Discover Computing

RESEARCH Open Access

Cyber-physical anomaly detection a deep adversarial fusion of sensor and network data

Andrea Pinto[1*], Luis-Carlos Herrera[2], Yezid Donoso[1] and Jairo A. Gutierrez[3]

*Correspondence:
Andrea Pinto
ya.pinto10@uniandes.edu.co
[1]Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, Bogotá 111711, Colombia
[2]Faculty of Fundamental Sciences, Department of Informatics Engineering, Vilnius Gediminas Technical University, Sauletekio al. 11, Vilnius LT-10223, Lithuania
[3]Networking and Security Research Centre, Department of Computer Science and Software Engineering, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand

Abstract
In critical infrastructure, the convergence of physical systems with digital networks forms complex Cyber-Physical Systems (CPS), that are vulnerable to threats compromising both data and physical operations. Traditional security systems, often focused solely on network traffic, create a significant security gap by neglecting the rich contextual data provided by physical sensors. To address this issue, the paper introduces a novel unsupervised multimodal framework that synthesizes data from these dual sources for holistic anomaly detection. The proposed architecture combines pre-trained Variational Autoencoder-Long Short-Term Memory (VAE-LSTM) networks to model temporal dependencies with a dual cross-attention mechanism for deep fusion of latent representations. To enhance the detection of subtle, low-observability threats, the model is further regularized through adversarial training using a discriminator that distinguishes between original and reconstructed data. Evaluated on the comprehensive SWAT dataset, the model successfully identifies 24 out of 26 relevant attack scenarios using 10-second time sequences and achieves an Area Under the Curve (AUC) of 0.87, outperforming unimodal benchmarks. This work validates the critical importance of deep data fusion and presents a more resilient, context-aware defense mechanism for modern CPS.

Keywords Unsupervised learning, Cybersecurity, Critical infrastructure, Anomaly detection, Multimodal

1 Introduction

As critical infrastructures (CIs) increasingly rely on interconnected and automated systems, the cybersecurity landscape has undergone significant changes [1]. These infrastructures extend beyond digital networks, integrating physical components such as sensors and actuators. This integration forms Cyber-Physical Systems (CPS), where cybersecurity threats can have both virtual and physical consequences. Cyberattacks on Cis frequently aim to disrupt physical operations [2], as demonstrated in recent incidents targeting water systems in the United States, Poland, and France. One illustrative example occurred in Muleshoe, Texas, where hackers managed to overflow a water tower, sending tens of thousands of gallons of water into streets and drainpipes [3].

Traditional security measures, which predominantly focus on a single data source, such as network traffic or log data, are often inadequate for protecting the entire ecosystem, as they overlook the physical indicators of cyber threats. Incorporating both network and physical data (from sensors and actuators) into cybersecurity strategies is therefore essential. While network data provides insights into communication patterns and potential breaches in digital traffic, physical data offers a real-time snapshot of the infrastructure's operational state. Anomalies in sensor readings or unexpected actuator behaviors can signal malicious activities or system failures that might not be evident from network data alone. This unimodal focus creates a critical blind spot: stealthy, coordinated attacks, where subtle anomalies in one domain are only interpretable with context from the other, remain undetected. Addressing this gap motivates the development of a sophisticated fusion mechanism capable of learning the direct, causal relationships between cyber commands and their physical manifestations.

Moreover, anomalies themselves do not always exhibit uniform behavior. As discussed in [4], anomalies can be categorized into three types: point anomalies, which are independent of time; contextual anomalies, which rely on surrounding values for identification; and collective anomalies, which occur over a sequence, showing abnormal behavior across a period. Although numerous methods have been proposed to detect these anomalies, no single approach performs optimally across all anomaly types when using various information sources.

Anomaly detection has emerged as a crucial machine learning technique for identifying deviations from normal patterns, making it particularly well suited for securing critical infrastructures against sophisticated threats, including zero-day attacks [5, 6]. Unlike signature-based methods that depend on known attack patterns, anomaly detection models establish a baseline of normal system operations, enabling them to identify novel and previously unseen anomalies. By utilizing both physical and network data, these models can detect subtle shifts in system behavior that may indicate the early stages of a zero-day attack. This dual-data approach significantly enhances the model's sensitivity to critical deviations, allowing for the timely detection of complex cyberattacks that could compromise essential operations [7]. The comprehensive nature of this method allows for a quick and effective response to emerging threats, positioning anomaly detection not only as a proactive defense mechanism but also as a key factor in preserving the integrity and availability of critical infrastructures.

In this paper, we introduce an advanced attention-based adversarial autoencoder for unsupervised anomaly detection in sensor and network data. The model fuses latent representations from pre-trained VAE-LSTM encoders using a cross-attention mechanism to capture the interdependencies between both data types. An attention mechanism further refines this combined latent space, enhancing the most relevant features. The model then reconstructs the sensor and network data separately, combining these outputs into a unified reconstruction, which is fed into a discriminator to evaluate its plausibility across both modalities. To optimize the discriminator's performance, hyperparameters are fine-tuned using Bayesian techniques, and it is trained with a binary cross-entropy loss function. The model's overall loss function merges reconstruction error with adversarial feedback, thereby improving its ability to detect subtle anomalies. This architecture specifically leverages the integration of sensor and network data, highlighting the importance of joint data processing for robust anomaly detection in CPS. The paper

details the development of this model, insights from its application on the Secure Water Treatment (SWAT) dataset [8], and its contribution to enhancing the resilience of critical infrastructures against cyber threats.

In the increasingly complex landscape of critical infrastructure, relying solely on a single data source is insufficient for comprehensive anomaly detection. Integrating multiple data sources provides a holistic view of system operations, capturing threats that might manifest differently across these domains. Building on this understanding, the present study poses the following research question: How can an AI model be designed to integrate multiple data sources, including physical data and network traffic, for the effective detection of anomalies related to cyberattacks on critical infrastructures?

The main contributions of this research to the cybersecurity domain for critical infrastructures are as follows:

- *A Novel Multimodal Fusion Architecture* We propose a new deep learning architecture that fuses heterogeneous sensor and network traffic data for CPS anomaly detection using a dual cross-attention mechanism. This novel approach enables the model to learn complex, time-dependent interdependencies between the cyber and physical domains, a capability fundamentally absent in prior unimodal or simply concatenated systems.
- *Enhanced Detection of Low-Observability Threats* We integrate an adversarial autoencoder framework to regularize the model's latent space. This forces the model to learn a highly precise and robust representation of normal system behavior, significantly improving its sensitivity to subtle zero-day anomalies that closely mimic legitimate operations and would otherwise go undetected.
- *A Robust Preprocessing and Evaluation Methodology* We develop a reproducible preprocessing pipeline to synchronize and engineer features from high-frequency sensor and network data. By retaining all physical device features and utilizing short 10-second analysis windows, our model demonstrates near real-time detection capabilities and improved generalization compared to methods that discard features or require longer observation periods for analysis.

The remainder of this paper is structured as follows: Sect. 2 provides background information; Sect. 3 explains the SWAT dataset; Sect. 4 reviews related work; Sect. 5 details the preprocessing steps; Sect. 6 describes the multimodal design; Sect. 7 presents the results; and Sect. 8 concludes the paper.

1.1 Methodology

Design Science Research Methodology (DSRM), as proposed in [9], has become an essential framework within the field of information systems research, providing a systematic and iterative approach to developing and evaluating innovative artifacts.

Unlike conventional research methodologies that primarily focus on observing or explaining phenomena, DSRM is proactive in designing solutions that address complex, real-world problems. It emphasizes the conception, implementation, and refinement of artifacts–whether models, frameworks, processes, or systems - that not only advance theoretical understanding but also offer tangible improvements to practice. At the core of DSRM is its problem-centric focus, where researchers begin by identifying a specific issue or gap in the existing knowledge base, thereby establishing a targeted research

agenda. The methodology then guides the researcher through a multi-phase process, starting with problem identification and motivation, moving to the rigorous design and development of the artifact, demonstrating its practical utility and finally evaluating its effectiveness in addressing the problem. The iterative nature of DSRM allows for continuous refinement, ensuring that the artifact evolves to meet both practical needs and academic rigor. By bridging the gap between theoretical exploration and practical application, DSRM not only deepens our understanding of complex systems but also drives the creation of actionable, solution-oriented research, thereby solidifying its role as a vital methodology for the advancement of information systems and related fields [9].

Given the complexities of integrating physical systems with digital networks in critical infrastructure environments, the research demands a methodology that not only provides theoretical insights but also facilitates the development of practical, adaptable solutions. In this context, the selected research approach focuses on creating an artifact that can tackle real-world challenges, such as identifying cyber-physical threats, while allowing for ongoing refinement as new attack vectors and vulnerabilities emerge. This methodology is particularly effective for capturing the intricate interplay between high-frequency physical and network data streams, a feature that traditional methods often overlook. By guiding the design, implementation, and validation phases in a systematic manner, the approach ensures that the resulting model is comprehensive, thereby meeting both theoretical rigour and practical utility. It emphasizes iterative evaluation and improvement, enabling the proposed anomaly detection model to adapt to dynamic and evolving cybersecurity landscapes in critical infrastructures, as shown in Fig. 1. This results in a more robust solution, designed to proactively respond to the ever-changing nature of cyber threats in industrial systems.

2 Background

In machine learning and deep learning, advanced techniques such as VAE and LSTM networks have been instrumental in addressing complex data representation and sequence modeling challenges. VAE encodes input data into a probabilistic latent space, enabling the generation of new data points, while LSTM networks effectively capture long-term dependencies in sequential data. The integration of these models, known as VAE-LSTM, combines their strengths for sequence prediction and reconstruction.

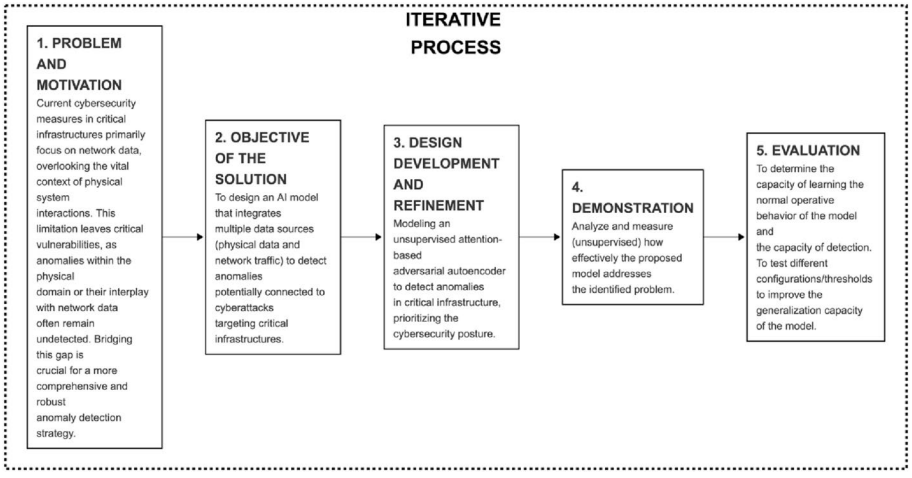

Fig. 1 DSRM methodology

Attention mechanisms further enhance these models by focusing on the most relevant parts of the input, improving accuracy and interpretability. Additionally, AAEs incorporate a discriminator to regularize the latent space, thereby boosting robustness, particularly in anomaly detection. The following sections will detail these techniques as integral components of the proposed model.

2.1 Variational autoencoders

Variational Autoencoders (VAE) are generative models that operationalize Bayesian inference to learn latent representations of complex data distributions in an unsupervised framework. A VAE conceptually differs from traditional autoencoders by representing input data through probabilistic latent variables rather than deterministic points. This representation is defined by the distribution parameters: means (μ) and variances (σ^2), which collectively describe a Gaussian distribution [10] .

