A Behavioral-based Forensic Investigation Approach for Analyzing Attacks on Water Plants Using GANs

Nataliia Neshenko[a,*], Elias Bou-Harb[b], Borko Furht[a]

[a]Department of Computer and Electrical Engineering and Computer Sciences, Florida Atlantic University, Boca Raton, USA
[b]The Cyber Center For Security and Analytics, University of Texas at San Antonio, San Antonio, USA

Abstract

With the continuous modernization of water plants, malicious, often state-sponsored attacks continue to create havoc in such critical realms. Motivated by this, this paper proposes an unsupervised data-driven approach to support cyber forensics in such unique setups. Specifically, the proposed approach aims at inferring and attributing cyber attacks using sensor readings and actuators states. The approach operates using attack-free data, which is attractive towards cyber forensics of such systems, where attack-related empirical data is rarely widely available due to security and privacy reasons. The proposed method also provides the capability to track and identify the attacked assets for prioritization purposes. The proposed approach exploits Bidirectional Generative Adversarial Networks (BiGAN) to fingerprint the behavior of the system under regular operation. It employs a combination of Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN) as a basis of its design components. The Energy Distance (ED) and Maximum Mean Discrepancy (MMD) are used to evaluate how firmly the model has learned the system's behavior. The approach also leverages the l_1-norm distance between unseen data and corresponding reconstruction to estimate the irregularity score representing cyber attacks. The relative importance of the obtained residual error for each sensor/actuator is put forward to attribute the attacked assets. To this end, we independently employ a regression tree technique, a game-theoretic concept known as Shapley values, and a model-wise approach, the KernelSHAP, as residual loss to identify the relation of each asset to the inferred anomaly. The results are then amalgamated to pinpoint the attacked asset. Empirical evaluations using data collected in a testbed representing a small-scale water treatment plant uncovered 32 out of the 36 cyber incidents; exceeding the performance of state-of-the-art. We also show that the proposed approach identifies the exploited sensors/actuators with more than 8-15% accuracy improvement over current available works. We postulate and stress the fact that such proposed methods significantly contributes towards the forensics of critical infrastructure.

Keywords: ICS forensics; anomaly detection; Generative Adversarial Networks (GAN); water treatment plants

1. 1. Introduction

With the embedding of the latest technologies varying from telecommunication-enablers to advances in data-driven artificial intelligence, Industrial Control Systems (ICS), which provide partial or fully automated control for critical infrastructure, are now accessible through network communications. Composed of various electrical and mechanical devices, computers, and human-machine interfaces, ICS are largely used in water treatment and distribution facilities.

The network-connected water infrastructure, however, can be used to put human population in danger. If an adversary gained remote control over the automated operations, the freshwater supply and quality can be seriously compromised. This threat indeed went beyond theoretical when back in 2000, police arrested a man who used a radio transmitter to take control over the waste-water system and released one million liters of untreated sewage directly in the waterway [1]. The threat of chemical attacks on a water systems is an ongoing concern. In 2016, attackers accessed a water company's industrial control system and altered the amount of chemicals entering the water supply [2]. Further, a partially successful cyber attempt to dangerously elevate the level of sodium hydroxide in the water supply recently threatened the health of around 15,000 individual [3]. Such security incidents, as well as many others, highlight the vulnerability of the network-accessible automated systems of water infrastructure and shed light into the reality of its cyber posture [4].

The increasing relevance and the lack of cyber forensics methods for ICS deployed in water plants paved the way for new research to define required methods supporting cyber forensics. Herein, we perceive that besides network forensics, which provides insights into the attacks that use computer networks as a malicious vehicle [5], the investigation of physical measurements of the water plants will discover the attacks that originated in the physical realm, where no change in network traffic patterns could be observed. Such an approach will also be valuable for the identification of the exploited ICS assets for cyber forensic prioritization purposes. Indeed, given the high number

*Corresponding author
Email addresses: nneshenko2016@fau.edu (Nataliia Neshenko), elias.bouharb@utsa.edu (Elias Bou-Harb), bfurht@fau.edu (Borko Furht)

of sensors and actuators in the system, it is imperative to prioritize the investigation of cyber misdemeanors by focusing on the attacked ICS assets first. Besides, such prioritization allows reducing the overhead associated with the digital triage investigation and evidence collection processes.

Existing methods in this area include control-theoretic approaches, various machine learning algorithms, and those aiming to leverage a particular scheme, known as Generative adversarial networks (GAN) [6]. The first approach assumes the availability of a mathematical model of the system, which is often an impractical assumption. Furthermore, control-theoretic models are highly specific to the system under investigation. Further, machine learning approaches can efficiently be scaled to support a modernized system or another ICS, which is a desirable quality in water facilities, where each deployed ICS can have its own peculiar characteristics. Machine learning approaches, however, often only focus on the classification task, omitting anomaly interpretation; knowing that the ability to provide the reason for the estimated irregularity score is quite essential in industrial settings [7].

With the goal to support cyber forensics for water plants, we introduce in this work an unsupervised method that fingerprints the behavior of the system under regular operations and then isolates the deviations representing cyber attacks. It further scrutinizes the inferred anomaly by distinguishing the attacked sensors/actuators (ICS assets). It departs from available works addressing this problem in three ways.

First, unlike works that require a precise understanding of the physical process, the proposed method learns systems' behavior autonomously using the observed/circulated empirical data. It avoids issues associated with obtaining and maintaining the process model, while possing the capability to infer a wide range of attacks. Second, unlike works that use the GAN-LSTM-RNN-based anomaly detection method, we enrich a classical GAN with the encoder and employ an architecture known as BiGAN [8]. We combine Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN) and leverage them as a foundation for the learning components of the attack inference approach. This architecture is advantageous for learning underlying data in-depth and improves the robustness of the model over a classical GAN-formulation. Finally, unlike methods that use the greatest value of the change ratio of prediction error for attack attribution, we uniquely combine three different algorithms scrutinizing feature-wise residual error, which is the difference between incoming and forecasted sequence of sensors' measurements and actuators' states. Such combination demonstrates better performance upon comparisons against state-of-the-art.

