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

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

(*) Rapid Analysis and Model Prediction

Full version (10 pages): L. PIPIPITONE - RAMP Python Project - Batadral

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[1] (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. Key components include:

[1] 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)

- **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.
- **Attack Scenarios List** (Attacks_TrainingDataset2.csv): This supplemental 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.

These files provide readings on water levels, flow rates, pressures, and binary statuses of pumps and valves across various points in the system. Each variable is identified by a prefix: “L_” for water levels, “S_” for pump and valve statuses, “F_” for flow rates, and “P_” for pressures. By capturing both operational and attack data, the dataset allows for an in-depth analysis of system behaviors and potential indicators of cyber intrusions.

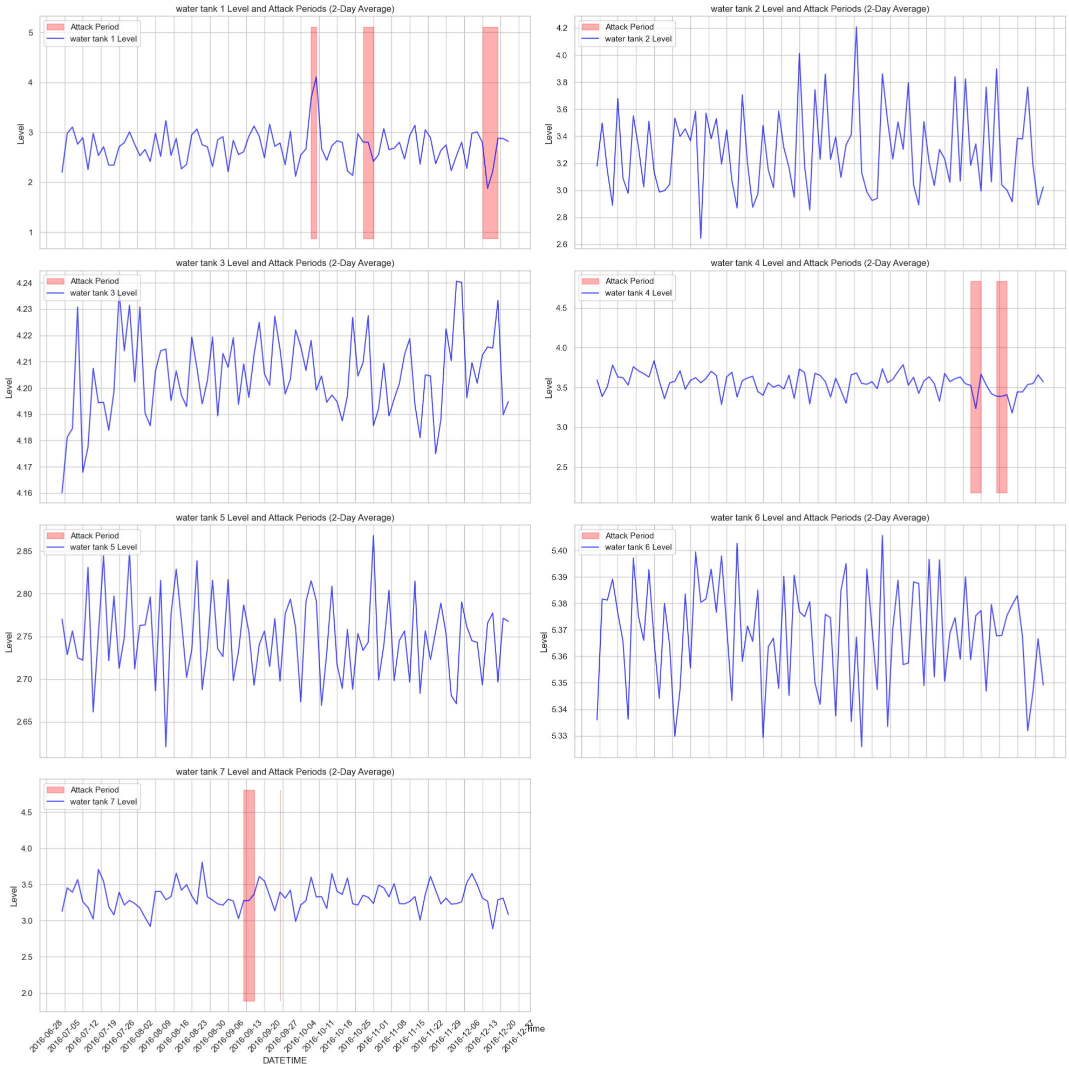
Simplification choices. To streamline the project, several simplifications were made. First, the CTOWN.INP file’s network structure was excluded, allowing algorithms to infer system interactions directly. Additionally, only Dataset 2 (including attack data) was used, focusing on labeled events for initial model training. These adjustments were made to keep the analysis focused and manageable.

