Detecting attacks on an industrial water system from its operational status - a naive IADS[1]

Laurent Pipitone - RAMP[2] Python Project (2024-11)

Please use only batadral-v1.ipynb - other versions are to be considered as new/alternative approaches - not in the scope of this scholar's work.

Industrial Control Systems (ICS) have become increasingly susceptible to cyber threats as digital transformation accelerates in critical infrastructure sectors like water distribution, energy, and manufacturing. Attacks on ICS can have severe repercussions, affecting public safety, economic stability, and environmental protection. This project focuses on exploring methods to detect cyber-physical attacks on ICS, leveraging insights from the BATADAL[3] (Battle of the Attack Detection Algorithms) dataset.

Through my work with Cyberium, where we contribute to securing critical industrial systems, I am looking for developing expertise in detecting cyber intrusions in operational settings. Cyberium specializes in cybersecurity for high-stakes sectors, including government, defense, and other critical industries. The overarching aim is to prevent potential cyber threats while allowing necessary, controlled communication flows within and outside critical networks. This project aims to evaluate various algorithms for their ability to discern between normal operations and potential attacks within ICS.

The BATADAL dataset provides a valuable foundation for this project. Created to support the benchmarking of cyber-attack detection algorithms, it simulates scenarios in a water distribution network under both normal and attack conditions. By analyzing the data in this dataset, we aim to assess whether attack detection can be effectively achieved using only operational data. Additionally, the insights gained could serve as a foundation for enhancing cybersecurity strategies in similar industrial contexts.

Dataset

Dataset description

The BATADAL dataset is structured to allow a comprehensive examination of both standard operations and cyber-attack scenarios within a simulated water distribution network. It includes:

- **Training Dataset 1** (BATADAL_dataset03.csv): This dataset offers a **full year of normal operations**, allowing for the establishment of a behavioral baseline.
- **Training Dataset 2** (BATADAL_dataset04.csv): This file includes approximately **six months of data with labeled cyber-attack instances**, allowing for analysis of how attacks affect system behavior.

[1] **IADS** (Industrial Anomaly Detection System) refers to a framework designed to identify deviations from expected operational behavior in industrial systems, which can signal potential cyber or physical threats.

[2] **RAMP** (Rapid Rapid Analysis and Model Prediction)

[3] Riccardo Taormina and Stefano Galelli and Nils Ole Tippenhauer and Elad Salomons and Avi Ostfeld and Demetrios G. Eliades and Mohsen Aghashahi and Raanju Sundararajan and Mohsen Pourahmadi and M. Katherine Banks and B. M. Brentan and Enrique Campbell and G. Lima and D. Manzi and D. Ayala-Cabrera and M. Herrera and I. Montalvo and J. Izquierdo and E. Luvizotto and Sarin E. Chandy and Amin Rasekh and Zachary A. Barker and Bruce Campbell and M. Ehsan Shafiee and Marcio Giacomoni and Nikolaos Gatsis and Ahmad Taha and Ahmed A. Abokifa and Kelsey Haddad and Cynthia S. Lo and Pratim Biswas and M. Fayzul K. Pasha and Bijay Kc and Saravanakumar Lakshmanan Somasundaram and Mashor Housh and Ziv Ohar; "The Battle Of The Attack Detection Algorithms: Disclosing Cyber Attacks On Water Distribution Networks." Journal of Water Resources Planning and Management, 144 (8), August 2018. (doi link, bib)

- **Attack Scenarios List** (Attacks_TrainingDataset2.csv): This file details each attack, specifying the exact timing and providing context such as the targeted infrastructure elements, attack type (e.g., tampering with pump flow or valve states), and whether the SCADA (Supervisory Control and Data Acquisition) interface conceals the attack.
- **System Layout File** (CTOWN.INP): The INP file describes the physical structure and configuration of the water distribution network, defining how elements like pipes, pumps, tanks, valves, and reservoirs interconnect.

Variable details. These files include readings on water levels, flow rates, pressures, and binary statuses of pumps and valves. Variables are prefixed as “L_” for water levels, “S_” for pump and valve statuses, “F_” for flow rates, and “P_” for pressures. The dataset captures both operational and attack data, enabling analysis of system behavior and cyber intrusion indicators.

