Machine Learning and User Interface for Cyber Risk Management of Water Infrastructure

ABSTRACT

With the continuous modernization of water plants, the risk of cyber attacks on them potentially endangers public health and the economic efficiency of water treatment and distribution. This paper signifies the importance of developing improved techniques to support cyber risk management for critical water infrastructure, given an evolving threat environment. In particular, we propose a method that uniquely combines machine learning, the theory of belief functions, operational performance metrics, and dynamic visualization to provide the required granularity for attack inference, localization, and impact estimation. We illustrate how the focus on visual domain-aware anomaly exploration leads to performance improvement, more precise anomaly localization, and effective risk prioritization. Proposed elements of the method can be used independently, supporting the exploration of various anomaly detection methods. It thus can facilitate the effective management of operational risk by providing rich context information and bridging the interpretation gap.

Keywords: operational risk; ICS cybersecurity; visual analytics; anomaly detection; water treatment plants

1. INTRODUCTION

The importance of reliability of computer-controlled for water infrastructure is well established (Coulbeck and Orr, 1993). The increasing global scarcity of water has led many providers towards developing smart water systems. These cyber-physical systems and their networked industrial control systems (ICS) face an increased risk of cyber attacks similar to one that was recorded over a decade ago in Australia, where around 120k people were put in danger by the water contamination due to cyber attack (Slay and Miller 2007). Another partially successful attempt to poison water was recently made in the U.S., threatening the health of around 15k individuals (Robles Frances and Perlroth Nicole 2021). Cyber attacks also can target water meters to manipulate the billing system and cause economic loss (Hassanzadeh et al. 2020). These are not isolated incidents but rather examples of numerous conducted attacks. According to DHS, the water sector demonstrates the fourth-largest number of incidents in the US (ICS-CERT 2015). Therefore, from both economic and public health perspectives, developing improved methods to support cyber risk management for critical water infrastructure is an imperative given an evolving threat environment.

We approach this imperative from a data-driven cyber incident identification and characterization perspective. The detection of cyber attacks against ICS deployed in water facilities is a complex task that can rarely be solved using only automatic data mining methods. Numerous methods rely on anomaly detection as a technique to discover attacks against ICS (Ahmed et al. 2016). This is aligned with a precursor analysis approach to risk management, especially in the context of infrastructure systsems (Guo and Haimes, 2016), and involves monitoring and analyzing patterns, behavior, and events related to ICS assets to detect potential precursors to cyber attacks. Cybersecurity and forensics analysts face several challenges when exploring the massive number of records generated by a large number of heterogeneous ICS assets. These challenges include but are not limited to the scarcity of attack-related data, spatial-temporal characteristics of the collected data, and the interpretability challenges (Ahmed et al. 2020). The latter inevitably hinders the transition of anomaly detection methods to operations.

Existing methods in this area include control-theoretic approaches (Bou-Harb et al. 2017; Busby, Green and Hutchison 2017), various machine learning algorithms (Elnour et al. 2020; Li et al. 2019), and visual analytics (Kotenko et al. 2018; Lohfink et al. 2020). The first approach assumes the availability of a mathematical model of the system, which is often an impractical assumption. On the other hand, machine learning approaches can be scaled efficiently to support a modernized system or another ICS, which is a desirable quality in water facilities, where each deployed ICS can have its unique characteristics. Further, the values of visual analytics for network monitoring and classification, incident and malware forensics have been largely recognized by research and operational communities (Fischer et al. 2008; Wagner et al. 2015).

Although these research endeavors provide sound methods for attack detection in ICS realms, they rarely offer the essential strategy to attribute assets and quantify the impact, which is a critical function towards cyber risk management. Without the ability to evaluate and determine the impact of the attack on ICS assets, efforts to transition risk inference methods to operations are hindered.

To address these challenges, we propose a visually supported method that can be integrated into attack detection and forensics tools to guide domain experts through a set of tasks to examine a detected anomaly. To achieve this, we first employ a deep-learning architecture to fingerprint ICS behavioral patterns and detect abnormal behavior. The result from the inference engine is then passed to the visual analytics module that models domain-aware analytical reasoning and maps it to the appropriate visual techniques, views, and interaction. The expert user can then analyze inferred anomaly at a global level, decide whether it represents a false alarm, or look for more information at ICS asset level if it is impossible to reach a confident conclusion. Hence, we propose increased robustness of incident inference and analysis by including a human decision maker-in-the-loop through the use of visual analytics.

The rest of this article is organized as follows. In the next section, we briefly review related works and in section 3, we discuss the proposed method of analysis and outline the design components. In Section 4, we describe the experiment and report the result of applying our method to the data collected using a testbed emulating working water treatment system, and present preliminary insights regarding the water plant response to cyber attacks. Finally, in Section 5, we highlight the contributions of our method and discuss limitations and future work.

2. A BRIEF LITERATURE REVIEW

In this section, we highlight relevant risk management and cyber incident detection methods in ICS realms, as well as studies that leverage visual analytics for cybersecurity.

2.1 Risk Management in Water Infrastructure

Risk management studies related to water infrastructure vary in their focus, from the identification of threats to an assessment of their potential consequences. In a survey of the state of cybersecurity research in water systems over the past two decades, Tuptuk et al. (2021) found that research has predominantly focused on detection models and analyzing the impact of attacks. For instance, Davis et al (2014) developed an analytic framework to estimate the impact of contaminants in water distribution systems under varying network conditions. Other studies evaluate various threat management strategies (Zechman 2011). Pate`-Cornell et al. (2018) estimated the effectiveness of protection measures for critical infrastructure against the full spectrum of attack severity levels. While their cases do not include water infrastructure specifically, their example of power infrastructure has many relevant parallels to water infrastructure risk management. Moraitis et al. (2020) described an approach to quantify the impact of cyber-physical attacks on water distribution networks. Shin et al. (2020) proposed a resilience metric to measure the ability of water systems to withstand, adapt and recover from cyber attacks. The urgent need to protect water infrastructure is also reflected in the growing recognition by policymakers to strengthen the cybersecurity resilience (Kott and Linkov, 2021) of water infrastructure through legislation. For example, in the U.S., the America's Water Infrastructure Act of 2018 and the Internet of Things Cybersecurity Improvement Act of 2020, at the federal level, and some legislation at the state level, are focusing on strategies for securing water infrastructure (You, 2022). Further, critical systems such as water are closely linked to social order and wellbeing. Emerging ideas in systems engineering of complex system of systems (Haimes, 2018) and social systems engineering (Scalco and Palmer, 2022) have potential to provide cyber and safety assurance as we continue to develop interconnected and interdependent infrastructure systems.

2.2 Cyber incident detection in Industrial Control Systems

A cyber incident renders abnormal behavioral patterns, and numerous inference methods are defined as anomaly discovery problems and leverage control-theoretical and machine-learning approaches.

The former method converts ICS architecture and process to a mathematical model and exploits it for incident identification and investigation. Some notable works include (Bou-Harb et al. 2017; Busby, Green and Hutchison 2017; Mo, Weerakkody, and Sinopoli 2015; Pasqualetti, Dörfler, and Bullo 2013; Chabukswar, Mo, and Sinopoli 2011; Henry and Haimes, 2009; Khanna and Liu 2008). Yet, the ICS realm challenges the widespread application of control-theoretic approaches. For instance, ICS can have an inconsistent structure, depending on their type, technology, and continuous modification and optimization level. Therefore, while rendering superior results, using the identical model-based attack detection technique for ICS deployed in different settings seems impractical. Moreover, control charts such as Multivariate Cumulated SUMs (Woodall et al. 1985) and Multivariate Exponential Weighted Moving Average (Lowry et al. 1992) monitor the mean and variance of a time series over time to detect abrupt changes in its statistical behavior. Coupled with machine learning, control charts can enable proactive detection and response to cyber threats.

Accordingly, the researchers consider data-driven methods to generalize the underlying system by employing deep learning algorithms operating on circulating ICS data. However, due to privacy and security issues, empirical data is rarely available. Moreover, these techniques require high computational power, while velocity and variety of data require advanced techniques to extract valuable insights. For instance, the methods that approach incident detection from data gathered by ICS sensors should consider multivariate time series. Prominent works herein include the application of Deep Neural Network (DNN) and Support Vector Machine (OSVM) (Inoue et al. 2017), and Dual Isolation Forest (DIF) (Elnour et al. 2020). Anomaly detection using Generative Adversarial Networks (GANs) are employed across infrastructure sectors such as power plant (Choi et al. 2020), water treatment (Li et al. 2019.) and distribution (Du et al. 2021), in-vehicle network (Seo et al. 2018), to name a few. However, most works omit bridging captured anomalies with the ICS assets that significantly contribute to it, while it is an essential function toward cyber risk management.