The proposed approach has a number of attractive characteristics. It is designed to work with multivariate data; therefore, it exploits the inherent dependencies between ICS assets deployed in water facilities, leading to better anomaly detection. Further, it can be trained using only attack-free instances; therefore, it manifests as a promising application in cyber forensics of ICS placed in water facilities, where attack-related data is rarely widely available due to security and privacy reasons. Finally, it pinpoints the attacked ICS assets, thus contributing towards the cyber forensic prioritization objective.

In this work, we seek to contribute towards the specific (somehow unique) area of cyber forensics for water systems. To this end, we frame the paper's contributions as follows.

- We complement existing methods supporting cyber forensics for water treatment plants. To this end, we introduce an unsupervised approach that infers maliciously modified processes and pinpoint attacked sensors. The technique is designed to work with multivariate data, where variables represent connected sensors and actuators of ICS deployed in a water treatment plant. It learns system behavior and does not require previous knowledge about its process model; therefore, it overcomes the issues associated with control-theoretic approaches, which assume the existence of a precise mathematical model representing system dynamics.

- We leveraged BiGAN, RNN, and CNN to fingerprint the behavior of ICS deployed in water plants, isolate anomalies and attribute the attack by pinpointing the exploited ICS assets. For many machine learning methods, anomaly interpretation is a well-known issue, while it is an essential feature towards prioritization for cyber forensics. For machine learning methods, this imperative task can be solved by exposing and scrutinizing feature importance (features/variables represent ICS assets hereafter). In this context, we explore a combination of three different algorithms that scrutinize the asset-wise residual error, which is the difference between incoming and expected data sequence for each sensor/actuator.

- We evaluate the approach on empirical data collected by a small-scale water treatment plant. The proposed method demonstrate state-of-the-art performance and improvement over several machine learning approaches tested on the same dataset. The approach successfully pinpointed 32 out of the 36 cyber incidents, exceeding state-of-the-art detection methods. Given the direct impact of such attacks on human health and life, this improvement is quite noteworthy. We also show that the proposed approach can provide an intuitive explanation and identify the exploited sensor/actuator with near 8-15% accuracy improvement over prior literature contributions.

The rest of this paper is organized as follows. In the next section, we review related works and demonstrate the added value of this work. In Section 3, we detail our method that infers and attribute cyber attacks conducted against smart water facilities. In Section 4, we describe the experimental setup and report on the results of applying our model to the data collected using a testbed from an operational water treatment system. We elaborate on the limitations of this work and pave the way for a few future endeavors in Section 5. Finally, we conclude the paper in Section 6.

2. Related work

This section elaborates on relevant endeavors leveraging control-theoretic and machine learning methods for discovering abnormal behavior of ICS. This section also explores the usage of Generative Adversarial Networks (GAN) for anomaly detection and corresponding methods for anomaly attribution.

2.1. Control-theoretic approaches for attack inference

By employing a system- and graph-theoretic approaches, Pasqualetti et al. [9] identified and characterized the vulnerabilities of power networks. The authors proposed an attack detection mechanism rooted in Luenberger-type detection filters. In the same vein, Mo et al. [10] centered their work on attack scenarios in which the adversary records regular system measurements to produce statistically identical yet malicious data for further injection into the system. To detect such attacks, the authors leveraged physical watermarking to authenticate the correct operation of the system. The injection of a known noisy input into a physical system, leading to controllable physical properties changes, was used to detect the attack. Chabukswar et al. [11] extended noisy control to multi-input, multi-output systems of chemical plants and microgrids.

Alternatively, Bou-Harb et al. [12] modeled a semantic behavioral graph consisting of malicious attack signatures, which are retrieved from active cyber threat intelligence, and data flows extracted from the physical layer. Significant similarities between semantic graphs represented indicators of ongoing malicious activities. Further, Khanna el al. [13] employed a control-theoretic Hidden Markov Model (HMM) for intrusion detection in ad hoc wireless networks. An observed deviation from predetermined rules which govern the system's behavior carries a higher probability of being the attack. To reduce the computational complexity of the approach, each node is observed in a periodic manner.

Although these methods demonstrate a remarkable performance of attack inference and attribution, they are computationally expensive and require a precise understanding of the underlying physical processes of ICS. The advantage of our approach over control-theoretic methods is that it efficiently generalizes the underlying system and can infer a wide range of attacks by employing learning algorithms operating on circulating ICS data.

2.2. Machine learning for attack inference in water plants

Several works utilized various machine learning approaches for inference of anomalies in water facilities. For instance, Inoue et al. [14] evaluated Deep Neural Network (DNN) and one class Support Vector Machine (OSVM) as a detection mechanism against ICS deployed in a water plant. The SVM classifier demonstrated a better attack detection rate though led to a higher false alarm rate than those derived using the DNN-based approach. Further, Elnour et al. [15] proposed a semi-supervised method rooted in the Dual Isolation Forest (DIF) model for attack detection in a water treatment plant. The approach comprised of two independently trained isolation forest models. The first was trained using the normalized raw data, while another leveraged a pre-processed version of the data using Principal Component Analysis (PCA). By modeling non-linear correlation among multiple time series, Li et al. [16] put forward an unsupervised GAN-based anomaly detection method in this context. The method employed Long Short Term-Recurrent Neural Networks (LSTM-RNN) for both the generator and discriminator and calculates scores to indicate the level of abnormality in the time series. Lin et al. [17] combined time automata learning and Bayesian neural network to infer the abnormal behavior of a water treatment plant. This combination led to the inference of wide range of attacks.

We complement these works by employing a BiGAN-based approach for inference and attribution of anomalies resulting from malicious manipulation of the ICS process. Unlike methods that do not pinpoint attacked sensors, this work infers anomalies and successfully addressed attack attribution with significant improvement over the state-of-the-art. In particular, the approach inferred and correctly attributed more attacks and their respective targets than previously proposed methods.