The architecture of a VAE encompasses two principal components: the encoder and the decoder. The encoder function, $q_{\emptyset}\,(z|x)$, parameterizes the latent space variables (z) conditioned on the observed data (x), thereby approximating the true posterior $p\,(z|\,x)$. The decoder, denoted as $p_{\theta}\,(z|x)$,reconstructs the input data from the sampled latent variables. The optimization objective integrates the reconstruction error, typically calculated as the negative log-likelihood of the observed data given the latent variables, and the Kullback-Leibler (KL) divergence. This divergence quantifies the discrepancy between the learned latent distribution $q_{\emptyset}\,(z|x)$ and a prior distribution $p\,(z)$, often assumed to be the standard normal distribution. Formally, the loss function is expressed as: $L(\theta, \varphi; x) = -E_{q\emptyset(z|x)}[\log p\theta(x|z)] + KL(q_{\emptyset}(z|x)||(z))$. The following pseudocode outlines the essential steps involved in encoding an input into a latent space and reconstructing it, using the reparameterization trick to allow for gradient descent through random processes [10].

Pseudocode: Variational Autoencoder (VAE)

Inputs:
    x - Input data
Outputs:
    reconstructed_x - Reconstructed output
    z - Latent vector
    mu - Mean of latent space
    log_var - Log variance of latent space
Procedure:
1. Encode:
    mu, log_var = encoder(x)
2. Reparameterization:
    epsilon = random_noise ()
    *z = mu + exp (0.5 * log_var) * epsilon*
3. Decode:
    reconstructed_x = decoder(z)

Return (reconstructed_x, z, mu, log_var)

VAEs facilitate the identification of deviations that signify potential security breaches or system malfunctions. This capability is invaluable in critical infrastructures such as power grids or telecommunications networks, where early detection of anomalies can prevent catastrophic failures and ensure operational continuity. Moreover, the

unsupervised nature of VAEs alleviates the burdensome requirement for labeled training data, which is often scarce or unavailable in real-world settings.

2.2 Long short-term memory networks

Long Short-Term Memory (LSTM) networks are a specialized form of Recurrent Neural Networks (RNN) engineered to effectively handle long-term dependencies within sequence data, addressing the challenges posed by vanishing and exploding gradients that are common in standard RNNs. LSTMs utilize a unique gating mechanism comprising input, forget, and output gates that orchestrate the flow of information. These gates apply specific mathematical operations to determine data retention and omission through the sequence: the forget gate $f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$decides which parts of the cell state to discard, the input gate $i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$ selects new data to update the state, and the output gate $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ filters the output from the current cell state $h_t = o_t * tanh\left(C_t\right)$ [11]. LSTMs are adept at processing and analyzing time-series data, a common format in cybersecurity logs and network traffic, where THE temporal sequence and context of events can be crucial for identifying malicious activities. The core strength of LSTMs lies in their ability to learn long-term dependencies, enabling them to recognize complex patterns and anomalies that unfold over extensive periods—a feature indispensable for detecting sophisticated cyber threats that may not be immediately apparent. Furthermore, LSTMs mitigate one of the significant challenges in cybersecurity anomaly detection: the high rate of false positives. By effectively learning normal behavioral patterns over time and discerning subtle deviations, LSTMs can significantly reduce false alarms, which are a common issue with less sophisticated detection methods. The following pseudocode outlines the key steps involved in LSTM network architecture, detailing the sequence processing operations within the network.

```
Pseudocode: Long Short-Term Memory (LSTM) Network

Inputs:
        X - Sequence of input vectors
        H_prev - Previous hidden state
        C_prev - Previous cell state
Outputs:
        H - Current hidden state
        C - Current cell state after updates
Procedure:
        // Initialize gate weights

        W_f, W_i, W_c, W_o - Weight matrices for forget, input, candidate,
        and output gates

        foreach input x_t in X do:
                combined = [H_prev, x_t]

         // Gate operations
         f_t = activation_function(W_f * combined) // Forget gate decision
         i_t = activation_function(W_i * combined) // Input gate decision
         C_hat_t = activation_ function (W_c * combined) // New candidate
         values

         // State update
        C_t = f_t * C_prev + i_t * C_hat_t // Update cell state
        o_t = activation_function(W_o * combined) // Output gate decision

         // Hidden state output
        H_t = o_t * activation_function (C_t)

         // Update states for next timestep
        H_prev = H_t
        C_prev = C_t

return H, C
```

2.3 Variational autoencoder long short-term memory

Variational Autoencoder Long Short-Term Memory (VAE-LSTM) networks merge the generative capabilities of VAE with the sequential data processing strength of Long Short-Term Memory (LSTM) networks. This integration equips VAE-LSTM to effectively encode and decode complex temporal sequences into a learned latent space, making them particularly adept for tasks that require modeling intricate time-dependent dynamics. The LSTM components in both the encoder and decoder of the VAE-LSTM handle the sequence input and output, respectively, while the VAE structure ensures that the model captures the probabilistic characteristics of the data. This dual functionality enables VAE-LSTM to identify anomalies by detecting deviations from learned normal sequences. In Fig. 2, the architecture of a VAE-LSTM is shown.

2.4 Attention and fusion mechanisms

The concept of attention mechanisms was introduced in 2014 and was designed to address limitations in the encoder-decoder framework commonly used in sequence-to-sequence models. Specifically, it aims to improve the model's ability to remember long sequences without losing context, which was a challenge with earlier models that relied on a fixed-length context vector [12].

In the field of cybersecurity, especially within critical infrastructure networks, the adaptive capabilities of attention mechanisms represent a significant leap forward. By enabling anomaly detection systems to focus on specific intervals where unusual patterns may indicate potential security breaches, these mechanisms adaptively prioritize segments of data that deviate from established norms. Fusion in the context of machine learning, particularly when applied to multimodal data, facilitates a comprehensive understanding by combining features from disparate sources at various stages of the model architecture [13, 14]. Early fusion might involve the direct integration of raw data streams, enabling interaction at the initial layers of the network, whereas late fusion typically focuses on combining higher-level features or decisions near the output layer, thereby enhancing final decision-making accuracy [15].

For example, in multimodal learning scenarios such as environmental monitoring systems, fusion allows for the integration of diverse sensor inputs, ranging from visual to acoustic signals, ensuring that the resultant predictions are based on a holistic view of the environment [16]. Particularly, cross-attention fusion is a technique employed to integrate information from multiple data modalities or sequences, enabling one sequence to attend to another. Unlike self-attention, where the same sequence serves

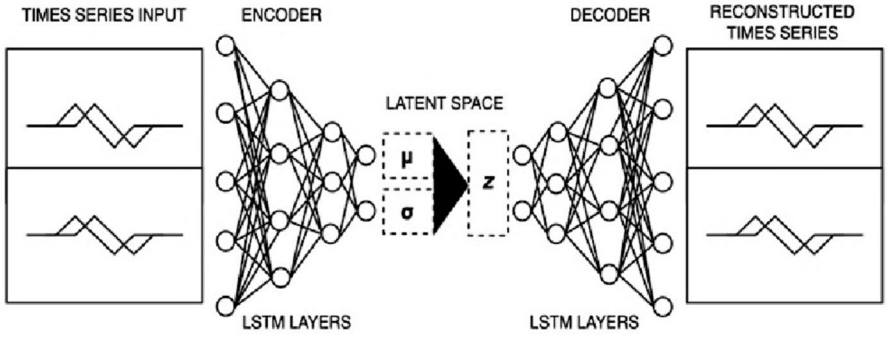

Fig. 2 VAE-LSTM architecture

as the source for queries, keys, and values, cross-attention derives the query from one sequence and the key and value from another. This mechanism allows the model to effectively focus on relevant parts of the second sequence based on the context provided by the first sequence, making it particularly useful in multimodal learning scenarios [17, 18].

For anomaly detection, combining sensor data and network traffic data through cross-attention fusion facilitates the extraction of complementary features, capturing intricate dependencies across these modalities. Sensor data might reveal anomalies in physical operations, such as unauthorized access or tampering, while network traffic data can expose digital threats like distributed denial-of-service (DDoS) attacks, malware communication, or data exfiltration attempts. This integration is crucial because it provides a comprehensive view of both physical and digital behaviors, significantly enhancing the detection of sophisticated cyber threats that might otherwise go unnoticed if each data type were analyzed in isolation [15].

2.5 The adversarial autoencoder

The Adversarial Autoencoder (AAE) is a model that combines autoencoding with adversarial training to enforce structured and meaningful latent representations, as shown in Fig. 3. In an AAE, the encoder compresses input data into a lower-dimensional latent space, while the decoder reconstructs the original data from this latent representation. The innovation of the AAE lies in its use of a discriminator network, which distinguishes between latent vectors produced by the encoder and those sampled from a prior distribution, such as a Gaussian distribution. This regularization of the latent space ensures that the representations are meaningful and structured in a way that captures the underlying patterns of the input data [19]. When applied to tasks such as anomaly detection in cybersecurity, this well-regularized latent space allows the model to more effectively identify deviations from normal behavior, [as anomalies typically result in latent representations that differ from the learned distribution, making them easier to detect].

In an AAE, the discriminator is trained to maximize the log-likelihood of correctly classifying real latent vectors z sampled from the prior $p(z)$ while minimizing the log-likelihood of incorrectly classifying latent vectors generated by the encoder $q_{\varphi}(z|x)$. This objective is formalized as a minimax optimization problem, expressed as:

$$\min_{\phi} \max_{\psi} \left[E_{z\sim p(z)} \left[\log D_{\psi}(z) \right] + E_{x\sim p(x)} \left[\log \left(1 - D_{\psi} \left(q_{\phi}(z|x) \right) \right) \right] \right]$$

where D_{ψ} denotes the discriminator's output, ψ represents the discriminator's parameters, and φ represents the encoder's parameters [20, 21]. The following pseudocode presents the training procedure of the discriminator in an AAE, outlining how it learns to differentiate between latent vectors generated by the encoder and those sampled from a prior distribution:

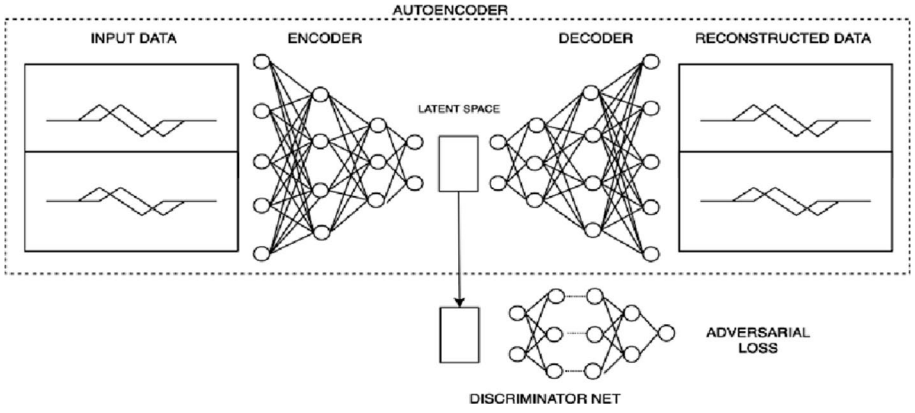

Fig. 3 AAE architecture

```
Pseudocode for the Discriminator in AAE

Inputs:

    Latent vector z from the encoder
    Latent vector z_prior from the prior distribution

Outputs:

    Discriminator loss L_D
    Updated discriminator parameters

Procedure:

    Sample Latent Vectors:
        z = Sample latent vector from encoder
        z_prior = Sample latent vector from prior distribution

    Discriminator Evaluation:
        D_real = Discriminator(z_prior) # Output for real latent
vector
        D_fake = Discriminator(z) # Output for fake latent vector
    Compute Discriminator Losses:
        L_real = BinaryCrossEntropy(1, D_real) # Loss for real latent
vector
        L_fake = BinaryCrossEntropy(0, D_fake) # Loss for fake
latent vector

    Calculate Total Discriminator Loss:
        L_D = L_real + L_fake

    Update Discriminator Parameters:
        Update discriminator parameters using gradients of L_D
```

Thus, our model integrates VAE-LSTM architectures, attention mechanisms, and AAE components to form a comprehensive framework for anomaly detection within critical infrastructure systems. Each of these techniques contributes distinct capabilities, from capturing complex temporal dependencies to identifying variations and leveraging adversarial training to improve robustness. The incorporation of both physical data and network traffic enables a holistic analysis of system behavior, facilitating the detection of deviations that may indicate cyber-attacks. This multimodal approach significantly enhances the model's ability to generalize across diverse data sources, addressing the intricate and evolving nature of cybersecurity threats in CPS.

