Artificial Intelligence in Cybersecurity: Activities of the CINI-AIIS Lab at University of Naples Federico II

Roberto Canonico[1], Annalisa Navarro[1], Simon Pietro Romano[1], Giancarlo Sperlí[1,*] and Andrea Vignali[1]

[1]*University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy*

Abstract

Integrating Information and Communication Technologies in industrial environments has transformed traditional industrial plant into Cyber-Physical Systems (CPSs), thereby expanding their vulnerability to cyber-attacks with potentially severe consequences. This paper presents the research activities of the CINI-AIIS Laboratory at the University of Naples Federico II targeting the challenge of anomaly detection in CPSs. Our study also presents an empirical evaluation of unsupervised deep learning techniques trained on either sensor/actuator or network traffic data, tested on synchronized real-world CPS datasets. These research activities underscore the importance of multimodal data integration in advancing cyber-security for critical infrastructures.

Keywords

Anomaly Detection, Cyber-Physical Systems, Artificial Intelligence

1. Introduction

Cyber–Physical Systems (CPSs) have evolved rapidly, blending digital technologies with physical processes in industrial settings [1], introducing new layers of complexity to CPSs [2]. As systems grow more interconnected, they become more exposed to cyber threats—from both internal actors and external adversaries [1]. These threats can lead to physical damage or digital compromise, allowing attackers to disrupt operations or infiltrate organizations [3].

The literature has presented various taxonomies for classifying cyberattacks targeting CPSs, reflecting the multidimensional nature of security threats in such environments. Canonico and Sperlì [4] propose a structured taxonomy for classifying these studies along three primary dimensions:

Surveys focused on attack vectors encompasses reviews that examine the methodologies and technologies employed to compromise CPSs. For instance, several works have concentrated on network-oriented attacks, particularly emphasizing intrusion detection mechanisms [5]. Additional surveys have addressed adversarial strategies targeting sensor data integrity through the injection of false measurements to manipulate the behavior of control systems [6].

Surveys oriented toward countermeasures concern research has focused on identifying effective defense strategies. For example, [7] discusses protective measures aimed at mitigating physical damage to infrastructure, whereas [8] investigates defense strategies from a physics-aware perspective, addressing both theoretical and practical challenges in CPS security design.

Surveys categorized by application domains involve organizing surveys according to CPS application areas. The work in [7] offers a longitudinal overview of major CPS-related incidents over the past two decades, analyzing their financial and operational impacts. Reviews in this dimension address domain-specific vulnerabilities and protective strategies across sectors such as smart grids, power systems, and water management infrastructures.

This study presents the research initiatives undertaken by the CINI-AIIS Laboratory at the University of Naples Federico II, specifically targeting the challenge of anomaly detection in CPSs [9, 10]. These efforts are motivated by the persistent and growing need to address cybersecurity threats within CPSs

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 22-23, 2025, Trieste, Italy

*Corresponding author.

roberto.canonico@unina.it (R. Canonico); annalisa.navarro@unina.it (A. Navarro); spromano@unina.it (S. P. Romano); giancarlo.sperli@unina.it (G. Sperlí); andrea.vignali@unina.it (A. Vignali)

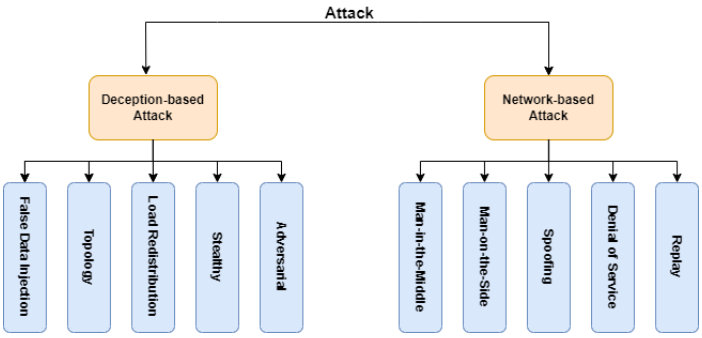

Figure 1: Cyber-attacks taxonomies in CPSs based on the operational methodologies employed in their execution.