2.3. GAN-based methods for anomaly inference

The GAN architecture demonstrated its ability to detect anomalies across different domains. For instance, one of the first applications of GAN-based anomaly detection was proposed by Schlegl et al. [18]. To infer and quantify disease markers in image data, the authors first trained a generator and a discriminator of the GANs using only normal data and then fixed the weights. Furthermore, this trained model was mapped to a latent vector that represented the distribution of the data. By assigning a higher anomaly score to more irregular images, the framework pinpointed the level of abnormality of each instance. Further, Ravanbakhsh et al. [19] employed conditional GANs to detect abnormal video events. They trained the model based on only normal events. The authors hypothesized that usual frames should have low reconstruction loss, whereas anomalous frames should be poorly reconstructed. To examine the abnormality in crowded scenes, the image-to-image translation was utilized. In an alternative work, Zheng et al. [20] trained an adversarial denoising autoencoder, which combined the properties of both a discriminator and a generator of a GAN. The trained model was then used to calculate the probability that the financial transaction is abnormal. The model demonstrated compelling results in the form of detection of more than 300 fraudulent cases during 12 weeks in two major banks in China. Moreover, the framework has a remarkably low misclassification rate. Further, Jiang et al. [21] adopted an encoder-decoder-encoder three-sub-network and DCGAN (Deep Convolutional Generative Adversarial Network) for the generator and discriminator, respectively. Only normal samples are used as an input to the training stage. Abnormal samples received higher anomaly scores compared to normal samples in the testing stage.

We extend the application of GAN-based anomaly inference methods towards cyber forensics for water facilities. The GAN is designed to work with multivariate data, which is a desired characteristic for water facilities with inherent dependencies between ICS assets. We exploit both the trained discriminator and generator for inferring irregularities and outliers in the water treatment process; this combination provides reliable anomaly inference in the systems with multiple dependent components.

2.4. Anomaly attribution

To support the prioritization objective in cyber forensic triage and gather relevant evidence, the inference method should reveal the ICS assets connected to anomalies [22]. To this end, two alternative approaches are presented in the literature. The first outlines model-agnostic frameworks, while the second is very specific to the methods that use reconstruction error as an indicator of anomalies.

Ribeiro et al. [23] introduced a model-agnostic, generic framework, dubbed as LIME, that explains the prediction of any classifier. The model identified a set of most important futures that contribute to the prediction. In an alternative work, Lundberg and Lee employed game theory and offered a unified framework, namely SHAP [24], for interpreting predictions based on feature importance in supervised settings. Later, Giurgiu and Schumann [25] leveraged SHAP to provide additive explanation for anomalies detected by GRU-based autoencoder. In particular, the authors used influence weighting to locate informed neighborhood to compute values per each signal of EEG for their further investigation.

In the alternative approach, Wang et al. [26] used the reconstruction error's change ratio as an anomaly localization technique. The authors noticed that the complexity of the system requires a sounder technique for better attribution. Further, Shalyga et al. [27] used the distance between the forecasted and the actual value of sensors' measurements to locate attacked tags at certain point of time. The greatest error value in the prediction pointed out the attacked sensor. We compare the attribution accuracy with these two methods later in this work.

We complement these contributions by proposing a unique combination of three different methods of deriving feature importance. In particular, we employ a regression tree method in addition to leveraging a game-theoretic concept, known as Shapley values, and model-wise approach proposed in [24]. The proposed attack attribution technique renders a higher inference rate when compared to prior research as evaluated on the same dataset.

3. Proposed approach

This section details the approach for achieving the following goals: *(i)* infer system instability at a certain time, and *(ii)* identify attacked ICS assets.

3.1. Methodology

The three core components - behavior fingerprinting, attack inference, and attack attribution – and the underlying architecture of the approach are illustrated in Figure 1.

The first component learns to fingerprint the typical behavior of the system by leveraging the BiGAN-based learning approach. The second component estimates anomaly score by comparing incoming data sequences for each sensor/actuator with the fingerprinted system behavior. This score indicates of any plausible ongoing attack. The attribution function evaluates the contribution of each ICS asset to the anomaly score to extract the affected sensors/actuators to prioritize cyber forensics.

We now detail the architectural design of each component.

3.1.1. Behavior fingerprinting

To fingerprint system behavior under normal operation, we employ a GAN-based model. In the classic formulation, a GAN is a generative and discriminative deep learning architecture that consists of two competing neural network models, namely a generator (G) and a discriminator (D). As a first step, G receives noise z to learn a distribution $p(z)$. Based on perceived distribution, G produces data samples and passes them to the discriminator D. In its turn, D determines the distribution between real and fake data, and back-propagates the probability of data authenticity to G, which adapts its parameters based on received gradient information and passes new samples back to D. The learning goal of a generator is to produce more realistic instances, while the discriminator aims at improving its ability to distinguish fake data from the real.

To unlock the potential of GAN and address their challenges, several extensions of the original framework have been proposed in the literature. The main difference between these developments is the adjusted architecture of GAN in the form of additional networks. The design components of the proposed behavioral fingerprinting (Figure 1) are inspired by BiGAN [8]. This architecture was originally employed for anomaly detection in images. In our work, we extend its application to data with temporal components and adjust the underlying learning networks accordingly.

In addition to the generator G, the architecture includes an encoder E which learns a feature space of underlying data $x = \{x_1, \dots x_t\}$, where x_t denotes a m-dimensional vector $\{x_t^1, \dots x_t^m\}$ representing sensor readings and actuator states at certain points of time. E maps x to latent variable space z. D discriminates data space x versus $G(z)$, and also $(x, E(x))$ versus $(G(z), z)$. In this architecture, an encoder E should learn to invert the generator G. The encoder cannot discover the output of the generator and vice versa. The simultaneous training of an encoder with generator and discriminator improves the robustness of the model.

The building blocks of the learning components incorporate Recurrent Neural Network (RNN) and Convolution Neural Network (CNN). This combination leverages the ability of RNN to effectively grasp time series since it keeps track of previous data points and can perceive the long-term pattern. Besides, it exercises the power of CNN to learn features of time series and extract behavioral patterns.