2.3 Visual Analytics

A most relevant to our work direction of visual analytics in decision support is to visualize anomaly and multivariate data over time. A classical design to display anomalous trends in temporal data is statistical diagrams such as line charts, and histograms (Laskov et al. 2005). However, multiple lines in the same space reduce anomaly visibility; therefore, a glyph-based design has gained more popularity due to its practical usage of screen space. For instance, it captures the behavior of individual users based on their communication activities over time (Cao et al. 2015); it is used for computer network monitoring (Kotenko et al. 2018). Another notable technique - spiral plots - demonstrated its promising application for monitoring ICS assets (Lohfink et al. 2020) over time and allowing visual analysis of individual assets. Further, dashboards, similar to one proposed in (Bakirtzis et al. 2018), promote interactive security analysis to provide different views largely centered around the system and its associated attack vector space.

In contrast to available contributions that allow visual analysis of detected attacks at individual assets, we offer an effective tool to explore anomalies simultaneously at all system levels: from generalized information for the entire system to business process and ICS asset level. The proposed model reasons the anomalies by allowing dynamic visual exploration and providing rich context information valuable for risk management.

3. DEVELOPMENT OF THE METHOD

This section first contextualizes the challenges of data-driven risk management methods. It then derives desired properties for such methods, followed by a detailed design of the proposed method.

3.1. Application Domain and Challenges

We consider a risk as a triplet $R = \{\langle s_i,\ p_i,\ i_i \rangle\}$ (Kaplan and Garrick 1981), where s_i is an undesirable scenario identification, p_i denotes a probability of the scenario, and i_i is an outcome of an adverse scenario. Contextualized in the ICS cybersecurity domain, a set of triplets can be identified from past events, simulations, and the detection of ongoing cyber incidents and their digital forensics.

3.1.1. Cyber Threat Model

An analysis of the cyber incidents in critical sectors that combines those that reported in (Hemsley and Fisher 2018) and in an open-source literature renders the following observations regarding threat scenarios s_i and outcomes i_i (Table I). We further confirmed the consistency of our observations with a public knowledge base ATT&CK for ICS (Alexander, Belisle, and Steele 2020).

Table I. Cyber attack scenarios and their outcomes

Identifier i	Cyber attack scenario s_i	Cyber attack outcome i_i
1	An attacker injects false measurements, which are dynamically calculated by ICS during usual operation process	An incorrect response of the PLC and lead undesired system state It can cause the reduced/elevated chemical injection and reduce the quality of purified water Undesirable state can lead to property damage and threaten safety.
2	An attacker can inject false measurements without violating the control-flow integrity of ICS (evasion)	All consequences of the scenario 1, without ability to detect the attack promptly. An operator can make incorrect decision based on corrupted data.
3	An attacker sends a direct command to the actuator to maliciously manipulate of the current state	Overflow/underflow in the tanks can cause the waste of resources. Damage of the critical components of ICS. Reduced amount of distributed potable water.
4	An adversary can send command messages to perform actions outside of their intended functionality	Activated system response outside the conditions and boundaries (open water pumps, activate alarms, etc.)
5	An adversary can plant malware to encrypt or delete critical data	Loss of data availability. Loss of the ability to perform intended functions. Financial loss due to inability to provide resources.

When an attacker has opportunity, capability, and motivation to conduct the attack, the latter can be considered probable. ICS nowadays are connected to the network, providing plenty of opportunities for the attack to gain remote control over their operations. There is also an abundance of capability out there. The tools to accomplish attacks are available on the dark web at large, making it possible to gain access to them at little cost (Samtani, Chai, and Chen 2022). Moreover, many attacks are conducted by using the access permissions of current or former employees. Some incidents aim to demonstrate attacker capabilities; another is an act of retaliation by disgruntled employees (Slay and Miller 2007); it can also aim to disrupt another country's critical infrastructure (Case 2016). Given the above considerations and the mounting number of attempts to gain control over the critical infrastructure, we regard cyber attacks in this area as highly probable and concentrate efforts in this work on identifying the incidents s_i and their impact i_i.

3.1.2 Challenges of Data-driven Approach

A risk management aims to focus the efforts on the significant threat scenarios and use appropriate techniques to convert data into valuable analytics for decision-makers (Bier 2020). As we define cyber incident detection as an anomaly inference problem and approach to the solution from a data-driven perspective, it is essential to acknowledge the corresponding challenges. Visibility and spatio-temporal characteristics of data generated by ICS introduce unique challenges for attack detection methods that solely rely on empirical data (S. Wang, Cao, and Yu 2020). Further, mapping data anomalies to the actual incident and confirming it is non-trivial due to the inherent dependencies between ICS devices (Ahmed, MR, and Mathur 2020). Moreover, given the high number of connected components, it is imperative to prioritize the remediation of cyber crises and further development by addressing the affected ICS assets first. Besides, we observe a lack of techniques to convert the results of anomaly detection methods to narratives that allow decision-makers to interpret the outcomes (Neshenko 2020).

Data availability. The primary reason for the lack of available data is its intrinsic sensitivity. Any leakage of the operational patterns can be used for crafting highly stealthy attacks that can lead to catastrophic consequences on operations of critical infrastructure and the human population. Another cause of data scarcity is the rarity of reported attacks.

Spatio-temporal data. ICS data contains spatial and temporal attributes with complex correlations. For instance, water distribution can show a cyclical pattern in time depending on usage by consumers during the day and nighttime. Extracting insightful patterns from spatial datasets introduces additional challenges due to the complexity of spatial data types and relationships (Rashid 2012). Overlooked relations, however, can lead to inadequate attack detection accuracy and localization.

Model scalability. Depending on their type, ICS deployed in various realms can have an incomparable structure. For instance, domestic and industrial wastewater treatment requires different filtering and disinfecting technology and advanced treatment such as oil separation and removing toxic dissolved organic ingredients, to name a few (EPA n.d.). Furthermore, ICS assets are prone to continuous modification demanding advanced scalable methods that can be directly employed for different systems.

Attack localization. Nonlinear and dynamic nature of water facilities, particularly water plants, causes evolving data characteristics at different levels of ICS process. In addition, attack techniques constantly evolve to evade detection. In this context, data-driven methods, which explore behavioral patterns of a single element of the system, explicitly detect irregular behavior of the underlying asset; however, they fail to detect the coordinated incidents that affect whole system. At the same time, the techniques that engage multivariate data display a high level of false localization. Therefore, a balance between different perspectives to reduce false alarms and enhance operating efficiency is paramount need.

Incident investigation. Empowering risk managers to drive consistent investigations and make more precise conclusions is one of the crucial considerations for transitioning the research to operation and must be noticed (Sommer and Paxson 2010). Unlike traditional rule-based or knowledge-based approaches, machine learning algorithms are designed to automatically learn from data and make decisions based on this learning. The latter results in highly accurate conclusions but can lead to a need for more transparency in decision-making. The proper incident investigation curated by the system will reduce dwell time, increase efficiency, and provide actionable insights.

3.2 Required Properties

Well-designed cyber risk management for ICS deployed in water plants should address mentioned challenges, support accurate inference and interpretation of the detected threats. We define the following required properties to meet this goal.

Property 1: Ability to discover boundaries and dependencies between ICS assets without previous knowledge about its process model.
Property 2: Ability to illustrate the status of the ICS system at each process level and individual asset, for the defined period.
Property 3: Ability to display incident alarms with relevant information, including detection results, impact, and data overview.
Property 4: Ability to display special warnings for the critical (predefined) ICS assets should an incident affect them.
Property 5: Ability to evaluate the potential impact of the disruption and display it as an alarm.
Property 6: Ability to allow configuration, including the maximum acceptable level of incident impact parameters.
Property 7: Ability to facilitate of dynamic visual data comparison.
Property 8: Ability to support browsing of relevant raw data.

Table II associates the properties to the corresponding challenges.

Table II. Mapping system properties to challenges

Challenges	Requirements
Data availability	Property 1
Spatio-temporal data	Property 1
Model scalability	Property 1
Attack localization	Property 2, Property 3
Incident investigation	Property 2 – Property 8

3.3 Design of the Method

As numerous techniques are proposed to improve cyber resilience, including those that focus on adaptive response, analytical monitoring, and dynamic representation (Ross et al. 2019), these techniques and the imperative need for cyber resilience quantification become a design foundation for our approach.

In summary, the proposed approach (Fig. 1) consists of four core segments, which incorporate machine learning methods and interactive visualization (together known as visual analytics) to infer cyber incidents and their potential effect on directly and indirectly connected components of water treatment plants.