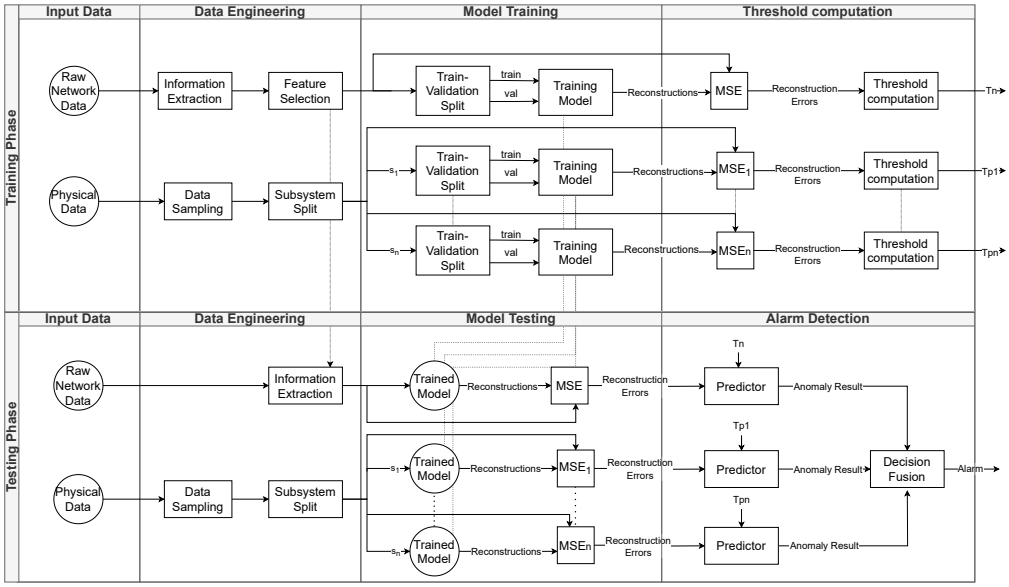

Figure 2: The proposed workflow for detecting cyber-threats by combining both network and sensor data.

due to the continuous escalation in the frequency and sophistication of cyber-attacks [11], which have been systematically categorized by Canonico and Sperlì [4] using the taxonomy depicted in Figure 1. At a high level, these attacks can be broadly classified into two main categories: *deception attacks*, which involve the injection of falsified or misleading data into the system to manipulate its behavior; and *network-driven* attacks, which target the communication infrastructure by injecting, dropping, or modifying network packets to disrupt or compromise data transmission and system coordination.

2. Methodology

This section outlines the methodology developed for detecting anomalies through the analysis of deviations from learned historical patterns, as schematized in Figure 2. The approach integrates unsupervised deep learning techniques (elaborated in Section 2.2) to process both cyber (network traffic) and physical (sensor-based) data streams, as introduced in Section 2.1. To enhance detection performance, a decision-level fusion mechanism is applied—described in detail in Section 2.3—which consolidates the outputs of the most effective models corresponding to each data modality.

2.1. Data Processing

Physical data preprocessing The system architecture is composed of multiple functionally distinct subsystems, each governed by a dedicated Programmable Logic Controller (PLC) and equipped with its

own suite of sensors and actuators. Sensor and actuator data are partitioned accordingly and routed to individual anomaly detection modules. The collected data, representing time-evolving measurements from heterogeneous sources, are treated as multivariate time series, in alignment with the modeling frameworks described in [12, 13]. Prior to anomaly detection, each subsystem's data is normalized using a min-max scaling procedure to ensure consistency in input ranges.

Network data preprocessing In this phase, the system initiates the process by capturing raw network packets in real-time through two phases. In the first phase, the packet processing phase involves the extraction of key flow parameters, specifically the five-tuple (FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, Protocol), along with supplementary information such as payload length and packet timestamps, which are crucial for computing biflow statistics. Following the extraction of per-flow metrics, packets are aggregated into biflow records by pairing packets that correspond to both the quintuple and its reverse counterpart. These biflows capture statistics for both directions of communication (i.e., forward and backward), including the total number of packets transmitted in each direction and overall metrics, such as the mean inter-arrival time for all packets across both directions. The direction of subsequent packets in each flow is determined based on the first packet, which establishes the forward (source to destination) or backward (destination to source) orientation. In TCP communication, flow termination is triggered by the exchange of a FIN (Finish) packet, while for UDP flows, termination is defined by an arbitrary timeout. In summary, the preprocessing module outputs a collection of tuples, each uniquely identified by its flow-specific quintuple, and includes over 80 network traffic-related features. These features are subsequently normalized for each biflow, as detailed in [12, 13].