3.1.2. Anomaly inference

Given the simultaneous training of D and G, it is advantageous to exploit both networks for attack detection, as suggested in [18]. First, we utilize the ability of trained D to distinguish the normal operational behavior of ICS. To this end, D determines the distribution between fingerprinted behavior and

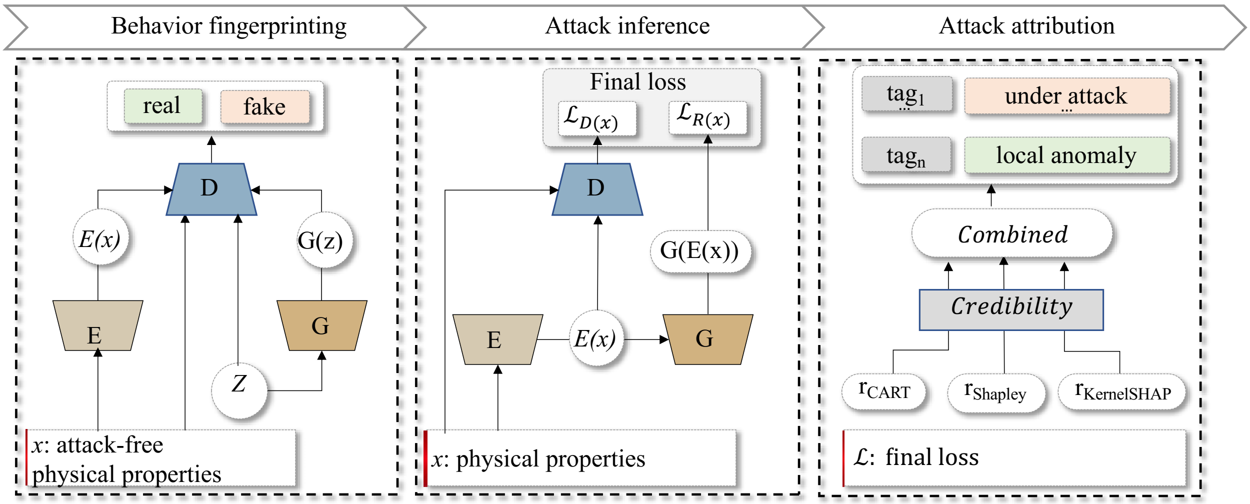

Figure 1: The components of the proposed approach

a latent representation of incoming data $E(x)$, and returns the abnormality score, namely the discriminator loss $\mathcal{L}_D(x)$ (Eq. (1)).

$$\mathcal{L}_D(x) = cross\_entropy(D(x, E(x)), 1) \tag{1}$$

Further, a network G was trained to capture the complexity of the data distribution and generate ICS assets' measurements and states at a certain point in time. Therefore, G can take a latent representation of incoming data obtained from the encoder and reconstruct the expected ICS behavior ($G(E(x))$). The trained model then evaluates l_1-norm distance, namely residual loss, between actual (x) and expected $G(E(x))$ physical measurements for all ICS assets (Eq. (2)).

$$\mathcal{L}_R(x) = \parallel x - G(E(x)) \parallel_1 \tag{2}$$

We further obtain the irregularity score $\mathcal{L}_F(x)$ for each time point as a weighted sum of $\mathcal{L}_R(x)$ and $\mathcal{L}_D(x)$ (Eq. (3)) .

$$\mathcal{L}_F(x) = (1 - \omega) * \mathcal{L}_R(x) + \omega * \mathcal{L}_D(x) \tag{3}$$

where ω is a weighting parameter controlling the impact of the loss on the anomaly score. A larger score $\mathcal{L}_F(x)$) denotes the time points representing physical properties of ICS that do not align with the fingerprinted behavior, therefore, portrays an anomaly.

3.1.3. Asset attribution

The high number of connected sensors and actuators in the system challenges discovering the attacked ICS asset, though this recognition is quite beneficial for forensics' prioritization. For data-driven methods, the attack attribution can be made with feature (ICS assets) selection algorithms [28]. To this end, we leverage three algorithms to derive the relative feature importance from the inference model. We further aggregate their predictions ($pr_i, i = 1, 2, 3$) for optimal output. The rationale behind combining different algorithms is that the independent results can be misinterpreted [28], while the combination would likely provide a better understanding of ICS assets' relation to anomaly score.

The first algorithm, namely Classification and Regression Trees (CART), creates a binary tree to obtain a representative feature (variable) based on an appropriate impurity criterion [29]. CART identifies the most significant variable and effectively handles outliers.

The next approach uses Shapley values [30], which is a concept from game theory. The variable values are represented as the players in a coalition. Shapley values describe how each variable contributed to anomaly prediction.

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\} - v(S)) \tag{4}$$

where ϕ_i is a Shapley value for t_i asset of interest; n is number of such assets; $v(S)$ is a payoff for coalition S; and $N \setminus i$ denotes all possible coalitions that do not consists of i.

Finally, we use KernelSHAP [24], which is a model-agnostic method that uses Shapley values and the Local Interpretable Model-agnostic Explanations (LIME) technique [23].

We finally use a fusion of the above methods by using a degree (deg_j) of a belief that the method j correctly identified an asset under attack. Formally, it is defined as follows.

$$score_i = \sum_{j=1}^{k} deg_j \cdot pr_j \tag{5}$$

were k is a number of combined methods, deg_i is calculated based on the amount of relative support [31] representing a distance between predictions made by the methods above:

$$deg_j = Support(pr_j) / \sum_{j=1}^{k} Support(pr_j) \tag{6}$$

$$Support(pr_j) = \sum_{m=1, j \neq m}^{k} (1 - d(pr_m, pr_j)) \tag{7}$$

We declare a sensor/actuator as an asset under attack if its aggregated coefficient $score_i$ is in the 75^{th} percentile.

3.2. Performance evaluation

To measure the performance, we evaluate the viability of the model to fingerprint system behavior, the ability to infer attacks and pinpoint the attacked ICS assets. Our approach heavily relies on the capability of the network to capture the complexity of the data distribution. The ultimate usage of the generated sequences is to train a model to detect anomalies in the physical measurements of a water treatment plant. To evaluate how firmly our model does so, we employ a discrepancy measure, known as Energy Distance (ED) [30], a statistic that is based on the idea that observations should have zero potential energy only if they originate from the same underlying distribution. Mathematically,

$$ED(x,z)^2 = (2\mathbb{E}|X - Z| - \mathbb{E}|X - X'| - \mathbb{E}|Z - Z'|) \qquad (8)$$

where X, X' and Z, Z' are independent random variables of original and forecasted physical properties, and x and z are their respective distribution. In addition, we use the Maximum Mean Discrepancy (MMD), which has proven to be well-suited to evaluate the quality of generated GAN samples in multivariate data [32]. Formally, MMD is defined as follows:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{j \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{j \neq j} k(z_i, z_j) - \frac{2}{n^2} \sum_{j,j} k(x_i, z_z) \qquad (9)$$

A lower ED and MMD measures indicate the lower discrepancy between two probabilities and suggest, after stabilization, the high quality of behavioral fingerprinting.