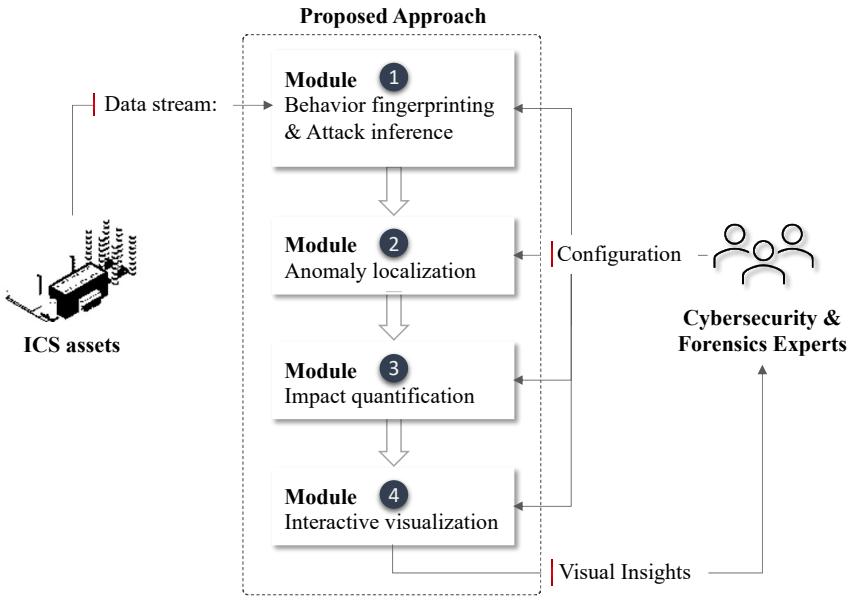

Fig 1. The proposed method to assist cyber risk inference for water treatment plants

As we define cyber incident inference as an anomaly discovery problem, the first module employs deep learning architecture to explore the boundaries and dependencies of the system and estimate the anomaly score. A score that exceeds a predefined threshold indicates the potential attack. Further, the anomaly localization module derives the exploited assets operating a belief function theory and various feature importance algorithms. Subsequently, the module "impact quantification" estimates the potential production loss that would indicate the severity of the incident impact. The role of cybersecurity and forensics experts is to *(i)* determine the initial parameters of the system and *(ii)* examine the validity of the output. To support evidence-based decision-making, module 4 maps domain-specific analytical strategies and results from the inference model to appropriate visual techniques, views, and interactions. We now adopt a granular perspective, highlighting the architectural design of each component.

3.3.1. Behavior Fingerprinting and Attack Inference

To discover boundaries and dependencies between ICS (*Property 1*) and infer incidents (*Property 2*), we employ a generative and discriminative deep learning architecture known as a Generative Adversarial Network (GAN), inspired by the architecture proposed in (Donahue, Krähenbühl, and

Darrell 2016). The framework utilizes three deep learning networks: discriminator D, generator G, and encoder E; each plays a distinct role in the training and continuous working process.

Let x_i represent the physical measurement of the ICS sensor (or state of the ICS actuator) i; $x = \{x_1, \dots x_m\}$ stands for an attribute vector, where m is the number of ICS assets (variables). During the training phase, network E receives attack-free behavioral vector x and maps it to latent variable space. Simultaneously, the network G models the data distribution as a fixed latent synthetic attribute vector $G(z) = \{x'_1, \dots x'_m\}$. The network D then receives the tuples $(x, E(x))$ and $(G(z), z)$, discriminates them, and assigns the labels $P(y) = \{0,1\}$, where 1 is a label for "real" data, and 0 is for "generated" data. To improve the accuracy of generated attribute vector, the error is back propagated to the network G and the training process repeats.

This architecture offers the salient capability for anomaly detection. First, trained network D distinguishes the normal operational behavior of ICS. The network D takes the incoming attribute vector and extracts the distribution between fingerprinted behavioral patterns and a latent representation of incoming vector $E(x)$. It then returns the loss $\mathrm{L}_{D(x)}$ calculated based on sigmoid cross entropy (Eq.(1)).

$$\mathrm{L}_{D(x)} = cross\_entropy\big(D\big(x, E(x)\big), 1\big) \quad (1)$$

Second, network G generates a synthetic attribute vector with the same probability distribution as vector x. The difference between incoming and expected ICS behavioral patterns is calculated as a distance (l_1-norm) between actual and reconstructed instances and represents a residual loss $\mathrm{L}_{R(x)}$ (Eq. (2)).

$$\mathrm{L}_{R(x)} = \| x - G(E(x)) \|_1 \quad (2)$$

The significant loss value indicates the large difference between the incoming and expected behavioral vectors. Further, an anomaly score is obtained as a weighted linear combination of $\mathrm{L}_{R(x)}$ and $\mathrm{L}_{D(x)}$ (Eq. (3)).

$$a\_sore = (1 - \alpha)\mathrm{L}_{R(x)} + \alpha \mathrm{L}_{D(x)} \quad (3)$$

where α is a weighting parameter indicating the priority of the loss. If $a\_score \geq \theta$, where θ is predefined severity threshold, the detected cyber incident requires attention of cyber operator.

3.3.2 Anomaly Localization

To mitigate uncertainties in detecting the location of anomalies, this module approaches the anomaly detection problem from two distinct perspectives: *(i)* multivariate and *(ii)* individual.

Multivariate perspective. Feature importance algorithms can serve as a medium for anomaly localization for the methods that solely rely on empirical data (Taormina et al. 2018a). The estimations of employed feature importance methods are recorded as $f = \{f_1, \dots f_j\}$, where $f_j = \{f_j^1, \dots f_j^m\}$ denotes a vector representing importance of attribute m for detected anomaly.

To render better anomaly localization, we combine the independent outcomes produced by each method. To achieve the latter, we use a degree of belief (b_j) that technique j precisely estimated the relevance of the ICS assets to the anomaly. To this end, relative support (Yong et al. 2004), RS, representing a distance between the methods' estimations, is used to estimate the degree b_j.

$$b_j = RS(f) / \sum_{j=1}^{k} RS(f_j) \quad (4)$$

$$RS(f_j) = \sum_{i=1, j \neq i}^{k} (1 - d(f_i, f_j)) \quad (5)$$

Finally, we record a score s_m denoting a degree of relevance of the variable m to the anomaly.

$$s_m = \sum_{j=1}^{k} b_j \cdot f_m \quad (6)$$

If s_m is in the 75th percentile, we declare asset m as an asset that is affected by the incident. The rationale behind the selection of the 75th percentile is as follows. We aim to determine a threshold that enables the identification of sensors under distress and minimizes false negatives**.** To this end, we empirically evaluated various percentile thresholds and found that the 75th percentile provides the optimal result. Our determination is critical for identifying affected sensors and can contribute to developing reliable systems.

Localization based on individual variable. Inherent dependencies among ICS prevent the precise identification of exploited assets. To filter out the potential false positives, we verify the abnormalities in the measurements of individual assets. To this end, $z - score$ for variables with s_m in 75th percentile is calculated. Formally, it is defined as

$$z - score_m = \frac{a\_score(x) - \mu}{\sigma} \quad (7)$$

where μ is the mean (Eq. (8)) and σ is a standard deviation of data during the time window (Eq. (9)); N stands for a number of datapoints in the respective time window.

$$\mu = \frac{1}{N} \sum_{i=1}^{n} x_i \quad (8)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2} \quad (9)$$

Datapoints with $z - score_m$ that is beyond three standard deviations of a given data sample is classified as abnormal (Aggarwal 2017) and as a potential false positive otherwise.

Final localization. We further obtain a final localization score l_m for each ICS asset m by employing weighted linear combination of s_m and $z - score$ (Eq. (10)) .

$$l_m = (1 - \alpha) \cdot s_m + \alpha \cdot z - score_m \quad (10)$$

where α is a weighting parameter governing the effect of each perspective on the final localization score. A larger score l_m highlights the variables representing ICS assets, whose measurements influence the anomaly score, therefore, portrays the affected assets.

3.3.3 Impact Quantification

To quantify the adverse effect of the cyber incident s_i (*Properties 3 and 5*), we estimate performance loss P_{loss} as a function of production and quality loss (Wei and Ji 2010), Pr_{loss} and Q_{loss}, respectively.

$$P_{loss}(s_i) = f(Pr_{loss}, Q_{loss}) \qquad (11)$$

Although the attack intentions vary, in this work, we focus on quantification of the threat of malicious regulation of water levels in tanks. To enrich the incident investigation and characterization for the production loss, we employ the indices suggested in (Taormina et al. 2017.) We employ the limited number of metrics to evaluate the feasibility and effectiveness of the designed system before comprehensive implementation.

The first index, T_{under} assesses the amount of time during which an attacked asset led to tank underflow:

$$T_{under} = \sum_{t=1}^{T} l_t \Delta t \qquad (12)$$

where l_t is an indicator defined as follows.

$$l_t = \begin{cases} 1, & h_t < l, \\ 0 & otherwise \end{cases} \qquad (13)$$

where h_t is a water level of the attacked tank, l is lower acceptable water level.

