Detection-Performance Tradeoff for Watermarking in Industrial Control Systems

Hengye Zhu, Mengxiang Liu, *Student Member, IEEE,* Chongrong Fang, *Member, IEEE,* Ruilong Deng, *Senior Member, IEEE,* Peng Cheng, *Member, IEEE*

***Abstract*—The watermarking method, which adds unique watermarks to data, has been widely used for integrity attack detection in industrial control systems (ICSs). Existing literature generally design watermarking mechanisms without considering the existence of noises, which cannot be trivially applied to the realistic ICS scenarios in the presence of strong noise interference. On one hand, the low-intensity watermarking will be ineffective under the strong noise environment; while on the other hand, the oversized watermarking can possibly degrade the control performance or even destabilize the system. Therefore, the intensity of watermarks plays a fundamental role in balancing the tradeoff between detection effectiveness and control performance, which, to the best of our knowledge, has never been thoroughly analyzed yet. To this end, in this paper, we for the first time propose an optimal watermarking design method for ICSs considering the detection-performance tradeoff. To begin with, we shift the watermark container from data points to segments and update the detection metrics to reduce the noise impact. Then, we formulate an optimization problem to determine the strength of watermarks to balance the detection-performance tradeoff. Meanwhile, the detection effectivenss and control performance metrics are analytically modeled and theoretically analyzed considering the discrepancy between added watermarks and noises, signal quality, detection latency, as well as estimation of detection metrics. Finally, extensive numerical simulations and systematical experiments based on a practical Ethanol Distillation ICS are conducted to validate the theoretical analysis and demonstrate the outperformance of our proposed watermarking method in comparison with related works.**

***Index Terms*—watermark-based detection system, data integrity verification, industrial control systems**

I. Introduction

INDUSTRIAL control systems (ICSs) are intelligent brains that control and automate industrial processes in critical infrastructures like power generation facilities and chemical plants [1]. With the implementation of the Industrial Internet of Things (IIoTs), real-time channels transfer the data from smart sensors and other information sources frequently over multiple layers, helping industrial devices and infrastructures autonomously take actions [2]. It also breaks the physical isolation between industrial local area networks and the internet, which increases the exposures and risks of critical infrastructures. In recent years, the security accidents against ICSs have been increasing dramatically, and the attacker is becoming more and more intelligent to hide their malicious footprint from engineers [3]. The diversity of attack strategies makes it difficult for the corresponding defense approaches to keep up. For example, some attackers attempt to compromise upper-layer standards such as TCP connections among engineer stations [4] or control commands sent by programmable logic controllers [5], while other threats include the injection of false data into real-time field-layer protocols like PROFIBUS [6] and CAN [7]. Both the mentioned attacks can cause severe damage by deviating the system states from normal operating conditions. The TRITON malware found in Saudi Arabia shut down several petrochemical plants and caused a wide-area impact [8].

There are some prevention methods from the information technology (IT) domain like network communication encryption [9], firewalls [10], and multi-factor authentication [11] that can effectively protect the data integrity in ICSs. These approaches will result in nontrivial time delay, which is not acceptable in the real-time communication channels of ICSs. Moreover, intelligent attackers often bypass them and modify the information without triggering any alarm. Hence, by integrating the cyber and physical properties of ICSs, numerous intrusion detection systems (IDSs) have been proposed to restrict the implications of intelligent attackers when they bypass the security mechanisms from the IT domain [12]. Mainstream detection methods in ICSs can be categorized as physical-model-based [13]–[20] and machine-learning-based [21]–[26]. Model-based approaches obtain and monitor physical invariants from the plant to be protected, and in machine-learning-based methods such hidden relationships can be easily extracted by intelligent algorithms. However, there are limitations: 1) Learning-based approaches require large normal and abnormal historical samples [27] to train data-driven models, while time-consuming data cleaning and model selection play important roles on accuracy [28]; 2) Many model-based approaches rely on manually extracted physical relationships and accurate physical parameters; 3) Finally, due to the hardware resource constraints, the edge IIoT devices are incapable of completing detection tasks promptly with the sophisticated models of both kinds of methods [29].

Watermark-based detection technology has been recently developed in ICSs to address the mentioned issues. This concept is comparable to the digital watermark that has been widely used in images and videos to protect data integrity. The principle of watermarking is to inject some crucial information into the measurements, which can only be identified by a certain algorithm, and any mismatch indicates that the

H. Zhu, M. Liu, R. Deng and P. Cheng are with the College of Control Science and Engineering, Zhejiang University, Hangzhou 300027, China (e-mail: {zhuhycse, lmx329, dengruilong, lunarheart}@zju.edu.cn).

C. Fang is with the Key Laboratory of System Control and Information Processing, Ministry of Education of China; Shanghai Engineering Research Center of Intelligent Control and Management; Department of Automation, Shanghai Jiao Tong University, Shanghai, China (e-mail: crfang@sjtu.edu.cn).

integrity has been compromised [30]. It's important to note that fingerprinting, another frequently used technology, differs from watermarking technology. Fingerprinting is used to generate content-based compact fingerprints without modifying the original information, while the watermarking technology embeds covert information into sensor signals. Different from the direct fingerprint comparison adopted in the fingerprinting technology, watermarking uses the statistical features of the data before/after removing watermarks for anomaly perception. In [31], Mo *et al.* proposed a Gaussian watermarking scheme for the detection of replay attacks in linear discrete-time systems. The watermarks are proactively added into the control signal such that system state and measurement would both contain the watermark information. If the attacker is unaware of the added watermarks, then the Kalman filter based detector can easily verify the integrity of measurements by checking the existence of the watermarking information. After that, researchers focused on the optimal watermarking generation scheme to balance the tradeoff between the control performance and the detection effectiveness [32]–[34]. However, the methods mentioned above are designed and validated based on ideal mathematical models, which are difficult to be applied to realistic ICS scenarios due to their high complexity. Song *et al.* [35] first applied the watermarking scheme to enhance the channel integrity in practical scenarios. The proposed lightweight recursive watermarking method (RWM) generates difficult-to-tamper-with watermarks on measurements that reliably perceive the malicious modifications on time. To meet the real-time requirement, sensors directly transfer the lightweight watermarked signals to controllers for command calculation without removing watermarks. It means watermarks can affect legitimate operations of the ICS.

There still exist gaps in the application of the watermark-based detection method to realistic ICS scenarios. First, it does not take the existence of measurement noises into account in practice, so watermarks may be invalidated when the strength is much smaller than that of the noise. Second, it ignores the effect of the added watermark on the control performance. Intuitively, if the added watermark distorts the original signal, the control performance may be largely degraded. Finally, the detection metrics are computed based on data sampling points from a sliding time window (STW). The STW with a long length will cause nontrivial detection delay, and the irrationally large detection metric resulting from the data sampling number may cause a large number of false alarms.

To address the above issues, this article proposes an enhanced watermark-based detection method for practical applications. The proposed method includes three parts: 1) Modeling the measurement noise to minimize its effect on watermarks, to make the differences reflected on detection metrics before and after the attack more significant; 2) Formulating an optimization problem with practical constraints to determine the strength of watermarks and the length of STW; 3) Normalizing the detection metrics. The main contributions of this article are as follows:

1) To the best of our knowledge, this is the first work that proposes a watermark-based data integrity verification method for ICSs considering realistic relationships between watermarks and noises, as well as detection performance requirements.
2) The watermark generation and intrusion detection algorithms are more noise-resistant than those of the RWM, since watermark containers are transferred from fixed data points to variable-length segments inspired by the Allan Variance (AVAR), and detection metrics are normalized at the end.
3) The minimum watermark strength, as well as the proper length of STW, are determined by solving the formulated optimization problem with practical constraints. The constraints are given considering the watermark-noise differences, the signal quality, the estimation accuracy of detection metrics, and the detection delay requirements.
4) The experimental results based on the realistic platform show the effectiveness and superiority of our method. A comprehensive comparison with other data integrity protection methods is conducted.

The remainder of this article is organized as follows. The system model and problems are formulated in Section II. We design an enhanced watermark-based detection method in Section III. Section IV demonstrates the simulation and experiment results that show the effectiveness and superiority of our method. We conclude this paper with future work in Section V.

II. Preliminaries and Problem Formulation

A. ICS and Attack Models

The simplified control loop of ICS is shown in Fig. 1, where the controller compares the sensor measurement s with the set point and calculates a control output u. The u acts on the actuators like valves and pumps to adjust the controlled variables. The communication channel between the sensor and controller should satisfy the real-time property of industrial scenarios [36], while the research [37] theoretically analyzes the negative effects of time delay on control system performance, which is introduced during sampling and additional operations.

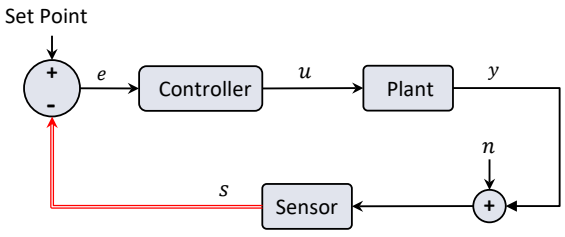


Fig. 1. The simplified control loop of ICS.

The sensor measurements $s[k]$ at time slot k can be decomposed as the legitimate part $y[k]$ and the noise $n[k]$, i.e.,

$$s[k] = y[k] + n[k], \tag{1}$$

where the former part contains the original continuous measurements and the latter part is probably introduced by physical environment or analog-to-digital conversion [38]. Our mission is to ensure the data integrity of the $s[k]$.