We determine the efficiency of the attack inference function by testing its ability to pinpoint the attacks (with acceptable false alarms). To this end, we first measure the sensitivity (or recall), which is the proportion of the correctly inferred attack points to the total number of cases when the ICS is under attack. It is formally defined as $tp/(tp + fn)$. Hereafter, tp and tn stand for the number of true positives and true negatives, respectively; fp and fn denote the number of false positives and negatives. We further evaluate the precision, which measures the proportion of correctly identified attack of all the points that are classified as positive; it is formally defined as $tp/(tp + fp)$.

To evaluate the capacity of the inference function to avoid false alarms, we estimate the specificity, which is the proportion of the correctly recognized non-attack points to all cases when the ICS is operating under a regular/normal operation cycle. Mathematically, it equates to $tn/(tn + fp)$. To test the accuracy of the model, we employ the typical F-measure, which represents the harmonic mean of the sensitivity and precision, and takes into account both false positive and false negative rates.

To assess the ability of the model to pinpoint the affected asset, we compare qualitative results with ground truth annotations from the attack log (in the empirical dataset). In addition, we compare precision, recall, and F-measure metrics obtained using the proposed approach and those which have been reported in related prior works; including works from [26] and [27] .

4. Empirical evaluation

In this section, we seek to answer for following questions (Q

4.1. Dataset

Fig. 1: Actual Photograph of SWaT testbed

To answer these questions, we evaluate the proposed approach on the dataset collected using a fully operational small-scale water treatment (SWaT) [33] plant, similar to those found in small cities. The testbed (Figure 2) consists of six stages of the water treatment process (P1 through P6), which are controlled by a dedicated PLC. The sensors, connected over the network, accumulate the water level and water flow by interacting with the physical environment.

research in the design of secure and safe CPS. SWaT has six main proc corresponding to the physical and control components of the water treat facility. It has the following six-stage filtration process, as shown in Figure

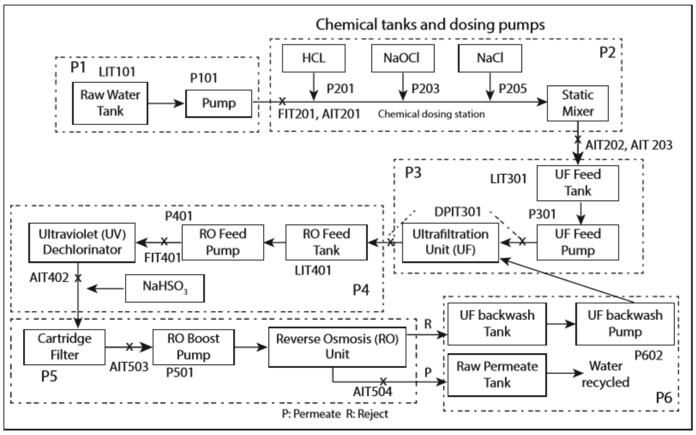

Figure 2: SWaT testbed process overview [33]. P1 though P6 pinpoint the six stages of the treatment process. ICS assets notation comprises types (letters), stage (first number), and sequential number. In particular, FIT***: flow meters, LIT***: level indicator and transmitter, AIT***: property indicator and transmitter, DPIT: differential pressure indicator and transmitter, P***: pump

Fig. 2: SWaT testbed processes overview

2.1 Water treatment process

The SWaT dataset consists of collected physical properties that include sensor measurements (continuous variables) and actuators' states (discrete variables). The dataset covers seven days of attack-free operation and four working days under different types of cyber attacks targeting one or more ICS assets. Moreover, the combination of attacked assets can be found on one, or multiple process stages P1-P6. The duration, as well as the objectives of these attacks, vary. The time system requires to recover from the attack depends on the attack scenario; though the exact time is not provided in the dataset description.

The process (P1) begins by taking in raw water and storing it in a tank. then passed through the pre-treatment process (P2). In this process, the

Table 1: The characteristics of SWaT dataset

Characteristics	Value
Variables	51
Instances in dataset	
Attack-free operation	496, 800
Data with attacks	449, 919
Number of attacks	36
Attack duration	100sec-10hrs

In Table 1, we summarize the main characteristics of the used dataset.

We conducted the following data preprocessing procedure. First, we trimmed the initial 6 hours from the training (attack-free) dataset, representing system stabilization which can affect the behavioral fingerprinting task. We then scaled the original dataset to optimize computation so that every attribute had a mean value of 0 and a standard deviation of 1.

4.2. Q1: Behavioral fingerprinting

As a prerequisite for anomaly detection, the proposed approach should carefully model system behavior under regular operation. Figure 3 shows that the largest decrease in both ED and MMD value occurred after 32 epoch. It is reasonable to suggest that the quality of the generated data is stabilized after this epoch.

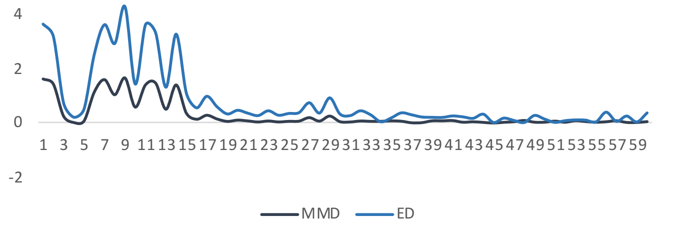

Figure 3: Discrepancy metrics: maximum mean discrepancy (MMD) and energy distance (ED) as a function of epoch for the proposed architecture

To evaluate how accurately does the proposed model perceive sensor readings' distribution in 2-dimensional space, we apply t-SNE [34] analysis on both the original sensor measurements and actuators' states. In Figure 4, we observe the significant overlap between the original (blue color) and the perceived data (orange color), which indeed indicates that the proposed model was able to grasp the system behavior quite closely.

