Future Generation Computer Systems 180 (2026) 108382

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

LOTL-hunter: Detecting multi-stage living-off-the-land attacks in cyber-physical systems using decision fusion techniques with digital twins

Carol Lo [a,*], Thu Yein Win [b], Zeinab Rezaeifar [a], Zaheer Khan [a], Phil Legg [a]

[a] Computer Science Research Centre, University of the West of England, Bristol, BS16 1QY, UK
[b] Faculty of Technology, University of Sunderland, Sunderland, SR1 3SD, UK

ARTICLE INFO

Keywords:
Digital twins
ICS testbed
Cyber physical systems
Big data analytics
Decision fusion
Threat hunting
Threat simulation
Advanced persistent threats
Multi-stage attacks
Living off the land attacks

ABSTRACT

The integration of smart sensors and actuators in industrial environments has expanded the cyber-physical attack surface, making it increasingly difficult to distinguish anomalies caused by cyberattacks from those due to mechanical or electrical faults. This challenge is exacerbated by stealthy, multi-stage attacks leveraging Living off the Land (LOTL) techniques, which often evade conventional anomaly detection or intrusion detection systems (IDS).

This study presents a Digital Twin-based testbed for safe, repeatable simulation of multi-stage cyber-physical attacks targeting Cyber-Physical Systems (CPS) and Industrial Control Systems (ICS). We propose a two-level decision fusion method that aggregates and aligns anomalies across network, process, and host domains in synchronized 1-minute intervals. The first-level fusion improves OT-layer detection by applying confidence-aware decision logic to outputs combined from (a) a supervised deep learning model (LSTM-FCN) for process anomalies, (b) an unsupervised model (Isolation Forest) for OPC UA network anomalies, and (c) process alarm signals. The second-level fusion integrates these results with host-based anomalies, computed through point-based scoring of Wazuh alerts, to provide comprehensive IT/OT situational awareness. Experimental results demonstrate improved detection of stealthy, multi-stage APT attack behaviours. Additionally, Large Language Models (LLM) provide summarization of the integrated IT/OT anomaly logs into human-readable insights, enhancing interpretability and supporting cyber threat hunting.

1. Introduction

The increasing adoption of TCP/IP-based communication protocols in industries is enhancing interoperability between Information Technology (IT) and Operational Technology (OT), enabling more efficient monitoring and control of physical processes [1]. However, this IT/OT convergence also expands the cyber-physical attack surface, exposing critical infrastructure to sophisticated threats [2]. For instance, Modbus TCP is unencrypted, making physical processes vulnerable to interception, spoofing, and replay attacks [3].

A major challenge is distinguishing between anomalies caused by legitimate faults (e.g. mechanical degradation) and those triggered by cyberattacks [4]. Misdiagnosing a cyber-induced disruption can delay incident response and recovery [5]. Conventional anomaly detection or intrusion detection systems (IDS) typically analyse on either network data or physical process data alone, limiting their detection capability [6].

This problem is intensified by Advanced Persistent Threats (APTs), which use stealthy, multi-stage attacks over prolonged campaigns. Living off the Land (LOTL) techniques, which abuse trusted tools or protocols already present in the target environment, are increasingly used to evade detection by blending into normal behaviours [7–10]. For instance, APT Group Sandworm Team remotely used operators' Human Machine Interface (HMI) to intermittently open and close substation breakers, disrupting Ukraine's power grid (MITRE ATT&CK T0823) [11].

Cyber-physical attacks typically span two domains: beginning with network-based attacks via IT infrastructure, followed by attempts to disrupt physical processes [12] through lateral movement. Early detection and situational awareness are critical to prevent irreversible damage and timely response to multi-stage attacks [13]. As such, the cross-domain nature of cyber-physical attacks necessitates an integrated detection approach that correlates IT/OT anomalies.

A Digital Twin is a virtual replica of a physical system or process [14], and has emerged as a viable platform for simulating threat scenarios in a safe and scalable industrial environment [15]. It enables evaluation of not only the impact of cyber-physical threats targeting Cyber-Physical Systems (CPS) and Industrial Control Systems (ICS), but also

* Corresponding author.
E-mail address: carol.lo@uwe.ac.uk (C. Lo).

https://doi.org/10.1016/j.future.2026.108382
Received 6 June 2025; Received in revised form 6 December 2025; Accepted 16 January 2026
Available online 20 January 2026

validation of detection capabilities [16]. This study leverages a Digital Twin-based testbed to simulate stealthy, multi-stage attacks and assess the performance of multimodal anomaly detection techniques across IT and OT domains.

1.1. Research questions (RQ)

Based on the challenges outlined above, we hypothesize that correlating physical anomaly data (captured by a Digital Twin) with network and host anomalies can enhance the visibility across the IT/OT domain, thereby facilitating the early detection of stealthy, multi-stage cyber-physical attacks. To investigate this hypothesis, we propose a two-level decision fusion approach that combines anomalies across heterogeneous data modalities to offer a holistic view of cross-domain attack behaviours. This approach is evaluated using data captured from a Digital Twin-based testbed that recently simulated an APT attack chain. Accordingly, this study addresses the following research questions:

RQ1: How can we model the stages of a LOTL attack chain using a Digital Twin simulated environment?

RQ2: How effective is the two-level decision fusion between OT network anomalies and physical process anomalies in improving the detection of stealthy cyber-physical threats compared to single-source models?

RQ3: How effective are continuous host-based security analytics within a Digital Twin-based testbed of a CPS in detecting multi-stage cyber-physical threats?

These research questions guide the design, implementation, and evaluation of the proposed method.

1.2. Novelty and contributions

The novelty of this research lies in the use of a coordinated, Digital Twin-based testbed that enables safe, repeatable evaluation of cross-domain anomaly detection strategies using realistic IT/OT attack scenarios. Unlike prior studies that focus on individual data sources, this work demonstrates the feasibility of correlating heterogeneous anomalies within a synchronised framework. The testbed integrates industrial-grade ICS/OT applications, the Open Platform Communications Unified Architecture protocol (OPC UA), and open-source tools such as Zeek [17], Wazuh [18] and PostgreSQL [19], allowing for efficient acquisition and analysis of multimodal data. It also reduces the hardware and infrastructure cost and complexity associated with maintaining physical testbeds, while also addressing the scarcity of APT-specific datasets, which are difficult to obtain in real-world industrial systems. Using this platform, the study offers two key contributions:

1. **Multimodal anomaly fusion for OT-centric threat detection**: We develop the first-level fusion strategy combining supervised time-series classification of process anomalies with unsupervised detection on OPC UA network anomalies and process alarm signals. A confidence-aware decision logic is used to handle missing data and conflicting predictions (Section 4.3). This fusion approach breaks the data silos, and OT-layer threat detection, outperforms individual modalities.
2. **IT/OT anomaly correlation for early detection of stealthy, multi-stage APT attack behaviours**: We use the second-level fusion strategy to correlate OT anomalies with host-based anomalies, using a point-based binary scoring system to quantify and visualise attack progression over time (Section 4.1.3). The resulting threat summary integrates IT/OT anomalies, and provides transparent scores and explanations of triggered rules (refer to Fig. 10 for an example of anomaly record and Fig. 11 for log summarisation by LLM). The dataset and code are released on GitHub for reproducibility.

The Digital Twin plays an indispensable role in our proposed anomaly detection methodology. It serves as a high-fidelity data synthesizer, enabling the controlled creation of multi-modal, labelled datasets for OT process and network behaviours. This fidelity is essential for training and evaluating our multi-class anomaly detection model, whose outputs feed directly into the first-level decision fusion strategy and support LLM-based threat summarisation for continuous monitoring, proactive detection and enhanced cyber situational awareness - a critical application in Digital Twin ecosystems and industries. The Digital Twin is also a structural component of the coordinated testbed that enables evaluation of the proposed two-level decision fusion strategy. Without the Digital Twin, neither cross-domain simulation nor rigorous evaluation of the proposed IT/OT detection approach would be possible.

1.3. Structure of the paper

Section 2 reviews related work. Section 3 presents the Digital Twin-based testbed. Section 4 describes the detection methodology. Section 5 reports the experimental results. Section 6 discusses benefits and limitations. Section 7 concludes the paper and outlines future work.

2. Related work

2.1. Digital twin as a testbed for threat detection

ICS/OT testbeds, including physical, virtual, and hybrid setups, have long supported cybersecurity research by enabling controlled environments for simulating attacks, testing detection capabilities, and generating datasets for intrusion or anomaly detection [12,20,21].