3 Dataset

The Secure Water Treatment (SWAT) dataset provides a detailed snapshot of the operations and security challenges in a scaled-down but fully functional water treatment facility. The system was designed as a multi-stage process that mimics a real-world water treatment plant with six main stages: raw water storage and pre-treatment, chemical dosing, ultrafiltration, dechlorination, reverse osmosis, and a final water storage stage. Each stage serves a specific function in the purification process, from initial filtration and chemical adjustment to more intensive cleansing methods like reverse osmosis, ensuring the production of clean and safe drinking water. This setup not only provides a controlled environment for research on water treatment but also serves as a testbed for cybersecurity measures, given its integration of physical processes with networked control systems. The SWAT dataset includes network traffic information and sensor data: 51 sensors, actuators, and PLC control devices, among other components. In its A2 version, data were collected over an 11-day period, consisting of 7 days of normal operation and 4 days of attack scenarios. SWAT dataset al.so includes a diverse range of both cyber and physical attack scenarios - the attacks were conceived on the premise that they were initiated by either a nation-state actor or an insider [22], with a total of 41 distinct attacks executed to test system resilience. The list of attacks is in Table 1. However, five attacks (attacks number 5, 9, 12, 15, and 18) had no physical impact [8]. The attacks varied in duration and had diverse impacts on the process dynamics, necessitating varying stabilization times for the system. In cases where multiple attacks were executed sequentially, the system was unable to stabilize between these incidents [23].

4 Related work

In [2], the authors proposed a detection technique based on a neural network (NN) with a one-class objective function and an additional regularization term. The proposed method was evaluated using only the SWAT physical dataset (actuators and sensor data). This work demonstrated superior recall values in 15 of the 36 attack scenarios that had a physical impact. However, the authors included attacks that, according to the official documentation, did not have any physical impact, such as attacks number 5, 9, 12, 15, and 18. Additionally, attack number 4, targeting a motorized valve named MV-504, could not be analyzed as information about this valve was not included in the official dataset. Using the same dataset, researchers in [24] employed two unsupervised machine learning techniques to detect anomalies in water treatment systems. They utilized a Deep Neural Network (DNN) featuring feedforward layers, and a One-class Support Vector Machine (OCSVM). While the OCSVM detected a higher number of anomalies, the DNN resulted in fewer false positives. Nonetheless, the recall rates for both models were reported to be under 0.7, suggesting significant room for improvement in anomaly detection capabilities within such critical systems. This underscores the ongoing need to develop and refine machine learning approaches that can more effectively identify potential threats in industrial control environments. Although the authors of both studies recognize that the attack scenarios are representative of typical network-based attacks on cyber-physical systems, neither of them included the network traffic data to train their models.

Based on the spatiotemporal correlation and dependencies between parameter values of sensors and controllers in cyber-physical systems (CPS), a method for detecting

abnormal states through the calculation of statistical deviations was proposed in [25]. The core concept of this anomaly detection approach involves using a 1D Convolutional Neural Network (1D_CNN) and a Gated Recurrent Unit (GRU) to predict the values for the next moment or sequence based on the previous period's data flow. The method then assesses anomalies by comparing the predicted values with the actual values. In this experiment, attack scenarios 4, 10, 11, and 24 were not detected, which aligns with those scenarios that either do not have a physical impact on the water treatment system or were unsuccessful.

Another CNN-based method was introduced in [23], where the authors proposed an algorithmic design for detecting cyber-attacks on communication links between smart devices. Rather than suggesting a one-size-fits-all deep learning architecture, their approach was tailored to autonomously discover architectures with a relatively small number of parameters using training data. They introduced a host-based intrusion detection system that employs univariate regression, creating separate models for each transmitted signal. For model construction, only 22 features representing sensors were used. In their method, data shuffling was employed, which is appropriate for memory-less DNN architectures (e.g., CNNs) where the sequence of training pairs is not critical. They demonstrated the importance of sensor data in CPS anomaly detection. For example, using data from only three sensors (FIT201, AIT401, and LIT301), their host-based approach detected 15 attack scenarios in the SWAT dataset.

In study [8], researchers implemented a modified version of the DenseNet architecture, incorporating an undersampling technique to address the imbalance between minority and majority classes—a common issue in anomaly detection datasets. A key characteristic of DenseNet is the direct connectivity between layers within each block. The authors reported an F1-score of 1, with a precision of 0.9997 and a recall of 0.999. They noted perfect recall in attack scenario number 4. However, this scenario, which involved a motorized valve named MV-504, could not be fully evaluated due to the absence of specific valve data in the official dataset. The authors also reported superior scores in attacks that were deemed unsuccessful by SWAT developers, such as attack numbers 13, 14, and 29.

A multi-layer perceptron (MLP) anomaly detection model was introduced in [22], designed to capture the sequential dependencies among observations and forecast plant behavior. Additionally, the established Cumulative Sum (CUSUM) method was applied to identify significant deviations between actual and predicted sensor readings, thereby facilitating anomaly detection. This model was evaluated exclusively on six isolated attacks affecting individual sensors of the SWAT dataset. Similarly, in [26], the authors found that the model was more sensitive to anomaly detection when working with a small group of sensors. The methodology presented employs a neuroevolutionary-based approach, wherein neural network architectures evolve over successive generations. The process begins with the initialization of a diverse population of neural networks, each varying in structural parameters such as the number of layers, neurons per layer, and activation functions. The method then focuses on optimizing the architecture and hyperparameters. The authors down sampling with a ratio of 5 and used the labels provided by the SWAT developers to evaluate their results, without specifying which attacks were detected. The best result achieved a precision of 99.35, a recall of 68.12, and an F1 score of 0.81. However, these studies [22] and [26] do not account for the

Table 1 Attacks scenarios from [8]

Attack number	Attack point	Start state	Attack
1	MV-101	MV-101 is closed	Open MV-101
2	P-102	P-101 is on whereas P-102 is off	Turn on P-102
3	LIT-101	Water level between L and H	Increase by 1 mm every second
4	MV-504	MV-504 is closed	Open MV-504*
5	No physical impact attack		
6	AIT-202	Value of AIT-202 is > 7.05	Set value of AIT-202 as 6*
7	LIT-301	Water level between L and H	Water level increased above HH
8	DPIT-301	Value of DPIT is < 40kpa	Set value of DPIT as > 40kpa
9	No physical impact attack		
10	FIT-401	Value of FIT-401 above 1	Set value of FIT-401 as < 0.7
11	FIT-401	Value of FIT-401 above 1	Set value of FIT-401 as 0
12	No physical impact attack		
13	MV-304	MV-304 is open	Close MV-304*
14	Mv-303	MV-303 is closed	Do not let MV-303 open*
15	No physical impact attack		
16	LIT-301	Water level between L and H	Decrease water level by 1 mm each second
17	MV-303	Closed	Do not let MV-303 open
18	No physical impact attack		
19	AIT-504	Value of AIT-504 < 15 uS/cm	Set value of AIT-504 to 16 uS/cm*
20	AIT-504	Value of AIT-504 < 15 uS/cm	Set value of AIT-504 to 255 uS/cm*
21	MV-101, LIT-101	MV-101 is open; LIT-101 between L and H	Keep MV-101 on continuously; Value of LIT-101 set as 700 mm
22	UV-401, AIT-502, P-501	UV-01 is on; AIT-502 is < 150; P-501 is open	Stop UV-401; Value of AIT502 set as 150; Force P-501 to remain on
23	P-602, DIT-301, MV-302	DPIT-301 is < 0.4 bar; MV-302 is on; P-602 is closed	Value of DPIT-301 set to > 0.4 bar; Keep MV-302 open; Keep P-602 closed
24	P-203, P-205	P-203 is on; P-205 is on	Turn off P-203 and P-205*
25	LIT-401, P-401	Value of LIT-401 < 1000; P-402 is on	Set value of LIT-401 as 1000; P402 is kept on
26	P-101, LIT-301	P-101 is off; P-102 is on; LIT-301 is between L and H	P-101 is turned on continuously; Set value of LIT-301 as 801 mm
27	P-302, LIT-401	P302 is on, LIT401 is between L and H	Keep P-302 on continuously; Value of LIT401 set as 600 mm till 1:26:01
28	P-302	P302 is on	Close P-302
29	P-201, P-203, P-205	P-201 is closed; P-203 is closed; P-205 is closed	Turn on P-201; Turn on P-203; Turn on P-205*
30	LIT-101, P-101, MV-201	P-101 is off; MV-101 is off; MV-201 is off; LIT-101 is between L and H; LIT-301 is between L and H	Turn P-101 on continuously; Turn MV-101 on continuously; Set value of LIT-101 as 700 mm; P-102 started itself because LIT301 level became low
31	LIT-401	Water level between L and H	Set LIT-401 to less than L
32	LIT-301	Water level between L and H	Set LIT-301 to above HH
33	LIT-101	Water level between L and H	Set LIT-101 to above H
34	P-101	P-101 is on	Turn P-101 off*
35	P-101; P-102	P-101 is on; P-102 is off	Turn P-101 off; Keep P-102 off
36	LIT-101	Water level between L and H	Set LIT-101 to less than LL
37	P-501, FIT-502	P-501 is on; FIT-502 in normal range	Close P-501; Set value of FIT-502 to 1.29 at 11:18:36
38	AIT-402, AIT-502	In normal range	Set value of AIT402 as 260; Set value of AIT502 to 260*
39	FIT-401, AIT-502	In normal range	Set value of FIT-401 as 0.5; Set value of AIT-502 as 140 mV
40	FIT-401	In normal range	Set value of FIT-401 as 0
41	LIT-301	Water level between L and H	Decrease value by 0.5 mm per second

*Unsuccessful cyberattack - limited or no impact on CPS

interconnections between networked sensors and actuators prevalent in cyber-physical systems (CPS). Typically, cyber-attacks aim to disrupt critical infrastructure operations, potentially impacting multiple sensors simultaneously due to a cascading effect.

To further contextualize our work, it is important to consider recent advanced architectures. The MTS-DVGAN model proposed by Sun et al. [6] represents a state-of-the-art approach, using a dual variational generative adversarial network for anomaly detection in CPS. This method's sophisticated adversarial framework is highly effective for learning the patterns within a single data modality. However, it was not designed to perform cross-modal fusion between sensor and network data. Consequently, it cannot learn the direct correlations between a specific cyber command and its subsequent physical impact, which constitutes a critical limitation when modelling tightly coupled cyber-physical interactions. This limitation highlights the need for approaches capable of jointly modelling multimodal dependencies, which forms the central novelty of our model. Other advanced methods have sought to model the inherent structure of industrial systems, however they remain constrained to unimodal analysis or limited feature interactions, thereby leaving room for more integrated cyber-physical modelling frameworks.

The work in [27] proposes a hybrid model combining a Graph Convolutional Network (GCN) with a Gated Recurrent Unit (GRU) to capture both the spatial correlations between sensors and the temporal dynamics of the system. Similarly, in work [28], the authors introduce a spatiotemporal-driven autoencoder to learn a comprehensive representation of normal industrial processes. While these approaches are powerful, their reliance solely on sensor data means they may not explicitly capture the originating cyber commands from network traffic that trigger anomalies. Our work, by contrast, is designed to directly fuse the network and physical data streams to learn this direct cyber-to-physical causality. Furthermore, another significant trend in time-series analysis is the adoption of Transformer-based architectures. Models like TranAD [29] have shown strong performance by using attention-based encoders to capture complex dependencies across the entire time sequence simultaneously. While powerful, these models often require larger datasets for effective training and their global attention mechanism can be less computationally efficient for the highly localized and repetitive patterns typical of ICS operations. Our choice of an LSTM-based architecture provides a more lightweight and targeted approach for modeling the direct, sequential causality between a cyber command and its immediate physical manifestation.