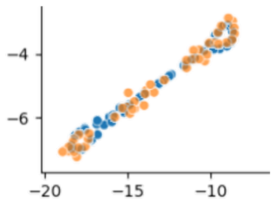

Figure 4: t-SNE visualization. Blue color denotes the original data, orange color indicates those that were perceived/projected by the model

4.3. Q2: Anomaly inference

The proposed method inferred 32 out of the 36 labeled attacks. We focus on two attack scenarios with their inference results for illustration purposes and then provide a comprehensive comparison for all attack scenarios against the related prior research works proposed in [14, 15, 17]. We selected two attack scenarios to demonstrate the examples for which the proposed approach *(i)* shows equivalent performance as [14, 15, 17], and *(ii)* outperforms the respective literature.

In the first scenario (Figure 5), three ICS assets deployed on two different process stages (P4 and P5) are attacked. A regular operation cycle is as follows. Water from the reservation unit at process stage P4 moves to the process stage P5 via ultraviolet (UV) and cartridge filter. An attack (Figure 5 highlights the actual attack window using red color) increased the amount of water produced at AIT502 and turns off the actuator UV. The false data forced PLC to stop pump P501.

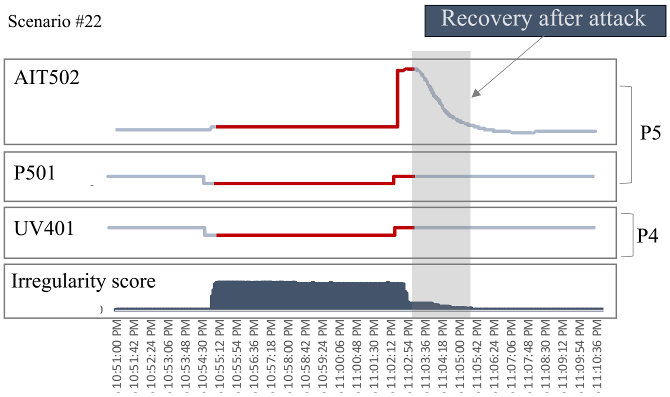

Figure 5: An example of sensor measurements and actuators' states during an attack (red), and their inference. A red color indicates an attack window. In this scenario, the amount of water produced at AIT502 is maliciously increased and the actuator UV401 is turned off. False data forced PLC to stop pump P501. The high irregularity score, which represents a false alarm after the attack, is caused by system instability representing a recovery period after the attack. In addition, the dataset labels the start of the incident with a 22-second delay, leading to false positives produced by the inference method at the beginning of the incident.

The inference mechanism precisely isolates the incident by assigning a high irregularity score. It also shows that the attack's start is 22 seconds earlier than recorded in the attack log. We manually confirmed that the dataset labeled an attack with this specific delay. In addition, the dataset marks recovery time as a regular operation; however, it is essential to infer a stabilization period for further investigation of a potential impact on the system since it may take a long time (sometime hours) for ICS to return to regular operation. These discrepancies lead to false positives produced by the inference method at the beginning and after the attack.

In the next scenario (Figure 6), the measurement of LIT301 was maliciously decreased by 1mm each second, while the actual water level was increasing, leading to tank overflow. The proposed method inferred the incident by assigning a high irregularity score 21 seconds earlier than the incident was recorded in the attack log. The reason behind this early inference is that the actual data change appeared earlier than it was labeled in the dataset (also manually confirmed).

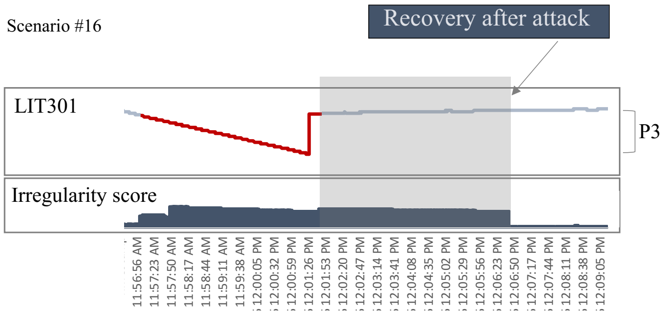

Figure 6: An example of sensor measurements and actuators' states during an attack (red), and their inference. A red color indicates an attack window. In this scenario, the water level at LIT301 is maliciously decreased by 1mm each second leading to tank overflow. The high irregularity score, which represents a false alarm after the attack, is caused by system instability representing a recovery period after the attack. In addition, the dataset labels the start of the incident with a 21-second delay, leading to false positives produced by the inference method at the beginning of the incident.

The thorough comparisons with existing literature is summarized in Table 2. Please note, that we intended to keep the dataset as close as possible to the actual operational cycle so that we did not eliminate the stabilization time from the test dataset as it was suggested in [15]. The stabilization period uncovers the impact of the attack on system operation and, thus, will contribute to cyber forensics' completeness. This discrepancy brings forth a higher number of false positives produced by the proposed approach, leading to a lower precision and F-measure. Thus, a direct comparison of the metric should only be made with this in mind.

Table 2: Performance across different anomaly detection methods. The results of DNN, SVM, DIF, MAD-GAN, and TABOT methods are taken from respective publications (rounded to the nearest hundredths)

Model	Precision	Recall	F-measure
DNN [14]	0.98	0.68	0.80
SVM [14]	0.93	0.70	0.80
DIF [15]	0.93	0.84	0.88
MAD-GAN [16]	0.99	0.64	0.77
TABOR [17]	0.86	0.79	0.82
This work	0.81	0.84	0.83

The relative performance for each attack is illustrated in Table 3, which compares the recall achieved by our approach and that reported by different detection methods. Please note that the dataset's attack scenarios consist of several attacks (5, 9, 12, 15, and 18) that do not affect physical properties and, thus, are irrelevant to this work. We keep the original scenario numbers in the table while we excluded such attacks. The results of DNN, SVM, Dual-Isolation-Forest (DIF), and TABOR methods are taken from their respective publications and rounded to the nearest hundredths. The bold font in the table portrays the higher recall measure and validates that our approach achieved a comparative state-of-the-art performance, while surpassing available methods for several attacks.