We assume that attackers have bypassed the prevention mechanisms from the IT domain and are able to read and write data arbitrarily, but they can not obtain the system time clock and the TrustZone [39] of the embedded devices. Let k_1 and k_2 be the beginning and ending timestamps of the attack, respectively, the data injection process is described as

$$s_a[k] = \begin{cases} a[k], & k_1 < k < k_2, \\ s[k], & k \leq k_1, k \geq k_2, \end{cases}$$

where $a[k]$ can be any signal and $a[k] \neq s[k]$.

B. Watermark-based Detection Method

As shown in Fig. 2, the method is divided into three steps: watermark-generation, watermark-decoding, and χ^2 testing validation conducted by transmitters 1), receivers 2) and detectors 3), respectively [35].

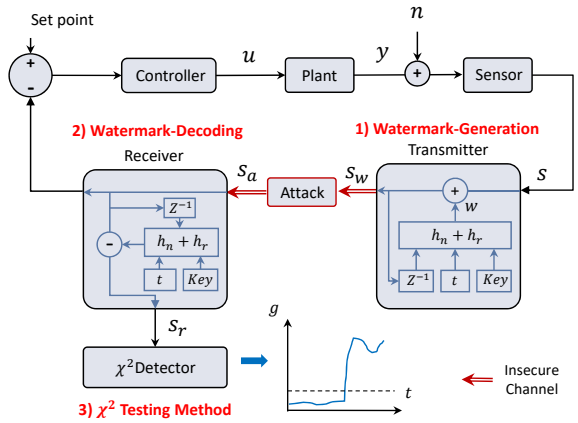


Fig. 2. The block diagram of the RWM.

1) Watermark-generation: The watermark at time k is defined as $w[k]$ which is directly added to the $s[k]$ in the discrete time domain, the process is completed by the transmitter as

$$s_w[k] = s[k] + w[k]. \tag{2}$$

We define two hashing functions $h_n(k, \mathcal{K})$ and $h_r(s_w[k-1], \mathcal{K})$. To save computational resources, the hash tables are established in advance and the key $\mathcal{K}$ is stored in the secure storage area like TrustZone. The first hashing method receives the current time k as input, whereas the second one takes the previous watermarked data $s_w[k-1]$. The $w[k]$ value at time slot k is determined by

$$w[k] = \begin{cases} 0, & k = 0, \\ h_n(k, \mathcal{K}) + h_r(s_w[k-1], \mathcal{K}), & k \geqslant 1. \end{cases} \tag{3}$$

The hashing functions map the timestamp imformation and the measurement causality information to some special system-tolerable perturbations in a lightweight way, which makes watermarks unique and hard to be completely erased. Additionally, unlike public MD5 or SHA algorithms used in fingerprinting technologies, the goal of our private algorithms is to ensure that the watermarks added in sensor measurements are hard to be analyzed and cracked by malicious adversaries.

2) Watermark-decoding: The receiver conducts two operations: it first directly sends received sequence to the controller for the real-time control requirement in the belief that the small watermark strength barely affects the normal control performance. Meanwhile, it independently executes the removal of watermark from duplicated received signals for further checking. Under normal conditions, the received signal is watermarked sequence s_w. It is trivial to obtain the initial recovered sample $s_r[0] = s_w[0] = s[0]$, then the receiver can decode the watermarked signal to the original signal s recursively with the initial signal $s_w[0]$ as follows

$$s_r[k] = s_w[k] - h_n(k, \mathcal{K}) - h_r(s_w[k-1], \mathcal{K}) = s[k]. \tag{4}$$

The removed parts are identical with the added watermarks since the inputs of the hashing function, i.e., synchronized timestamp k, received watermarked $s_w[k-1]$, and the key $\mathcal{K}$ are all the same.

However, when the communication channel is compromised before $k-1$ such as clock asynchronism and watermarked signal modification, the influence will propagate to the following hashing values. The process

$$s_r[k] = s_a[k] - h_n(k, \mathcal{K}) - h_r(s_a[k-1], \mathcal{K}) \neq s[k] \tag{5}$$

fails to recover the original signal using wrong hashing inputs, i.e., $w_a[k] \neq w[k]$. The special statistical properties of the watermark will remain in the original signal and can be easily recognized by the detector.

3) χ^2 testing method: In the early 1970's, Mehra *et al.* defined the innovation sequence for the calculation of whiteness, mean as well as covariance via hypothesis testing [40], which has been widely used for intrusion detection in the control society. We define $g[k]$ as the intrusion indicator, which is determined by the square of the error drawn from data samples in a sliding time window as follows

$$g[k] = \frac{1}{W} \sum_{k=1}^{W} e[k] \sigma_n^{-1} e[k], \tag{6}$$

where σ_n is the standard deviation of noise, and the predicted error $e[k]$ is the difference between recovered signal $s_r[k]$ and predicted signal $\hat{s_r}[k|k-1]$. Based on the principle of linear approximation prediction, the first derivative is taken as the increment of the next sampling point, with which we have

$$\begin{aligned} e[k] &= s[k] - \hat{s}[k|k-1] \\ &\approx s[k] - (s[k-1] + (s[k-1] - s[k-2])) \\ &= s[k] - 2s[k-1] + s[k-2]. \end{aligned} \tag{7}$$

According to (2), current $s[k]$ is the sum of measurement $y[k]$, noise $n[k]$ and watermark $w[k]$, which are all independent and identically distributed variables. We can rewrite the $e[k]$ as

$$e[k] = e_y[k] + e_n[k] + e_w[k], \tag{8}$$

where the $e_y[k]$, $e_n[k]$, and $e_w[k]$ are the respective discrete second-order derivatives like $e[k]$.

C. Detection Metrics With and Without Attacks

In this subsection, we show the detection metrics with and without attacks, where the expected means of $g[k]$ are denoted by $\mathbb{E}\{g[k]|\bar{A}\}$ and $\mathbb{E}\{g[k]|A\}$, respectively.

Without Attack: In normal cases, the watermark $w[k]$ can be totally removed from the watermarked data, i.e.,

$$s_r[k] = s[k] = y[k] + n[k]. \tag{9}$$

Lemma 1: Assume that continuous legitimate data $y[k]$ is smoothing and the noise $n[k]$ follows normal distribution $\mathcal{N}(0, \sigma_n^2)$, we have

$$\begin{aligned}\mathbb{E}\{g[k]|\bar{A}\} &= \sigma_n^{-1}\mathbb{E}\{{e_k}^2\} = \sigma_n^{-1}\mathbb{E}\{(e_y + e_n)^2\} \\ &= \mathbb{E}\{e_n^2\} = 6\sigma_n.\end{aligned}$$

Proof: $\mathbb{E}\{(e_y + e_n)^2\}$ can be regarded as the cumulative sum of $\mathbb{E}\{e_y^2\}$, $\mathbb{E}\{e_n^2\}$ and $\mathbb{E}\{2e_y e_n\}$. The first part e_y has an alternative expression $e_y = y'' \cdot T_s^2$, where T_s^2 is the sampling time of the sensor and it is in microsecond range due to the high sampling rate. Hence $\mathbb{E}\{e_y\}$ tends to be zero, under which we have $\mathbb{E}\{e_y^2\} = 0$. According to the assumption that y and n are independent, we can derive $\mathbb{E}\{2e_y e_n\} = 0$. We expand the last part $\mathbb{E}\{e_n^2\}$ to $\mathbb{E}\{(n_k - 2n_{k-1} + n_{k-2})^2\}$, as noise points in every time slot are also independent. the expectation of noise further reduces to $\mathbb{E}\{{n_k}^2 + 4{n_{k-1}}^2 + {n_{k-2}}^2\} = 6\sigma_n^2$. ■

With Attack: When the attacker modifies the measurements starting from a random point, the input pairs of hash functions would be altered accordingly, under which the watermark cannot be removed, i.e.,

$$s_r[k] = s_a[k] - w_a[k] \neq s[k]. \tag{10}$$

The expected value of $g[k]$ is determined by the modified data $s_a[k]$ and the incorrectly removed watermark $w_a[k]$, which is formally represented as

$$\mathbb{E}\{g[k]|A\} = \sigma_n^{-1}\mathbb{E}\{e_{sa}^2\} + \sigma_n^{-1}\mathbb{E}\{e_{wa}^2\}. \tag{11}$$

Since $\mathbb{E}\{e_{sa}^2\} \geqslant 0$, the detection metric $\mathbb{E}\{g[k]|A\}$ with attack will deviate significantly from the one without attack once $\mathbb{E}\{e_{wa}^2\}$ is far larger than ${6\sigma_n}^2$. An intuitive method to enlarge $\mathbb{E}\{e_{wa}^2\}$ is to choose the watermark with a large strength. Besides, as every single sample $g[k]$ fluctuates wildly around the theoretical expected value due to the stochasticity, enough $g[k]$ samples should be adopted to estimate the overall mean. Hence, the detector has to determine a proper length of STW W to decrease the fluctuation degree and reflect the actual level of $\mathbb{E}\{g[k]\}$.