Finally, as anomaly detection models become more complex, the need for transparency and interpretability is paramount for their adoption in critical operational environments. The emerging field of Explainable AI (XAI) is critical for building trust in these automated security systems. The comprehensive survey [30] on XAI in cybersecurity outlines the key methods and challenges in making black-box models understandable. While our current work focuses on the detection model itself, the challenges highlighted in this survey motivate future directions to integrate techniques that can pinpoint the specific sensors or network patterns responsible for a detected anomaly, thereby providing actionable intelligence to system operators. Recent advancements in 2024 have further expanded the scope of Cyber-Physical security toward collaborative and federated paradigms. Specifically, new research in [31] utilizes collaborative SRU networks with dynamic behavior aggregation to enhance explainability, while work in [32] addresses

data imbalance in large-scale networks through Federated-Boosting. Additionally, the framework in [33] introduces a federated reinforcement-based fusion model to secure distributed environments. While these studies focus on collaborative privacy and distributed intelligence, they reinforce the critical necessity of multi-source data synthesis and adversarial regularization techniques implemented in our proposed framework. Unlike previously developed models, our unsupervised approach integrates multiple data sources to enhance the model's generalization capacity in detecting anomalies from individual sources or their combination, utilizing attention mechanisms.

In this framework, physical data provides contextual information for network traffic data, and while network data similarly informs physical processes. To improve reconstruction accuracy, we incorporated a discriminator, enabling the model to detect subtle anomalies that closely resemble normal behavior. During processing, we retained all features related to physical devices to ensure a precise architectural representation. Mechanisms were also incorporated to handle sensor-specific issues, such as noise. For network traffic data, we extracted features from Modbus packets to accurately capture the behavior of the control system. To address the challenge of identifying attack sources, particularly in deep networks, the autoencoder performs separate reconstructions of physical and network data, enabling CI operators to determine whether an anomaly originated from a specific source or both.

5 Preprocessing

To build the multimodal model, both physical and network traffic data from the SWAT dataset were used. The initial six hours of data from both collections were omitted as this duration is necessary for the hydroelectric plant to stabilize [23]. Additionally, due to the longer stabilization needs of specific sensors such as AIT201 and AIT203, an extra 98,200 samples were discarded along with their equivalent samples in the network traffic data. Moreover, all numerical data were standardized using the `StandardScaler()`. Finally, both datasets (physical and network) were reshaped to feed the LSTM layers with a timestep of ten (10), as they are both high-frequency data. This means the model analyzes each 10-second window to determine if there is an anomaly within that period.

5.1 Preprocessing of physical (sensor and actuator) data

The SWAT A2 dataset is a detailed time-series collection featuring high-frequency data from 25 sensors and 26 actuators, for a total of 51 features. It includes 449,919 samples from four days of simulated attacks and 495,000 samples from normal operational periods [8]. Although it is standard practice to remove features with negligible variance from the training set, this study opted to retain these features. Firstly, from a cybersecurity perspective, each feature could potentially serve as an attack vector; thus, understanding its normal behavior is essential, even if its variance is low. Secondly, in Industrial Control System (ICS) datasets, correlations between features often reflect the interconnected operations of system components. For example, readings from a sensor indicating a full water tank typically trigger corresponding adjustments in related pumps, either by stopping the inflow or starting the outflow. These correlations, while expected, are vital for anomaly detection as they define the system's expected operational patterns; therefore, all features were retained for model development despite their correlations. Additionally, many sensor outputs are not normally distributed, and there are noticeable variations

between the training and testing datasets distributions, especially evident in sensor AIT201, AIT 203, P201, PIT 502. Previous studies on the SWAT dataset al.so noted this issue, with some research excluding up to 22 sensors [23, 28, 34]. These characteristics could pose challenges from a data science perspective but also make the model more generalizable.

Feature extraction techniques were applied to the original 25 numerical sensor readings to identify cyclical patterns through autocorrelation. We extensively evaluated multiple feature extraction configurations, including Fourier transformations and various windowing settings; however, these alternative approaches consistently yielded inferior detection performance compared to the proposed model. Consequently, our analysis led to the generation of two types of derived features: (1) Rolling Window Statistical Features (RWSF), such as mean, median, standard deviation, maximum, and minimum, calculated over a fixed 120-second window to facilitate data smoothing and noise reduction, and (2) Time-shifted Features (TF) derived from specific lags of 10, 20, and 30 seconds to capture temporal trends and dependencies. Despite standard practices often discarding features with low variability to reduce noise, we opted to retain all physical features, as our results confirm this holistic inclusion does not introduce detrimental noise. Instead, it preserves the integrity of the 'normal' operational baseline, enabling the cross-attention mechanism to learn the subtle, static relationships between controllers that are frequently disrupted during stealthy, single-point attacks. The resulting physical dataset contains 115 features: 51 original, 40 RWSF, and 24 TF. Finally, all data were synchronized into 10-second analysis sequences using mean aggregation to capture the impact of anomalous patterns within high-frequency data streams.

5.2 Preprocessing of network traffic data

Llevel 1 of the SWAT network, as illustrated in Fig. 4. This data captures the communication between the SCADA system and the six PLCs. The network protocol implemented was the Common Industrial Protocol (CIP) over EtherNet/IP. The dataset is composed of 18 features - Date, Time, Origin, Type, Interface, Direction, Source IP, Destination IP, Protocol, Proxy Source IP, Application Name, Modbus Function Code, Modbus Function Description, Modbus Transaction ID, SCADA Tag, Modbus Value, Destination Port, Source Port -. The network data was divided into several CSV files, each containing a maximum of 500,000 packets. However, since the data capture occurred at one-second intervals, instances of overlap arose where multiple rows displayed different activities but share the same timestamp. The network data was also captured during the 11-day period that includes the last 3 days of attack scenarios [6].

The preprocessing steps applied to the network data were as follows: (1) invalid data, such as NaN values, were systematically removed; (2) features exhibiting low or zero variance, including Origin, Type, Interface Name, Interface Direction, Protocol, Application Name, Modbus Function Code, and Service, were excluded from the training dataset to enhance model performance; (3) measurements from physical devices, originally grouped under the feature 'Modbus Value,' were encoded in a little-endian, single-precision floating-point format. To facilitate further analysis, these values were parsed to extract the physical readings and subsequently converted to integer format; (4) leveraging the documentation provided by the SWAT developers, the 'Modbus Value' feature

was disaggregated into multiple columns corresponding to the specific physical devices within the scaled-down hydroelectric plant, resulting 24 distinct features; (5) to address missing values, a linear interpolation technique was employed, which is consistent with the typical incremental or decremental behavior of sensor data.

Finally, there is a one-to-many correspondence between the data collected from a physical device and the network traffic data, as each timestamp is associated with a single report from physical devices but multiple reports from network traffic. To synchronize the high-frequency network packets into 10-second analysis windows, we evaluated various aggregation techniques, including mean, median, min, max, standard deviation, and sum. While robust aggregators like the median are often preferred for filtering stochastic noise, mean aggregation proved superior for this anomaly detection task due to its inherent sensitivity to outliers. In industrial cybersecurity, anomalous spikes or malicious network commands—even if transient—are critical indicators of a potential breach. Median aggregation tends to filter out these short-duration signals as noise, whereas the mean is disproportionately affected by even brief deviations in the 24 network features, such as those found in the 'Modbus Value' fields. This sensitivity ensures that malicious traffic significantly increases the reconstruction error, thereby enhancing the model's ability to identify low-observability threats that robust aggregators might otherwise miss.

Following the preprocessing and feature engineering steps, the final datasets consist of 115 features for the physical data and 24 features for the aggregated network traffic data. To capture temporal dynamics for the LSTM layers, we generated input sequences using a sliding window technique. A window of length $T = 10$ timesteps (equivalent to 10 s) was moved across the synchronized time series with a stride of 1 timestep, creating overlapping sequences. This method ensures that the temporal dependencies within every possible 10-second interval are captured and analyzed. The final input for the model is composed of two distinct tensors for each 10-second window: a physical sequence with dimensions (10, 115)—comprising 51 original, 40 RWSF, and 24 TF features—and a network sequence with dimensions (10, 24) derived from aggregated Modbus traffic. These

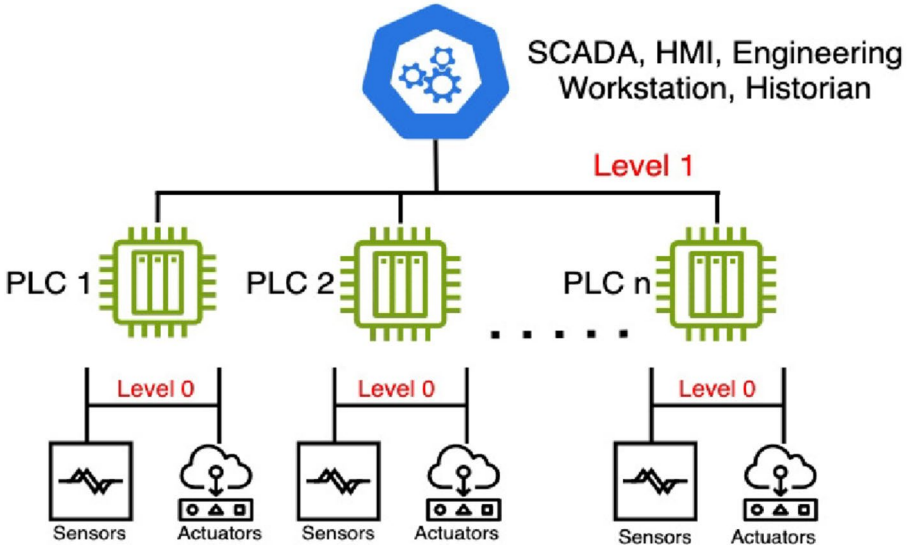

Fig. 4 SWAT process overview based on [8]

dimensions are reflected in the input layer shapes of the sensor input and network input, respectively, ensuring the model processes the full context of the engineered feature set. These paired sequences are fed simultaneously into their respective VAE-LSTM encoders to generate the latent representations used in the fusion stage.

6 Multimodal design

The proposed model introduces an advanced architecture that integrates multi-head attention mechanisms with adversarial training achieving a high level of robustness and accuracy in the reconstruction of sensor and network data within an unsupervised learning framework, as shown in Fig. 5. This model leverages pre-trained variational autoencoder (VAE) encoders to derive latent representations from both sensor and network inputs.

The encoders for both physical and network modalities utilize a parallel architecture with a latent dimension of 64 for each modality, comprising two stacked LSTM layers, each with 128 hidden units, followed by dense layers to output the latent space parameters μ and $\log(\sigma)^2$. Following the dual cross-attention mechanism, a subsequent multi-head attention layer with a latent dimension of 128 further refines the interactions among features. This is followed by a dense transformation layer consisting of 304 units with a dropout rate of 0.2, which enhances the model's ability to capture intricate patterns within the integrated latent spaces before the reconstruction phase.

The latent representations are subjected to a dual cross-attention mechanism, wherein the latent features of each modality are used as queries, keys, and values within a multi-head attention layer. This process enables the model to learn and leverage complex intermodal correlations, thereby enhancing its ability to capture nuanced relationships between different data sources. The cross-attended latent features are subsequently processed through an additional

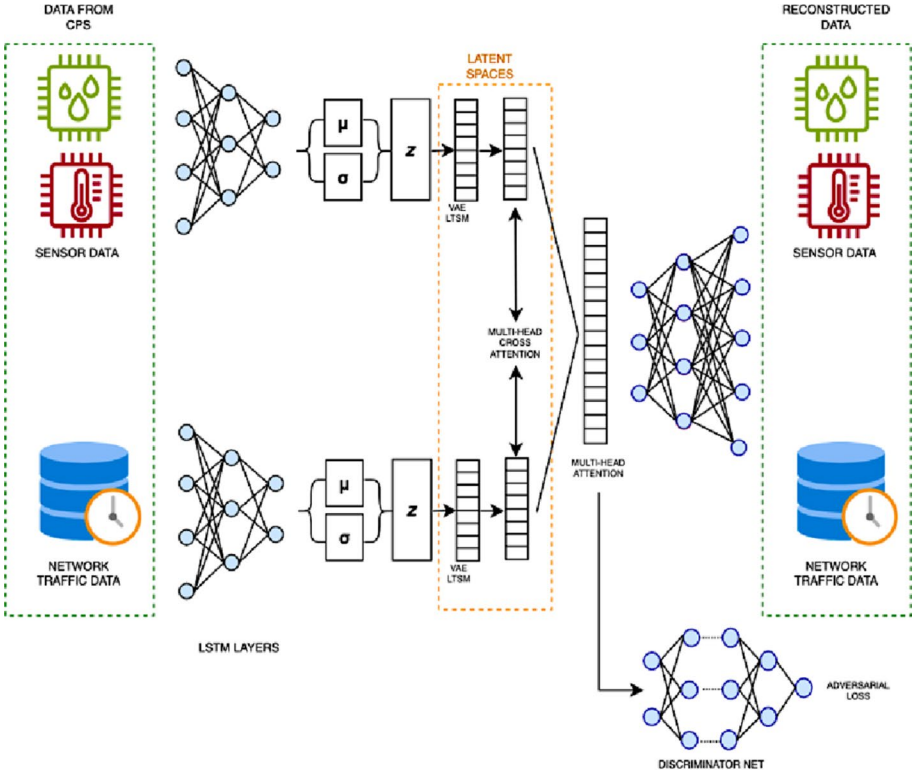

Fig. 5 Proposed model architecture

multi-head attention layer, which further refines the interactions among features, ensuring that the model can capture intricate patterns within the integrated latent spaces. The refined latent representation is then propagated through dense layers, culminating in the reconstruction of the original sensor and network data. This reconstruction is optimized using a mean squared error (MSE) loss function, to ensure the fidelity of the reconstructed outputs.