Table 3: Recall across different anomaly detection methods. The results of DNN, SVM, DIF, and TABOT methods are taken from their respective publications (rounded to the nearest hundredths)

Attack scenario	DNN [14]	SVM [14]	DIF [15]	TABOR [17]	This work
1	-	-	0.01	**0.05**	-
2	-	-	0.29	**0.93**	0.79
3	-	-	**1.00**	-	0.69
4	-	-	-	**0.33**	-
6	0.72	0.72	**1.00**	**1.00**	0.97
7	-	0.89	**1.00**	-	0.40
8	0.93	0.92	**1.00**	0.61	0.28
10	**1.00**	0.43	**1.00**	0.99	0.99
11	0.98	**1.00**	**1.00**	**1.00**	**1.00**
13	-	-	-	-	-
14	-	-	0.06	-	**0.35**
16	-	-	0.55	-	**0.91**
17	-	-	**0.64**	0.60	0.33
19	0.12	0.13	**0.45**	0.01	0.31
20	0.85	0.85	0.45	**1.00**	0.94
21	-	0.02	-	**0.08**	0.03
22	0.99	**1.00**	**1.00**	**1.00**	**1.00**
23	0.87	0.88	0.82	-	**0.99**
24	-	-	0.34	-	**0.39**
25	-	0.01	**1.00**	-	**1.00**
26	-	-	0.17	**1.00**	0.18
27	-	-	-	0.20	**0.97**
28	0.94	0.94	**1.00**	**1.00**	**1.00**
29	-	-	**1.00**	-	**1.00**
30	-	-	-	**1.00**	0.60
31	-	-	**1.00**	-	0.24
32	-	0.91	**1.00**	-	-
33	-	-	0.43	**0.89**	0.11
34	-	-	-	**0.99**	0.47
35	-	-	0.95	0.26	**1.00**
36	-	0.12	**0.93**	0.89	0.86
37	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
38	0.92	0.93	1.00	**1.00**	0.98
39	0.94	-	**1.00**	0.37	0.91
40	0.93	0.93	1.00	**1.00**	0.77
41	-	0.36	**0.63**	-	0.35

4.4. *Q3: Attack attribution*

We now answer the question about how well does the approach can attribute the attack. To reduce the computation load, we apply the attribution methods to five ICS assets that display the higher reconstruction loss for each attack. Please recall that we label the ICS asset as an attack point only if the proposed method assigned a score which falls into the 75^{th} percentile. Although, we conduct the investigation of each inferred attack, we illustrate two representative examples and further compare the overall performance with available prior works.

In the first scenario, the incident log registered the following attack points: valve UV401, water level indicator AIT501, and pump P501. As noted in Table 4, the proposed method correctly attributes the attack to the respective ICS assets and outperforms the previously proposed methods.

In the second scenario (Table 5), the incident log recorded that the level indicator LIT301 as the attack point. The pro-

Table 4: The result of attack attribution for attack scenario 22. Bolt font indicates the correctly attributed ICS assets. Attack points for [26] and [27] are taken from their respective papers.

Model	Reported attack points
This work	**UV401, AIT502, P501**
Wang et al. [26]	**UV401, P501**, FIT504
Shalyga et al. [27]	DPIT301, MV302

posed method marked two sensors as attacked, while it assigned a score for LIT101 into 83^{rd} percentile. In contrast, the score for the correctly pinpointed level indicator LIT301 falls into the 100^{th} percentile.

Table 5: The result of attack attribution for attack scenario 16. Bolt font indicates the correctly attributed ICS assets. Attack points for [26] and [27] are taken from the respective papers.

Model	Reported attack points
This work	**LIT301**, LIT101
Wang et al. [26]	-
Shalyga et al. [27]	MV301, MV303

The inherent correlation of the elements in the water treatment ICS can cause false alarms using data-driven methods. Despite this false alarm, the proposed method outperformed competing algorithms for the single-point attack.

We examine in Table 6 the attack attribution performance of our approach in contrast to those reported in [26] and [27].

Table 6: A comparison of attack attribution across different methods

Model	Precision	Recall	F-measure
This work	**0.38**	**0.49**	**0.43**
Wang et al. [26]	0.30	0.43	0.35
Shalyga et al. [27]	0.22	0.21	0.21

Indeed, the results produced by the proposed method isolates more ICS assets under attack with fewer false alarms in contrast to those that were reported in prior literature.

5. Limitations and future perspectives

Broadly, the results of the empirical evaluations demonstrated the capability of the proposed approach to infer and attribute the attacks with equivalent or higher performance in comparison with the state-of-the-art. Thus, this work presents a motivating step towards cyber forensics in water treatment facilities.

Although we leveraged a testbed that replicates a small-scale water treatment plant to fingerprint its behavior, the testbed might not consist of all required assets and, thus, does not represent the entire dynamics of a realistic, complex system. The proposed approach would benefit from an extended evaluation using different variants of ISC deployed in real-world water facilities, including pH sensors, consumers tanks, and return water grid. While the proposed approach is scalable from a data analysis perspective, it can only be practical after broader evaluation (the same conclusion applies to literature methods). In addition, we evaluated it on a system that has a quite stable behavior model. Therefore, we plan to investigate the effect of pattern instability (e.g., the variation of water consumption). To this end, we are working towards evaluating the approach on other empirical data, including those noted in [28].

Further, the identification of attacked ICS assets remains challenging. Inherent dependencies among sensors/actuators prevent precise identification of exploited assets. To address this challenge, we are currently investigating supplementary information fusion approaches to further calibrate attack point identification accuracy. Last but not least, in accordance with ICS security and forensic tactics, we intend to also explore tailored remediation methods, given knowledge about the exact attacked assets from this work.

6. Conclusion

Aiming to support cyber forensics in the context of ICS deployed in smart water facilities, we complement available contributions and introduce an unsupervised approach that infers cyber attacks and reveals the attacked assets for forensics' prioritization. The model is strengthened by BiGAN, RNN, CNN, and Shapley values of residual error. The former component is advantageous for learning underlying data in-depth. RNN is valuable for time series processing, while CNN can learn features and extract patterns. Finally, Shapley values provide the ability to trace the potential attack points from both class-wise and model-wise perspectives.

The approach has three distinguishing characteristics: *(i)* it operates with multivariate data, which is highly desired in ICS deployed in water treatment plants; *(ii)* it is trained using only attack-free instances, therefore avoid the problem of imbalanced data, *(iii)* it identifies the attack points and thus can reduce forensics' overhead in critical infrastructure realms.