2.2. Anomaly Detection

To address the anomaly detection task, we utilized two separate data streams—one for sensor data and the other for biflow data—employing distinct deep learning models for each. Both models were trained exclusively on normal system behavior, encompassing network traffic as well as physical data. For this purpose, we adopted deep learning architectures as proposed in [14], which are particularly suited to tackling common challenges in unsupervised anomaly detection.

2.3. Fusion

Integrating sensor and biflow data presents a challenge due to their differing collection methods: sensor data is periodically sampled, while biflow data is non-periodically aggregated. Both data sources yield an anomaly score and a binary label indicating whether a sample is anomalous. To align these two streams, we aggregate predictions for biflows within a given time interval $[t_s; t_s + k]$, summing the number of anomalous biflows. This aggregation allows for a unified prediction within the interval. The prediction process involves shifting the time window by k and collecting data from sensor subsystems at each step. At each time window W, the system determines whether to trigger an alarm. Anomalies are flagged if the number of anomalous biflows exceeds a defined threshold Θ within the interval, or if any physical sensor detects an anomaly. An alarm is triggered when the number of anomalous k values reaches the time window size, or if at least one subsystem reports an anomaly, enabling the detection of anomalies within the specified window.

3. Experimental analysis

This section is divided into two parts: the first will detail the experimental protocol employed (see Section 3.1), which will be used to evaluate the proposed detection approach against various baselines (see Section 3.2).

3.1. Experimental Protocol

The objectives of the experimental protocol are outlined as follows: i) *Model Comparison*, in which the performance of the proposed models will be compared to those found in existing literature (i.e., *USAD* [15] and *GDN* [16] models) for identifying the models that exhibit the best performance in analyzing both network and physical data; and ii) *Assessing fusion approach*: The final step involves assessing the effectiveness of the fusion strategy utilized within the framework.

The experimental evaluation is conducted using the SWaT dataset, which is a scaled-down simulation of a real-world water treatment plant, designed to replicate the operations of an actual water treatment process consisting of six main phases [17]. For this study, both the 2015 and 2019 versions of the SWaT dataset are utilized.

Performance evaluation is carried out using standard metrics such as $F1$, $Recall$, and $Precision$. Additionally, we compute $F1_{An}$, which is calculated specifically for the anomalous class. The framework is implemented on Google Colaboratory, utilizing an Intel Xeon processor with 2 virtual CPUs (vCPUs) and 13GB of RAM.

The network dataset comprises $4,231,290$ biflow records associated with nominal system behavior and an additional $72,127$ biflows representing malicious activity. These attack biflows are divided into two types: $72,095$ for the first attack type and 32 for the second. The first $442,173$ biflows, corresponding to the initial hour of normal operation before attacks commence, are used for training and validation, with the remaining biflows (including both attack and normal instances) forming the test set. The **physical dataset** contains A total of $12,301$ benign instances and 901 malicious instances—exclusively linked to the second attack scenario in the 2019 dataset release—were reserved strictly for evaluation during the testing phase. Conversely, the 2015 release, which includes $396,800$ normal samples, is utilized for model training.

3.2. Results

Table 1 presents the evaluation results of the considered models on the test datasets, using standard metrics: $F1$, $F1_{An}$, $Precision$, and $Recall$. In the case of the first network-based attack, the Autoencoder (AE) model consistently outperforms other architectures, while AE integrated with LSTM cells, Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs) demonstrate reduced effectiveness. Conversely, the second type of network attack yields suboptimal performance across all models, indicating its challenging detection characteristics.

For the physical dataset, derived from sensor and actuator readings, GANs achieve the highest $F1$ score, surpassing advanced detection methods such as USAD and GDN. These results underline the necessity of model specialization for different data modalities.

To account for this heterogeneity in detection capabilities, the proposed framework strategically combines AE-based models for identifying network anomalies and GAN-based models for detecting more nuanced deviations in physical data. This dual-model approach enhances detection accuracy by aligning model strengths with the specific characteristics of cyber-physical data sources.

The findings reported in Table 2 provide a comparative overview of detection performance across different data integration strategies. Notably, the fusion-based method delivers superior overall performance, achieving the highest scores in both $F1$ and $F1_{AN}$ metrics. Additionally, it outperforms the physical-only approach in terms of Precision, although a marginal decline in Precision is observed when compared to network-only predictions. This reduction is attributable to the propagation of false positives introduced by the physical anomaly detector. Conversely, the post-fusion strategy markedly enhances Recall, successfully detecting all anomalous operational cycles, thereby demonstrating its robustness in capturing diverse attack manifestations across the system.