However, physical testbeds are costly to setup, subject to safety constraints, and difficult to replicate [22]. Virtual testbeds, while more accessible, are often criticised for limited realism in process simulation [23]. Hybrid testbeds seek to balance these trade-offs, but still require specialized hardware knowledge and careful planning for hardware/software integration [24].

Digital Twin technology has recently gained traction in securing industrial systems as a safe option. The NIST SP 800-82 Rev. 3 guideline [25] explicitly recommand the use of Digital Twins for real-time anomaly detection by replicating and analysing data from operational environments. Several studies, including [4,26,27], have demonstrated the Digital Twins-driven anomaly detection using techniques ranging from heuristics or deep learning.

Beyond replicating industrial processes for anomaly detection, Digital Twin also offer a risk-free platform for simulating cyber-physical attacks [28]. For instance, Dietz et al. [29] suggested integrating simulated incidents into Security Information and Event Management (SIEM) systems, while Empl and Pernul [30] proposed generating security insights from Digital Twin simulations.

However, these efforts do not explicitly address stealthy, multi-stage APT attack behaviours leveraging LOTL techniques - a critical gap in current literature. Recent survey studies [31,32] revealed that the primary focus of LOTL detection has been on techniques involving PowerShell and fileless malware. Detection approaches that consider multi-stage APT scenarios targeting OT-layer cyber-physical threats remain lacking. To address this gap, our study develops a Digital Twin-based testbed integrating HMI and Engineering Workstations to simulate APT-style attacks that traverse IT and OT boundaries using LOTL techniques.

Recent Digital Twin-based detection frameworks have different threat scopes. For example, Alcaraz and Lopez [33] focused on Modbus protocol misuse. Different from their work, when simulating the abuse of standard OT network traffic as per MITRE ATT&CK Technique T0869 [11], we focused on OPC UA - a widely adopted machine-to-machine protocols in Industry 4.0 environments.

Despite its design for secure communication, the security of OPC UA is heavily dependent on proper configuration. For instance, *'None'* security policy disables encryption or authentication, making systems vulnerable to interception and tampering [3]. Dahlmanns et al. [34] surveyed Internet-exposed OPC UA servers, and reported that 92% exhibit insecure configurations, revealing a widespread false sense of security.

Although OPC UA is widely used, its security remains under-explored in research community. Our work addresses this gap through focused simulation and analysis of OPC UA-based threats within a Digital Twin-based testbed.

2.2. Anomaly detection in OT network and systems

In response to the evolving threat landscape, IDS are commonly implemented to detect anomalies across networks and systems. IDS can be classified into *network-based* (NIDS) or *host-based* (HIDS), which monitor network traffic and host-level activities [12,35].

Among NIDS, signature-based systems like Snort [36] and Suricata [37] are widely used to detect known attack patterns using predefined signatures [38,39]. They have low false positives but struggle to detect novel attacks and require frequent rule updates [35].

OT traffic typically exhibits periodic, predictable patterns due to cyclic polling. In stable ICS/OT environments, communication patterns tend to be consistent and predictable [40]. However, when adversaries abuse standard OT protocols using LOTL techniques (e.g., via HMI or OPC UA clients already present), the traffic may closely resemble legitimate client/server polling. This makes it difficult for signature-based NIDS to distinguish benign variation from subtle malicious activity.

Retuerta et al. [41] observed that OPC UA encryption limits Suricata's ability to inspect packet contents effectively. They proposed supplementing NIDS with HIDS deployed at endpoints, capable of analysing network packets before encryption, or after decryption. Nonetheless, NIDS can still be valuable for detecting anomalous traffic patterns, such as sudden increases in data flow, even without decrypting packet content.

Examples of open-source HIDS include Wazuh [18] and OSSEC [42], which monitor host-level telemetry such as system logs, file integrity, and user activity. Sysmon [43], a native Windows utility, records detailed system-level events, such as process creation, PowerShell script execution, and network connections. These tools are particularly useful in detecting LOTL behaviours that exploit native binaries and command-line utilities within the Windows operating systems.

Machine learning-based anomaly detection is gaining popularity because it does not require attack signatures and can adapt to dynamic environments [35]. However, the lack of high-quality labelled datasets remains a challenge for model training [39].

Despite the increasing adoption of OPC UA, research on machine learning-based intrusion or anomaly detection specifically targeting OPC UA traffic remains limited. This highlights the need for investigation into OPC UA-specific anomaly detection strategies, particularly stealthy attacks leveraging LOTL techniques.

2.3. Anomaly detection in physical process data

Detecting anomalies in physical process data - such as sensor readings, actuator states, and PLC control signals - is a crucial aspect of fault detection and CPS security, contributing to overall system safety and availability. Process-level anomalies could be monitored by operators using HMI or SCADA software. For basic anomaly detection, threshold-based alarms can be configured to alert operators when sensor/ actuator values fall outside acceptable ranges. However, these approaches struggle to identify anomalies that involve small perturbation within the tolerable range [44].

Unlike threshold-based system, anomaly-based systems can detect otherwise undetectable attacks [45]. The knowledge base of the system could be the fundamental physical laws of CPS (e.g., fluid dynamics, Newton's laws, electromagnetic laws). To detect potential false control commands, or false sensors readings, the monitoring system observes sensor readings and actuator outputs to detect deviations from physical laws, such as unexpected movement trajectories, or mismatch between control commands and system responses. While this method can detect obvious anomalies, it may not detect subtle or stealthy attacks that gradually and incrementally drive the system toward unsafe states [7].

Recent research increasingly applies machine learning to develop anomaly detection systems, driven by two main factors: (1) the growing availability of TCP/IP-enabled process data, and (2) reduced reliance on expert-defined detection rules. Unsupervised machine learning is often used to formulate the baseline of the anomaly detection system, through observing live or historical process data over time. Alert is triggered when there are deviations from baseline [35].

In earlier work, a container-based Digital Twin was set up for capturing data for training unsupervised machine-learning anomaly detection, focusing on detecting subtle deviations on physical process at an early stage [46]. However, the detection was limited to binary outcomes (normal vs. abnormal), without distinguishing between different types of attack or process failure. Moreover, although unsupervised machine learning can detect unknown attacks, it comes with the cost of many false positives [7,35], hindering adoption [47].

To reduce false positives, Urbina et al. [23] proposed to consider the attack impact and tune the detection algorithm to increase the chance of detecting high-risk attacks. Umsonst and Sandberg [48] used a game theory-based approach, but focused on detecting stealthy attacks targeting sensors only.

Compared to unsupervised models, supervised machine learning achieves higher detection accuracy and more precise classification. To address the limitations of binary detection and high false positives, supervised learning approaches have been explored. Our earlier work demonstrated the capability of detecting and predicting multi-class classification of different attack scenarios (e.g., sensor manipulation, actuator failure, normal operations), offering fine-grained insights into the nature of process anomalies, allowing better situational awareness for operators and incident responders [49].

However, supervised machine may struggle to detect novel or zero-day attacks. To address the issues, Faramondi et al. [50] suggested to build a better knowledge base about the system dynamics by extracting temporal features indicating specific cyber- or physical anomalies. However, variability in industrial processes and the limited availability of labelled datasets for attack scenarios further complicate the application of supervised machine learning models to detect physical process anomalies.

Overall, significant challenges remain in confirming whether physical anomalies are caused by physical degradation or from cyberattacks [4]. Ahmed et al. [47] suggested to have more detailed feature extraction from process data, and separate detection scope between physical anomalies and cyber/network anomalies. Contextual information is also required to distinguish faults and attacks.

2.4. Multimodal data fusion for anomaly detection

Defined by Lahat et. al. [51], a data modality refers to a dataset produced through a specific acquisition framework. Data fusion is the analytical process of integrating multiple modalities to extract insights unattainable from a single source. This technique is particularly relevant for detecting cyber-physical attacks spanning IT and OT domains, where information silos remain a significant challenge in industrial environment.

Fusion strategies are classified into early and late [52]. Early (feature-level) fusion, combines raw or preprocessed data before model training, while late (decision-level) fusion preserves the uniqueness of each modality by combining model outputs. This work adopts late fusion to retain modality-specific interpretability, and support explainable, actionable predictions for identifying cyber-physical threats.

While most CPS anomaly detection approaches rely on either network data or physical process data alone, such as [53] focused on detecting physical anomalies in robotic arms without correlating network or host-level anomalies. Some recent studies show the benefit of combining both. For example, Canonico et al. [6] demonstrated that decision

fusion of network and process data can detect subtle attacks. Building upon this direction, our work focuses on detecting multi-stage APT attack behaviours across IT and OT domains. To achieve this, we incorporate host-level anomalies and process alarms signals in addition to process and network data.