To augment the generative capabilities of the model, an adversarial discriminator is incorporated. The discriminator is tasked with distinguishing between real and reconstructed data, compelling the autoencoder to generate outputs that are nearly indistinguishable from the authentic data. The training of the discriminator is guided by a binary cross-entropy loss function, while the adversarial process ensures that the reconstructed data maintains a high degree of fidelity. The following pseudocode explains the main steps of the developed model [23]:

Model Overview

Input:
X_s*: Raw sensor data*
X_n*: Raw network data*

Output:

$\widehat{X_s}$*: Reconstructed sensor data*
$\widehat{X_n}$*: Reconstructed network data*
$D(\hat{X})$*: Discriminator's assessment of the reconstruction authenticity*

Procedures:

1. *Latent Representation Extraction*
   Procedure:
   Pass X_s *through the pre-trained encoder for sensors to obtain latent representations:*
   $$z_{\mu_s}, z_{\sigma_s}, z_s$$
   where:
   z_{μ_s}*is the mean of the latent space,*
   z_{σ_s}*is the log variance,*
   $z_s = z_{\mu_s} + \epsilon \cdot \exp\left(\frac{z_{\sigma_s}}{2}\right)$ where $\epsilon \sim \mathcal{N}(0,1)$

   Similarly, process $\widehat{X_n}$ *to obtain:*
   $$z_{\mu_n}, z_{\sigma_n}, z_n$$

2. *Cross-Attention Mechanism*
   Procedure:
   Compute the cross-attention between the sensor and network latent spaces:

$$A_{s\to n} = \text{softmax}\left(\frac{Q_s K_n^T}{\sqrt{d_k}}\right) V_n$$
$$A_{n\to s} = \text{softmax}\left(\frac{Q_n K_s^T}{\sqrt{d_k}}\right) V_s$$

Where Q_s, K_n, V_n *are the query, key, and value matrices derived from* z_s, z_n*,respectively.*

3. *Concatenation of Latent Spaces*

   Procedure:
   Concatenate the cross-attended latent spaces to form a combined latent representation:
   $$z_{\text{combined}} = \text{concat}(A_{s\to n}, A_{n\to s})$$

4. *Multi-Head Attention on Combined Latent*

Procedure:
Apply multi-head attention to z_{combined}*:*

$$A_{\text{combined}} = \text{MultiHead}(Q_{\text{combined}}, K_{\text{combined}}, V_{\text{combined}})$$

5. *Dense Layer Transformation*

Procedure:
Pass A_{combined} *through a series of dense layers:*

$$H_1 = \phi(W_1 A_{\text{combined}} + b_1)$$
$$H_2 = \phi(W_2 H_1 + b_2)$$

6. *Reconstruction*

Procedure:
Reconstruct the original sensor and network data:

$$\widehat{X_s} = W_s H_2 + b_s$$
$$\widehat{X_n} = W_n H_2 + b_n$$

7. *Adversarial Discriminator*

Procedure:
Evaluate the authenticity of $\widehat{X_s}$ *and* $\widehat{X_n}$ *using a discriminator:*

$$D(\hat{X}) = \sigma\left(W_D \cdot \text{concat}(\widehat{X_s}, \widehat{X_n}) + b_D\right)$$

8. *Model Compilation and Training*

Procedure:
Optimize the model using loss functions:

$$L_{\text{sensor}} = \text{MSE}(X_s, \widehat{X_s})$$
$$L_{\text{network}} = \text{MSE}(X_n, \widehat{X_n})$$
$$L_{\text{adv}} = \text{BinaryCrossentropy}\left(1, D(\hat{X})\right)$$
$$L_{\text{total}} = \lambda_s L_{\text{sensor}} + \lambda_n L_{\text{network}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{KL}} L_{\text{KL}}$$

7 Results

The ML procedures were developed using an Intel(R) Xeon(R) Silver 4310 CPU@ 2.10 GHz, with the operating system Ubuntu 22.04.3 LTS and NVIDIA-SMI 525.147.05, (Driver Version: 525.147.05), and CUDA Version: 12.0. For deep neural networks, TensorFlow 2.11.0 and Keras 2.11.0 were utilized, in conjunction with Python 3.10.12. The performance of our unsupervised model is evaluated based on its ability to distinguish between normal and anomalous data by analyzing its reconstruction error. For each 10-second input sequence, the anomaly score S was calculated as the mean squared error (MSE) between the original data and the model's reconstruction.

The anomaly detection threshold τ was established through an analysis of reconstruction error distributions, which exhibited clear peaks concentrated near zero for normal operational traffic. We evaluated various threshold candidates and selected a value that maximized attack identification while minimizing false positives—a critical requirement in industrial cybersecurity—ultimately selecting a threshold that resulted in an anomaly classification rate of 9% for the test traffic. This is technically consistent with the original SWAT dataset characteristics (~ 12% anomaly rate); the difference reflects our focus on sustained physical disruptions. Consequently, the labels utilized for performance metrics represent a technical refinement of the original SWAT ground truth rather than arbitrary modifications. Based on original SWAT documentation and precise attack timestamps, we only labeled periods as anomalous when a real attack successfully affected the system's control measures and physical state. This refinement ensures that the

calculated metrics—TP: 39,406; TN: 382,673; FP: 7,017; FN: 10,487—accurately reflect the model's capacity to detect operational impacts, resulting in a robust Area Under the Curve (AUC) of 0.87.

To evaluate our model, we selected the SWAT dataset due to its relevance to the types of attack vectors that critical infrastructures commonly encounter. This dataset is particularly valuable as it includes both physical and network traffic data from normal and attack scenarios, providing a comprehensive view of the system's behavior under different conditions. The dataset was partitioned chronologically to prevent data leakage and simulate a real-world scenario. The initial 7-day period of normal operations was used for training and validation, with a 90/10 split respectively. The subsequent 4-day period containing the attack scenarios was held out as the unseen test set. Due to the realistic nature of the dataset, some attacks scenarios were unsuccessful attacks or had an unplanned outcome. The list of attacks is provided in Table 1. However, five of these attacks (numbers 5, 9, 12,15, and 18) had no physical impact. Attack number 4, which targeted a motorized valve designated as MV-504, was not analyzed due to the absence of data regarding this valve in the official dataset. Some attacks, such as numbers 24 and 34 had negligible or no effect on system performance. Additionally, attacks numbered 13, 14, and 29 were unsuccessful. Cyberattacks numbered 6, 19, 20, and 38 targeted inactive chemical sensors. These attacks were anticipated to affect other sensors, but various malfunctions prevented this, making these cyberattacks undetectable. Therefore, there are 26 attacks in total that require detection.

We assume that an attacker can access field devices through the network and fully understands the target system. Our approach is fully unsupervised, meaning that it does not rely on labeled data for training or testing. Moreover, the labels used were based on the times when the attacks were launched and ended, rather than when the physical systems were affected or disrupted. Additionally, industrial control systems (ICS) require time to stabilize after an attack; the more severe the attack, the longer it takes to return to normal operation. This recovery time was not considered during labeling, the time intervals between attacks vary, ranging from minutes to hours. These conditions can destabilize the system or generate cascading effects due to interdependencies within the system. This theory is supported by sensor data, such as from AIT201 and AIT203, which, despite not being direct attack points, exhibited different distributions between the training and testing sets. This underscores the importance of including all system components in the model to gain a comprehensive understanding of how the CPS functions under non-attack conditions.

The primary objective of the autoencoder during training is to learn a compressed representation of normal operational data from which it can accurately reconstruct the original input. Figure 6 provides a visual validation of this process, showing a representative sample from the training set alongside the model's reconstruction. For the high-frequency sensor data, the reconstructed signals closely mirror the original inputs. This high fidelity demonstrates that the model has successfully learned the underlying patterns and temporal dependencies of normal system behavior, establishing a robust baseline. The model's ability to detect anomalies, therefore, is predicated on its performance when presented with unseen data containing attacks, which is examined next. The reconstructed network data (Fig. 7), which exhibits a highly periodic and regular pattern with sharp transitions, mirrors the periodic nature of the original signal but

shows a noticeable smoothing effect, particularly in the sharp peaks and troughs. This smoothing indicates that the model may struggle with accurately reconstructing abrupt changes in the signal. Nevertheless, the general characteristics of both data sources are well learned by the model, and more precise reconstructions could lead to overfitting, potentially compromising the model's generalization capacity.

The histograms in Fig. 8 illustrate the reconstruction errors for sensor and network data from the training set, demonstrating the model's performance. The sensor reconstruction errors, as depicted in the left histogram, are highly skewed toward zero, with most errors falling below 0.5, indicating that the model has effectively learned and captured the underlying patterns in the sensor data. Similarly, the network reconstruction errors, shown in the right histogram, are also skewed to the left, with most errors concentrated below 0.1. However, the network data exhibits a slightly wider error distribution compared to the sensor data, suggesting that while the model performs well overall, it encounters slightly more variability when reconstructing network data. This slight difference in error distribution may indicate the presence of more complex patterns within the network data. Overall, the low reconstruction errors across both datasets affirm the model's efficiency in learning from the training data, with minimal errors indicating strong generalization within the scope of the training set.

To visually validate this detection principle, Fig. 9 shows that most points for both modalities -sensor and network data- are tightly clustered near zero error, indicating the model's accurate reconstruction of normal operational behavior. The high error points (outliers) are the sequences flagged as anomalous by our detection threshold, confirming that the reconstruction error is a reliable metric for identifying deviations in both the physical and cyber domains.

While the previous figure illustrates the statistical separation of anomalies, Fig. 10 demonstrates the model's detection process in a temporal context. The purple line represents the calculated reconstruction error—the difference between the original and reconstructed data—for each sequential data point in the test set. The plot shows that for normal operational data, the error remains consistently close to zero. However, it also clearly visualizes several distinct events where the error signal spikes to a high magnitude. These spikes represent the moment when the model identified anomalous behavior, confirming that high reconstruction error is a direct and effective indicator of an attack. This spike-based detection is the mechanism that underlies the results presented in the subsequent attack scenario analyses.

To provide a deeper, qualitative analysis of the model's detection capabilities, we now examine its performance on several representative attack scenarios from the test set. These examples demonstrate the model's versatility in identifying anomalies with diverse temporal characteristics, from abrupt, high-magnitude changes to more subtle, gradual deviations from the established baseline. The following figures visualize these successful detections, highlighting how the model's reconstruction error serves as a reliable indicator of malicious activity across different attack vectors.

The model is highly effective at identifying anomalies that represent abrupt changes which modify or disrupt the normal operating behavior of physical devices. Figure 11 (upper), which visualizes attack scenario 11, shows that a sensor value is artificially forced to zero. Similarly, Fig. 11 (lower) depicts attack scenario 37, which involves a sudden, malicious change to a different sensor's reading. In both cases, the model, which

was trained only on normal operational patterns, fails to accurately reconstruct these unnatural, flat-line states. This failure results in a clear and immediate spike in the reconstruction error, demonstrating the model's high sensitivity to this class of attack.