We demonstrated the effectiveness of the proposed approach by employing it to data collected by a testbed representing a small-scale water treatment plant. Most of the attacks presented in the empirical data were detected with high sensitivity, while the inference method maintained a high irregularity score after the threat has passed. The latter is affected by the time system required for recovery after an attack. The results of the empirical evaluations demonstrate the capability of the proposed approach to infer and attribute the attacks with equivalent or better performance over state-of-the-art methods.

There are several concerns to examine in future work. Given the direct impact of water on our well being, cyber forensics would benefit from evaluating the cascading attacks' impact on various water plant assets. Another aspect that we are currently pursuing is the calibration of the approach to improve the inference of the attacked ICS assets. Further, we plan to investigate how water consumption variability would affect the performance of the proposed method.

Acknowledgements

The authors would like to express their sincere gratitude to the anonymous reviewers and PC members for their constructive feedback. This work was partially supported by the Florida Center for Cybersecurity (Cyber Florida) and U.S. National Science Foundation (NSF) (Office of Advanced Cyberinfrastructure (OAC)).

References

[1] J. Slay, M. Miller, Lessons learned from the maroochy water breach, in: International conference on critical infrastructure protection, Springer, 2007, pp. 73–82.

[2] B. Verizon, Data breach digest. Scenarios from the field., https://enterprise.verizon.com/resources/reports/data-breach-digest.pdf.

[3] Robles Frances, Perlroth Nicole, 'Dangerous Stuff': Hackers Tried to Poison Water Supply of Florida Town, The New York Times.
URL `https://www.nytimes.com/2021/02/08/us/oldsmar-florida-water-supply-hack.html`

[4] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, N. Ghani, Demystifying IoT security: an exhaustive survey on IoT vulnerabilities and a first empirical look on internet-scale IoT exploitations, IEEE Communications Surveys & Tutorials 21 (3) (2019) 2702–2733.

[5] E. Bou-Harb, M. Debbabi, C. Assi, Behavioral analytics for inferring large-scale orchestrated probing events, in: 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2014, pp. 506–511.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[7] K. Amarasinghe, K. Kenney, M. Manic, Toward Explainable Deep Neural Network Based Anomaly Detection, in: 2018 11th International Conference on Human System Interaction (HSI), 2018, pp. 311–317.

[8] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, arXiv preprint arXiv:1605.09782.

[9] F. Pasqualetti, F. Dörfler, F. Bullo, Attack detection and identification in cyber-physical systems, IEEE transactions on automatic control 58 (11) (2013) 2715–2729, publisher: IEEE.

[10] Y. Mo, S. Weerakkody, B. Sinopoli, Physical Authentication of Control Systems: Designing Watermarked Control Inputs to Detect Counterfeit Sensor Outputs, IEEE Control Systems Magazine 35 (1) (2015) 93–109. doi:10.1109/MCS.2014.2364724.

[11] R. Chabukswar, Y. Mo, B. Sinopoli, Detecting integrity attacks on SCADA systems, IFAC Proceedings Volumes 44 (1) (2011) 11239–11244, publisher: Elsevier.

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

[13] R. Khanna, H. Liu, Control theoretic approach to intrusion detection using a distributed hidden Markov model, IEEE Wireless Communications 15 (4) (2008) 24–33. doi:10.1109/MWC.2008.4599218.

[14] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, J. Sun, Anomaly detection for a water treatment system using unsupervised machine learning, in: 2017 IEEE international conference on data mining workshops (ICDMW), IEEE, 2017, pp. 1058–1065.

[15] M. Elnour, N. Meskin, K. Khan, R. Jain, A Dual-Isolation-Forests-Based Attack Detection Framework for Industrial Control Systems, IEEE Access 8 (2020) 36639–36651. doi:10.1109/ACCESS.2020.2975066.

[16] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.-K. Ng, MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks, in: International Conference on Artificial Neural Networks, Springer, 2019, pp. 703–716.

[17] Q. Lin, S. Adepu, S. Verwer, A. Mathur, 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, 2018, pp. 525–536.

[18] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 146–157.

[19] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, in: Image Processing (ICIP), 2017 IEEE International Conference on, IEEE, 2017, pp. 1577–1581.

[20] P. Zheng, S. Yuan, X. Wu, J. Li, A. Lu, One-class adversarial nets for fraud detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 1286–1293, issue: 01.

[21] W. Jiang, Y. Hong, B. Zhou, X. He, C. Cheng, A GAN-based anomaly detection approach for imbalanced industrial time series, IEEE Access 7 (2019) 143608–143619.

[22] N. Neshenko, C. Nader, E. Bou-Harb, B. Furht, A survey of methods supporting cyber situational awareness in the context of smart cities, Journal of Big Data 7 (1) (2020) 1–41, publisher: SpringerOpen.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, " Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.

[25] I. Giurgiu, A. Schumann, Additive Explanations for Anomalies Detected from Multivariate Temporal Data, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2245–2248.

[26] C. Wang, B. Wang, H. Liu, H. Qu, Anomaly Detection for Industrial Control System Based on Autoencoder Neural Network, Wireless Communications and Mobile Computing 2020, publisher: Hindawi.

[27] D. Shalyga, P. Filonov, A. Lavrentyev, Anomaly detection for water treatment system based on neural network with automatic architecture optimization, arXiv preprint arXiv:1807.07282.

[28] R. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, A. Ostfeld, D. G. Eliades, M. Aghashahi, R. Sundararajan, M. Pourahmadi, M. K. Banks, others, Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks, Journal of Water Resources Planning and Management 144 (8) (2018) 04018048, publisher: American Society of Civil Engineers.

[29] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and regression trees, CRC press, 1984.

[30] L. S. Shapley, A value for n-person games, Contributions to the Theory of Games 2 (28) (1953) 307–317.

[31] D. Yong, S. Wenkang, Z. Zhenfu, L. Qi, Combining belief functions based on distance of evidence, Decision support systems 38 (3) (2004) 489–493, publisher: Elsevier.

[32] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. J. Smola, A kernel method for the two-sample-problem, in: Advances in neural information processing systems, 2007, pp. 513–520.

[33] J. Goh, S. Adepu, K. N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, in: International Conference on Critical Information Infrastructures Security, Springer, 2016, pp. 88–99.

[34] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE., Journal of machine learning research 9 (11).