To address the latency and alignment issues often encountered in distributed IDS environments, Abid et al. [54] proposed a cloud-based, big data streaming approach. In contrast, our study adopts a simpler time-based aggregation strategy, aligning multimodal anomalies into one-minute time intervals to reduce computational overhead and complexity, as further discussed in Section 3.2. While this resolution offers practical benefits, it may reduce visibility of short-lived or transient anomalies.

Finally, as noted by [51], several types of data uncertainty complicate the data fusion process: (1) missing data caused by logging failures or asynchronous sampling can disrupt alignment across modalities, (2) conflicting predictions between modalities, (3) noise such as errors in real-world dataset hindering evaluation. Our methodology explicitly considers these challenges to mitigate the risk of false correlations and improve the robustness of the detection process.

3. Experimental testbed design for decision fusion

This section states our decision fusion approach for multimodal anomaly detection and outlines the design and implementation of a Digital Twin-based testbed.

3.1. Overview of two-level decision fusion

Real-world multi-stage APT threats often originate as network-based intrusions and escalate to physical process attacks targeting CPS. Accordingly, our decision fusion approach spans multimodal data from IT and OT systems across Levels 0 to 3 of the Purdue Reference Model [55], with further details of the testbed design presented in Section 3.3.

As illustrated in Fig. 1, a two-level decision fusion strategy is proposed to correlate anomalies detected across diverse sources. Rather than replacing IDS, this strategy serves as a decision support layer to enhance cyber situational awareness by integrating heterogeneous anomalies into a unified, explainable output.

First-level decision fusion operates within the OT domain. It combines predictions from two classifiers with process alarm signals using confidence-aware decision logic. This logic serves two key purposes: (1) to bridge the data gaps due to logging failures or limited sensor visibility in one modality, and (2) to resolve conflicting or uncertain predictions across classifiers.

Second-level decision fusion correlates IT-centric host anomalies with the outputs of the first-level OT fusion. The final output is a cyber-physical threat summary that consolidates cross-domain anomalies in chronological order, improving visibility into multi-stage APT attack behaviours.

3.2. Aggregation and alignment strategy

Temporal aggregation and alignment are used to merge anomalies from all modalities in one-minute intervals, enabling time-based synchronisation across heterogeneous data sources - including host logs, network traffic, process alarms, and physical process telemetry. The one-minute window was chosen as a time anchor for two key reasons:

1. Practical data engineering - it simplifies data extraction and synchronisation by avoiding the complexity of aligning sub-second events, handling variable sampling rates, latency issues, and timestamp inconsistencies across devices.
2. Level of granularity - it sufficiently supports the detection of low-and-slow APT behaviours, while also facilitating intuitive visual analysis by amplifying significant or sustained anomalies and enhancing visibility into stealthy attack progression.

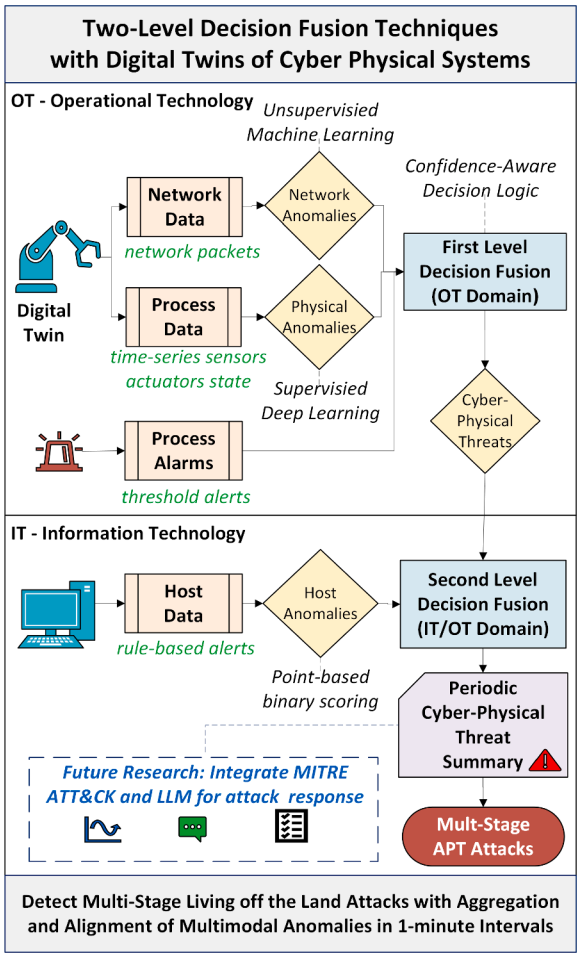

Fig. 1. An overview of two-level decision fusion.

A batch processing strategy was selected over real-time processing for two practical reasons. First, it mitigates alert fatigue issues by reducing the volume of less significant or transient alerts commonly seen in real-time IDS. Second, it provides a temporal buffer to accommodate data quality issues possibly caused by transmission latency, while also reducing computational overhead for processing data from multi-sources.

While this temporal resolution offers practical benefits for detecting stealthy, low-and-slow APT-style attack behaviours, it may come at the cost of averaging out short-lived, transient anomalies. Nonetheless, this trade-off was considered acceptable to prioritise computational efficiency and deployment simplicity.

3.3. Testbed design and implementation

Our detection methodology reflects the adoption of Secure-by-Design principles in Digital Twin ecosystems engineering. In the current stage, the Digital Twin is intentionally operated in a 'disconnected state' [56] to enable iterative refinement and safe early-stage testing of the detection pipeline. The Digital Twin and associated software components support future hardware-in-the-loop (HITL) validation, providing a clear progression toward further testing and eventual deployment within the Digital Twin lifecycle.

Our Digital Twin-based testbed spans IT and OT systems within Level 0 to Level 3 of the Purdue Reference Model [55], as illustrated in Fig. 2, enabling the simulation of real-world multi-stage APT threats that typically originate from network-based intrusions and escalate to physical process attacks targeting CPS.

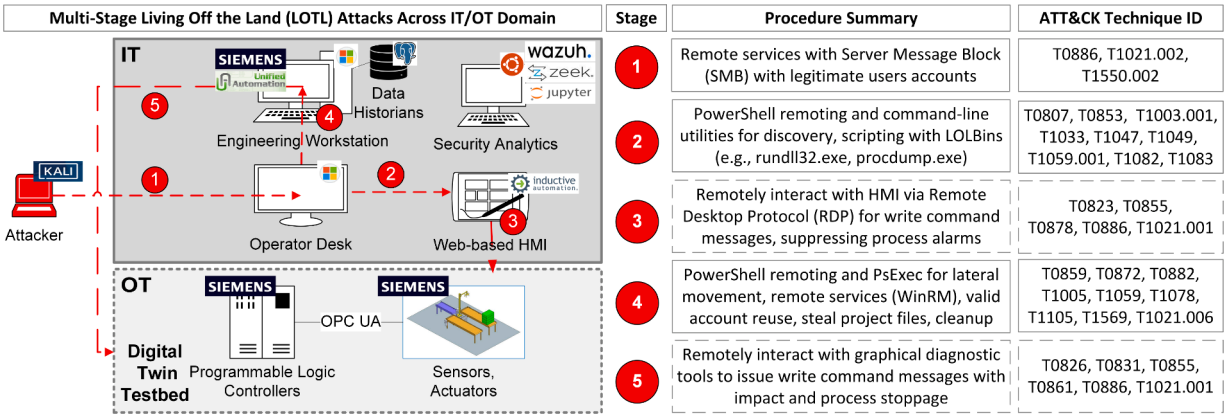

Fig. 2. Digital twin testbed for multi-stage threat simulation.

The simulated attack sequence in Fig. 2 comprises five LOTL-based steps: (1) initial access via Operator Desk, (2) discovery and execution through command-line tools, (3) manipulation of the physical process via remote HMI access, (4) lateral movement to Engineering Workstation, and (5) process manipulation using diagnostic tools. A detailed analysis of how these steps were simulated and detected is provided in Section 5.5. Furthermore, while we have selected a set of commonly used attack techniques that are well-suited to the current case study, we believe the framework is flexible enough to accommodate a broader range of techniques. To explore this further, a separate case study is currently being developed to include Debian-based attacks. However, its details fall outside the scope of this paper.

This setup enables cross-domain anomaly detection across process telemetry, network traffic and host activity, providing a holistic view of stealthy, multi-stage cyber-physical threat scenarios. It offers a safe and repeatable environment for evaluating attack progression, detection performance, and decision fusion under realistic conditions, while avoiding operational disruption and preserving fidelity for experimental validation.