In contrast to abrupt attacks, the model also proves capable of identifying anomalies that gradually increase or decrease sensor levels over time. Figure 12 illustrates this with attack scenario 22, where the sensor level drifts away from the normal baseline. While the initial deviation is subtle, the reconstruction error accumulates as the drift continues, eventually crossing the detection threshold and demonstrating the model's ability to detect slower-moving, persistent threats.

The model partially identified some attacks where the disruption to the hydroelectric plant was subtle or the attack duration was short, as seen in attack scenario number 33, where the attack was detected by our model only when the disruption reached high levels. However, if the tolerance for false positives is higher, as is often the case in cybersecurity for CIs, the attack could be detected sooner. As illustrated in Fig. 13, the red points represent detected anomalies. The framework successfully identified 24 out of 26 relevant attacks, as presented in Table 2. A granular post-hoc analysis indicates that Attacks 1 and 8 remained undetected because they targeted a single attack point with a low magnitude that did not cause the reconstruction error to exceed the established 9% threshold. Similarly, the detection delay observed in Attack 33 (less than 30 s) was a direct consequence of the gradual nature of the sensor drift induced by the attack. The anomaly was only flagged once the cumulative deviation became statistically significant against the learned baseline. This highlights the inherent trade-off in industrial cybersecurity: maintaining a threshold that prevents excessive false positives may lead to slight delays in detecting slow, incremental drifts.

It is important to highlight that our approach is fully unsupervised, limiting the applicability of conventional performance metrics such as precision and recall. To address this, we conducted an extensive review and analysis of the entire time series data to determine which attack scenarios were successfully detected. This methodology, while different from traditional evaluation measures, offers a nuanced understanding of the model's performance in an unsupervised context. Furthermore, direct comparison with similar studies is challenging, as, to the best of our knowledge, no other work has analyzed the same set of cyber-attacks while integrating both physical and network data from the SWAT dataset. Most existing works, such as those referenced in [2, 23], and [26], have primarily focused on physical data alone. Moreover, previous studies have often suggested excluding numerous features during preprocessing based on their varying distributions between the training and test sets. However, these differences likely arise from the fact that cyber-attacks are present exclusively in the test set. In contrast to earlier research that discarded features with low variability, our results confirm that retaining the full 115-feature set does not introduce detrimental noise, rather, it provides a more granular and resilient operational baseline. This decision was validated through extensive comparative testing, where alternative configurations—such as Fourier-transformed features—consistently failed to achieve the same level of discriminative power as the proposed RWSF and TF combination. By prioritizing a holistic inclusion of all physical device data, the model effectively utilizes the cross-attention mechanism to monitor the subtle, static interactions between components that are often the primary targets of zero-day attacks.

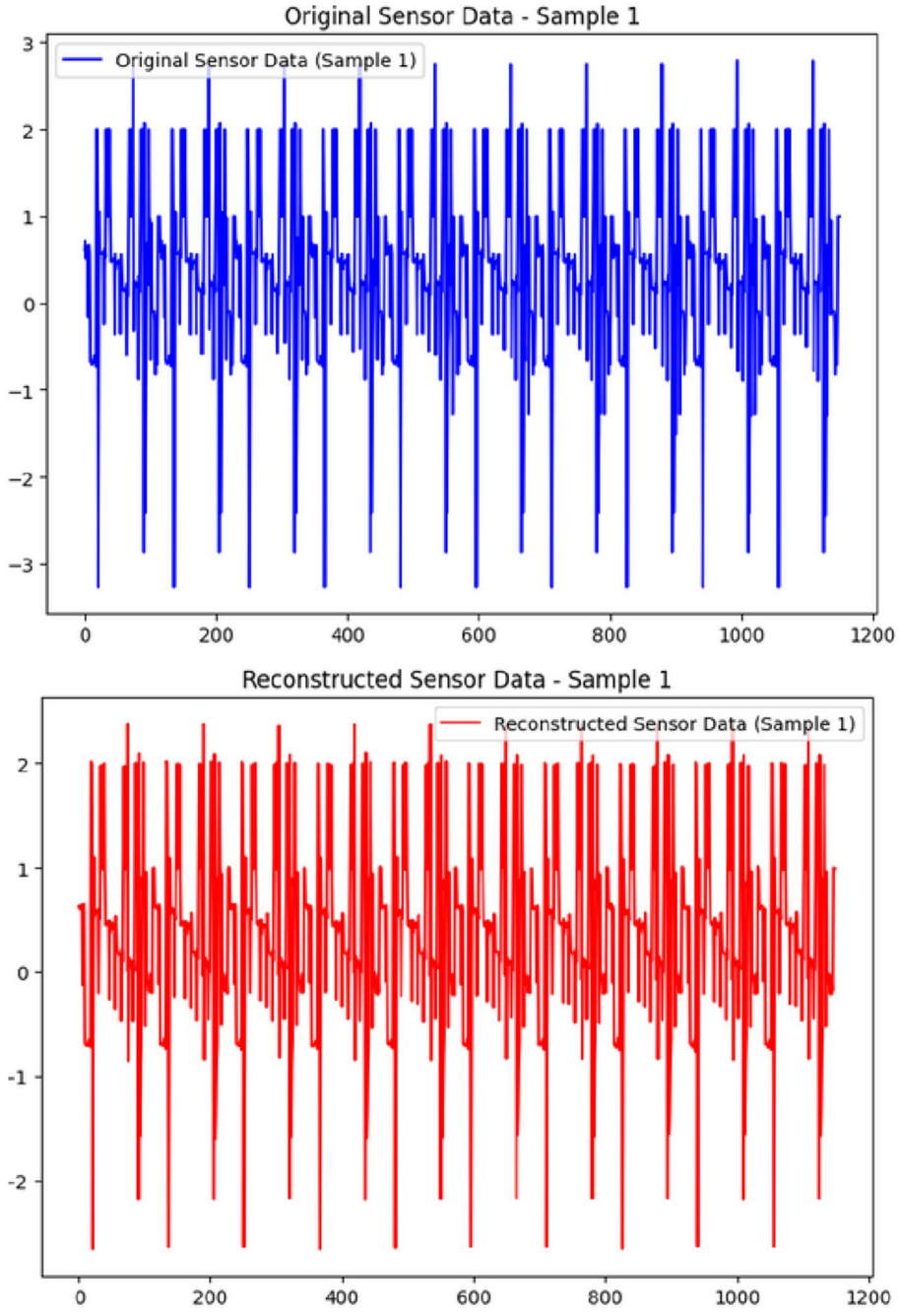

Fig. 6 Original and reconstructed samples (train set – sensor data)

7.1 Comparative analysis

A direct numerical comparison between models evaluated on the SWAT dataset is nuanced due to variations in experimental setups, data subsets, and the specific attack scenarios analyzed. Our proposed multimodal framework demonstrates strong classification performance, achieving an Area Under the Curve (AUC) of 0.87, which indicates robust detection capability across all thresholds. At the specific operating point used for our analysis, the model achieved an F1-score of 0.82 and successfully identified 24 out of 26 relevant attack scenarios. In contrast, earlier unsupervised methods that relied exclusively on physical data reported varied and often limited success. For instance, the

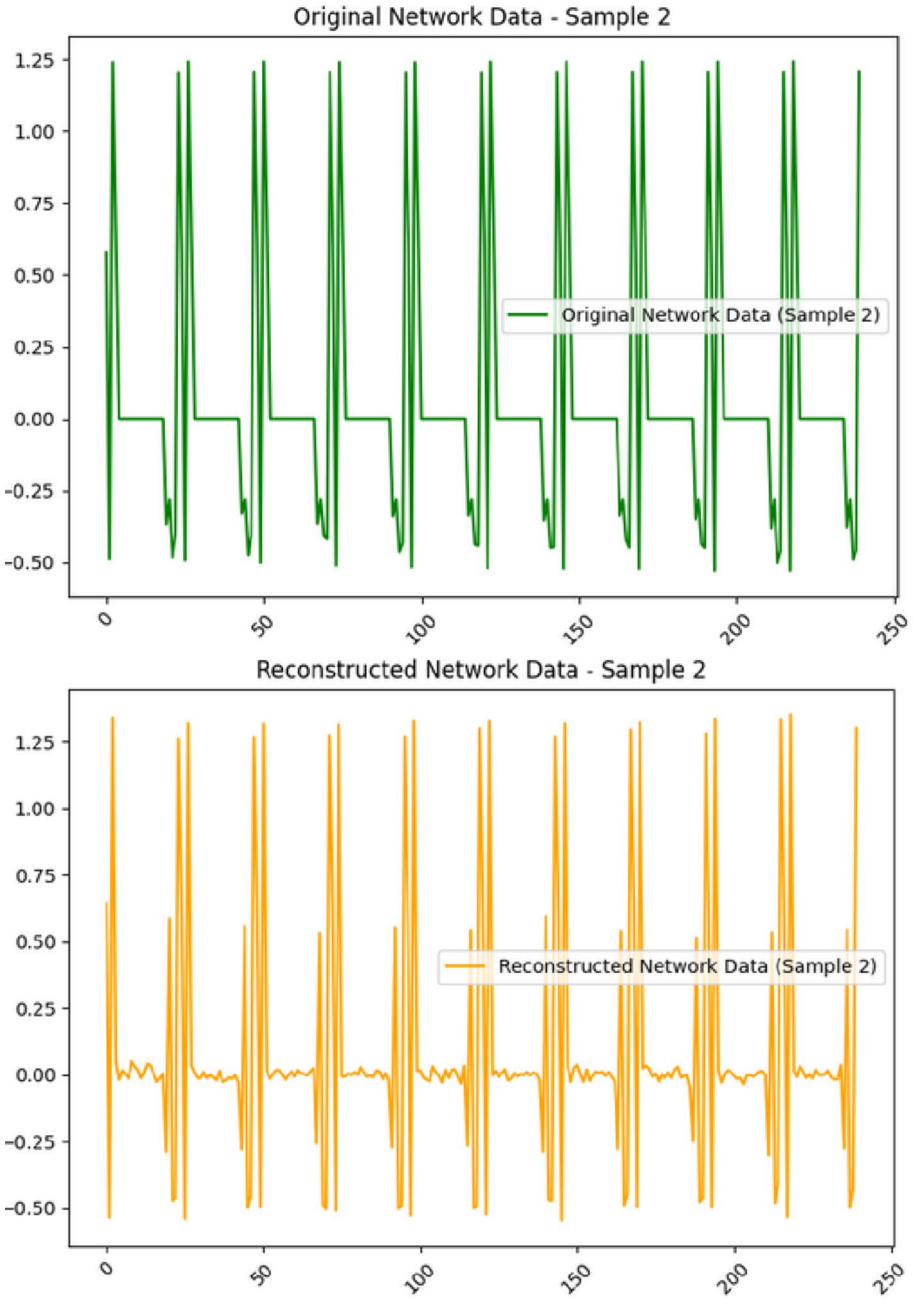

Fig. 7 Original and reconstructed samples (train set – network data)

work of Inoue et al. [24], using DNN and OCSVM models, reported recall rates significantly lower than 0.7, while Faber et al. [26] achieved an F1-score of 0.81 with their neuroevolutionary approach. The one-class neural network proposed by Boateng et al. [2] showed strong performance on a different set of attacks, identifying 15 of the 36 scenarios they analyzed. Similarly, the CNN-based methods by Nedeljkovic & Jakovljevic [23] and Xie et al. [25], also focused solely on sensor data, with the former detecting 15 attack scenarios using a reduced feature set. A significant limitation noted across these studies is their unimodal nature. Furthermore, some works like Raman MR et al. [22] evaluated their MLP model on only six isolated attacks, which does not account for the

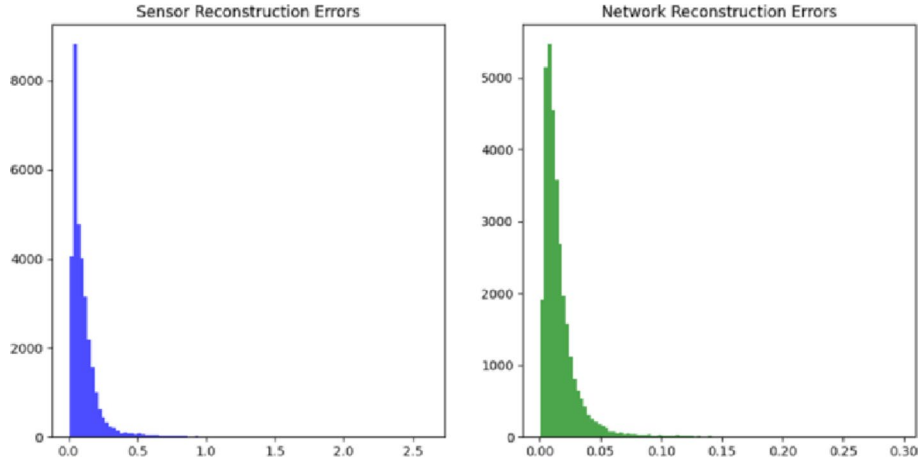

Fig. 8 Reconstruction errors (train sets)