D. Motivation and Problem Formulation

The RWM attempts to find an appropriate threshold ξ to automatically judge anomalies by comparing $g[k]$ with ξ. If there is an intrusion, the indicator $g[k]$ appears to show a significant rise exceeding the ξ. However, without considering the measurement noise, the arbitrary setting of the watermark strength may make $\mathbb{E}\{e_{wa}^2\}$ comparable or smaller than ${6\sigma_n}^2$, under which the impact of an attack is indistinguishable from that of the measurement noise. As depicted in Fig. 3, a piece of signals with $\sigma_n = 0.02$ is intercepted and the attack is launched at $t = 40$s. Three kinds of watermarks with different variances $\sigma_w = 0.01, 0.02, 0.08$ are considered. The results indicate that the small watermark strength ($\sigma_w \leq \sigma_n$) is submerged by the measurement noise, and it is difficult to set up an appropriate threshold. In comparison with a large watermark strength ($\sigma_w > \sigma_n$), the range of threshold falls in the range of $g[k] \in [0.2, 0.4]$, from which we can choose one feasible value to detect the anomaly.

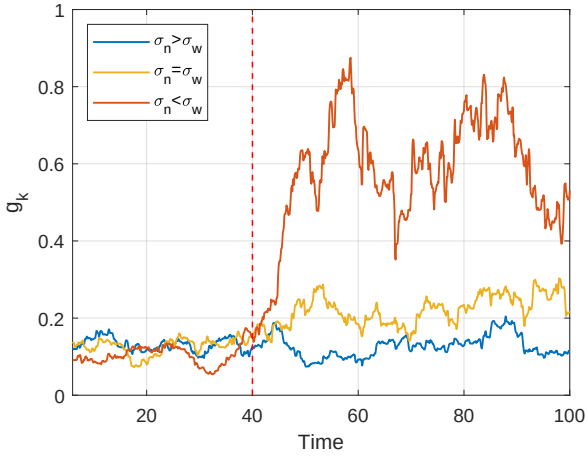
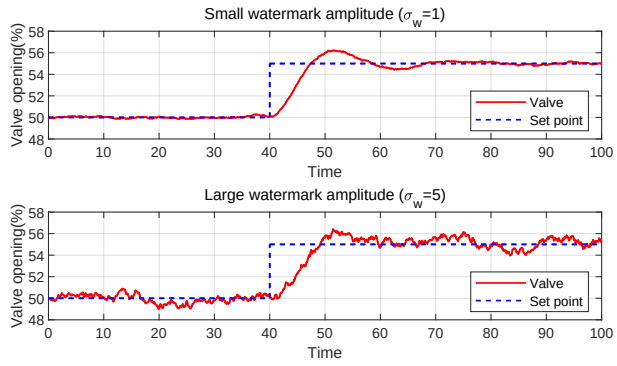


Fig. 3. The performance of watermarks with different strengths.

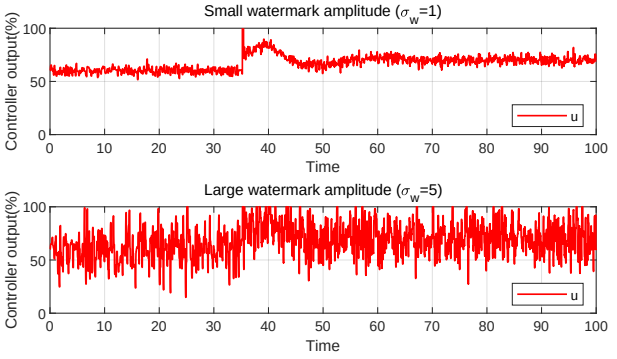
Although increasing the watermark strength can effectively improve the positive detection rate, there should be an upper bound for the watermark strength since the watermarked data is directly transferred to the controller for control command calculation, to satisfy the real-time requirement. The fluctuations introduced by too large watermarks may make the system deviate severely from the set point. We apply two watermarks with $\sigma_w = 1$ and $\sigma_w = 5$ to a typical heating system simulation in the distillation process, where the controller controls a fuel valve of the boiler by receiving temperature feedback. Fig. 4 shows that the large watermark strength causes a significant variation in both control performance and controller outputs, which could destroy the valve and lead to a PID controller's integrated windup problem.

The above two scenarios show that the noise scale and watermark strength have a substantial impact on the intrusion indicator $g[k]$ as well as the legitimate measurement. We conclude the problems as follows:

- Most approaches for noise removal are offline [41]. There is no suitable method for online modeling of the noise to weaken the impact of which on the performance of the watermark-based integrity verification.
- The relationship between $\mathbb{E}\{{e_{wa}}^2\}$ and $\mathbb{E}\{{e_{sa}}^2\}$ cannot be quantified before we determine the characteristics of watermarks, and the impact of watermarks on legitimate signals is also not modeled.
- The distributions of $g[k]$ samples before and after the attack are determined by the length of STW W in the equation (6). It is significant to find a description of the distance of the two distributions to guide the choice of W.



(a) Control performance comparison



(b) PID controller output comparison

Fig. 4. Large watermark strength leads to a system fluctuation.

III. The Enhanced Watermark-Based Detection Method

The proposed detection method will first gather sensor data series from the historical database to estimate the characteristics of noises. This step helps determine the proper length of the data segment to carry watermarks which weaken the effect of noises. Then, an optimization problem is formulated to find the minimum watermark strength under the condition that the control performance and detection performance are both well satisfied. The detection performance is specifically summarized as four mathematical constraints the difference between noises and watermarks, data quality, detection latency, and the difference between detection metrics before and after the attack. The optimal watermark strength with appropriate STW length can be directly applied to the original watermarking-related modules (e.g., RWM). Finally, the method uses a normalized threshold-based detection rule.

A. Noise Modeling and Smoothing

Intuitively, modeling the noise is the first step to eliminate its effect on watermarked signals. It is difficult to give an accurate description of the noise in realistic industrial sensors since factors like the external environment and equipment quality can affect the measurements. Many studies on watermarking methods assume that the noise is normally distributed with known expected value and variance, to gain more insights on the relations between the noise strength and watermarking performance. After we collect a large number of system noise samples in the targeted industrial sensors, we find the noises are very close to the normal distribution, so we also assume that the noise follows the normal distribution for subsequent theoretical analysis. Moreover, with the perspective that the sensor noise exhibits different features from a micro to a macro time scale, we introduce Allan Variance (AVAR) [42], which was first developed in the 1960s for oscillator frequency stability investigation in a clock system and has been recently applied to reveal the error characteristics of sensor measurements, to provide stability information on the types and strength of various noise terms [43].

The AVAR's basic principle is to divide the time series into non-overlapping continuous blocks, each including a segment of measurements that spans the length determined by the correlation time τ, and the value of τ is an integer multiple of the sampling interval. Such segments are regarded as new units for the calculation of statistical features at the current time scale. To illustrate the process explicitly, the signal sequence $S = \{s_1, s_2, ..., s_N\}$ containing N sampling points is divided into K clusters as

$$\underbrace{s_1, \ldots, s_m}_{k=1}, \underbrace{s_{m+1}, \ldots, s_{2m}}_{k=2}, \ldots, \underbrace{s_{N-m+1}, \ldots, s_N}_{k=K}. \tag{12}$$

Each cluster contains m values and the τ equals to m/f_s, where the f_s is the sampling frequency.

The mean value of each cluster forms an innovative sequence Θ. We use

$$\theta[k] = \frac{1}{m}\sum_{i=1}^{m} s_{(k-1)m+i} \tag{13}$$

to represent the kth element. This step is quite similar to the mean filtering method in that it smooths down noise within each block while maintaining the fluctuating features among all blocks.

Under a proper block length, the noise's impact on watermarks can be effectively reduced when the sequence Θ is used to calculate the new intrusion indicator $g_{new}[k]$, which is defined as

$$g_{new}[k] = \frac{1}{W}\sum_{i=k-W+1}^{k} e_\theta[k]\sigma_\theta^{-1}e_\theta[k], \tag{14}$$

where

$$e_\theta[k] = \theta[k] - 2\theta[k-1] + \theta[k-2]. \tag{15}$$

Remark 1: In Fig. 5, the black points are original sampling points $s[k]$ while the short red lines are mean values $\theta[k]$ of 32 points. It is observed that the discrete points fluctuate between -0.25 to 0.25, while these lines are smoother and lie within the bound $[-0.05, 0.05]$. As the red zone depicted in Fig. 5, adding a watermark directly to a single sampling point requires a larger strength than those to a segment of sampling points to achieve a similar detection performance. That is, adding watermarking identifiers to θ instead of s helps reduce the watermark strength. On this account, both the signal quality and control system stability can be guaranteed.

In the following, we use $\mathbb{E}\{g_{new}[k]|\bar{A}\}$ and $\mathbb{E}\{g_{new}[k]|A\}$ to separately represent $g_{new}[k]$'s mean values before and after the attack. Compared with the original $g[k]$, $g_{new}[k]$ stays below a lower level with the same noise.

Theorem 1: If the noise $n[k]$ of the original sequence S subjects to normal distribution $\mathcal{N}(0, \sigma_n^2)$, we have

$$\mathbb{E}\{g_{new}[k]|\bar{A}\} = \frac{1}{\sqrt{m}}\mathbb{E}\{g[k]|\bar{A}\} = \frac{6}{\sqrt{m}}\sigma_n \tag{16}$$

where we define the deviation of Θ is $\sigma_\theta = \sigma_n/\sqrt{m}$.