This integrative approach enhances both detection granularity and resilience, yielding a reported increase in F1-score of approximately 10% for network-based threats and 30% for anomalies in physical data. Furthermore, the system achieves threat detection latency within 2–3 seconds, enabling timely intervention to contain or mitigate damage in operational CPS environments.

Dataset	Model	$F1$	$F1_{An}$	$Precision$	$Recall$
Network First attack	**USAD**	$0.199_{\pm 0.007}$	$0.128_{\pm 0.001}$	$0.534_{\pm 0.000}$	$0.578_{\pm 0.004}$
	GDN	$0.755_{\pm 0.001}$	-	$0.682_{\pm 0.020}$	$0.835_{\pm 0.003}$
	AE	$0.939_{\pm 0.018}$	$0.886_{\pm 0.035}$	$0.898_{\pm 0.028}$	$0.992_{\pm 0.002}$
	AE$_{LSTM}$	$0.485_{\pm 0.002}$	$0.0006_{\pm 0.0004}$	$0.474_{\pm 0.004}$	$0.498_{\pm 0.001}$
	VAE	$0.484_{\pm 0.002}$	$0.0007_{\pm 0.0004}$	$0.474_{\pm 0.004}$	$0.498_{\pm 0.001}$
	GAN	$0.485_{\pm 0.001}$	0	$0.471_{\pm 0.003}$	0.5
Network Second attack	**USAD**	$0.132_{\pm 0.007}$	$0.0005_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.476_{\pm 0.005}$
	GDN	$0.001_{\pm 0.0005}$	-	$0.001_{\pm 0.000}$	$0.517_{\pm 0.177}$
	AE	$0.497_{\pm 0.003}$	$0.003_{\pm 0.002}$	$0.501_{\pm 0.0005}$	$0.891_{\pm 0.280}$
	AE$_{LSTM}$	$0.498_{\pm 0.002}$	$0.0005_{\pm 0.001}$	$0.500_{\pm 0.0003}$	$0.513_{\pm 0.047}$
	VAE	$0.498_{\pm 0.002}$	0	$0.500_{\pm 0.001}$	$0.513_{\pm 0.047}$
	GAN	$0.500_{\pm 0.001}$	0	$0.500_{\pm 0.001}$	$0.500_{\pm 0.001}$
Physical Data	**USAD**	$0.562_{\pm 0.160}$	$0.31_{\pm 0.186}$	$0.586_{\pm 0.089}$	$0.768_{\pm 0.291}$
	GDN	$0.492_{\pm 0.263}$	-	$0.349_{\pm 0.206}$	$0.917_{\pm 0.144}$
	AE	$0.658_{\pm 0.232}$	$0.397_{\pm 0.303}$	$0.643_{\pm 0.130}$	$0.734_{\pm 0.224}$
	AE$_{LSTM}$	$0.563_{\pm 0.132}$	$0.240_{\pm 0.200}$	$0.589_{\pm 0.155}$	$0.600_{\pm 0.081}$
	VAE	$0.645_{\pm 0.177}$	$0.428_{\pm 0.380}$	$0.612_{\pm 0.147}$	$0.718_{\pm 0.262}$
	GAN	$0.797_{\pm 0.001}$	$0.638_{\pm 0.002}$	$0.734_{\pm 0.001}$	$0.958_{\pm 0.0003}$

Table 1
Model performance was evaluated on the test dataset by reporting the mean values and corresponding 95% confidence intervals for key metrics.

Data source	$F1$	$F1_{An}$	$Precision$	$Recall$
Network Data	0.8316(+10.17%)	0.7104(+21.99%)	0.9426(-6.25%)	0.7781(+23.88%)
Physical Data	0.7078(+29.44%)	0.5048(+71.67%)	0.7449(+18.63%)	0.6859(+40.49%)
Fusion (Both Data)	**0.9162**	**0.8666**	**0.8837**	**0.9639**

Table 2
Effectiveness analysis by comparing the proposed fusion approach w.r.t. the single-layer analysis.

4. Conclusion

This study presents an integrated anomaly detection framework designed to enhance the resilience of CPSs against a wide spectrum of cybersecurity threats. The proposed approach systematically assesses and selects the most effective deep learning models tailored to the unique characteristics of both network traffic and physical process data. By adopting a dual-stream analysis and applying a decision-level fusion strategy, the framework synthesizes complementary insights from heterogeneous data sources to deliver more robust and accurate anomaly detection. Experimental evaluations demonstrate that the combined use of specialized models significantly enhances detection performance and minimizes false positives, a common challenge in CPS security monitoring. Furthermore, the framework achieves low-latency detection, with anomalies being identified within a 2–3 second window—an operationally critical timescale for prompt threat response and containment.