The testbed integrates industrial-grade Digital Twin technologies at Level 0 to Level 1, including Siemens NX Mechatronics Concept Designer (NX MCD) [57] for simulating physical process and network behaviour, and Siemens S7-PLCSIM Advanced [58] for PLC logic and control. These tools were selected for their high-fidelity modelling of OT processes and networks using software widely deployed in industry, while supporting future hardware-in-the-loop scalability. The emulated process involves a robot gripper and conveyor belts representing a typically pick-and-place operation. Virtual sensors and actuators operate autonomously, without manual intervention, as illustrated in Fig. 3.

At Level 2, a Windows-based Operator Desk runs Ignition HMI (Maker Edition) [59] for supervisory control. At Level 3, Engineering Workstation hosts Siemens TIA Portal [60] for PLC project management, PostgresSQL [19] for threshold-based alarm logging, and UA Expert [61] for diagnostic inspection. Both tools can observe and influence the system state, enabling operator control during simulation. This choice of components reflects the prevalence of Windows-based HMI and engineering tools.

To simulate and detect LOTL techniques - including remote access via legitimate services, command-line abuse for discovery, lateral movement, data exfiltration, and physical disruption - the testbed integrates several open-source security tools. Wazuh agents (host-based IDS) [18] are installed on both the Operator Desk and Engineering Workstation to forward host telemetry to Wazuh SIEM in Ubuntu for centralised logging and monitoring. Also, Wireshark [62] and Zeek with an OPC UA plug-in [17] are used for enriched network traffic logging and packet analysis.

3.4. Threat simulation

For modelling stealthy APT threats that leverage LOTL techniques, the attack scenario is inspired by the operations of the Sandworm Team which conducted various cyber-physical attacks on Ukraine's energy system. We also made reference to the CISA guideline [63] and MITRE ATT&CK [11] framework and open-source penetration testing tools within Kali Linux [64] for emulating Tactics, Techniques, and Procedures (TTP) employed by the real-world APT groups.

The five-step cyber-physical attack chain in Fig. 2 was performed using Kali Linux [64] on our Digital Twin-based testbed, over a time period of 2 hours and 30 minutes. Out of the five steps shown in Fig. 2, **Steps 1, 2, and 4** involved activities confined to hosts and TCP/IP network layers between hosts, without affecting actuators states. In contrast, **Steps 3 and 5** directly manipulated actuators in the pick-and-place process, and were expected to exhibit anomalous actuator behaviours detectable in physical and network data.

To identify potentially disruptive risk events in the pick-and-place process, a what-if analysis was conducted to assess the physical impact and its severity. The risk assessment involved changing the states of sensors and actuators individually to visualise its physical effects in the Digital Twin. Based on their potential to cause operational disruption, four threat scenarios were identified, as shown in Fig. 3, involving actuator-level manipulation of either the conveyor belts or the robot gripper. This evaluation guided the development of supervised models for classifying physical anomalies.

Sensor-level attack scenarios were excluded following the what-if analysis showed that persistent false sensor injection is infeasible in the current pick-and-place process. Any changed sensor value is automatically restored by the PLC using real-time feedback from the Digital Twin. Therefore, our study focused on actuator-level manipulation, including subtle changes in conveyor belt speed, suction, rotation angle, and the Z-axis speed of the robot gripper.

3.5. Data collection

To enable a holistic view of multi-stage cyber-physical attack scenarios, data collection was systematically conducted across three aspects from the Digital Twin testbed: process, network, and host.

3.5.1. Collecting process data

During the simulation of the pick-and-place process in Digital Twin, sensors and actuators' state values were extracted using the *Export* function within *Runtime Inspector* at a sampling rate of 0.03 seconds. Each export file produces a time-series dataset (csv files), with each row representing a timepoint.

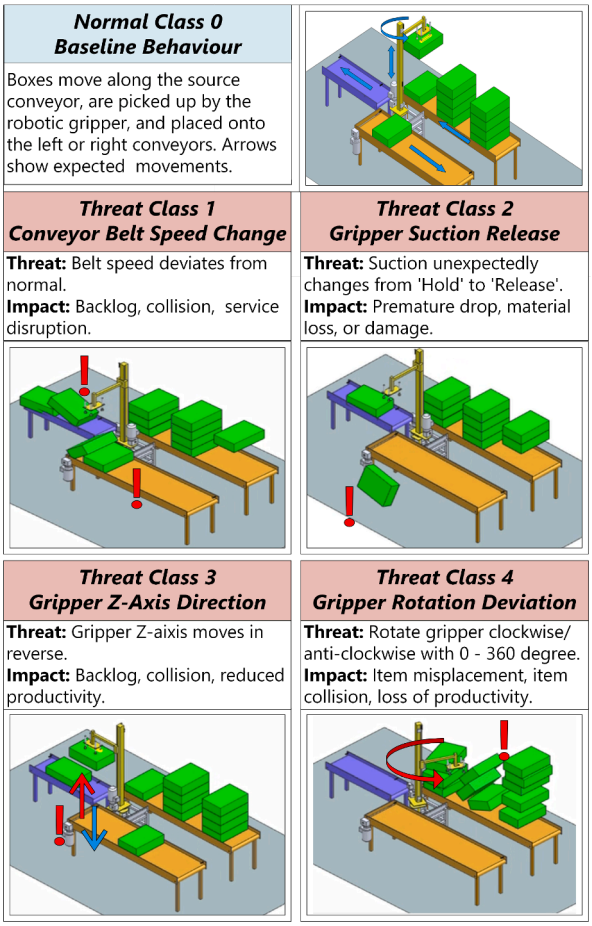

Fig. 3. Normal and threat scenarios simulated in digital twin.

Two categories of process data were collected - one for model development in Section 4.1.1 and the other for first-level decision fusion in Section 4.2.

1. Data for Model Development: For creating a balanced time-series dataset suitable for supervised model development, each normal and threat class mentioned in Fig. 3 were equally simulated ten times for more than one minute. Sixty csv files were generated in total.
2. Data For Model Inference: During the simulation of the five-step attack chain described in Fig. 2, the same Export function in NX MCD was used to extract process data after performing each step, spanning five csv files. Although immediate export was made after each step during the simulation period, missing time intervals were observed between files, likely due to NX MCD's memory constraints. The problem was resolved by reconstructing the timestamps by using the file creation time (i.e. export completion time) as factual reference points. The corresponding start time of the simulation were deduced based on the number of rows in each csv file and known sampling rate (0.03 seconds). After reconstructing the timestamps across the five csv files, it was noted that 50 of 148 minutes were missed out during the simulation of the multi-stage attack. Therefore, the remaining 98 minutes were used for model inference. Nevertheless, the missing data issue was handled by the confidence-aware decision logic in Section 4.3.

3.5.2. Collecting network data

Network traffic was captured using Wireshark [62] during the simulation. Similar to collection of process data, two categories of network data were collected - one for model development in Section 4.1.2 and one for model inference in Section 4.2.

1. Data for Model Development: Each scenario was run and recorded as a standalone simulation lasting over 60 seconds. These simulations produced 60 pcap files for unsupervised model development. Because each file was explicitly started and stopped after the full simulation had completed, no data loss or timestamp misalignment was observed.
2. Data for Model Inference: Network traffic related to OPC UA network communications was continuously captured throughout the full five-step attack simulation using Wireshark [62], with no file breakage or missing records.

3.5.3. Collecting data in hosts

Mimicking industrial practice for operational and compliance needs, process alarms were configured in Ignition gateway [59] on Engineering Workstation. Alarms were triggered when actuator states became intolerable (e.g. excessive conveyor speed), displayed on Operator's HMI, and automatically stored in PostgreSQL database [19]. During multi-step attack simulation, alarms were triggered and then suppressed in Step 3. Alarm records were exported as csv files using SQL queries to facilitate first-level decision fusion.

Additionally, for second-level decision fusion, host-level telemetry were captured using native Windows event audit logging features, alongside System Monitor (Sysmon) [43]. Specific advanced audit policies were enabled to capture command-line activity, PowerShell script block execution, user activities, process activities, and network connections.

To facilitate centralised logging and analysis of events, Wazuh agents were installed on Operator Desk and Engineering Workstation, functioning as a host-based IDS. The Wazuh agents forward Windows events to Wazuh SIEM hosted on Ubuntu operating system for analysing using Wazuh's community-driven detection rules. Event data was extracted from the Threat Hunting module of the Wazuh dashboard. Specific data filters were applied to extract metadata, such as MITRE ATT&CK tactics and techniques, to enhance the explainability of the second-level fused results.