Similar index, T_{over}, is employed to estimate the amount of time during which an attacked ICS asset lead tank overflow:

$$T_{over} = \sum_{t=1}^{T} l_t \Delta t \qquad (14)$$

where

$$l_t = \begin{cases} 1, & h_t > u, \\ 0 & otherwise \end{cases} \qquad (15)$$

and u is upper acceptable water level.

Further, T_{under} and T_{over} are compared with predefined threshold θ_u and θ_o, to assign a corresponding level of the severity of the production loss.

$$severity_{under} = \begin{cases} 1, & T_{under} < \theta_u, \\ 0 & otherwise \end{cases} \qquad (16)$$

$$severity_{over} = \begin{cases} 1, & T_{over} > \theta_o, \\ 0 & otherwise \end{cases} \qquad (17)$$

The respective mitigation procedure should be employed for the cases with severity level 1.

3.3.4 Visual Analytics

The result from the inference and impact quantification modules is passed to the interactive visualization module, where the domain expert can analyze inferred anomalies. To achieve this, we use visual analytics (VA), a transdisciplinary field involving several components that maximize the human capacity to perceive, interpret, and reason complex data and events. These components

include *analytical reasoning* to provide insights that support decision-making, *visual interaction techniques* to enable the interpretation of extensive data, *data transformation methods* to promote analysis, and various *methods for result dissemination* to communicate analytical results to diverse audiences (Thomas and Cook 2006).

Some unique aspects of analysis in the ICS settings include *(i)* inherited dependencies among subsystems and individual assets of ICS. It signifies that monitoring individual ICS assets is insufficient. There is a need to *(ii)* separate the facts from false alarms. Finally, it is essential *(iii)* to generate insights into the impact on the physical environment. To this end, the following strategies and their respective visual representation support the VA for anomaly investigation in ICS operations.

Strategy 1: Comprehensive monitoring. Situational awareness and preliminary triage of anomalies in ICS assets require simultaneous monitoring of many connected assets. Essential inquiries for a complete understanding of an incident may include but are not limited to: Which ICS assets are exploited and to what extent? What type of exploitation is taking place? Which incident requires immediate attention? How quickly can the system recover from an attack? *(Property 2)*

To this end, the *detection view* (Fig. 2) shows the distribution of anomalies across ICS in the feature space for a selected period.

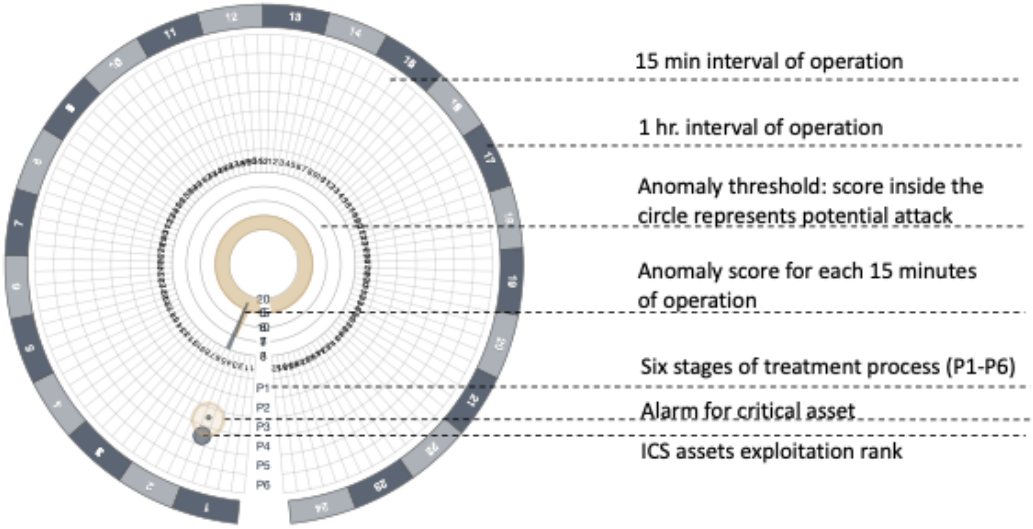

Fig 2. A layout of detection view

The view is a multilevel circular layout, each level of which represents the system process level. The circle is divided into 96 regions expressing 15 minutes intervals of operational hours; every four intervals grouped in one-hour periods to clearly communicate the cyber state of ISC at each point in time. The anomaly score is represented as a bar in the inner circle, and the anomaly threshold is a highlighted area (*Property 6*). The position of the bar in the shaded area indicates the anomaly and requires investigation. In the process level context, the visual model shows anomaly ranking for ICS assets for each hour. The representation also includes the confidence level that the attack has affected each ICS asset. Moreover, most critical assets are clearly defined to capture the immediate attention of cyber analysts or investigators. To this end, we employed the

pulsation technique and the color like the threshold to stress the importance and separate anomaly score (*Property 4.*)

Strategy 2: Inspection at process level. The *inspection view* (Fig. 3) allows a closer look at the details of the anomaly score chosen from the detection view.

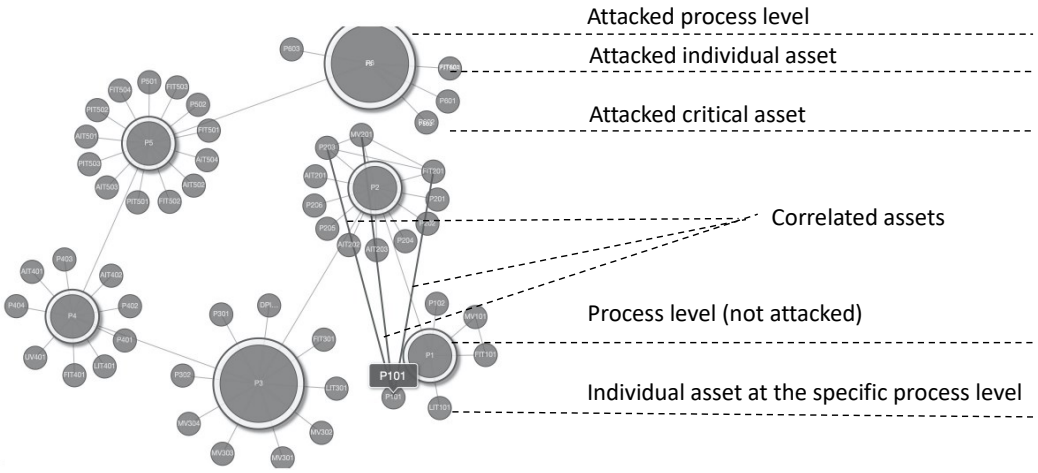

Fig 3. A layout of inspection view

Specifically, the view is a network graph, the nodes of which symbolize ICS assets. Central nodes represent ICS process levels and consolidate their elements. The edges that connect various assets depict the correlation between individual sensors and actuators. The graph illuminates the edges only between strongly correlated assets. This representation provides supplementary information for investigating threat impact. The dark color contrasts the attacked assets, while the pulsation lightens the most critical attacked assets to properly prioritize the analysis (*Properties 2, 4, and 5.)*

Strategy 3: Prioritization. By selecting an individual asset from the detection or inspection view, the expert navigates to the ranking view (Fig. 4), which provides numerous ranks that support evident-based prioritization for remediation and risk management.

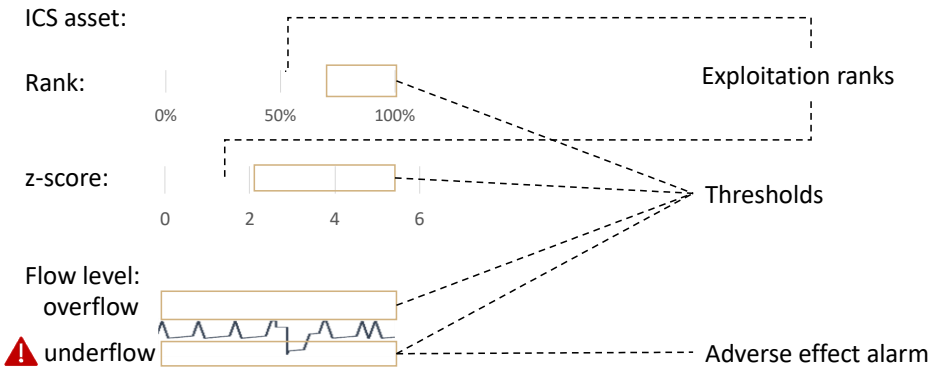

Fig 4. A layout of ranking view

The *ranking view* brings attention to the following rank types for ICS asset investigation:

- a degree to which ICS asset contributes to the anomaly given the relationship between system components (estimated as s_m by the localization algorithms)
- degree of abnormality at the individual asset level ($z - score$)
- an impact of the attack on ICS operation (estimated based on predefined indices)

The shaded areas represent predefined thresholds: if not meaningful, the rank value will not hit this area (*Properties 3-6*).