Proof: $s_1, s_2, \ldots, s_m$ are independent with each other. According to the additive property of normally distributed random variables, the sum of $s_1, s_2, \ldots, s_m$ construsts a new variable Σ subjecting to $\mathcal{N}(0, m\sigma_n^2)$. As $\Theta = m^{-1}\Sigma$, we can derive $m^{-1}\Sigma \sim \mathcal{N}(0, m^{-1}\sigma_n^2)$. The variable Θ follows the *Lemma 1*. ∎

Remark 2: The theoretical result demonstrates that this approach is proven to be effective in weakening the impact of noise on the initial expected value of the χ^2 detector if the noise distribution is approximately regarded as a normal distribution. For other general noise distribution types like completely random distributions, qualitatively speaking, our approach still works, but the explicit expressions regarding weakened degrees need to be studied in future works.

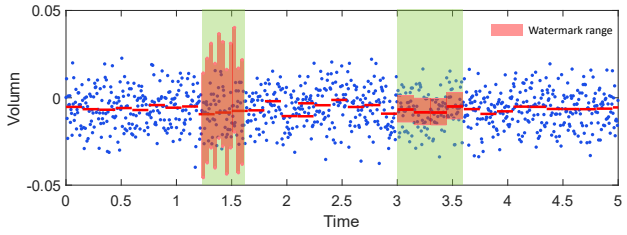


Fig. 5. Different strength requirement for watermark addition.

It becomes another significant issue to select the appropriate value of m to make the $\mathbb{E}\{g_{new}[k]\}$ less than a setpoint E_{max}, meanwhile, we also set an upper bound τ_{max} for m to ensure it will not be too large to affect the detection performance. According to the noise type, we divide it into two cases:

1) For the Gaussian noise, we can calculate the unbiased estimation of standard deviation σ_n from a stationary sampling interval by

$$\sigma_n = \sqrt{\frac{\sum_{i=1}^{N}(s_i - \bar{s})}{N-1}}. \tag{17}$$

According to the *Theorem 1*, we have

$$\frac{36{\sigma_n}^2}{{E_{max}}^2} \leqslant m \leqslant \tau_{max}, \tag{18}$$

and the optimal length m^* takes the smallest integer in the range.

2) For other difficult-to-model noises, we propose a heuristic solution which uses the binary search algorithm in the range $[0, \tau_{max}]$. At each step of execution, it calculates the expectation $\mathbb{E}\{g_{new}[k]|\bar{A}\}$ based on the equations (13)-(15) with the obtained signal samples until the detection value drops below the threshold.

Algorithm 1 illustrates the above process. The function first picks a set of sensor measurements to check whether the noise is normally distributed. If the result is true, the algorithm directly calculate the suitable m, otherwise, it searches for the value of m that will effectively lower the watermark's strength range below the threshold by calculating the variance of the innovative series from micro to macro time scales. The transmitter releases the optimal m^* for further usage. This step is intermittently activated to handle the drift of noise characteristics.

Algorithm 1: Block Size Determination Algorithm

Input: Offline signal set S; Target E_{max}; Limitation τ_{max}
Output: Block size m

Set initial $m_i = 1$ and $m_{max} = \tau_{max}$;
case = IfWhiteNoise(S);
if *case=True* **then**
  Calculate $\sigma_n = SDestimation(S)$;
  return $m_i = MinInteger(36{\sigma_n}^2/E_{max}^2, \tau_{max})$
else
  Truncate signal time series S_i to nearest integer power of 2 to yield data length = $2^p (p \in \mathbb{Z}^+)$;
  while $m_i \leqslant m_{max}$ **do**
    Initialize block number $k = 1$;
    Set number of data blocks, $K_i = 2^p/m_i$;
    while *block number* $b \leqslant K_i$ **do**
      Find mean for $\theta_k^i = \frac{1}{m_i}\sum_{j=1}^{m_i} s_{(k-1)m_i+j}$;
      $k \leftarrow k + 1$;
      if $k \geqslant 3$ **then**
        Calculate $e_\theta^i[k] = \theta_k^i - 2\theta_{k-1}^i + \theta_{k-2}^i$;
      end
    end
    Calculate expected $E_{ob} = \frac{1}{K_i - 2}\sum_{j=3}^{K_i} {e_\theta^i}^2[k]$;
    if $E_{ob} \leqslant \mathbb{E}\{g_{new}\}$ **then**
      return m_i
    end
    $m_i \leftarrow 2m_i$
  end
end

Remark 3: Clearly, when the exact noise distribution knowledge is available, the determination of m only requires to calculate equations (17) and (18) once. While the search of m without knowing the noise distribution needs to calculate equations (13)-(15) for multiple times, which is time-consuming and induces much more computation burden.

B. Mathematical Model of the Optimization Problem

At the beginning, we give a full description of the watermark strength. Equation (3) shows that $w[k]$ at time slot k consists of two hash values generated by $h_n(k, \mathcal{K})$ and $h_r(s_w[k-1], \mathcal{K})$, respectively. The $h_n(k, \mathcal{K})$ and $h_r(s_w[k-1], \mathcal{K})$ separately map the discrete timestamp sequence and the watermarked measurement sequence to fixed-size values from $[-\mu/2, \mu/2]$. According to the uniformity property of the hash function, every hash value in the output range should be generated with roughly the same probability, which means both two hashing parts follow identical uniform distributions $\mathcal{U}(-\mu/2, \mu/2)$. So we define the

upper bound of watermark strength as $|w[k]|$, and

$$\max_{k\geqslant 0}|w[k]| = \frac{\mu}{2}+\frac{\mu}{2} = \mu. \tag{19}$$

Since the carrier of watermarks changes from sequence S to Θ, our goal is to choose the minimum μ on Θ which satisfies the detection and control requirements. Inspired by the LQG controller [44], we adopt the quadratic objective function $\mathcal{J}(\mu)$ to evaluate the strength of watermarks. Four significant constraints describe 1) the sensitivity of the detector to watermarks in the presence of noise, 2) the impact of watermark and noise on the signal quality, 3) the detection latency, and 4) the distribution distinction of $g_{new}[k]$. We describe them in detail in the following. The optimization problem is formulated as

$$arg\min_{\mu\geq 0} \quad \mathcal{J}(\mu) = \frac{1}{2}\mu^2, \tag{20}$$

$$\text{s.t.} \quad R(\mu) \geq r_l, \tag{21}$$

$$\text{SNWR}(\mu) \geq \eta, \tag{22}$$

$$T_{set} - T(W) \geq 0, \tag{23}$$

$$\Psi(\mu, W) \geq 0. \tag{24}$$

First, we have noticed that watermarks with small strength will be concealed by large noises and rendered useless, so we aim to set a fixed obvious level gap of $\mathbb{E}\{g_{new}\}$ between non-attack and attack scenarios, which is called the detection sensitivity. We quantify the logarithmic ratio of $\mathbb{E}\{g_{new}|A\}$ to $\mathbb{E}\{g_{new}|\bar{A}\}$ above the detection sensitivity.

Detection Sensitivity Constraint: The objective of feasible watermark strength is to make sure that $\mathbb{E}\{g_{new}[k]\}$ can exceed the agreed new threshold. The lower bound of $\mathbb{E}\{g_{new}[k]|A\}$ is calculated as

$$\begin{aligned}\mathbb{E}\{g_{new}[k]|A\} &= \sigma_\theta^{-1}\mathbb{E}\{e_{\theta a}^2\} + \sigma_\theta^{-1}\mathbb{E}\{e_{wa}^2\} \\ &\geqslant \sigma_\theta^{-1}\mathbb{E}\{e_{wa}^2\},\end{aligned} \tag{25}$$

where $e_{\theta a}^2$ is the unknown calculation residue caused by the injected false data.

Theorem 2: The lower bound of $\mathbb{E}\{g_{new}[k]|A\}$ is represented as a polynomial function $E(\mu)$

$$\sigma_\theta{}^{-1}\mathbb{E}\{e_{wa}^2\} = E(\mu) = \sigma_\theta{}^{-1}\mu^2. \tag{26}$$

Proof: The detailed proof can be found in Appendix A. ■

We use the function

$$R(\mu) = \ln\frac{\sigma_\theta{}^{-1}\mathbb{E}\{e_{wa}^2\}}{\mathbb{E}\{e_\theta^2\}} = \ln\frac{\mu^2}{6\sigma_\theta{}^2} \tag{27}$$

to quantify the level-stepping degree of $\mathbb{E}\{g_{new}[k]\}$ caused by the attack. The larger the watermark strength μ, the greater the difference in $g[k]$ before and after the attack. The $\sigma_\theta{}^2$ item in the denominator of (27) shows that the strong noise can invalidate the inappropriate watermarks with small strength. $R(\mu)$ needs to be greater than the detection sensitivity r_l as (21).

Signal Quality Constraint: The second key point is that the oversized watermark strength can degrade the signal quality, just like the distortion caused by measurement noises. The premise of active attack detection is that the added watermarks cannot severely interfere with the functionalities of the original data. Hence, the watermarked measurements fed back to the controller should not make the controlled variables deviate sharply from the set points. Inspired by the Signal-to-Noise Ratio (SNR) which describes the influence of existing noises on sensor measurements quantitatively, we introduce a similar concept as the Signal-to-Noise-plus-Watermark Ratio (SNWR), to mathematically describe the impact of watermarks and measurement noises on the original data. The SNWR is defined as the ratio of the mean value of signals to the standard deviation of the watermarked measurement, i.e.,

$$\text{SNWR}(\mu) = \frac{\frac{1}{N}\sum_{k=1}^{N} s[k]}{\sqrt{\sigma_n^2+\sigma_w^2}} \geq \eta. \tag{28}$$

Notably, the mean value of the sampling points represents the signal level, and the variances of the noises and watermarks contribute to the disturbance. The σ_n^2 is estimated by sampling points, not the segments. The σ_w^2 equals $\mu^2/6$, and the η is the minimum SNWR that channels can accept.