4. Detection methodology

Building upon the data collected through the Digital Twin-based testbed, this section describes the methodology for detecting and fusing anomalies across process, network, and host layers.

4.1. Model development

Different models were used to identify anomalies in process, network, and hosts, considering the data nature and complexity for recognising anomalous patterns. The end-to-end decision fusion - including data preprocessing, feature extraction, detection model development and visualisation - was primarily conducted using Python [65] as separate scripts and Jupyter Notebooks [66] for reproducibility and verification.

4.1.1. Process data - supervised deep learning

To detect abnormal behaviours in physical processes, a supervised deep learning time series classifier was developed using the *LSTM-FCN* architecture from the *sktime* library [67]. LSTM-FCN was selected as it outperformed CNN, InceptionTime, MCDCNN, and ResNet in our previous validation work [49]. The model classified five process states (1 normal and 4 attack scenarios), as shown in Fig. 3.

Data preprocessing was conducted prior to model development. All 60 csv files collected as discussed in Section 3.5.1 underwent systematic pre-processing to ensure structural consistency, completeness, and data integrity. This included checks for column headers, missing values, and uniform row counts. The time-series datasets were standardized to 2000 timepoints per instance, with consistent formatting of instance identifiers, timestamps, and labels. All files were then merged into a unified dataset in one csv file.

For feature extraction, a custom Python script was developed and used on the merged csv file to extract five domain-specific features. These features reflected abnormal discrete values and irregular temporal actuator behaviours. Anomaly flags were created to support the model development, including flags to indicate abnormal conveyor belt speed, abnormal gripper rotation, unexpected gripper Z-axis directional flips. Raw signals related to gripper suction, including MCD_GetBoxDone and GetBox_signal, were also used for model training.

For model training and validation, a 70:30 train-test split was applied to the first 50 labelled instances (label 0 with no attacks, and labels 1 to 4 with single-vector attacks). The model training achieved 73% accuracy during validation. Evaluation on the withheld unseen instances (label 5 with multi-vector attacks) resulted in 100% classification accuracy. The trained model was subsequently used for inference in Section 5.2.

The supervised classifier outputs probabilities (value from 0% to 100%) and class labels (label 0 for normal case, and label 1 to 4 for threat cases). In first-level decision fusion, the class probabilities serves as the process anomaly scores. Instances with probabilities below 0.2 are considered confidently normal, above 0.8 as confidently anomalous, and values in between are treated as marginal. These thresholds were empirically chosen to convert the 5-class outputs into a standardised anomaly score for fusion. This margin is adjustable depending on the tolerance level for uncertainty - a wider margin (e.g., 0.1 - 0.9) enforces stricter confidence filtering.

4.1.2. Network data - unsupervised machine learning

To detect abnormal write command requests via OPC UA clients, an unsupervised anomaly detection model was developed to analyse OPC UA network traffic using *IsolationForest* model from *scikit-learn* library [68]. The detection objective was to develop a baseline of normal behaviours, such that any deviations from the baseline would be distinguished as anomalies.

Isolation Forest was selected for its robustness in detecting rare and subtle anomalies in unlabelled OPC UA network traffic. It natively generates anomaly scores based on how easily rare cases can be isolated. This unsupervised model complements supervised learning on process data, helping to reduce the risk of missed detections from unknown sources issuing unauthorised write commands that impact the physical process.

Data preprocessing was conducted prior to model development. All 60 .pcapng files collected as discussed in Section 3.5.2 underwent systematic pre-processing to convert and structure for consistency across instances. Raw .pcapng files were first converted into Zeek with icsnpp-opcua-binary parser [17] to extract metadata related to OPC UA protocol. Two specific Zeek logs (opcua_binary.log and opcua_binary_write.log) were correlated and parsed into 60 structured csv files.

For feature extraction, semantic features were extracted to support anomaly detection. Features included message size anomalies, write request frequency, and write operation ratios. Additional metadata were extracted to support the detection of abnormal WriteRequest operations. For instance, source ports, indicator of message originator, and the identifier 673 (which denotes a WriteRequest in OPC UA protocol specification). The data were merged into a single csv file and segmented to fixed 1-minute intervals using the last 60 seconds of the packet capture.

For model training and validation, the unlabelled instances that represented normal operations (instance 1 to 10) were used for pattern recognition. The remaining instances (instance 11 to 60 - include only threat scenarios) were used for model evaluation. Contamination parameters for *Isolation Forest* were empirically tuned to minimize false positives while retaining precision. The model achieved 93% accuracy and 96% for F1-score (attack) during validation.

The anomaly score output by the model is a range from -1 to 1, where any values below 0 indicate anomalies, whereas values above 0 indicates normal behaviour. In first-level decision fusion, the instance is regarded as marginal case when its network anomaly score falls within the range from -0.01 and 0.01.

4.1.3. Host data - point-based binary scoring

A point-based binary scoring system was developed to quantify the number of anomalous host activities occurring in each one-minute interval. The detection objective is to provide anomaly scores with interpretability, leveraging Wazuh's native detection capabilities, to explain the cyber situations for better incident response.

To detect LOTL techniques, 23 features were extracted from Wazuh for scoring. Each feature corresponds to a specific behavioural indicator, including suspicious connection or logon, suspicious command line usage or scripts, lateral movement and file transfer.

Lower-confidence indicators were excluded to reduce false positives, as their legitimacy is difficult to verify without additional context, including generic account logon or process creation events. Yet, the features are extendible to provide customisable security analytics to suit the needs of different operating requirements.

To support interpretable frequency-based analysis of LOTL activity across time windows, a binary anomaly score of 1 was assigned when a feature has been observed. Scores were then aggregated within each 1-minute interval to generate a composite anomaly score for each minute, representing the frequency and intensity of potentially malicious host activity. The scores were used in the second-level decision fusion (refer to Section 3.2 for aggregation rationale).

4.2. Identification of cyber-physical threats with first-level decision fusion

To generate a unified view of OT-level anomalies within synchronised time intervals, the process anomalies are correlated with network anomalies (specially focusing on OPC UA anomalous write commands) to assist operators in determining whether observed process deviations on CPS are indicative of cyber-attacks.

The effectiveness of the detection relies on the assumption that changes on the physical processes via manual intervention are less frequent in automated industrial processes. If such changes are needed, operators usually trigger such change via the use of legitimate user interface, such as HMI and diagnostics tools.

In such cases, a corresponding write command request would be issued from the OPC UA client (such as HMI and diagnostic tools) to OPC UA server (such as PLC), a corresponding write command request could be found on the network traffic.

When process anomalies were observed by operators without corresponding network anomaly, it is likely that the process anomalies were not induced from write commands initiated from OPC UA client, but it was resulted from situations that did not initiate OPC UA write commands, such as wear and tear, equipment degradation or other mechanical or electric issues.

Furthermore, other than the predictions by the machine learning or deep learning models, process alarms are integrated into the first-level decision fusion as additional evidence to substantiate the existence of the cyber-physical threats. When the process alarm coincides with detected anomalies in either the physical process or network layer in CPS, it reinforces the likelihood of a true cyber-physical attack.

4.3. Confidence-aware decision logic for first-Level decision fusion

A confidence-aware decision logic was developed to handle potential issues during aggregation and alignment process, as discussed in Section 3.2. As shown in Fig. 4, the availability of process and network data is checked at the start of the evaluation. When process data is missing, network data becomes the primary signal for decision-making. When network data is unavailable, decisions default to process-level indicators and the process alarm records. This design ensures maximum coverage and interpretability, even under incomplete data conditions. When both process and network outputs are available, a confidence-aware decision logic is applied to align predictions and resolve conflicts.

In this case, the workflow considers two possible outcomes - aligned predictions, or conflicting predictions. When outputs of both models

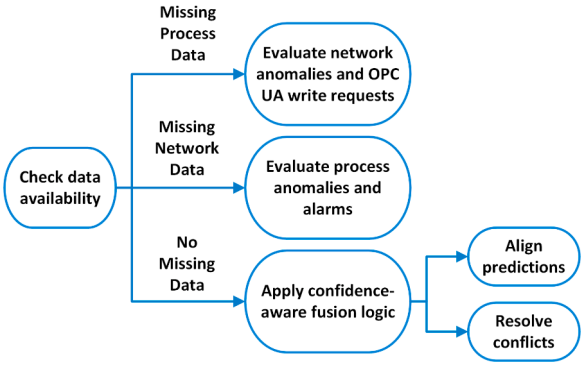

Fig. 4. Activity diagram for first-level decision fusion.

agree, the result is adopted directly. When predictions diverge, conflicts are resolved based on confidence scores. Specifically, when the process model predicts 'normal' and the network model indicates 'threat', the process model's threat class probability is compared against the network anomaly score to determine the final label.

