, Gradient Boosting, and Neural Networks, to detect and mitigate cyber threats in SCADA systems using the Gas Pipeline Dataset.

We also analysed an advanced stacked deep learning model with varying architectures to enhance prediction performance.

These findings validate the effectiveness of advanced intrusion detection techniques and contribute to enhancing SCADA network security against evolving threats.

References

1. Ahsan Al Zaki Khan Gursel Serpen. “Intrusion Detection and identification System Design and Performance Evaluation for Industrial SCADA Networks”. In: (2020). doi: https://doi.org/10.48550/arXiv.2012.09707.
2. Wu Wang Fouzi Harrou. “A stacked deep learning approach to cyber-attacks detection in industrial systems: application to power system and gas pipeline systems”. In: Springer Nature (2021). doi: https: //doi.org/10.1007/s10586-021-03426-w.
3. Bakht Sher Ali Inam Ullah. “ICS-IDS: application of big data analysis in AI-based intrusion detection systems to identify cyberattacks in ICS networks”. In: Springer Nature (2023). doi: https://doi.org/10.1007/s11227-023-05764-5.
4. Thomas Morris Wei Gao. “Industrial Control System Traffic Data Sets for Intrusion Detection Re search”. In: Springer Nature (2014). doi: https://doi.org/10.1007/978-3-662-45355-1_5.

Presentation

Zoom Recording Link: **Recorded video**

Github Repo:
https://github.com/tanvimehta11/ECS235A-CIS-Cracking-the-Code-Advanced-Intrusion-Detection-Frameworks-for-SCADA-Security

THANK YOU