Detection Latency Constraint: As equation (14) shows, we adopt the STW technique that samples finite items of $e_\theta{}^2$ to calculate $g_{new}[k]$. Some detection latency, which is related to the block size m and the length of STW W, would inevitably be caused. First, when we transfer the watermark carrier from single points $s[k]$ to segments $\theta[k]$, we need to wait $\tau = m/f_s$ seconds while assembling continuous m points to generate the follow-up segment. Second, when the attack happened at time slot k, the STW needs to contain segments from $\theta[k+1]$ to $\theta[k+W]$ to correctly estimate the $\mathbb{E}\{g_{new}[k]\}$. The latency function $T(W)$ should satisfy

$$T(W) = W\tau = \frac{mW}{f_s} \leq T_{set}, \tag{29}$$

where the maximum acceptable attack latency is set to be T_{set}.

Distribution Distinction Constraint: The watermark strength μ and the length of STW W jointly affect the distributions of $g_{new}[k]$ before and after the attack. Fig. 6 illustrates that the increased W will effectively reduce the variance of $g_{new}[k]$. The observed $g_{new}[k]$ before and after the attack would both approach the theoretical derived value $\mathbb{E}\{g_{new}[k]\}$. As Fig. 6 shows, large watermark strength μ can discriminate the two distributions. In the condition that μ is small, the two distributions would be nearly identical. When it comes that $g_{new}[k]$ falls into the overlapping areas of the two distributions, it will be hard to judge the existence of intrusions. The probability density function of a discrete random variable $\mathcal{X}$ is defined as

$$F_{\mathcal{X}}(x) = P(\mathcal{X} < x) = z, \tag{30}$$

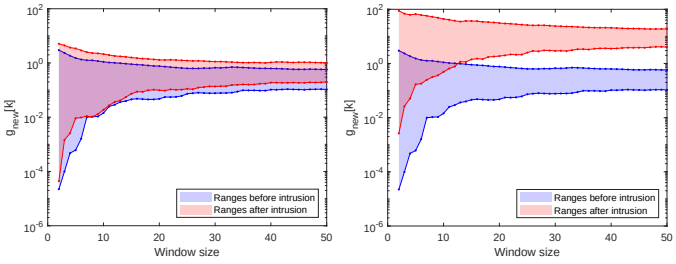
and its inverse function is

$$F_{\mathcal{X}}^{-1}(z) = x, \tag{31}$$

where z is the probability of the samples occurring with values less than the variable value x.

Random variables of $g_{new}[k]$ before and after the attack are defined separately as $\bar{\mathcal{G}}$ and $\mathcal{G}$, and there is an overlap area between distributions $\bar{\mathcal{G}}$ and $\mathcal{G}$. Function $\Psi(\mu, W)$

$$\Psi(\mu, W) = F_{\mathcal{G}}^{-1}(1-0.95) - F_{\bar{\mathcal{G}}}^{-1}(0.95) \tag{32}$$



(a) μ is small (b) μ is large

Fig. 6. The distribution of $g_{new}[k]$ over different (μ, W) pairs.

is defined as the distinction of the two distributions. $\Psi(\mu, W)$ ensures that 95% of the sample points are within the respective distinguishable intervals.

Fig. 7 is a longitudinal section with the fixed STW size in Fig. 6b, where the black-colored region contains 95% of the $\bar{\mathcal{G}}$ samples that are not attacked, and the same as the red region for $\mathcal{G}$. Constraint (24) greatly reduces the probability of $g_{new}[k]$ that appears in the green region. Qualitatively speaking, increased μ helps pull the two peaks of the distribution apart and extended W shrinks the distribution toward the expected mean.

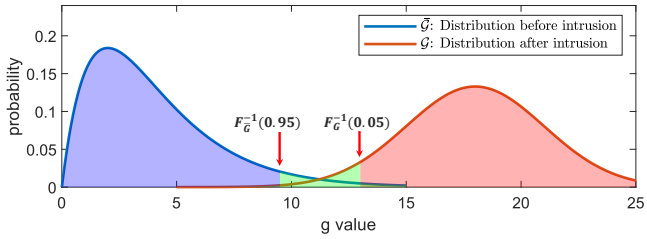


Fig. 7. Overlap area between the distribution of $\bar{\mathcal{G}}$ and that of $\mathcal{G}$.

Since we cannot obtain the analytic expression of $\Psi(\mu, W)$, we show the relationship between $\Psi(\mu, W)$ and the pair (μ, W) through the numerical results in Fig. 14. Algorithm 2 shows a numerical solver for problem (20). Considering that the detection sensitivity constraint $R(\mu)$, the signal quality constraint $\mathrm{SNWR}(\mu)$, and the detection latency constraint $T(W)$ all have explicit functions, the numerical solver directly solves the inequalities for a feasible range $[\mu_l, \mu_h]$ and the maximum length of STW W_{max}. Then, a numerical traversal search is performed for $\Psi(\mu, W)$ with the step $(\Delta\mu, \Delta W)$.

C. Attack Response Activating Function

The difference between $\mathbb{E}\{g_{new}[k]|\bar{A}\}$ and $\mathbb{E}\{g_{new}[k]|A\}$ is obvious after we choose the appropriate watermarks, but the detection metrics need to be normalized to avoid the false alarm caused by measurement noises with different strengths. For instance, under a low level of the measurement noise, $\mathbb{E}\{g_{new}[k]\}$ that changes from 1 to 10 indicates the existence of an attack, but when the noise level is large, only the $\mathbb{E}\{g_{new}[k]\}$ stepping from 10 to 100 indicates an attack. Another problem is that the traditional threshold-based detection may be confused by the sharp noise variation. Assuming that we set $\xi = 20$ in the second condition if the noise causes $g_{new}[k]$ to rise to around

Algorithm 2: Pair (μ, W) Searching Algorithm

Input: Constraints r_l, η, and T_{set}; Historical data S; Time τ
Output: Optimal (μ^*, W)

Solve $R(\mu) \geqslant r_l$ and $\mathrm{SNWR}(\mu) \geqslant \eta \Rightarrow [\mu_l, \mu_h]$;
Set $W_j \leftarrow W_{max} = T_{set}/\tau$;
Set $\mu_i \leftarrow \mu_l$;
Use S and W_{max} to generate $\bar{\mathcal{G}}$;
Add watermarks to S generated by μ_i;
Inject false data Δs to S;
Remove watermarks and get recovered samples S_a;
Use S_a and W_{max} to generate $\mathcal{G}$, create $\Psi(\mu_i, W_j)$;
while $\Psi(\mu_i, W_j) < 0$ **do**
  $\mu_i \leftarrow \mu_i +$ step size $\Delta\mu$;
  Update $\Psi(\mu_i, W_j)$ based on $line3 - 7$;
  if $\Psi(\mu_i, W_j) \geqslant 0$ *and* $\mu_i \leqslant \mu_h$ **then**
    return *Pair* (μ_i, W_j)
  else
    return *Null*
  end
end
while $\Psi(\mu_i, W_j) \geqslant 0$ **do**
  $W_j \leftarrow W_j -$ step size ΔW;
  Update $\Psi(\mu_i, W_j)$ based on $line3 - 7$;
  if $\Psi(\mu_i, W_j) < 0$ **then**
    return *Pair* $(\mu_i, W_j + \Delta W)$;
  end
end

25, the detector will incorrectly alarm, in spite of that only the raise of $g_{new}[k]$ to 100 indicates an attack.

An attack response activating function $D[k]$ standardizes the rule with noise tolerance to some extent. First, we introduce the neighbor contrast $V_{nc}[k]$

$$V_{nc}[k] = \frac{g_{new}[k]}{g_{new}[k - W]} \tag{33}$$

to calculate the relative distance between two observed points with a distance W, which considers the inherent calculation delay [41]. V_{nc} helps detectors quickly and precisely capture the rate of values changing. The two observed points are not adjacent to each other with W distance. The spikes triggered by attacks and noises are significantly different as shown in Fig. 8.

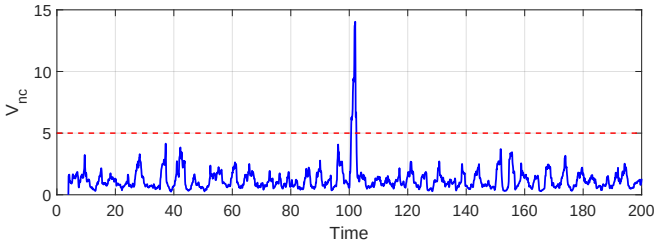


Fig. 8. The largest spike in V_{nc} caused by an attack.

$D[k]$ is with output 0/1 separately representing normal and abnormal situations. It needs to consider the previous state to

implement the flip, and the previous state reveals the level that $g_{new}[k]$ remains at. The detector will first initialize $D[0] = 0$ if there is no attack. In practical application, the new threshold α can be slightly less than e^{r_l} which we use as the detection sensitivity.

$$D[k] = \begin{cases} 1 & \{V_{nc} > \alpha\} \wedge \{D[k-1] = 0\} \\ 0 & \{V_{nc} < \alpha^{-1}\} \wedge \{D[k-1] = 1\} \\ D[k-1] & \alpha^{-1} < V_{nc} < \alpha \end{cases} \tag{34}$$

IV. Evaluations

We apply the proposed method to a platform of the Ethanol Distillation System (EDS), where the sensors contain some inherent noises [45]. The EDS contains 3 feedback control loops of liquid level, cooling water flow, and tower temperature as Fig. 9 illustrates. We mainly focus on the liquid level loop. Water evaporation caused by the heater and water reflux caused by condensation can cause small level disturbances, and sensor sampling also has errors. In this section, We use a series of real level sensor readings to model the liquid level measurements with noise and search for the optimal watermark strength using the numerical solver.