Strategy 4. Inspection at individual level. To investigate anomalies in the individual ICS assets or confirm the regular operation of the respective asset, the *raw data* view (Fig. 5) illustrates measurements collected by a specific ICS asset at a selected time and visually compares them to regular system operations obtained from the fingerprinted behavior (*Properties 2, 7 and 8*).

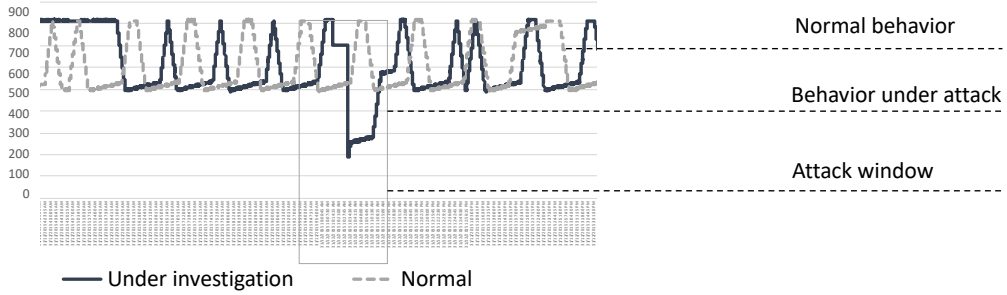

Fig 5. A layout of raw data view

The view illuminates the attack period (based on an anomaly score higher than the threshold) so that the investigator can adequately observe the difference or the absence thereof (false localization.)

Table III maps analytical strategies and coordinated views to functional requirements and visual techniques.

Table III. Mapping analytical strategies and required properties to coordinated views and visual techniques

Strategy	Required properties	View	Visual technique	Description
Comprehensive monitoring	Properties 2, 4-6	Detection	Radar diagram	24 hrs of ICS operation aggregated by 15 minutes and six treatment processes (P1-P6)
			Bar chart	Anomaly score for each 15 minutes of operation
			Pulsation	Displays warnings for critical ICS assets defined in settings
			Bubble chart	ICS assets exploitation rank aggregated by process level (P1- P6)
			Shaded area	Anomaly threshold: score inside the shade represents potential attack
Inspection (ICS level)	Properties 2, 4, 5	Inspection	Force graph	Display ICS assets at six treatment processes (P1-P6)
			Color	Indicates the exploited ICS assets
			Pulsation	Displays warnings for critical ICS assets defined in settings

Prioritization	Properties 3-6	Ranking	Bar diagram	Displays rank of abnormality
			Line chart	Shows the measurements/state of selected ICS asset during the attack
			Attention icon	Indicates the adverse impact on ICS assets
			Text	Displays estimated impact indices
			Shaded area	Thresholds: ranks, score, and data inside the shaded area indicates the exploitation or its negative effect
Inspection (individual level)	Properties 2, 7, 8	Raw data	Shaded area	Highlight the period of operational disruption
			Line chart	Shows the measurements/state of selected ICS asset under normal operation
			Line chart	Shows the measurements/state of selected ICS asset during the attack

The dynamic nature of the visual module enables filtering data and examining individual ICS assets at the lowest data level, facilitating incident investigation from general to specific characteristics. Each view can be used separately, allowing for hypothesis testing and identifying false negatives while generating collective knowledge.

4 APPLICATION OF THE METHOD TO WATER TREATMENT

4.1 Dataset

We evaluated proposed approach using data collected by a small-scale water treatment plant that encompasses six stages of the water treatment process (P1 through P6) controlled by a dedicated PLC (Fig. 6) (Goh et al. 2016). The sensors, connected over the network, accumulate the water level and water flow by interacting with the physical environment.

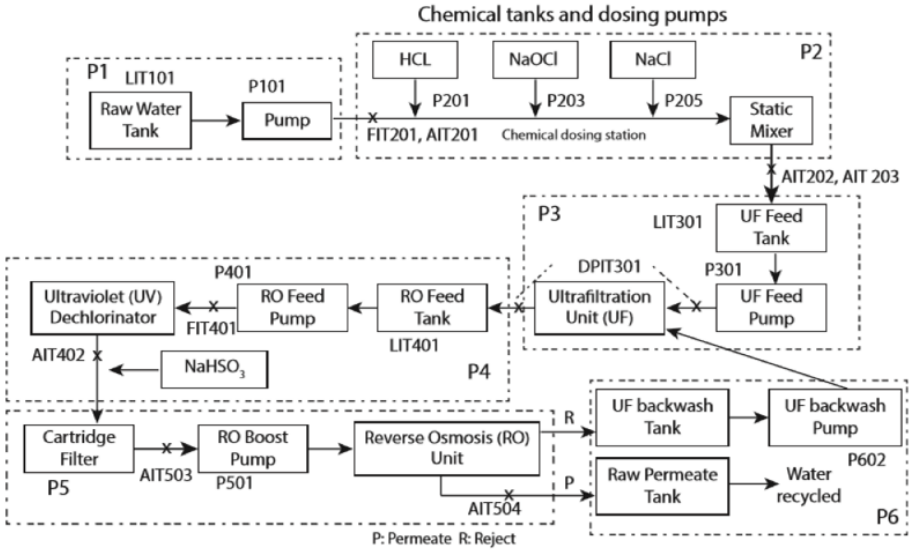

Fig 6. SWaT testbed (Goh et al. 2016) process overview. ICS consists of six stages of the treatment process (P1 though P6).

The dataset retains sensor measurements (25 continuous variables) and actuators' states (26 discrete variables). It constitutes seven days of attack-free operation and four working days under different cyber attacks.

4.2 Threat Model

The following classes of the attacks are carefully labeled in dataset.

- Single Stage Single Point (SSSP) attack targets exactly one ICS asset
- Single Stage Multi Point (SSMP) attack aims at several ICS assets deployed on one process level
- Multistage Single Point (MSSP) attack is performed on multiple process levels and targets exactly one asset at each level
- Multistage Multi Point (MSMP) attack is performed on two or more ICS process levels and targets multiple assets at each of them

The attack intentions vary from overflow/underflow of the tank, reduced water quality, and system malfunctioning. The duration of these attacks varies from 100 sec to 10 hrs. While system requires time to restore normal operation after an attack, the exact time for such recovery is not provided in the dataset description.

4.3 Results

To demonstrate the strengths of the proposed approach, over further discussion, we consider data related to an operational day in which all types of attacks (described in Section 5.1.1) have been administered. In this operational day, the detection algorithm identified three incidents with the high recall; in addition, two abnormalities were detected with the low recall level. The comparison with existing literature is summarized in Table V.

Table V. Performance (recall) across different anomaly inference models. A, B, C, D, E denote the attack scenarios (in the time order) administrated in the selected day

Scenarios	(Inoue et al. 2017)	(Elnour et al. 2020)	(Lin et al. 2018)	(Xie et al. 2020)	(C. Wang et al. 2020)	This work
A	-	-	0.20	1.00	-	0.97
B	0.94	1.00	1.00	1.00	0.98	1.00
C	-	1.00	-	-	-	1.00
D	-	-	1.00	1.00	0.96	0.60
E	-	1.00	-	0.21	-	0.24

The anomaly localization function takes the result of the attack inference module as a set of measurements indicating the probability that the system is exploited at a particular time. As the localization algorithm allows a combination of several methods, for validation purposes, we selected the following feature importance techniques: Classification and Regression Trees (CART) (Breiman et al. 2017), Logistic Regression (Kleinbaum et al. 2002), and Shapley values (Shapley 1953). We contrast in Table VI the attack localization performance of the proposed approach and those reported in (C. Wang et al. 2020) and (Shalyga, Filonov, and Lavrentyev 2018).

Table VI. A comparison of attack localization across different methods

Model	Precision	Recall	F-measure
This work	**0.39**	**0.52**	**0.45**
(C. Wang et al. 2020)	0.30	0.43	0.35
(Shalyga, Filonov, and Lavrentyev 2018)	0.22	0.21	0.21

The proposed method demonstrates the improved performance over the available methods.

Table VII provides a deeper look at the classification of the attack rendered by the detection engine.

Table VII. Deeper look into inferred attack scenarios

Scenarios	Attack window	Process	Exploited asset	Attack type
A	15 min	P3, P4	P302, LIT401	MSSP
B*	9.5 hrs.	P3, P4	P302, LIT401, FIT401	MSMP
C	2 min	P2	P203, P204	SSMP
D**	19 min	P1, P2	P101, MV201	MSSP
E***	6 min	P1	MV101	SSMP

Table notes: *An attack log indicates only pump P302 as exploited, rendering original attack class SSSP. The discrepancy has occurred due to the adverse effect of the attack on the water level in the tank. **The anomaly localization method did not classify LIT101 as an asset under attack, rendering the attack class MSSP, while the original scenario is classified as MSMP. ***The anomaly localization method did not identify the attacked asset.