However, when the process model predicts 'threat' and the network model indicates 'normal', the process model's prediction is prioritised - in part due to its supervised nature and higher precision when detecting unseen threat cases, as demonstrated in the model evaluation in Sections 4.1.1 and 4.1.2.

4.4. Second-level decision fusion for multi-stage APT attack detection

The second-level decision fusion correlates host-based anomalies with the OT-level fused results to identify stealthy, multi-stage APT attacks, which typically leverage compromised machines in IT domain as a pivot point, and then propagate into OT domain for physical process manipulation and disruption.

Correlating multimodal anomalies across IT and OT domains facilitates earlier detection of multi-stage APT attack. Host-based IDS generally lacks visibility into OT process anomalies. Therefore, anomalies detected in hosts - such as suspicious abuse of PowerShell-based reconnaissance and other IT-centric anomalous behaviours - provide Indicators of Attacks (IOA) and Indicators of Compromise (IOC) as upstream evidence of adversary presence.

Without periodic cross-domain correlation, a complete understanding of attack progression across IT/OT boundary remains fragmented. Therefore, under the assumption that host anomalies do not contradict OT anomalies, host anomalies are merged with first-level OT fused results. This provides a more comprehensive view of threat activity, enhancing the visibility across IT/OT domains and reducing the effects of data silo.

5. Experimental result

The effectiveness of the two-level decision fusion approach on detecting multi-stage LOTL attacks in CPS was assessed using the data collected during the simulation of the five-step cyber-physical attack chain shown in Fig. 2. **Steps 1, 2, and 4** were **threats in IT domain** such as PowerShell-based execution and lateral movement, whereas **Steps 3 and 5** were **threats in OT domain** related to unauthorised OPC UA write commands that impacted the physical process.

5.1. Result of network anomaly detection model

The performance of the unsupervised model is summarised in Table 1. The model is more sensitive but less precise, achieving 45% precision and 65% recall for attack cases. Many normal instances were misclassified as attacks, leading to an overall 76% accuracy.

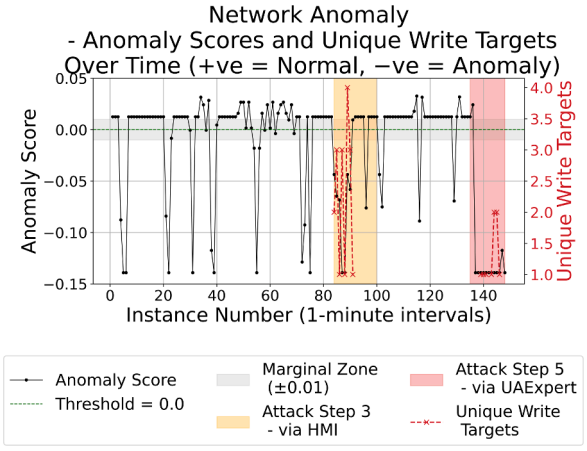

Fig. 5. Network anomaly scores and write commands count.

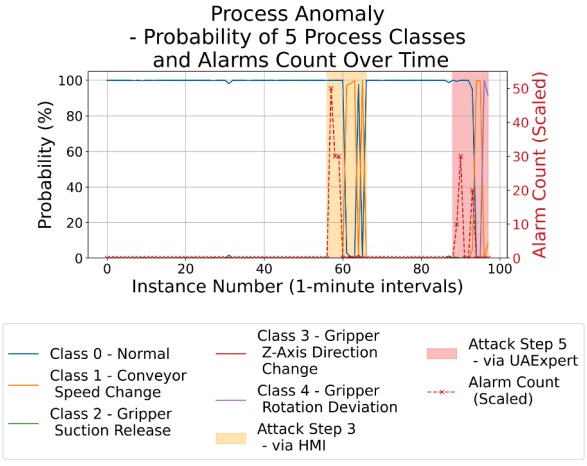

Fig. 6. Process anomaly scores and alarm count.

Fig. 5 shows network anomaly scores over 148 one-minute intervals. The left Y-axis represents the network anomaly score, where values below 0 indicate anomalies, and values above 0 suggest normal behaviour. The right Y-axis tracks the number of unique write targets observed per minute.

The orange-shaded region denotes Attack Step 3 (manipulation via HMI), the anomaly scores largely remain above 0 (normal) with a few anomaly scores with values below 0 (anomaly). This indicates that the model struggled to identify subtle changes originating from the HMI. In contrast, the red-shaded region denotes Attack Step 5 (manipulation via diagnostics tool), shows a cluster of negative anomaly scores, clearly detecting command injection via OPC UA diagnostics tool. Non-shaded regions do not have adversarial activity over the OPC UA network. Sporadic false positives suggest occasional misclassification causing false alarms, while twelve borderline cases (8%) fall into marginal zone (-0.01 and +0.01) in grey-shaded area.

5.2. Result of process anomaly detection model

The performance of the supervised model is summarised in Table 1. It was highly conservative, achieving 100% precision and 38% recall for attack cases. It only predicts the attack cases when highly confident, but misses 62% of attacks. Overall accuracy was 87%.

Fig. 6 populates the model inference results for the 98 instances. The coloured lines (blue, orange, green, red, and purple) represent the predicted probabilities for five classes. The right Y-axis shows the number of alarm triggered, scaled by a factor of 10 for visual clarity.

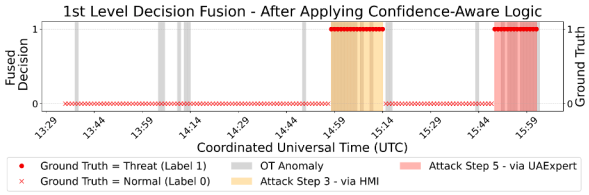

Fig. 7. Result of first-level decision fusion.

Table 1
Performance of individual modalities and fused decision.

Metric	Network Only	Process Only	1st-Level Fusion
True Normal (TP)	93	77	107
False Positive (FP)	24	0	10
False Negative (FN)	11	13	11
True Attack (TP)	20	8	20
Precision (Attack)	45%	**100%**	67%
Recall (Attack)	**65%**	38%	**65%**
Accuracy	76%	**87%**	86%
F1-score (Attack)	53%	55%	**66%**
Macro F1-score	69%	74%	**78%**

The probabilities of the four threat classes are close to zero for most instances, causing their lines to appear visually suppressed beneath the dominant normal class (Class 0) curve. No borderline cases with probabilities below 80% for the four threat cases. Notably, the model not only flagged the anomalies but also provided interpretable class labels, such as conveyor speed manipulation (orange line), enhancing situational awareness.

5.3. Result of first-level decision fusion

The confidence-aware logic discussed in Section 4.3 is applied on the merged network and process data to handle three critical cases. Firstly, to resolve the uncertainty due to missing process data mentioned in Section 4.1.1. Secondly, to resolve the borderline cases in network data. Lastly, to resolve conflicting predictions between network and process models. Fig. 7 shows the result of first-level decision fusion. Ground truth labels are shown as red circles and crosses, while light red bars denote fused decisions. The fused decisions effectively covered majority of the Attack Steps 3 and 5.

To evaluate the impact of first-level OT fusion, we compared detection performance across network-only, process-only, and fused detection models altogether in Table 1. The result shows that the first-level decision fusion effectively balances precision (67%) and recall (65%) for attack cases, reducing false alarms while improving detection compared to individual modalities.

The outcome of the first-level decision fusion highlights the complementary nature of the models as different detection models exhibit varying performance characteristics. The network model is sensitive but noisy, while the process model is conservative but precise. The confidence-aware decision logic achieves a balanced detection strategy, improving overall F1-score by 11% or 13% over individual models, showcasing it is beneficial to fuse the predictions. Although the fused result slightly reduced accuracy (from 87% to 86%), this trade-off is justified in detecting rare but potentially devastating incidents. As such, recall becomes a more critical metric than precision - failing to detect an attack poses a greater risk than tolerating occasional false positives for early detection of stealthy attack behaviours.

5.4. Result of host anomaly detection

Point-based binary scoring is used to calculate the anomaly scores on the event logs generated from Wazuh SIEM [18]. As shown on Fig. 8,

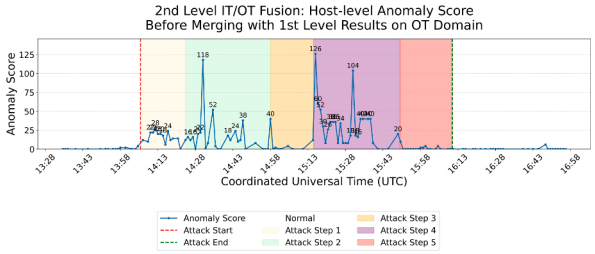

Fig. 8. Host anomaly score.