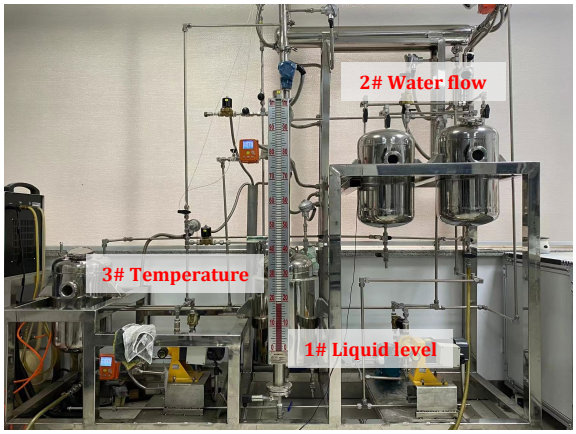


Fig. 9. The EDS platform.

A. Numerical Results for Watermark Optimization

We extract a typical piece of historical liquid level measurements to determine the block size. Fig. 10 shows that the liquid level fluctuates between 520mm and 540mm with a few spikes, due to sensor accuracy limitations and oscillations during system operation. The expected mean is 534mm with variance $\sigma_n^2 = 2.26$, and the sampling rate f_s is 50Hz. As Fig. 11 shows, the green solid line represents the theoretical $\mathbb{E}\{g_{new}[k]\}$'s downward trend following the expression $6\sigma_n/\sqrt{m}$. The red dotted line is calculated by automatic simulated signals with the normally distributed noise $\mathcal{N}(0, \sigma_n^2)$, and the black dotted line is drawn by the practical data. Both the simulation and experimental results prove that the decreasing of $\mathbb{E}\{g_{new}[k]\}$ with the block size increasing follows an approximated inverse proportion. As the target value of the detection metric is 5,

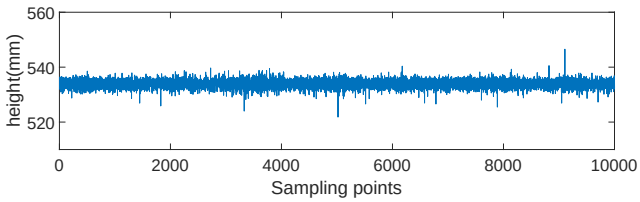


Fig. 10. A segment of real data from the liquid level sensor.

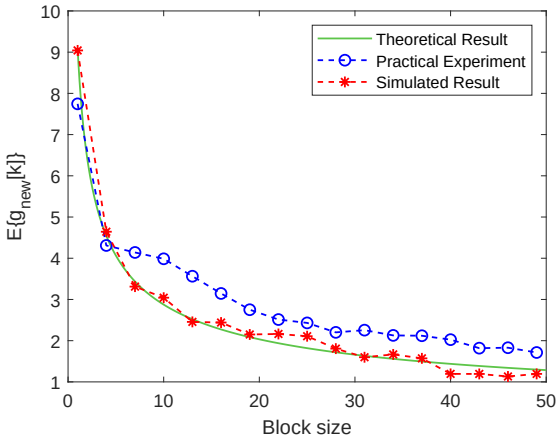


Fig. 11. The decreasing of $\mathbb{E}\{g_{new}[k]\}$ with the increasing block sizes.

we choose $m = 4$ to simplify the calculation of the following optimization problem.

Before we try to solve the watermark optimization problem for the liquid level, we conduct simulations to verify the lower bound of $\mathbb{E}\{g_{new}[k]\}$ after the attack derived in equation (26). It is worth emphasizing that this transformation is effective to enhance the difference between watermarks and noises when we model the detection sensitivity constraint. We progressively generated three typical attack vectors to test whether the $\mathbb{E}\{g_{new}[k]\}$ is above the envelope line $\sigma_\theta^{-1}E(\mu)$.

- *Ordinary attackers* directly inject expected malicious signals, although the noise characteristics of threats are far different from the original data. They generate signals without noise in this case.
- *Imitated attackers* construct an attack vector by imitating the statistical noise features of original signals, in order to bypass the manual inspection.
- *Replay attackers* record a segment of the original sensor measurements and replace the current sensor signal at a scheduled moment, in order to bypass the advanced IDS.

As Fig. 12 illustrates, it is obvious that all three attacks are captured by the red area above $E(\mu)$. Furthermore, other different shapes of attack vectors will also fall into the red area. This justifies the transformation of the detection sensitivity constraint.

For this liquid level data series, the numerical solver first attempts to choose a feasible watermark strength range within the constraints of detection sensitivity and signal quality. It calculates the original noise variance as $\sigma_n^2 = 2.26$ and the mean liquid level of the current conditions 533mm. Since the innovative sequence Θ is created with $m = 4$, $R(\mu)$ and

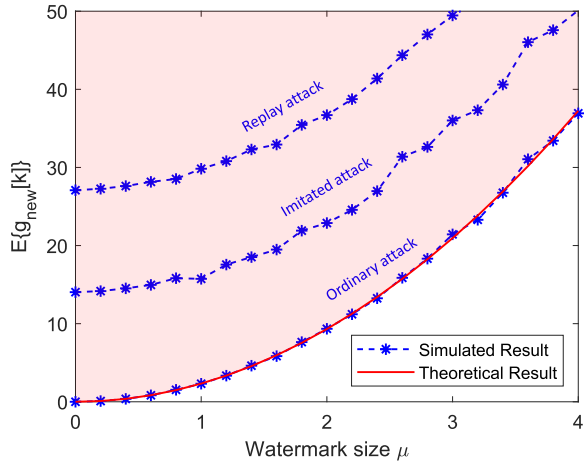


Fig. 12. The lower bound of $\mathbb{E}\{g_{new}[k]|A\}$ when μ varies.

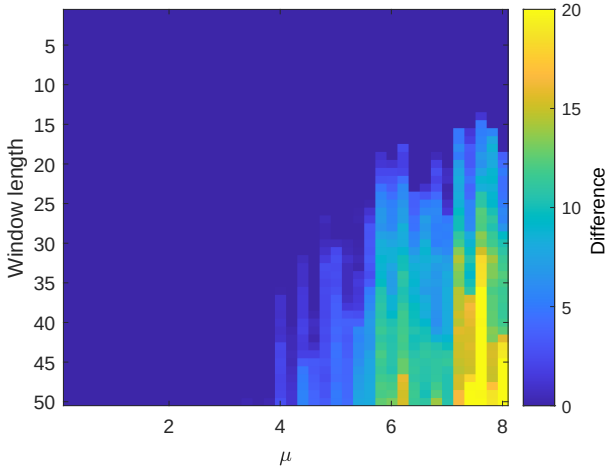


Fig. 14. The $\Psi(\mu, W)$ heat map.

SNWR(μ) are drawn in Fig. 13. We input the lower bounds $r_l = 3$ as well as the minimum SNWR $\eta = 150$, and the feasible range $[\mu_l, \mu_h]$ is $[4.90, 7.90]$. In addition, under some strict constraints, we cannot find a suitable range of μ and we do not suggest using the watermarking technique to protect data integrity because it is conflicting that the chosen watermarks have a good performance.

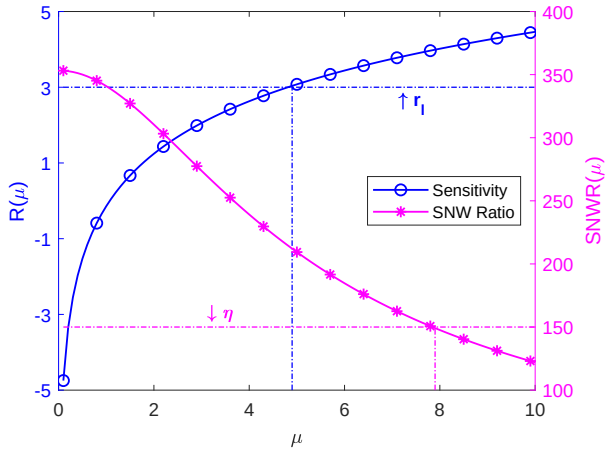


Fig. 13. Numerical results of $R(\mu)$ and SNWR(μ).

The next step considers the detector-side constraints and finally retrieves the optimal pair (μ^*, W). As we have mentioned that $\Psi(\mu, W)$ dose not have the explicit expression, we draw a heat map Fig. 14 to reveal the relationship between the dependent variable and pairs (μ, W). The numerical searching is conducted on the map. The minimum detection time delay is $m/f_s = 0.08s$, if we set $T_{set} = 2s$, the upper bound of the STW length is 25. We can search an (μ^*, W) in the region that satisfies $\mu \in [4.90, 7.90]$, $W \leqslant 25$ and $\Psi(\mu, W) \geqslant 0$. The optimal (μ^*, W) is determined as $(5.8, 20)$.