The classification demonstrates discrepancies with the attack log. In scenario B, for example, the attack log indicates only pump P302 as exploited, suggesting attack class SSSP. However, the incident affected the water level in the tank of the subsequent process; therefore, the localization algorithm codes level indicator LIT401 as under attack. Incident A may exacerbate this impact, as previously suggested.

The final result of the incident inference and localization is passed to a visual analytics component in the form of JSON files. Concurrently, the visualization module receives operational performance indices to convert them into visual representation to focus the attention of the risk manager.

In the visual module, the user employs the analytical strategies by navigating through the coordinated views depicted in Fig. 7 and configuring parameters (detection model, critical assets, thresholds) as needed.

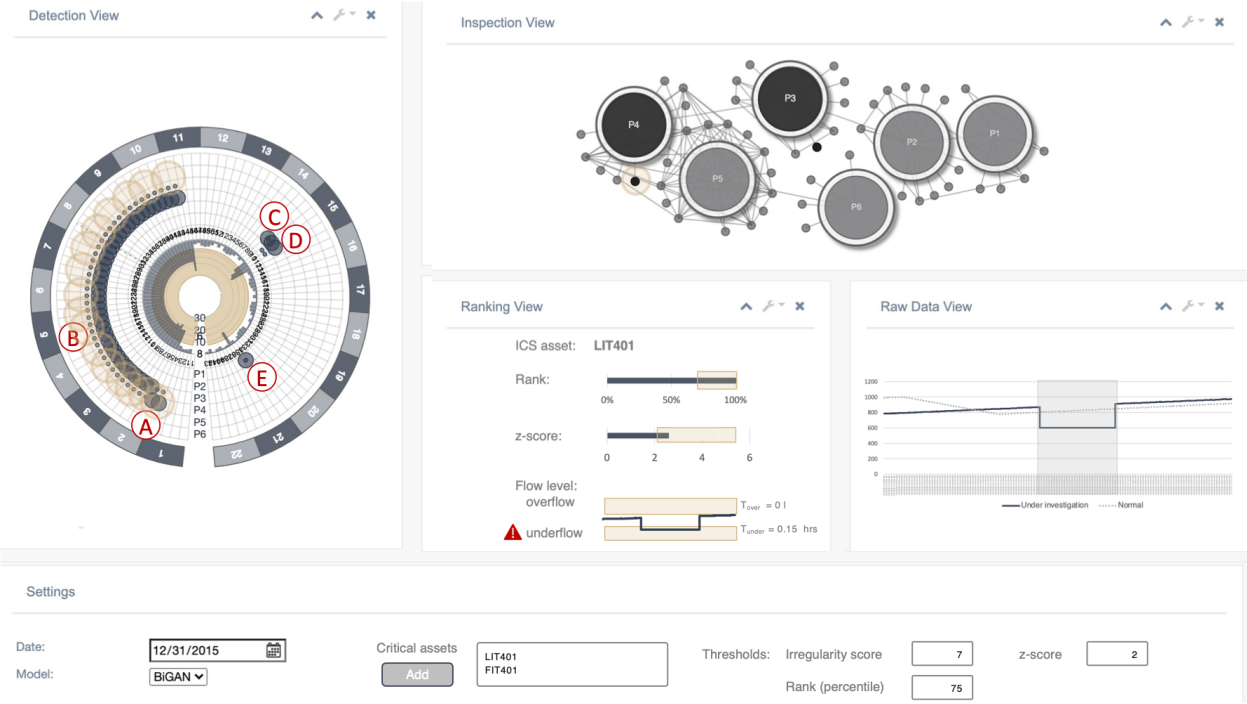

Fig 7. Interactive analytical dashboard. A, B, C, D, E denote cyber incidents. The dashboard illustrates the scenario A with a level indicator LIT401 under investigation

The *detection view* (Fig. 7) illustrates 24 hours of the system under investigation. An anomaly score (bar in the inner circle) for five scenarios appears inside the highlighted area, indicating alleged attacks. For further reference, we denote these incidents as A, B, C, D, and E. The duration of the incidents varies from short for A, C, D, and E to lengthy (more than 9 hrs.) for incident B. The *detection view* shows the attacked ICS assets (bubbles at the process levels P1-P6) for each incident. The bubble size indicates the rank (percentile) that the localization algorithm assigned to a particular ICS asset. The pulsation technique indicates that critical assets listed in the settings view are involved in incidents A and B. Noble, the attacks have not targeted process levels P5 and P6. Further, since incidents A and B appeared one after another and affected the same process levels, P3 and P4, it is reasonable to expect the trace of incident A in the following incident B. Same is relevant for attacks C and D.

For illustration purposes, we provide visual insights for one scenario since it consists of the critical sensors defined by an expert.

Incident A. Figure 7 illustrates a detailed view for the detected attack A (only one sensor is detailed for illustration purposes). In this attack scenario, the measurements of the water level indicator are maliciously set to the level below minimal and kept this way through the incident duration. At the same time, pump P302 kept on, allowing continuous actual water flow. The *inspection view* reflected this incident and pinpointed that the critical asset defined in the settings is exploited.

Ranking view. Both multivariate and individual (z-score) ranks reached the shaded area (the threshold), confirming the attack localization. The flow level indicator demonstrates the signs of the underflow: the line plot reached the shaded area; the index T_{under} confirmed this assertion. In

the *raw data view*, the grey shaded area visually segregates the attack window, obtained from an irregularity score centered around the same assets and time. It further visually compares the actual and expected behavior of the ICS asset, showing the sign of an ongoing incident considering a significant apparent deviation.

4.4 Usability test

To evaluate the capability and usability of the visualization module to support threat identification and initial analysis of the incident impact, we asked a focus group of six participants to apply defined analytical strategies and complete questionnaires assessing the success of the investigation and usability of the system. A focus group consisted of two cybersecurity experts, one user experience expert, and the rest had no mentioned expertise. The rationale behind the group composition is to support system evaluation by both cybersecurity experts and non-experts. The summary of the assessment is presented in Table VIII.

Table VIII. Fraction of successful answers and standard deviation (SD)

Question	Strategy	Answers	SD
Identify exploited ICS assets in the selected time frame	Comprehensive monitoring	100%	-
	Inspection at ICS process level	100%	-
Identify the process levels involved in cyber incidents	Comprehensive monitoring	100%	-
	Inspection at ICS process level	100%	-
Identify a timeframe of a selected incident	Comprehensive monitoring	83%	0.41
	Inspection at the individual level	100%	-
Identify the incident that requires prompt attention	Comprehensive monitoring	83%	0.41
	Inspection at ICS process level	83%	0.41
	Prioritization	100%	-
Identify the severity rank of the incident	Prioritization	100%	-
Identify possible false alarm	Prioritization	67%	0.52
Identify recovery time after an attack	Comprehensive monitoring	50%	0.55
	Inspection at the individual level	100%	-
For a selected incident, identify remediation priority	Comprehensive monitoring	50%	0.55
	Inspection at ICS process level	50%	0.55
	Prioritization	83%	0.41
Identify the impact of the incident	Prioritization	100%	-

The focus group found identifying and analyzing cyber incidents effective. As a recommendation for improvement, the cybersecurity experts mentioned the improvement of visualization prioritization based on the impact ranking early at the detection and inspection views. The latter will allow the experts to focus their investigation not only on the critical assets (current implementation) and the severity of the incident's effect. Further, the experts mentioned the importance of incident trend analysis to provide more insights to develop better remediation strategies. Specifically, the experts pinpointed two possibilities for trend analysis: the anomalies by the same time for different days and by the device type or vendor.

Further, we asked the focus group to assess the usability of the visualization system using the System Usability Scale (SUS) Score (Brooke 1996). The questions are composed of 10 statements altered between positive and negative and scored on a 5-level Likert scale from 1 (Predominantly Disagree) to 5 (Predominantly Agree) by participants. We slightly modified the original questionnaire to put it in the context of the proposed visual system. In addition, we linked the

questions to the categories recommenced in ISO 9241-210: 2019 (ISO 2019). The summary of the evaluation is presented in Table IX.

Table IX. Summary of usability assessment per question in the usability questionnaire. For questions with Target ⇑ high values are good results, while with Target ⇓ - low values are good results.