Table 2
Detection delays reduced by second-level fusion.

Attack Step	Step Start	1st-Level Fusion	2nd-Level Fusion	Delays Reduced
1	14:06	**14:10**	**14:07**	3 Min
2	14:24	**14:49**	**14:24**	25 Min
3	14:58	14:59	14:59	Nil
4	15:15	15:16	15:16	Nil
5	15:49	15:51	15:50	Nil

significant anomaly scores are concentrated almost entirely within the defined attack window, from 14:00 to 16:00 UTC time.

Anomaly scores in normal operation stage (non-coloured areas) are near zero. Then, emerging anomaly scores are observed in Attack Step 1 (light yellow). Spikes of high anomaly scores are observed in Step 2 (light green) and Step 4 (purple), demonstrating the current scoring mechanisms effectively detect IT-centric threats related to LOTL techniques.

5.5. Result of second-level decision fusion

Referring back to the multi-stage attack chain in Fig. 2, the following analysis discusses how each stage was simulated and detected. Fig. 9 overlays host anomaly scores with first-level OT fusion results, incorporating process alarms, OPC UA write requests, and OT anomalies. Table 2 shows that second-level fusion detects earlier than first-level fusion, reducing delays by 3 minutes (Step 1) and 25 minutes (Step 2), which is critical for detecting cyber-initiated physical threats.

Attack Step 1 (Initial Access to Operator Desk) is mainly dominated by host anomaly scoring, reflecting stealthy IT-centric attacks. Indeed, this step simulates attackers leveraging Windows-native Server Message Block (SMB) remote services (T0886) via CrackMapExec and smbmap on Kali Linux [64]. The fused anomaly log highlights abuse of NTLMSSP authentication service.

Attack Step 2 (Discovery and Execution) triggers elevated host anomaly scores despite no observations in OT anomalies. The spike reflected unsuccessful attempts to download LOLBins (e.g., procdump.exe) using Invoke-WebRequest cmdlet. Except for this spike, the point-based binary scoring was able to detect stealthy information extraction from Operator Desk to Kali Linux, involving PowerShell, command-line scripts (T0807) and Impacket (S0357).

Attack Step 3 (Manipulation of Physical Process via HMI) shows overlapping spikes in host anomaly scores, process alarm counts, and OT anomalies, reinforcing this step as a critical crossover point. A spike on host anomaly score was observed at the start of Attack Step 3, which was related to remote login using xfreerdp3 from Kali Linux. Upon successful remote access to Operator Desk, subsequent manipulations via HMI were detected through OT anomalies. The spikes on threshold-based process alarms and OPC UA write requests confirmed safety issues and cyber-physical threats, even when process alarms were immediately suppressed during simulation.

Attack Step 4 (Lateral Movement to Engineering Workstation) again shows elevated host anomalies, with no OT anomalies observed. This

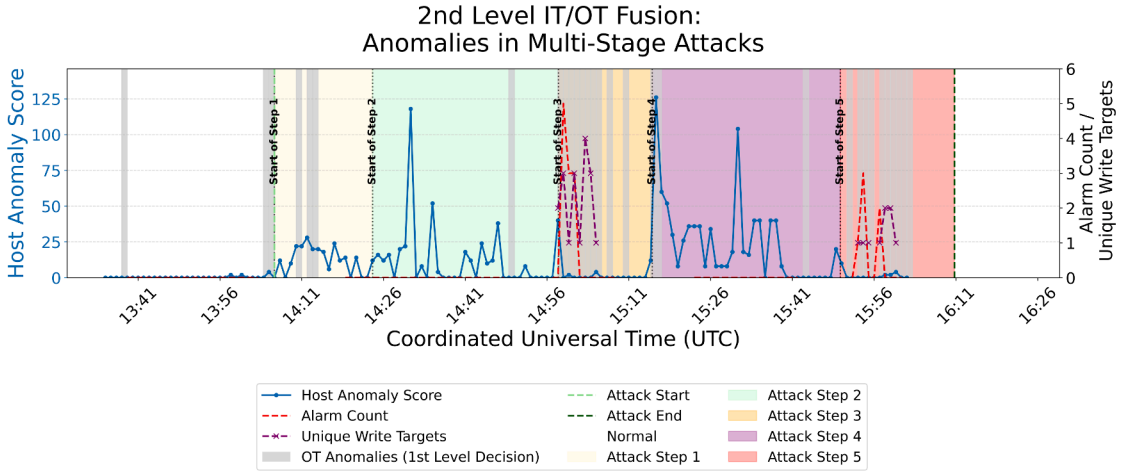

Fig. 9. Result of second-level decision fusion.

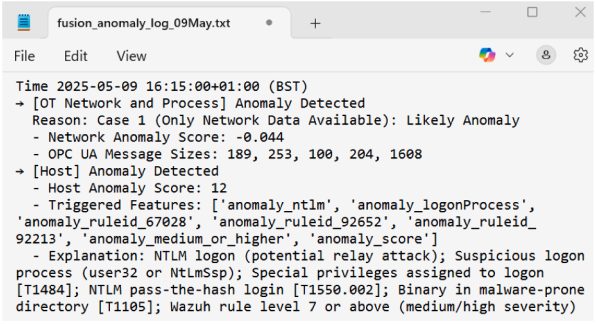

Fig. 10. One record extracted from the fused anomaly log.

aligned with the expectation that the attack simulated during the period were mostly related to reconnaissance and exfiltration of project files in Engineering Workstation using PowerShell on compromised Operator Desk.

Attack Step 5 (Manipulation of Physical Process via Diagnostics Tool) displayed multiple indicators, including multiple instances of OT anomalies, increased write commands and process alarms. The result confirmed remote manipulation using OPC UA diagnostics tools was done to read tags and issue write commands, resulting in observable disruptions to physical processes.

Crucially, no significant false positives in normal time windows (before 14:00), highlighting the precision of the fusion logic. The temporal aggregation and alignment strategy in the two-level decision fusion approach showed not only an understanding of **what** occurred, but also **when** and **how** each attack stage progressed from IT-based attacks to physical manipulation.

Overall, this approach support early detection, enables contextual linking of adversarial activities across IT/OT domains, and enhances situational awareness of multi-stage threats, supported by data visualisation and cyber-physical threat summary. The outcome of the experimentation reinforces the value of multimodal correlation in detecting stealthy, multi-stage cyber-physical attacks that leverages LOTL techniques.

To aid interpretation, a fused anomaly log integrates detections into unified, time-aligned records with interpretable justifications. A representative case is shown in Fig. 10: host anomaly score of 12 reflects NTLM relay abuse and privilege escalation, while the OT anomaly was detected via network score of -0.044. Associated MITRE ATT&CK [11] techniques such as T1550.002 (Authentication by Pass the Hash) and T1105 (Ingress Tool Transfer) are shown.

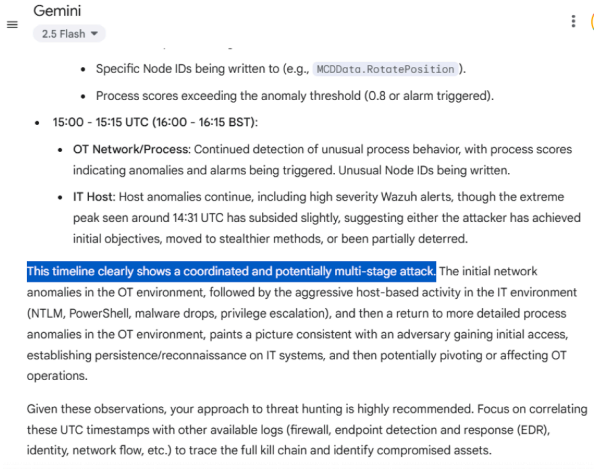

Fig. 11. LLM-generated text related to Fig. 10.

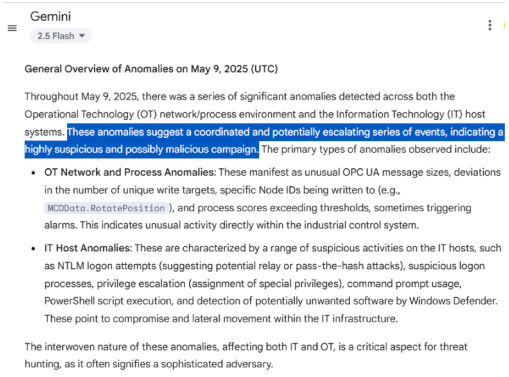

Fig. 12. LLM-generated summary for entire anomaly log.