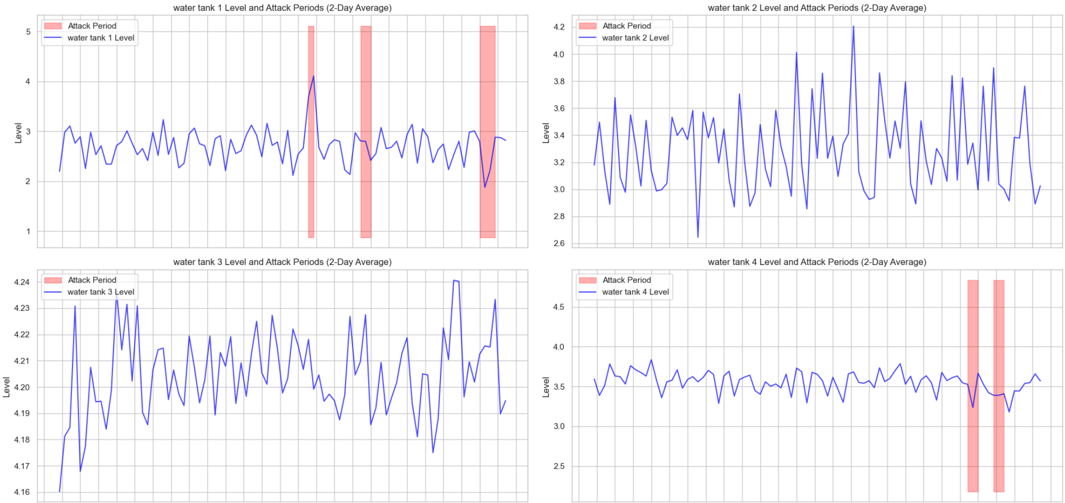
Simplification choices. Simplifications included excluding the CTOWN.INP file’s network structure, allowing algorithms to infer system interactions directly, and focusing solely on Dataset 2 with labeled attack data for initial model training. These adjustments streamlined the analysis.

Attack Data Completion. To address inconsistencies in the ATT_FLAG column, missing or ambiguous labels (e.g., -999 values) were completed using the attack scenario file, raising attack representation from 5.24% to 5.43% and improving dataset reliability.

Water Tank-specific Attack Flags. Tank-specific flags (T1_ATT_FLAG to T7_ATT_FLAG) were added to identify attacks targeting specific tanks, enabling granular analysis of whether anomalies in tank levels corresponded to targeted attacks, potentially informing more focused detection mechanisms.

Visuals and Analysis. Visual analysis of tank levels during attack periods did not reveal clear patterns linking anomalies to known attacks, reflecting the limitations of SCADA data, which attackers often manipulate to obscure disruptions. This highlights the challenge of relying on visual inspection for cyber-physical anomaly detection.

The figure below shows water tank levels (T1 to T4) versus attack periods (red histograms). Similar results were observed for tanks T5 to T7.



Classification Approach

In the classification approach, the objective is to distinguish attack states (ATT_FLAG = 1) from normal operations (ATT_FLAG = 0) using supervised learning techniques. Following the scikit-learn algorithm cheat sheet. This algorithm was selected due to its suitability for datasets with fewer than 100,000 samples and its proven efficiency with binary classification problems.

The Linear SVC was optimized using GridSearchCV to identify the best hyperparameters. The grid search identified the optimal parameters. Using these parameters, the Linear SVC achieved an accuracy of 97% on the test set. However, the results revealed significant limitations in attack detection performance:

- Precision for attacks (class 1): 89% – indicating that most predicted attacks were correct.
- Recall for attacks (class 1): 46% – meaning that the model missed more than half of the actual attacks.
- F1-score for attacks: 61% – reflecting the imbalance between precision and recall.

The recall of 46% highlights the difficulty in detecting attacks in a highly imbalanced dataset, where attacks represent only 5% of the entries. Despite the high accuracy, the model struggled to generalize well for the minority class, limiting its reliability for practical use. In the context of cyber-physical security, recall is prioritized over other metrics like precision or F1-score because the primary business objective is to identify all potential attacks. Missing an attack (false negatives) could have significant operational, financial, and safety consequences. While an increase in false positives (low precision) may lead to inefficiencies, it is preferable to err on the side of caution by ensuring that all potential threats are flagged for further investigation.