Attack Data Completion. Initial inspection of the dataset revealed inconsistencies in the ATT_FLAG column, with some attack periods either missing labels or marked ambiguously (e.g., with -999 values). To ensure comprehensive and accurate labeling, I cross-referenced the dataset with the attack scenario file to complete missing ATT_FLAG entries. This adjustment increased the representation of attack data from 5.24% to 5.43%, thereby improving the dataset’s accuracy and reliability for subsequent analyses.

Water Tank-specific Attack Flags. To examine the relationship between individual water tanks and potential attack impacts, I generated additional flags: T1_ATT_FLAG through T7_ATT_FLAG. These tank-specific indicators denote whether each attack specifically targeted a particular tank. While the ATT_FLAG column provides a general indication of an attack, these specific flags enable a more granular analysis, making it possible to correlate water tank levels with targeted attacks. This approach allows us to evaluate whether anomalies in tank levels correspond to incidents targeting that particular tank, potentially informing more targeted detection mechanisms.

Visuals and Analysis. Initial visual exploration of the dataset sought to determine if observable patterns in tank levels or other parameters aligned with known attack periods. However, the visual analysis did not reveal clear or consistent correlations between attack times and distinct changes in tank levels. This outcome underscores the challenge of relying solely on visual inspection for detecting cyber-physical anomalies, particularly since SCADA data might be compromised during attacks. Attackers often manipulate SCADA data to obscure operational disruptions, thereby complicating the identification of attacks through straightforward pattern recognition.

This figure shows each water tank level (from 1 to 7) vs targeted attack periods (in red histograms)



Classification Approach

In the classification approach, I aimed to distinguish attack states (ATT_FLAG = 1) from normal operations using supervised learning techniques. However, the dataset's class imbalance—where only about 5% of entries indicate an attack—posed a challenge for conventional classifiers. Without addressing this imbalance, the classifier achieved an overall accuracy of 97% and a high precision for attack detection (93%), but recall for attack instances was low (56%), meaning many attacks were missed.

After adjusting for class imbalance, recall for attack detection decreased slightly to 50%, with minimal effect on overall accuracy. This outcome underscores the challenge of reliably detecting attacks in a highly imbalanced dataset. Despite balancing efforts, the classifier struggled to detect attacks consistently, suggesting that additional methods are necessary to enhance sensitivity to minority classes.

Results: Using optimized parameters, the classifier achieved a recall of approximately 68% for attack detection, a notable improvement yet still limited by data imbalance and possible SCADA manipulation.

Statistical Approach - Unsupervised Learning

The objective of this section was to implement an anomaly detection method based on statistical modeling without relying on labeled attack data. I employed a tuned Isolation Forest model to identify anomalies. By learning typical operational patterns, the model flagged deviations that could potentially represent cyber-physical attacks. To enhance its predictive power, I included previous system states and changes between consecutive readings, creating a stateful model that captures the dynamic nature of the water distribution system.

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

This project illustrates the complexity of detecting cyber-physical attacks in ICS environments, especially when data may be compromised by the attack itself. While the supervised classification model provided insights into key features, it struggled with class imbalance and required labeled data. The unsupervised anomaly detection approach, using an Isolation Forest, achieved similar recall performance without labeled attack instances, showing potential for detecting unusual behaviors indicative of attacks.
The limitations of these approaches 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.

To go further: competition high-level results by approach and more information.