We verify the effectiveness of the designed watermark strategy under the three types of attacks mentioned above. The results are shown in Fig. 18. The ordinary attacker mimics liquid level dropping without noises, the malicious data is a segment of the ramp function. The imitated attacker tries to change the stable measurement to a periodic signal with similar characteristics to the disturbance. Both the two attacks are launched at $t = 75$s. The replay attacker just retained the characteristics of the original sensor measurement, it started sending the recorded signal at different level heights at $t = 75$s and ends at $t = 160$s. The attack response activating function $D[k]$ caught all three types of threats with clear start and end times.

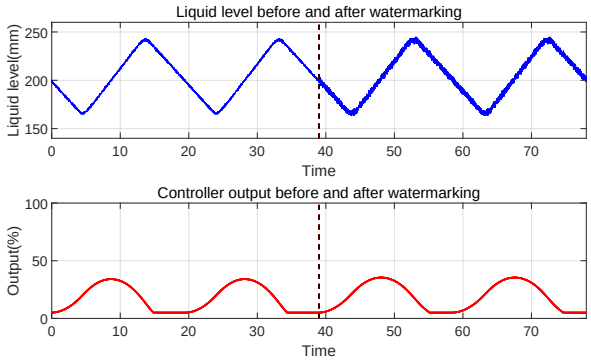


Fig. 15. Control performance before and after watermarking in EDS.

We applied our detection system to the EDS to protect the liquid level data. The programmable logic controller (PLC) is running a proportional-integral (PI) control algorithm where the coefficients for the proportional term is 1 and the integration time is $100s$. The controller keeps the liquid level at about 200mm with a specified triangular wave. The optimal watermark strength $\mu = 3.1$mm. Fig. 15 shows that the watermarked data does not affect the performance of the controller in practical applications while still ensuring the sensor data's integrity.

B. Comparison with Other Approaches

We compare the performance of the proposed method with that of RWM on the dataset collected by three different sensors in our platform. What we investigate includes signal quality SNWR(μ) and detection time delay $T(W)$. We assume that the two techniques have similar positive detection rates provided that the upper bound of $\bar{\mathcal{G}}$ and the lower bound of $\mathcal{G}$

TABLE I
PARAMETERS OF TWO METHODS

Methods	Sensor 1	Sensor 2	Sensor 3
RWM	(7.20, 50)	(0.40, 50)	(8.40, 50)
Our strategy	**(5.10, 20)**	**(0.24, 10)**	**(5.80, 20)**

distribution are clearly discriminated. The mean and variance (e, σ_n^2) of the three dataset are $(533.94, 2.26)$, $(2.55, 0.005)$, and $(666.17, 3.63)$. The sizes of the block that we get from the first process are 5, 10, and 5. Pairs (μ, W) are illustrated in Table I, and the STW length of RWM is always 50 according to the previous research.

First, we demonstrate the decrease rate of SNWR(μ) values under the conditions of no watermarking, using the proposed watermark-based method, and the conventional RWM. Although both our method and RWM have a nontrivial impact on the original signal, which is the necessary cost spent on protecting signal integrity, our method has better signal quality with the same detection performance $\Psi(\mu, W) = 0$. The result in Fig. 16 is normalized and we can find our method adds fewer fluctuations to data, as we use a smaller watermark strength μ to achieve similar detection results.

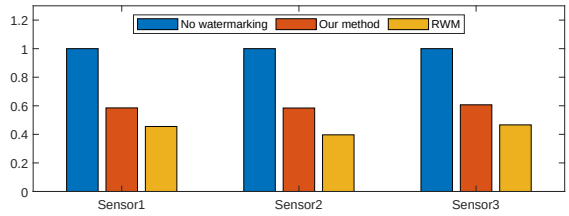


Fig. 16. The SNWRs among three sensors.

Second, the detection latency will be longer than those of the RWM in Fig. 17, but they can be controlled below the maximum detection latency requirement $2s$ as we have already modeled it in our optimization problem. In other words, our method reduces some measurement disruptions by tolerating such acceptable intrusion response delays.

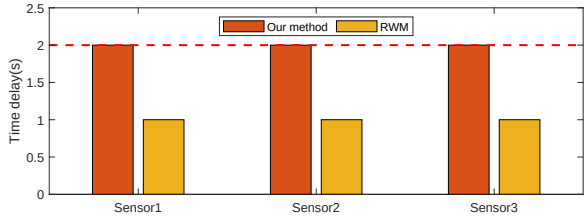


Fig. 17. The time delay among three sensors.

We qualitatively compare our watermark-based detection system with the traditional IDSs and some prevention mechanisms in three dimensions, where the former is generally based on machine learning or feature matching methods and the latter contains access control techniques, encryption algorithms, etc.

Prior knowledge of the process: Traditional IDSs require a large amount of abundant normal and anomalous operation data fed into well-designed models to obtain high accuracy, which means both learning-based and model-based IDSs need to have more system prior knowledge. System designers should pay more attention to feature engineering. Our watermark-based detection system is quite like some lightweight encryption algorithms that only focus on the pure data transferred in real-time channels, only involving data types, ranges, and other simple characteristics. Hence, our system can be easily migrated to other ICSs.

Computational and storage overhead: Most IDS solutions leverage large open-source datasets [46] or collect and use high-performance computers or servers with multiple GPUs to train models for several hours [47], which is a time-consuming task and needs sufficient computing resources. The use of firewalls or data encryption methods also affects the real-time performance of ICSs. However, our system can be directly deployed to IIoT devices with a 17KB total code size. The algorithm for adding watermarks has a low computational complexity which does not need to be removed when used for process control, it can search feasible parameters offline in seconds and operate online in microseconds.

Security policies: The watermark-based data integrity detection system ensures that sensor data remains consistent and trustworthy throughout its lifecycle, from transmitter to receiver. The watermarked signals are still directly available for the control output calculation. Unlike encryption techniques, watermarking does not ensure data confidentiality, and an attacker can still obtain legitimate information by the watermarked signal.

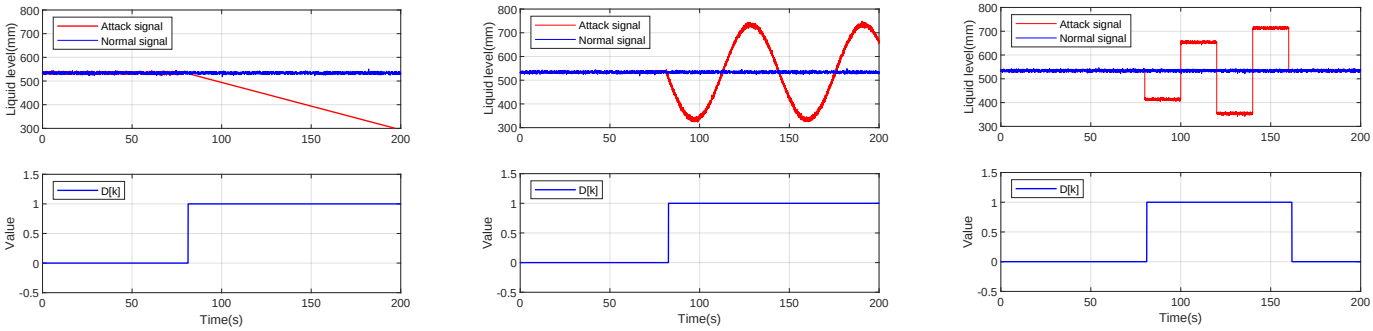
V. CONCLUSION AND FUTURE WORKS

The system uncertainty, latency, and other control objectives are difficult to model when using the watermark-based detection approach on a real-world ICS. We proposed an enhanced watermark-based detection method for practical applications. The detection performance was summarized as several mathematical expressions in a realistic way, and the formulated optimization problem determined the optimal watermark strength. The expansions of watermark-addition and intrusion response algorithms could significantly reduce the effect of realistic noises on detection performance. The experimental results based on the realistic platform validated the effectiveness and superiority of our method. In the future, we will develop more specific constraints to describe the realistic requirements of the detection/control performance or signal channels. We will also try to introduce the inherent dynamics of the controlled objects into the construction of watermarks that are much more difficult to be inferred, which can protect multiple control loops simultaneously.

APPENDIX A
PROOF OF *THEOREM 2*

The e_w can be rewritten as the composition of hash functions with different inputs like timestamps and watermarked signals

$$\begin{aligned} e_w[k] = \; & w[k] - 2w[k-1] + w[k-2] \\ = \; & h_n(k) - 2h_n(k-1) + h_n(k-2) + h_r(s_w[k-1]) \\ & - 2h_r(s_w[k-2]) + h_r(s_w[k-3]). \end{aligned}$$



(a) Ordinary attack (b) Imitated attack (c) Replay attack

Fig. 18. Effectiveness of the strategy for different types of attacks.