Further Directions

Despite extensive efforts to optimize supervised learning models, such as the Linear SVC (and other teste, such as KNeighborsClassifiers with similar results) with hyperparameter tuning and class balancing, the fundamental challenge potentially lies in the nature of the dataset itself. The SCADA data, while labeled with attack and normal states, is inherently unreliable due to potential manipulation during cyber-physical incidents. Attackers may alter sensor readings or system logs to obscure disruptions, rendering the labels less trustworthy and compromising the integrity of supervised learning approaches.

This inherent uncertainty in the dataset suggests that relying solely on labeled data may not fully capture the complexity of the problem, and even cause false interpretation (for instance, let’s imagine a value that has been hacked, and the ATT_FLAF = 1, then the algorithm will learn from this “false” value which has no meaning). As a result, transitioning to a statistical, unsupervised learning approach becomes a logical next step. By focusing on detecting anomalies without depending on labels, the following section explores how models like the Isolation Forest can leverage the system’s operational dynamics to identify deviations that may indicate potential attacks.

Statistical Approach - Unsupervised Learning

The objective of this section is to implement an anomaly detection method based on statistical modeling without relying on labeled attack data. For this purpose, I employed a tuned Isolation Forest model. This model was selected due to its efficiency in identifying anomalies in high-dimensional datasets and its ability to flag data points that deviate significantly from learned patterns. Unlike clustering-based methods (e.g., DBSCAN or K-Means), which assume homogeneous density or clusters, the Isolation Forest is particularly suited to identifying rare, sparse anomalies typical of cyber-physical attack scenarios. Moreover, its tree-based architecture scales well to datasets with diverse feature types and distributions, as is the case with SCADA data.

Feature Transformation: to capture the system's dynamic behavior, I included prior readings of each feature in the dataset, effectively incorporating stateful dependencies. This transformation allowed the model to account for both system state and recent trends, improving the detection of deviations from normal operations.

Results and Analysis: the model, with optimized parameters, achieved a recall of 68% for ATT_FLAG = 1, aligning closely with the recall of the supervised classifier. This result indicates that the unsupervised approach offers comparable detection capability, with the advantage of not requiring labeled data. However, the model's sensitivity to non-attack-related anomalies suggests that further refinement is needed to distinguish between routine anomalies and genuine attacks.

Conclusion

While it may seem unintuitive that the unsupervised model performed better than the supervised classifier, this outcome stems from the inherent challenges of the dataset. The attack labels are unreliable, as SCADA manipulation during attacks obscures true operational disruptions. This results in noisy or incomplete labels that degrade the performance of supervised models. Despite testing multiple supervised algorithms and hyperparameter tuning, these limitations persisted. By contrast, the unsupervised model focuses solely on detecting statistical deviations in the data, bypassing the need for labels. This allows it to identify potential attacks based on operational patterns rather than relying on potentially flawed annotations. From a business standpoint, given the challenges with label reliability in the SCADA dataset, the unsupervised approach hence appears more appropriate for this use case. It avoids dependence on flawed labels, ensures better generalization to unknown attack types, and aligns well with the goal of minimizing undetected attacks. Future efforts could explore hybrid approaches, combining the strengths of both methods to achieve more robust detection.

This project illustrates the complexity of detecting cyber-physical attacks in ICS environments, especially when data may be compromised by the attack itself. The limitations of these approaches also reveal the need for more sophisticated detection mechanisms. Moving forward, an **Industrial Intrusion Detection System (IIDS)** would provide a more robust solution, focusing not only on anomalies but also on specific attack patterns. Unlike the current **Industrial Anomaly Detection System (IADS)**, which relies on post-event detection, an IIDS could integrate threat intelligence and sequence-based models to identify attack tactics in real time. This progression from anomaly detection to intrusion detection reflects an essential evolution for enhancing ICS resilience against sophisticated, evolving cyber threats.