Question	ISO category	Mean	Median	Target
I was able to perform all defined tasks	Suitability for the task	4.3	4.5	⇑
I found the system unnecessarily complex	Suitability for the task	1.3	1.0	⇓
I found the navigation between visualization strategies intuitive	Self-descriptiveness	4.5	4.5	⇑
There are too much inconsistency in this system	Self-descriptiveness	1.0	1.0	⇓
I found the various functions in this system are well integrated	Self-descriptiveness	4.2	4.0	⇑
I think that I would need the support of a technical person to be able to use this system	Self-descriptiveness	1.0	1.0	⇓
I believe most people would be able to use this system very quickly	Suitability for the learning	4.7	5.0	⇑
It is required to learn a lot of things before working with the system	Suitability for the learning	1.0	1.0	⇓
I felt very confident using the system	Conformity with user expectations	4.2	4.0	⇑
I found the system very difficult to use	Conformity with user expectations	1.2	1.0	⇓

We received favorable scores concerning the suitability of the proposed visual module for the defined strategies and learning. Self-descriptiveness and conformity with the user expectations were also rated positively, implying that the system is intuitive, and screens are well integrated. The SUS score calculation considers the alteration of positive and negative statements; the reader can find a detailed calculation strategy in the original publication (Brooke 1996). The SUS score range is 0 to 100, with higher scores indicating better usability. The focus group assessed the usability score as 91 (with a median of 92.5), which is above the average score of 68 (Grier 2013). The result indicates that the selected visualization is well-designed to support threat identification and initial analysis of the incident impact.

5 DISCUSSION AND CONCLUSION

The method developed in this study integrates several techniques: *(i)* a deep learning architecture to efficiently generalize the underlying system by operating on circulating ICS data and identify notable deviation from the regular operation; *(ii)* a combination of theory of belief functions to aggregate the results of various feature selection methods and *z-score* for individual variables to address uncertainties in anomaly localization and derive attacked ICS assets; *(iii)* operational performance metrics to evaluate the attack effect; and *(iv)* dynamic visual analytics to extract valuable insights and facilitate operational efficiency. These unique combination complements available attack inference methods by broadening the perspective, which provides desired granularity for the investigation of attack impact. It offers an extensive scope of knowledge as opposed to solely evident indicators of malicious activity. Furthermore, our approach provides the cyber operators and digital investigators an effective tool to dynamically and visually interact,

explore and analyze heterogeneous, complex, at times, conflicting data, and provide rich context information. Such an approach is envisioned to facilitate the cyber incident investigation and support a timely evidence-based risk management process.

5.1 Contributions of the Method

This work offers three particularly salient capabilities for attack incident detection in the ICS setting. First, the proposed approach operates on heterogeneous empirical data with the rare availability of attack-related instances. To this end, it leverages Generative Adversarial Networks, a machine learning architecture widely used for anomaly detection across different domains such as healthcare, public safety, finance, and cybersecurity. This work extends the application of GAN-based anomaly inference methods toward cybersecurity and forensics in critical water infrastructure. As previous research suggests, GAN-based architecture effectively generalizes the underlying system (Choi et al. 2020, Li et al. 2019, Du et al. 2021, Seo et al. 2018.) It can infer a wide range of attacks by employing learning algorithms operating on circulating ICS data, regardless of the deployment domain.

Second, anomaly localization is a challenging task for multidimensional data. This work uniquely employs the degree of belief functions to address uncertainties in anomaly localization by converging the multivariate (operation of entire ICS) and individual assets perspectives. The empirical evaluation (Table VI) demonstrates that the proposed approach significantly increases the localization accuracy.

Finally, anomaly detection methods often have limited capability to generate extensive scope of knowledge regarding detected anomalies (e.g., investigation of incident impact), and insufficient abilities to communicate the result to the broad audience with diverse background to support evidence-based risk management. This work pays particular attention to impact quantification by employing the notion of performance loss as a function of production and quality loss. Moreover, through the visualization module, the proposed approach shows the behavior and state of ICS at both local (individual assets) and global (all assets simultaneously) levels. The module advances the visibility of cyber incidents by effectively utilizing a screen space; it promotes investigation and enriches analytical pivot by integrating extensive empirical analytics and interactive techniques. It aims to understand detected anomalies presented in empirical data and draw conclusions regarding their implication for the system to support proper risk management in a critical realm.

The study results draw an important preliminary conclusion regarding the response of ICS deployed in the water facilities to cyber threats. For instance, the correlation between ICS assets challenges anomaly localization accuracy for the methods relying solely on empirical data. However, the proposed visual exploration allowed an efficient investigation of falsely classified attacked ICS assets by providing an observable correlation between ICS assets and enabling subsequent analysis of the corresponding raw data. Further, the system can demonstrate the effect of the earliest attack with a delay. This effect, therefore, reveals incorrect incident localization since the system behavior does not correspond to the expected conduct. A combination of the proposed analytical strategies and their respective dynamic visualization provided evidence of such an effect and the basis for proper risk management decisions.

5.2 Limitations and Future Work

The proposed work has several limitations that lead to further research. For instance, the method does not consider the cause of the anomaly and the probability of such incidents. Further research in this area would include developing strategies to classify the various anomalies (including system faults) in empirical data and incorporating incident probability assessment to enable a transition to operational implementation through developing a suite of appropriate mitigation or response strategies.

In addition, employed operational performance indices support one threat scenario – malicious manipulation of the water level in the tank. Although numerous attack scenarios exist, their impact quantification is out of the scope of this work. Nevertheless, the proposed framework supports the extension, and we are working on identifying and gathering the required data to employ an extensive set of indices to quantify the impact of the system malfunctioning and water quality for further incorporation into the analytical framework.

Further, the proposed method is limited to the number of incidents, and neither analyzes incident trends nor provides sharing capabilities. The operational community would benefit from the attack trend analysis for targeted and more effective remediation strategies, including those related to the specific devices' vendors. We are now working on the methods for visualization of such trends and generation of the accumulated knowledge for further collaboration among the players of the operating community. The techniques under investigation include but are not limited to visual and semantic data analysis, the data structure for collaborative sharing, and the analytical strategies for collective knowledge.

Finally, the proposed method is also limited by its focus on ICS as an isolated unit. As water facilities become a part of the connected infrastructure of smart cities, the increased interdependence directly relates to the severity of cyber attacks through risk contagion. This has a multiplicative effect on impact, requiring future research regarding quantifying the impact on the entire smart city ecosystem. As we develop such cyber-physical-social systems, we need to design cyber and safety assured systems from the very start through system of systems engineering and design.

REFERENCES

Aggarwal, C. C. (2017). An introduction to outlier analysis. In Outlier analysis (pp. 1-34). Springer, Cham.

Ahmed, C. M., MR, G. R., & Mathur, A. P. (2020). Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems. In *Proceedings of the 6th ACM on cyber-physical system security workshop* (pp. 23-29).

Ahmed, C. M., Palleti, V. R., & Mathur, A. P. (2017). WADI: a water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks* (pp. 25-28).

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19-31.

Alexander, Otis, Misha Belisle, and Jacob Steele (2020). MITRE ATT&CK® for Industrial Control Systems: Design and Philosophy. *The MITRE Corporation: Bedford, MA, USA*.

Bakirtzis, G., Simon, B. J., Fleming, C. H., & Elks, C. R. (2018). Looking for a black cat in a dark room: Security visualization for cyber-physical system design and analysis. *In 2018 IEEE Symposium on Visualization for Cyber Security (VizSec)* (pp. 1-8). IEEE.

Bier, V. (2020). The role of decision analysis in risk analysis: a retrospective. *Risk Analysis,* 40(S1), 2207-2217.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

Bou-Harb, E., Lucia, W., Forti, N., Weerakkody, S., Ghani, N., & Sinopoli, B. (2017). Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence. *IEEE Communications Magazine*, 55(5), 198-204.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. Routledge.

Busby, J. S., Green, B., & Hutchison, D. (2017). Analysis of affordance, time, and adaptation in the assessment of industrial control system cybersecurity risk. *Risk Analysis*, 37(7), 1298-1314.

Cao, N., Shi, C., Lin, S., Lu, J., Lin, Y. R., & Lin, C. Y. (2015). Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on visualization and computer graphics*, 22(1), 280-289.

Case, D. U. (2016). Analysis of the cyber attack on the Ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388, 1-29.

Chabukswar, R., Mo, Y., & Sinopoli, B. (2011). Detecting integrity attacks on SCADA systems. *IFAC Proceedings Volumes*, 44(1), 11239-11244.

Choi, Y., Lim, H., Choi, H., & Kim, I. J. (2020, February). Gan-based anomaly detection and localization of multivariate time series data for power plant. In *2020 IEEE international conference on big data and smart computing (BigComp)* (pp. 71-74). IEEE.

Coulbeck, B., & Orr, C. H. (1993). Essential considerations in the computer control of water distribution systems. *Reliability Engineering & System Safety*, 42(1), 55-64.

Davis, M. J., Janke, R., & Magnuson, M. L. (2014). A framework for estimating the adverse health effects of contamination events in water distribution systems and its application. *Risk Analysis,* 34(3), 498-513.

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782.*

Du, B., Sun, X., Ye, J., Cheng, K., Wang, J., & Sun, L. (2021). GAN-based anomaly detection for multivariate time series using polluted training set. *IEEE Transactions on Knowledge and Data Engineering*.