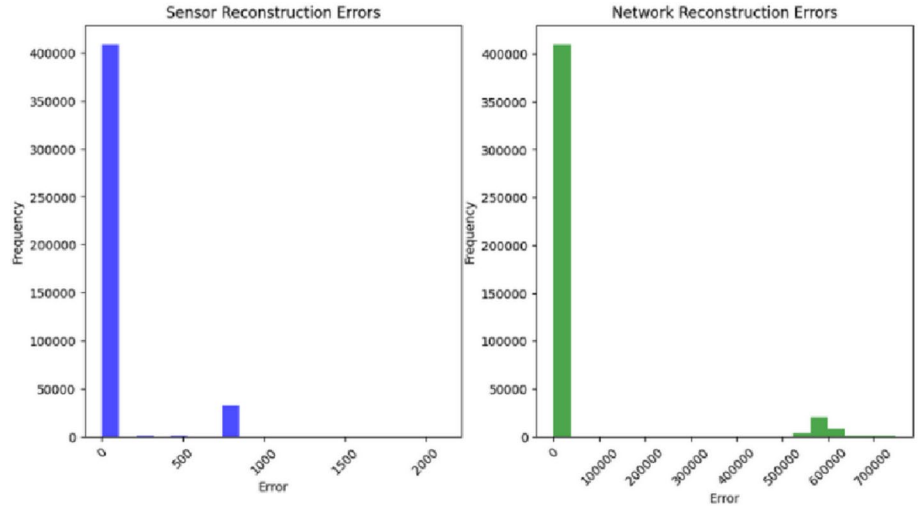

Fig. 9 Reconstruction errors (test set)

complex, system-wide interdependencies that our framework is designed to capture. These approaches, while valuable, are architecturally constrained in detecting threats that manifest primarily in network patterns, a critical gap addressed by our model.

Contemporary research also provides high-performance benchmarks from unimodal systems. For instance, a model using Hierarchical Temporal Memory (HTM) [35] reports a state-of-the-art F1-score of 0.92 on the SWAT dataset. Similarly, other work on lightweight adversarial autoencoders [36] emphasizes computational efficiency for ICS environments. While these models are highly effective, their strength lies in handling only a single data stream. The HTM model [37], for example, is not designed to explicitly fuse network traffic, and may therefore be blind to cyber-attacks that have not yet manifested as significant physical deviations. Our model, by contrast, is architecturally designed to learn the direct correlation between cyber and physical events, providing a more proactive security posture. The value of our more complex architectural choices is further supported by research on industrial time-series, which validates through ablation experiments that integrating attention mechanisms and adversarial training is crucial for robust performance on datasets like SWAT.

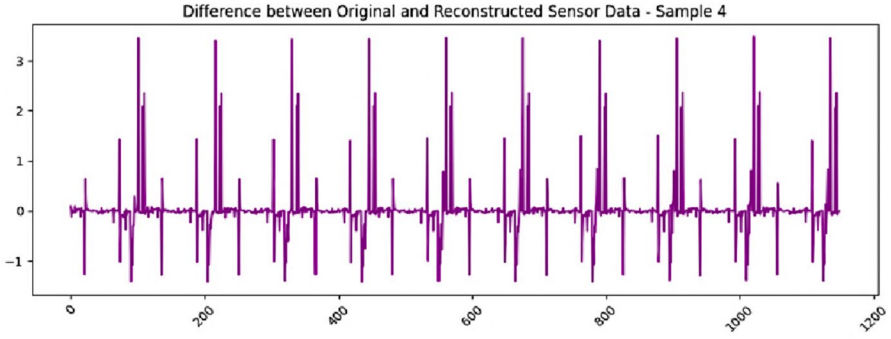

Fig. 10 Reconstruction errors (test set)

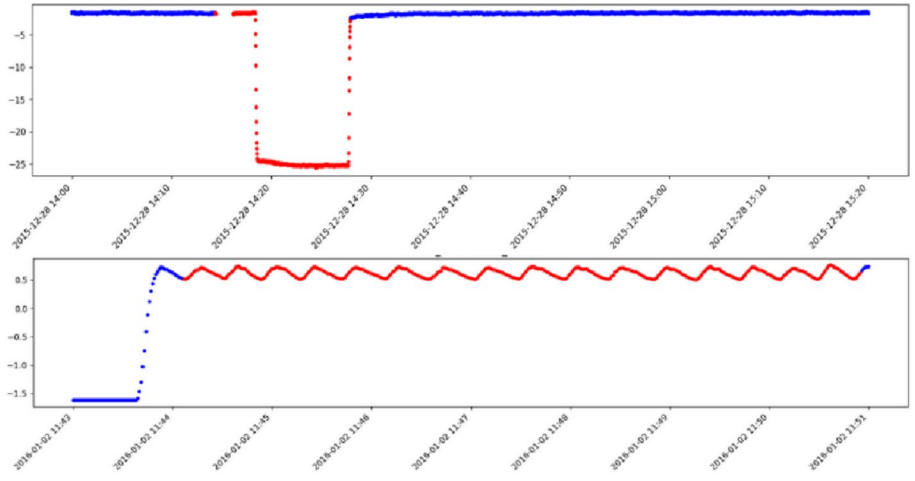

Fig. 11 Attack scenario number 11(upper) and attack scenario number 37 (lower)

Finally, it is important to address the work by Adepu et al. [8], which is notable for reporting an F1-score of 1.0. As noted in our analysis, these near-perfect scores were achieved while evaluating scenarios that included unsuccessful attacks or those targeting components with missing data.

The core contribution of our work is the novel multimodal fusion architecture, which addresses a more complex and realistic threat landscape. Our model is designed to detect coordinated attacks by learning the direct correlation between cyber and physical events—a capability fundamentally absent in unimodal systems. The robustness of the model is reflected in its 0.87 AUC on the unseen 4-day test set, which contains attack scenarios that the model had not encountered during training. While testing on additional datasets like WADI, which is a recognized standard for external validity, the SWAT dataset was prioritized for this study because it provides the synchronized, high-frequency multimodal streams (both network and physical) necessary to validate our dual cross-attention fusion architecture. The successful detection of 24 out of 26 relevant attacks—including those with gradual drifts—serves as a primary indicator that the model has learned the underlying physical laws of the plant rather than overfitting to stochastic noise.

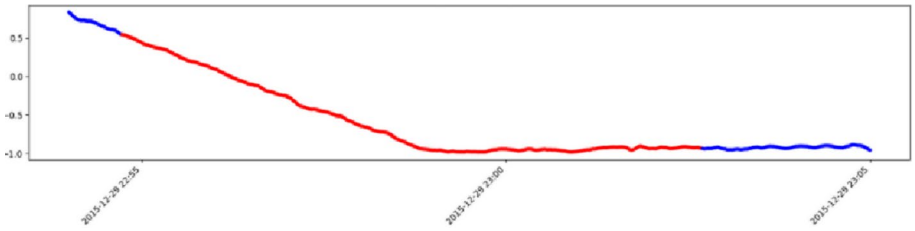

Fig. 12 Attack scenario number 22

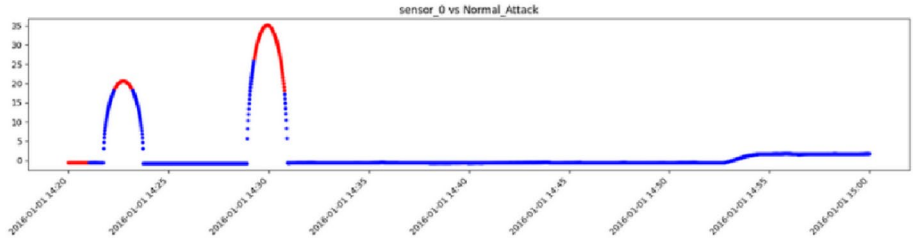

Fig. 13 Attack scenario number 33

7.2 Ablation study

To quantitatively validate our architectural design choices and demonstrate the contribution of each key component, we conducted an extensive ablation study. We compared the performance of our full proposed model against four simplified variants. Each variant was trained and evaluated under the exact same conditions, with performance measured by Area Under the Curve (AUC) and F1-score on the full attack test set. The variants are: (1) a unimodal model using only sensor data; (2) a unimodal model using only network data; (3) a multimodal model using simple concatenation for fusion instead of attention; and (4) a multimodal model without the adversarial training component.

The ablation study results, detailed in Table 3, quantitatively confirm that each architectural component is essential for the model's overall effectiveness. Specifically, replacing the dual cross-attention mechanism with simple concatenation led to a performance degradation, with the AUC dropping from 0.87 to 0.81. Similarly, the absence of the adversarial training component (No AAE) resulted in an AUC reduction to 0.84, highlighting the crucial role of latent space regularization in identifying subtle anomalies. These findings provide robust evidence that the synergy between deep data fusion and adversarial training is necessary to capture the intricate interdependencies between the 115 physical features and the cyber-domain data. Furthermore, this extends the validation methodology reported in [37], demonstrating that these mechanisms are particularly critical for achieving resilience in multimodal environments.

8 Conclusions and future work

This study demonstrated the critical need to advance from isolated, unimodal analysis to a holistic, multimodal fusion strategy for securing critical infrastructures. We introduced a novel, unsupervised framework that significantly enhances anomaly detection by deeply integrating physical process data with network traffic. The primary contribution is an architectural paradigm shift towards context-aware, cyber-physical security; by leveraging VAE-LSTM encoders, a dual cross-attention mechanism, and an Adversarial Autoencoder (AAE), our model learns the complex relationships between cyber commands and physical events.

Table 2 Attacks detected by proposed model

1	Attacks detected	Detection delay	Unexpected outcome reported by SWAT developers
	X		
2	X *	< 30	
3	X		
7	X*	30–60	
8			
10	X		UV did not shutdown; P-501 did not turn off.
11	X		
16	X*	30–60	
17	X *	30–60	
21	X		
22	X		P501 could not be kept on; Reduced output at FIT-502.
23	X*	30–60	
25	X		
26	X		
27	X*	< 30	
28	X		
30	X		
31	X		
32	X		
33	X*	< 30	
35	X*	< 30	
36	X		
37	X		P-501 did not turn off; FIT-502 decreased to 0.8; Speed of P-501 increased to 28.50 Hz from 10 Hz during attack.
39	X		UV did not shutdown.
40	X		P-402 did not close, both should be interlinked.
41	X		Rate of decrease in water level reduced after 1:33:25 PM.

This approach enables the detection of subtle anomalies indicative of zero-day attacks. The model's effectiveness was validated on the SWAT dataset, where it successfully identified 24 out of 26 relevant attack scenarios, proving the viability of detecting complex threats that manifest across both physical and digital domains.

A recognized limitation of this study is the evaluation on a single benchmark. While the SWAT dataset is highly comprehensive for multimodal research, confirming external validity across diverse industrial domains (e.g., WADI or power grids) remains a priority for future work. Future research will involve applying the proposed VAE-LSTM cross-attention architecture to these environments to empirically confirm its adaptability to different physical processes and network protocols.

Furthermore, two advanced research avenues are essential for moving this work toward practical deployment. First, the detection of anomalies that closely mimic normal system behavior remains a significant challenge. This underscores the need for models with enhanced sensitivity that do not simultaneously increase the rate of false alarms. Future research will focus on developing novel loss functions designed to amplify subtle deviations

Table 3 Ablation study results

Model variant	AUC	F1-score
Unimodal (Sensor only)	0.79	0.71
Unimodal (Network only)	0.75	0.68
No Attention (Simple Concatenation)	0.81	0.74
No Adversarial Training	0.84	0.78
Full Model (Proposed)	**0.87**	**0.82**

from the learned baseline. Second, to advance beyond modality-level attribution, we will investigate the integration of explainable AI (XAI) techniques. By analyzing the model's internal attention mechanisms, we aim to develop a methodology capable of identifying the specific sensors or communication patterns most indicative of a detected anomaly, providing actionable intelligence for incident response.