Acknowledgments

This work was supported in part by the Piano Nazionale Ripresa Resilienza (PNRR) Ministero dell'Università e della Ricerca (MUR) Project under Grant PE0000013-FAIR. We also acknowledge support from NextGenerationEU via PNRR - DM 352 (CUP: E66G22000400009).

Declaration on Generative AI

During the preparation of this work, the authors did not use generative AI tools.

References

[1] M. Segovia-Ferreira, J. Rubio-Hernan, A. Cavalli, J. Garcia-Alfaro, A Survey on Cyber-Resilience Approaches for Cyber-Physical Systems, ACM Comput. Surv. 56 (2024). doi:10.1145/3652953.

[2] J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, R. Candell, A Survey of Physics-Based Attack Detection in Cyber-Physical Systems, ACM Computing Surveys 51 (2018). doi:https://doi.org/10.1145/3203245.

[3] A. T. A. Ghazo, R. Kumar, Critical Attacks Set Identification in Attack Graphs for Computer and SCADA/ICS Networks, IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023) 1–0. doi:10.1109/TSMC.2023.3274613.

[4] R. Canonico, G. Sperlì, Industrial cyber-physical systems protection: A methodological review, Computers & Security 135 (2023) 103531. doi:https://doi.org/10.1016/j.cose.2023.103531.

[5] L. Cao, X. Jiang, Y. Zhao, S. Wang, D. You, X. Xu, A Survey of Network Attacks on Cyber-Physical Systems, IEEE Access 8 (2020) 44219–44227. doi:10.1109/ACCESS.2020.2977423.

[6] L. Cui, Y. Qu, L. Gao, G. Xie, S. Yu, Detecting false data attacks using machine learning techniques in smart grid: A survey, Journal of Network and Computer Applications 170 (2020) 102808. doi:https://doi.org/10.1016/j.jnca.2020.102808.

[7] T. Alladi, V. Chamola, S. Zeadally, Industrial Control Systems: Cyberattack trends and countermeasures, Computer Communications 155 (2020) 1–8. doi:https://doi.org/10.1016/j.comcom.2020.03.007.

[8] C. M. Ahmed, J. Zhou, Challenges and Opportunities in Cyberphysical Systems Security: A Physics-Based Perspective, IEEE Security & Privacy 18 (2020) 14–22. doi:10.1109/MSEC.2020.3002851.

[9] S. Schmidl, P. Wenig, T. Papenbrock, Anomaly detection in time series: a comprehensive evaluation, Proc. VLDB Endow. 15 (2022) 1779–1797. doi:10.14778/3538598.3538602.

[10] Y. Yang, C. Zhang, T. Zhou, Q. Wen, L. Sun, DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3033–3045. doi:10.1145/3580305.3599295.

[11] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, A. Liotta, Smart anomaly detection in sensor systems: A multi-perspective review, Information Fusion 67 (2021) 64–79. doi:https://doi.org/10.1016/j.inffus.2020.10.001.

[12] R. Canonico, G. Esposito, A. Navarro, S. P. Romano, G. Sperlí, A. Vignali, An anomaly-based approach for cyber-physical threat detection using network and sensor data, Computer Communications (2025) 108087.

[13] R. Canonico, G. Esposito, A. Navarro, S. P. Romano, G. Sperlí, A. Vignali, Empowered cyber-physical systems security using both network and physical data, Computers & Security (2025) 104382.

[14] G. Li, J. J. Jung, Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges, Information Fusion 91 (2023) 93–102. doi:https://doi.org/10.1016/j.inffus.2022.10.008.

[15] J. Audibert, P. Michiardi, F. Guyard, S. Marti, M. A. Zuluaga, Usad: Unsupervised anomaly detection on multivariate time series, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3395–3404. doi:https://doi.org/10.1145/3394486.3403392.

[16] A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time series, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 4027–4035. doi:https://doi.org/10.1609/aaai.v35i5.16523.

[17] J. Goh, S. Adepu, K. N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, in: Critical Information Infrastructures Security: 11th International Conference, CRITIS 2016, Paris, France, October 10–12, 2016, Springer, 2017, pp. 88–99. doi:https://doi.org/10.1007/978-3-319-71368-7_8.