Elnour, M., Meskin, N., Khan, K., & Jain, R. (2020). A dual-isolation-forests-based attack detection framework for industrial control systems. *IEEE Access*, 8, 36639-36651.

EPA. n.d. "Information about Public Water Systems." *EPA*. Environmental Protection Agency. Accessed June 5, 2022. https://www.epa.gov/dwreginfo/information-about-public-water-systems.

Fischer, F., Mansmann, F., Keim, D. A., Pietzko, S., & Waldvogel, M. (2008). Large-scale network monitoring for visual analysis of attacks. *In International Workshop on Visualization for Computer Security* (pp. 111-118). Springer, Berlin, Heidelberg.

Goh, J., Adepu, S., Junejo, K. N., & Mathur, A. (2016, October). A dataset to support research in the design of secure water treatment systems. *In International conference on critical information infrastructures security* (pp. 88-99). Springer, Cham.

Grier, R. A., Bangor, A., Kortum, P., & Peres, S. C. (2013). The system usability scale: Beyond standard usability testing. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 187-191). Sage CA: Los Angeles, CA: SAGE Publications.

Guo, Z., & Haimes, Y. Y. (2016). Risk assessment of infrastructure system of systems with precursor analysis. *Risk analysis*, *36*(8), 1630-1643.

Haimes, Y. Y. (2018). Risk modeling of interdependent complex systems of systems: Theory and practice. *Risk Analysis*, *38*(1), 84-98.

Hassanzadeh, A., Rasekh, A., Galelli, S., Aghashahi, M., Taormina, R., Ostfeld, A., & Banks, M. K. (2020). A review of cybersecurity incidents in the water sector. *Journal of Environmental Engineering,* 146(5), 03120003.

Hemsley, K. E., & Fisher, E. (2018). History of industrial control system cyber incidents (No. INL/CON-18-44411-Rev002). *Idaho National Lab.(INL), Idaho Falls, ID (United States)*.

Henry, M. H., & Haimes, Y. Y. (2009). A comprehensive network security risk model for process control networks. *Risk Analysis: An International Journal,* 29(2), 223-248.

ICS-CERT, NCCIC. (2015). "Ics-Cert Year in Review (2015)."

Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning. In *2017 IEEE international conference on data mining workshops (ICDMW)*(pp. 1058-1065). IEEE.

ISO. (2019). ISO 9241-210: 2019 Ergonomics of human-system interaction—*Part 210: Human-centered design for interactive systems.*

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, *1*(1), 11-27.

Khanna, R., & Liu, H. (2008). Control theoretic approach to intrusion detection using a distributed hidden Markov model. *IEEE Wireless Communications*, *15*(4), 24-33.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression* (p. 536). New York: Springer-Verlag.

Kotenko, I. V., Kolomeets, M., Chechulin, A., & Chevalier, Y. (2018). A visual analytics approach for the cyber forensics based on different views of the network traffic. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, *9*(2), 57-73.

Kott, A., & Linkov, I. (2021). To improve cyber resilience, measure it. *Computer*, 54(2), 80–85. doi: 10.1109/MC.2020.3038411

Laskov, P., Rieck, K., Schäfer, C., & Müller, K. R. (2005). Visualization of anomaly detection using prediction sensitivity. *Sicherheit 2005, Sicherheit–Schutz und Zuverlässigkeit*.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., & Ng, S. K. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks* (pp. 703-716). Springer, Cham.

Lin, Q., Adepu, S., Verwer, S., & Mathur, A. (2018, May). TABOR: A graphical model-based approach for anomaly detection in industrial control systems. In *Proceedings of the 2018 on asia conference on computer and communications security* (pp. 525-536).

Lohfink, A. P., Anton, S. D. D., Schotten, H. D., Leitte, H., & Garth, C. (2020). Security in process: Visually supported triage analysis in industrial process data. *IEEE transactions on visualization and computer graphics*, 26(4), 1638-1649.

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, *34*(1), 46-53.

Mo, Y., Weerakkody, S., & Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, *35*(1), 93-109.

Moraitis, G., Nikolopoulos, D., Bouziotas, D., Lykou, A., Karavokiros, G., & Makropoulos, C. (2020). Quantifying failure for critical water infrastructures under cyber-physical threats. *Journal of Environmental Engineering*, *146*(9), 04020108.

Neshenko, N., Nader, C., Bou-Harb, E., & Furht, B. (2020). A survey of methods supporting cyber situational awareness in the context of smart cities. *Journal of Big Data*, *7*(1), 1-41.

Pasqualetti, F., Dörfler, F., & Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE transactions on automatic control*, *58*(11), 2715-2729.

Paté-Cornell, M. E., Kuypers, M., Smith, M., & Keller, P. (2018). Cyber risk management for critical infrastructure: a risk analysis model and three case studies. *Risk Analysis*, *38*(2), 226-241.

Rashid, A. N. M. B., & Hossain, M. A. (2012). Challenging issues of spatio-temporal data mining. *Computer Engineering and Intelligent Systems*, 3(4), 55-63.

Robles Frances and Perlroth Nicole (2021). 'Dangerous Stuff': Hackers Tried to Poison Water Supply of Florida Town. *The New York Times*. https://www.nytimes.com/2021/02/08/us/oldsmar-florida-water-supply-hack.html.

Ross, R., Pillitteri, V., Graubart, R., Bodeau, D., & McQuaid, R. (2019). Developing cyber resilient systems: A systems security engineering approach. *National Institute of Standards and Technology.*

Samtani, S., Chai, Y., & Chen, H. (2022). Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model. *MIS Quarterly*, *46*(2), 911-946.

Scalco, A., & Palmer, E. (2022, July). Social Systems Engineering for Achieving Cyber Physical-Social System Multi-Concern Assurance. In *INCOSE International Symposium* (Vol. 32, pp. 30-41).

Seo, E., Song, H. M., & Kim, H. K. (2018, August). GIDS: GAN based intrusion detection system for in-vehicle network. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)* (pp. 1-6). IEEE.

Shalyga, D., Filonov, P., & Lavrentyev, A. (2018). Anomaly detection for water treatment system based on neural network with automatic architecture optimization. *arXiv preprint arXiv:1807.07282*.

Shapley, Lloyd S. (1953). A Value for N-Person Games. *Contributions to the Theory of Games* 2 (28): 307–17.

Shin, S., Lee, S., Burian, S. J., Judi, D. R., & McPherson, T. (2020). Evaluating resilience of water distribution networks to operational failures from cyber-physical attacks. *Journal of Environmental Engineering*, *146*(3), 04020003.

Slay, J., & Miller, M. (2007). Lessons learned from the Maroochy water breach. In *International conference on critical infrastructure protection* (pp. 73-82). Springer, Boston, MA.

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., & Ostfeld, A. (2017). Characterizing cyber-physical attacks on water distribution systems. *Journal of Water Resources Planning and Management*, *143*(5), 04017009.

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., Ostfeld, A., Eliades, D. G., ... & Ohar, Z. (2018). Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*, *144*(8).

Taormina, R., Galelli, S., Tippenhauer, N. O., Salomons, E., Ostfeld, A., Eliades, D. G., ... & Ohar, Z. (2018). Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*, *144*(8).

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE computer graphics and applications*, *26*(1), 10-13.

Tuptuk, N., Hazell, P., Watson, J., & Hailes, S. (2021). A systematic review of the state of cyber-security in water systems. *Water*, *13*(1), 81.

Wagner, M., Fischer, F., Luh, R., Haberson, A., Rind, A., Keim, D. A., & Aigner, W. (2015). A survey of visualization systems for malware analysis. In *Eurographics conference on visualization (EuroVis)* (pp. 105-125).

Wang, C., Wang, B., Liu, H., & Qu, H. (2020). Anomaly detection for industrial control system based on autoencoder neural network. *Wireless Communications and Mobile Computing*, *2020*.

Wang, S., Cao, J., & Yu, P. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*.

Woodall, W. H., & Ncube, M. M. (1985). Multivariate CUSUM quality-control procedures. *Technometrics*, *27*(3), 285-292.

Xie, X., Wang, B., Wan, T., & Tang, W. (2020). Multivariate abnormal detection for industrial control systems using 1D CNN and GRU. *Ieee Access*, *8*, 88348-88359.

Yong, D., WenKang, S., ZhenFu, Z., & Qi, L. (2004). Combining belief functions based on distance of evidence. *Decision support systems*, *38*(3), 489-493.

You, J. (2022). Strengthening Cybersecurity of Water Infrastructure through Legislative Actions. *JAWRA Journal of the American Water Resources Association*, *58*(2), 282-288.

Zechman, E. M. (2011). Agent-based modeling to simulate contamination events and evaluate threat management strategies in water distribution systems. *Risk Analysis: An International Journal*, *31*(5), 758-772.