The calculations of $h_n(k)$, $h_n(k-1)$, $h_n(k-2)$, $h_r(s_w[k-1])$, $h_r(s_w[k-2])$, and $h_r(s_w[k-2])$ in every round are independent. The corresponding random variables are defined as H_1, H_2, H_3, H_4, H_5, and H_6, which all identically follow uniform distributions $\mathcal{U}(-\mu/2, \mu/2)$. We have

$$H_i \times H_j = 0, \text{if } i \neq j.$$

The $\mathbb{E}\{{e_w}^2\}$ represents as

$$\begin{aligned}\mathbb{E}\{{e_w}^2\} = &\quad \mathbb{E}\{{H_1}^2 + 4{H_2}^2 + {H_3}^2 + {H_4}^2 + 4{H_5}^2 + {H_6}^2\} \\ = &\quad 12\mathbb{E}\{H^2\},\end{aligned}$$

where H is a unified description of H_i. Let $X = \frac{H+\mu/2}{\mu}$ which follows the distribution $\mathcal{U}(0,1)$, we update $\mathbb{E}\{H^2\}$:

$$\begin{aligned}\mathbb{E}\{H^2\} = &\quad \mathbb{E}\{(\mu X - \frac{\mu}{2})^2\} \\ = &\quad \mathbb{E}\{\mu^2 X^2 - \mu^2 X + \frac{\mu^2}{4}\} \\ = &\quad \frac{\mu^2}{3} - \frac{\mu^2}{2} + \frac{\mu^2}{4} = \frac{\mu^2}{12},\end{aligned}$$

where we have the property from [48]

$$\mathbb{E}\{X^k\} = \frac{1}{k+1}; k = 1, 2, 3, \cdots.$$

The lower bound of $\mathbb{E}\{g_{new}[k]|A\}$ is

$$\mathbb{E}\{g_{new}[k]|A\} \geqslant {\sigma_\theta}^{-1}\mu^2$$

References

[1] K. Stouffer, J. Falco, K. Scarfone *et al.*, "Guide to industrial control systems (ics) security," *NIST special publication*, vol. 800, no. 82, pp. 16–16, 2011.

[2] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (iiot): An analysis framework," *Computers in industry*, vol. 101, pp. 1–12, 2018.

[3] Z. Zhang, R. Deng, D. K. Yau, P. Cheng, and J. Chen, "Analysis of moving target defense against false data injection attacks on power grid," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2320–2335, 2019.

[4] E. Biham, S. Bitan, A. Carmel, A. Dankner, U. Malin, and A. Wool, "Rogue7: Rogue engineering-station attacks on s7 simatic plcs," *Black Hat USA*, 2019.

[5] L. Garcia, F. Brasser, M. H. Cintuglu, A.-R. Sadeghi, O. A. Mohammed, and S. A. Zonouz, "Hey, my malware knows physics! attacking plcs with physical model aware rootkit." in *NDSS*, 2017.

[6] T. Müller, A. Walz, M. Kiefer, H. D. Doran, and A. Sikora, "Challenges and prospects of communication security in real-time ethernet automation systems," in *2018 14th ieee international workshop on factory communication systems (wfcs)*. IEEE, 2018, pp. 1–9.

[7] L. Xiao, X. Lu, T. Xu, W. Zhuang, and H. Dai, "Reinforcement learning-based physical-layer authentication for controller area networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2535–2547, 2021.

[8] D. Formby and R. Beyah, "Temporal execution behavior for host anomaly detection in programmable logic controllers," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1455–1469, 2019.

[9] D. Fauri, B. de Wijs, J. den Hartog, E. Costante, E. Zambon, and S. Etalle, "Encryption in ics networks: A blessing or a curse?" in *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2017, pp. 289–294.

[10] J. Nivethan and M. Papa, "On the use of open-source firewalls in ics/scada systems," *Information Security Journal: A Global Perspective*, vol. 25, no. 1-3, pp. 83–93, 2016.

[11] A. S. Sani, D. Yuan, P. L. Yeoh, J. Qiu, W. Bao, B. Vucetic, and Z. Y. Dong, "Cyra: A real-time risk-based security assessment framework for cyber attacks prevention in industrial control systems," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2019, pp. 1–5.

[12] T. H. Morris, Z. Thornton, and I. Turnipseed, "Industrial control system simulation and data logging for intrusion detection system research," *7th annual southeastern cyber security summit*, pp. 3–4, 2015.

[13] B. Li, R. Lu, W. Wang, and K.-K. R. Choo, "Ddoa: A dirichlet-based detection scheme for opportunistic attacks in smart grid cyber-physical system," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, pp. 2415–2425, 2016.

[14] S. Ghosh, M. R. Bhatnagar, W. Saad, and B. K. Panigrahi, "Defending false data injection on state estimation over fading wireless channels," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1424–1439, 2020.

[15] D. Huang, X. Shi, and W.-A. Zhang, "False data injection attack detection for industrial control systems based on both time-and frequency-domain analysis of sensor data," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 585–595, 2020.

[16] Z. Yang, L. He, P. Cheng, J. Chen, D. K. Yau, and L. Du, "{PLC-Sleuth}: Detecting and localizing {PLC} intrusions using control invariants," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, 2020, pp. 333–348.

[17] Z. Yang, L. He, H. Yu, C. Zhao, P. Cheng, and J. Chen, "Reverse engineering physical semantics of plc program variables using control invariants," in *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 548–562.

[18] M. Liu, C. Zhao, Z. Zhang, R. Deng, P. Cheng, and J. Chen, "Converter-based moving target defense against deception attacks in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3984–3996, 2022.

[19] M. Liu, C. Zhao, J. Xia, R. Deng, P. Cheng, and J. Chen, "Pddl: Proactive distributed detection and localization against stealthy deception attacks in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 714–731, 2023.

[20] M. Liu, C. Zhao, Z. Zhang, and R. Deng, "Explicit analysis on effectiveness and hiddenness of moving target defense in ac power systems," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4732–4746, 2022.

[21] C. M. Ahmed, J. Zhou, and A. P. Mathur, "Noise matters: Using sensor and process noise fingerprint to detect stealthy cyber attacks and authenticate sensors in cps," in *Proceedings of the 34th Annual Computer Security Applications Conference*, 2018, pp. 566–581.

[22] S. Mokhtari, A. Abbaspour, K. K. Yen, and A. Sargolzaei, "A machine learning approach for anomaly detection in industrial control systems based on measurement data," *Electronics*, vol. 10, no. 4, p. 407, 2021.

[23] E. Hallaji, R. Razavi-Far, M. Wang, M. Saif, and B. Fardanesh, "A stream learning approach for real-time identification of false data injection attacks in cyber-physical power systems," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3934–3945, 2022.

[24] S. Ahmed, Y. Lee, S.-H. Hyun, and I. Koo, "Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2765–2777, 2019.

[25] M. Higgins, F. Teng, and T. Parisini, "Stealthy mtd against unsupervised learning-based blind fdi attacks in power systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1275–1287, 2021.

[26] W. Xu, M. Higgins, J. Wang, I. M. Jaimoukha, and F. Teng, "Blending data and physics against false data injection attack: An event-triggered moving target defence approach," *IEEE Transactions on Smart Grid*, pp. 1–1, 2022.

[27] J. M. Beaver, R. C. Borges-Hink, and M. A. Buckner, "An evaluation of machine learning methods to detect malicious scada communications," in *2013 12th international conference on machine learning and applications*, vol. 2. IEEE, 2013, pp. 54–59.

[28] S. K. Biswas *et al.*, "Intrusion detection using machine learning: A comparison study," *International Journal of pure and applied mathematics*, vol. 118, no. 19, pp. 101–114, 2018.

[29] M. Gaiceanu, M. Stanculescu, P. C. Andrei, V. Solcanu, T. Gaiceanu, and H. Andrei, "Intrusion detection on ics and scada networks," in *Recent Developments on Industrial Control Systems Resilience*. Springer, 2020, pp. 197–262.

[30] M. Chen, J. Fridrich, M. Goljan, and J. Lukás, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.

[31] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 2009, pp. 911–918.

[32] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2013.

[33] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "A new watermarking approach for replay attack detection in lqg systems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5143–5148.

[34] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Optimal periodic watermarking schedule for replay attack detection in cyber–physical systems," *Automatica*, vol. 112, p. 108698, 2020.

[35] Z. Song, A. Skuric, and K. Ji, "A recursive watermark method for hard real-time industrial control system cyber-resilience enhancement," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 1030–1043, 2020.

[36] W. Zhang, M. S. Branicky, and S. M. Phillips, "Stability of networked control systems," *IEEE control systems magazine*, vol. 21, no. 1, pp. 84–99, 2001.

[37] K. G. Shin and X. Cui, "Computing time delay and its effects on real-time control systems," *IEEE Transactions on control systems technology*, vol. 3, no. 2, pp. 218–224, 1995.

[38] K. Jerath, S. Brennan, and C. Lagoa, "Bridging the gap between sensor noise modeling and sensor characterization," *Measurement*, vol. 116, pp. 350–366, 2018.

[39] S. Pinto and N. Santos, "Demystifying arm trustzone: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[40] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971.

[41] Y. Liu, T. Dillon, W. Yu, W. Rahayu, and F. Mostafa, "Noise removal in the presence of significant anomalies for industrial iot sensor data in manufacturing," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7084–7096, 2020.

[42] D. W. Allan, "Statistics of atomic frequency standards," *Proceedings of the IEEE*, vol. 54, no. 2, pp. 221–230, 1966.

[43] N. El-Sheimy, H. Hou, and X. Niu, "Analysis and modeling of inertial sensors using allan variance," *IEEE Transactions on instrumentation and measurement*, vol. 57, no. 1, pp. 140–149, 2007.

[44] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Cost-effective watermark based detector for replay attacks on cyber-physical systems," in *2017 11th Asian Control Conference (ASCC)*. IEEE, 2017, pp. 940–945.

[45] Z. Yang, L. He, H. Yu, C. Zhao, P. Cheng, and J. Chen, "Detecting plc intrusions using control invariants," *IEEE Internet of Things Journal*, 2022.

[46] S. Potluri, S. Ahmed, and C. Diedrich, "Convolutional neural networks for multi-class intrusion detection system," in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2018, pp. 225–238.

[47] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial iot networks," *IEEE Internet of Things Journal*, 2021.

[48] L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*. Courier Corporation, 2012.