Appendix A

See Figure 14

Model: "model"

Layer (type)	Output Shape	Param #	Connected to
sensor_input (InputLayer)	[(None, 10, 115)]	0	[]
network_input (InputLayer)	[(None, 10, 24)]	0	[]
encoder_sensor (Functional)	[(None, 64), (None, 64), (None, 64)]	108160	['sensor_input[0][0]']
encoder_network (Functional)	[(None, 64), (None, 64), (None, 64)]	136064	['network_input[0][0]']
reshape (Reshape)	(None, 1, 64)	0	['encoder_sensor[3][2]']
reshape_1 (Reshape)	(None, 1, 64)	0	['encoder_network[3][2]']
multi_head_attention (MultiHeadAttention)	(None, 1, 64)	124384	['reshape[0][0]', 'reshape_1[0][0]', 'reshape_1[0][0]']
multi_head_attention_1 (MultiHeadAttention)	(None, 1, 64)	124384	['reshape_1[0][0]', 'reshape[0][0]', 'reshape[0][0]']
flatten (Flatten)	(None, 64)	0	['multi_head_attention[0][0]']
flatten_1 (Flatten)	(None, 64)	0	['multi_head_attention_1[0][0]']
combined_latent (Concatenate)	(None, 128)	0	['flatten[0][0]', 'flatten_1[0][0]']
reshape_2 (Reshape)	(None, 1, 128)	0	['combined_latent[0][0]']
multi_head_attention_2 (MultiHeadAttention)	(None, 1, 128)	247328	['reshape_2[0][0]', 'reshape_2[0][0]']
flatten_2 (Flatten)	(None, 128)	0	['multi_head_attention_2[0][0]']
dense (Dense)	(None, 304)	39216	['flatten_2[0][0]']
dropout (Dropout)	(None, 304)	0	['dense[0][0]']
sensor_reconstruction (Dense)	(None, 1150)	350750	['dropout[0][0]']
network_reconstruction (Dense)	(None, 240)	73200	['dropout[0][0]']
sensor_output (Reshape)	(None, 10, 115)	0	['sensor_reconstruction[0][0]']
network_output (Reshape)	(None, 10, 24)	0	['network_reconstruction[0][0]']
combined_reconstruction (Concatenate)	(None, 10, 139)	0	['sensor_output[0][0]', 'network_output[0][0]']
discriminator (Functional)	(None, 10, 1)	29897	['combined_reconstruction[0][0]']

Total params: 1,233,383
Trainable params: 959,262
Non-trainable params: 274,121

Fig. 14 Final model hyperparameters and configuration

Author contributions
Conceptualization, A.P., Y.D; methodology, A.P.; validation, Y.D., H.L.-C., and G.J.A; formal analysis, A.P. and Y.D.; investigation, A.P., Y.D., H.L.-C., and G.J.A.; resources, Y.D. and; data curation, A.P.; writing original draft preparation, A.P.; writing review and editing, Y.D., H.L.-C., and G.J.A.; visualization, A.P.; supervision, Y.D. and G.J.A. All authors have read and agreed to the published version of the manuscript.

Funding
Andrea Pinto, the corresponding author, has received research support from Universidad de los Andes. The other authors have no relevant financial or non-financial interests to disclose.

Data availability
The datasets generated and/or analyzed during the current study, along with the model code, are available in the Zenodo repository at https://doi.org/10.5281/zenodo.18717110.

Declarations

Ethics approval and consentt to participate
Not applicable.

Consent to publication
The authors provide their consent to publish this manuscript in Discover Computing.

Competing interests
The authors declare no competing interests.

Received: 21 July 2025 / Accepted: 13 March 2026
Published online: 30 March 2026

References

1. Herrera LC, Maennel O. A comprehensive instrument for identifying critical information infrastructure services. Jun 01 2019 Elsevier B V https://doi.org/10.1016/j.ijcip.2019.02.001
2. Aboah Boateng E, Bruce JW, Talbert DA. Anomaly Detection for a Water Treatment System Based on One-Class Neural Network. IEEE Access. 2022;10:115179–91. https://doi.org/10.1109/ACCESS.2022.3218624.
3. Roncone G et al. Apr., APT44: Unearthing Sandworm, 2024. Accessed: Sep. 25, 2024. [Online]. Available: https://cloud.google.com/blog/topics/threat-intelligence/apt44-unearthing-sandworm
4. Lin S, Clark R, Birke R, Schönborn S, Trigoni N, Roberts S. Anomaly Detection For Time Series Using VAE-LSTM Hybrid Model, IEEE, 2020.
5. Pinto A, Herrera LC, Donoso Y, Gutierrez JA. Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure, Mar. 01, 2023, *Sensors*. https://doi.org/10.3390/s23052415
6. Sun H, Huang Y, Han L, Fu C, Liu H, Long X, MTS-DVGAN. Anomaly detection in cyber-physical systems using a dual variational generative adversarial network. Comput Secur. Apr. 2024;139. https://doi.org/10.1016/j.cose.2023.103570.
7. Tushkanova O, Levshun D, Branitskiy A, Fedorchenko E, Novikova E, Kotenko I. Detection of Cyberattacks and Anomalies in Cyber-Physical Systems: Approaches, Data Sources, Evaluation. Algorithms. Feb. 2023;16(2). https://doi.org/10.3390/a16020085.
8. Adepu S, Junejo KN, Mathur A, Goh J. A Dataset to Support Research in the Design of Secure Water Treatment Systems. [Online]. Available: https://www.researchgate.net/publication/305809559
9. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research, *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007, https://doi.org/10.2753/MIS0742-1222240302
10. Kingma DP, Welling M. An Introduction to Variational Autoencoders. Jun. 2019. https://doi.org/10.1561/2200000056.
11. Staudemeyer RC, Morris ER. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.09586
12. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate, Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473
13. Nam J, Lee H, Ngiam J, Khosla A, Kim M, Ng AY. Multimodal Deep Learning, 2011. [Online]. Available: https://www.researchgate.net/publication/221345149
14. Dai Y, Yan Z, Cheng J, Duan X, Wang G. Analysis of multimodal data fusion from an information theory perspective. Inf Sci (N Y). Apr. 2023;623:164–83. https://doi.org/10.1016/j.ins.2022.12.014.
15. Pereira LM, Salazar A, Vergara L. Computers. Jan. 2024;13(1). https://doi.org/10.3390/computers13010013. A Comparative Study on Recent Automatic Data Fusion Methods †.
16. Boulahia SY, Amamra A, Madi MR, Daikh S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Mach Vis Appl. Nov. 2021;32(6). https://doi.org/10.1007/s00138-021-01249-8.
17. Liao TY, Wang W, Xing Y. A method for disturbance identification in power quality based on cross-attention fusion of temporal and spatial features. Electr Power Syst Res. Sep. 2024;234. https://doi.org/10.1016/j.epsr.2024.110560.
18. Yu K, Qin X, Jia Z, Du Y, Lin M. Cross-attention fusion based spatial-temporal multi-graph convolutional network for traffic flow prediction, Sensors, vol. 21, no. 24, Dec. 2021, https://doi.org/10.3390/s21248468
19. Lunardi WT, Lopez MA, Giacalone J-P. ARCADE: Adversarially Regularized Convolutional Autoencoder for Network Anomaly Detection, May 2022, [Online]. Available: http://arxiv.org/abs/2205.01432

20. Goodfellow IJ et al. Generative Adversarial Nets, Adv. Neural Inf. Process. Syst., vol. 27, 2014, [Online]. Available: http://www.github.com/goodfeli/adversarial
21. Zideh MJ, Khalghani MR, Solanki SK. An unsupervised adversarial autoencoder for cyber attack detection in power distribution grids. Electr Power Syst Res. Jul. 2024;232. https://doi.org/10.1016/j.epsr.2024.110407.
22. Raman G, Somu MRN, Mathur AP. A multilayer perceptron model for anomaly detection in water treatment plants. Int J Crit Infrastruct Prot. Dec. 2020;31. https://doi.org/10.1016/j.ijcip.2020.100393.
23. Nedeljkovic D, Jakovljevic Z. CNN based method for the development of cyber-attacks detection algorithms in industrial control systems. Comput Secur. Mar. 2022;114. https://doi.org/10.1016/j.cose.2021.102585.
24. Inoue J, Yamagata Y, Chen Y, Poskitt CM, Sun J. Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. Sep. 2017. https://doi.org/10.1109/ICDMW.2017.149.
25. Xie X, Wang B, Wan T, Tang W. Multivariate Abnormal Detection for Industrial Control Systems Using 1D CNN and GRU. IEEE Access. 2020;8:88348–59. https://doi.org/10.1109/ACCESS.2020.2993335.
26. Faber K, Pietron M, Zurek D. Ensemble Neuroevolution-Based Approach for Multivariate Time Series Anomaly Detection _ Enhanced Reader. Entropy. 2021;23. https://doi.org/10.3390/e23111466.
27. Wang Z, Wang J, Wang C, Zhang L. Anomaly Detection for Industrial Control System based on Heterogeneous Spatio-Temporal GCN, in 2024 *4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, 2024, pp. 79–84. https://doi.org/10.1109/RAAI64504.2024.10949539
28. Fährmann D, Damer N, Kirchbuchner F, Kuijper A. Lightweight Long Short-Term Memory Variational Auto-Encoder for Multivariate Time Series Anomaly Detection in Industrial Control Systems, *Sensors*, vol. 22, no. 8, Apr. 2022, https://doi.org/10.3390/s22082886
29. Shreshth Tuli G, Casale, Jennings NR. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data, *arXiv preprint arXiv:2201.07284*, 2022, Accessed: Oct. 13, 2025. [Online]. Available: https://arxiv.org/abs/2201.07284
30. Capuano N, Fenza G, Loia V, Stanzione C. Explainable Artificial Intelligence in CyberSecurity: A Survey. IEEE Access. 2022;10:93575–600. https://doi.org/10.1109/ACCESS.2022.3204171.
31. Khan IA, Razzak I, Pi D, Zia U, Kamal S, Hussain Y. A Novel Collaborative SRU Network With Dynamic Behaviour Aggregation, Reduced Communication Overhead and Explainable Features, *IEEE J. Biomed. Health Inform.*, vol. 28, no. 6, pp. 3228–3235, Jun. 2024, https://doi.org/10.1109/JBHI.2024.3352013
32. Khan IA, Pi D, Kamal S, Alsuhaibani M, Alshammari BM. Federated-Boosting: A Distributed and Dynamic Boosting-Powered Cyber-Attack Detection Scheme for Security and Privacy of Consumer IoT. IEEE Trans Consum Electron. 2025;71(2):6340–7. https://doi.org/10.1109/TCE.2024.3499942.
33. Khan IA, et al. Fed-Inforce-Fusion: A federated reinforcement-based fusion model for security and privacy protection of IoMT networks against cyber-attacks. Inform Fusion. 2024;101:102002. https://doi.org/10.1016/j.inffus.2023.102002.
34. Gómez ÁLP, Maimó LF, Celdrán AH, Clemente FJG. MADICS: A methodology for anomaly detection in industrial control systems. Symmetry (Basel). Oct. 2020;12(10). https://doi.org/10.3390/SYM12101583.
35. Malits R, Mendelson A. The Use of Hierarchical Temporal Memory and Temporal Sequence Encoder for Online Anomaly Detection in Industrial Cyber-Physical Systems †, *Water (Switzerland)*, vol. 17, no. 3, Feb. 2025, https://doi.org/10.3390/w17030321
36. Chen Y, et al. DBN-BAAE: Enhanced Lightweight Anomaly Detection Mechanism with Boosting Adversarial Autoencoder. Sensors. May 2025;25(10). https://doi.org/10.3390/s25103249.
37. Yang W, et al. Industrial multivariate time-series data anomaly detection incorporating attention mechanisms and adversarial training. Apr 04. 2024. https://doi.org/10.21203/rs.3.rs-4198335/v1.

Publisher's Note