To further improve interpretability and enhance cyber situational awareness, we experimented with a publicly available Large Language Model (LLM), Gemini 2.5 Flash [69], to automatically summarize fused anomaly logs. Fig. 11 highlighted **what** and **when** happened across IT/OT domains.

Furthermore, Fig. 12 shows an LLM-based threat summary. Scheduling periodic reports (e.g., hourly or daily) from the LLM could support continuous monitoring across domains, giving analysts situational awareness to identify stealthy LOTL campaigns that single-alert triage

might miss, and this approach may warrant future validation in operational settings.

6. Discussion

While it is possible to detect APT attack behaviours using traditional security monitoring approaches - such as host-based IDS, network traffic analysis, or rule-based event correlation - these methods often siloed and lack visibility into the physical impact of cyber attacks. In contrast, a Digital Twin-based testbed enables safe, repeatable simulation of IT/OT threats and facilitated evaluation of detection capability, i.e. multimodal anomaly correlation from heterogeneous data sources demonstrated in this study. The testbed added significant value by enabling validation of our two-level decision fusion approach, supporting proactive threat hunting and detection engineering without impacting physical operations in real world.

6.1. Responding to research questions (RQ)

This study hypothesised that correlating physical anomalies (captured using a Digital Twin) with network and host anomalies would facilitate early detection of stealthy, multi-stage cyber-physical attacks. The experimental results provide empirical support for this hypothesis. In light of our findings, we address the research questions stated in Section 1.1 as follows:

RQ1: Integrating commercial IT/OT software and open-source security tools enabled the creation of a realistic yet safe simulation environment, allowing for safe and repeatable simulation and detection of multi-stage LOTL techniques that cannot be tested in production environments due to safety and availability concerns.

RQ2: The one-minute temporal aggregation and alignment strategy proved effective in correlating OT network and process anomalies. While neither modality alone provided sufficient visibility of the attack progression, the confidence-aware decision logic in the first-level fusion successfully balanced precision and recall for attack cases, resulting in reduced false positives.

RQ3: The host anomaly scoring method effectively captured stealthy LOTL activity at a per-minute resolution. By correlating host and OT anomalies, a unified and interpretable anomaly log was generated. Additionally, LLM enabled the summarization of anomaly logs into human-readable insights, enhancing analyst interpretability and improving situational awareness.

Overall, our approach unifies host, network, and process data via time-synchronised decision fusion, offering a holistic, interpretable view of multi-stage APT activity that would otherwise remain fragmented.

6.2. Strengths of our approach

Holistic Visibility Across IT/OT Domains: The two-level decision fusion provides unified anomaly view across cyber-physical layers. The first-level fusion, which focused on process and OT network anomalies, has limited visibility over host-based intrusion such as malicious user activities (e.g. pass-the-hash). As shown in Table 2, the second-level fusion compensates for this gap by reducing detection delays for initial access and discovery steps, whereas later attack steps were sufficiently captured by first-level fusion. The one-minute aggregation amplifies stealthy, staged patterns, and helps trace attack progression across modalities, reducing the risk of misinterpreting isolated signals.

Explainability and Interpretability: The models for OT anomalies provides contextual anomaly scores, while the scoring method for host anomalies, leveraging existing open-source host-based IDS, maps detections to MITRE ATT&CK techniques. This approach avoids model training complexity while yielding actionable insights, supporting human-in-the-loop decision-making.

Flexibility and Extensibility: The host anomaly scoring system is extensible to include data from NIDS (e.g., Suricata [37]) or host telemetry (e.g., CPU, memory usage). Likewise, the confidence margins in the first-level fusion can be tuned based on the organisational risk appetite. This enables adaptable detection strategies tailored to different risk profiles.

6.3. Limitations and opportunities

Visibility of Short-Lived or Transient Anomalies: As outlined in Section 3.2, the one-minute aggregation strategy may reduce visibility into brief attacks. However, it was intentionally selected for balancing detection efficacy and deployment simplicity. Importantly, it remains suitable for detecting LOTL attacks, which tend to unfold gradually rather than brief bursts.

Limited Root Cause Analysis: The system highlights behavioural anomalies and indicators of staged attacks, but does not infer root causes. Future work will explore mapping detection results to MITRE ATT&CK mitigations [11] and incorporating this into a dashboard for proactive threat hunting and investigation support. Additionally, we aim to explore the critical question of *'When is the optimal time to alert?'* in multi-stage attack scenarios, with a focus on triggering alerts based on inferred root causes rather than isolated symptoms.

Scalability: The current implementation, focused on a pick-and-place process, run on a Windows 11 workstation (Intel Core i5 @ 2.6 GHz, 16 GB RAM) for 2.5 hours, with a Tesla T4 GPU for model training and inference. Results suggest that initial experimentation is feasible on a standard workstation running several virtual machines. The architecture based on Siemens NX MCD [57] and PLCSim Advanced [58] also supports hardware-in-the-loop configurations and can extend to multi-Digital Twin setups or physical assets via OPC UA and other industrial protocols. Future work will assess resource needs in large-scale industrial settings.

Generalisability: While the fusion strategy is demonstrated through a single detailed case study as a proof of concept, our work is deliberately designed as a technology- and model-agnostic framework that provides that a flexible foundation for incorporating other cross-domain LOTL attack techniques or physical processes. More generalisable studies can be built upon this detection framework without altering the core architecture, enabling detection across larger datasets or multi-process environments, provided thta the underlying detectors (e.g., for process, network, or host data) are appropriately trained and configured, and that outputs from multiple processes can be aligned to a common time anchor. While the framework is theoretically generalisable, further empirical validation through additional case studies is necessary to substantiate this claim and this lies beyond the scope of the current work.

Dataset Benchmarking Limitation: While our dataset benefits from high fidelity and repeatability, and captured a representative multi-stage LOTL attack chain, it is constrained to a single pick-and-place process, limiting generalisability. A further limitation is the loss of 50 minutes of process data during simulation due to Siemens NX MCD memory constraints. These gaps were reconstructed using timestamps (see Section 3.5.1) and mitigated through confidence-aware fusion logic (see Section 4.3). Another limitation is the lack of publicly available multimodal datasets combining IT and OT evidence - host logs, OPC UA network traffic, and process signals - for multi-stage APT detection. As such, direct benchmarking against datasets such as Secure Water Treatment (SWaT) dataset [20] is infeasible. Instead, we established internal baselines for evaluating detection performance across individual and fused modalities. To enhance realism and reproducibility, the testbed incorporated commercial-grade ICS applications, OPC UA, and emulated APT scenarios using security tools in Kali Linux [64]. For transparency, our dataset and code are released on GitHub.

Reliance on Expert-Defined Fusion Rules: The fusion logic is knowledge-driven, offering interpretability but reduced adaptability. Future work could explore supervised learning methods (e.g., decision tree-based fusion) to automate correlation rule learning once labelled fusion datasets are available.

7. Conclusion and future work

This study presented a two-level decision fusion approach using a Digital Twin-based testbed to detect stealthy, multi-stage cyber-physical attacks leveraging Living off the Land (LOTL) techniques in industrial CPS and ICS. To provide comprehensive, realistic experimentation, we integrated commercial-grade ICS software, OPC UA protocol, and security monitoring tools to simulate a five-step APT-style attack chain.

Experimental results shows that the first-level fusion strategy enhances OT-level visibility by combining supervised deep learning predictions on process data, unsupervised anomaly detection on OPC UA network traffic, and process alarm signals. The second-level fusion enhances situational awareness by integrating host-based anomaly scores derived from Wazuh alerts. Together, this layered fusion approach enables a holistic and interpretable view of cyber-physical threat progression, improving detection performance than models using single modality alone. Overall, we show multimodal anomaly correlation significantly improves early threat detection across IT and OT domains.

Future work will focus on extending this approach to real-world environments, through hardware-in-the-loop integration and validating the fusion strategy across diverse industrial settings. We also plan to explore automated decision fusion logic using supervised models, once sufficient labelled data are available. In addition, we aim to map fused alerts to MITRE ATT&CK mitigations and use an LLM-based playbook to guide analysts on both immediate responses (e.g., threat investigation) and longer-term root-cause actions, such as reviewing user account management on affected devices (e.g. ATT&CK mitigations M1018), helping to prevent recurrence and reduce reactive responses. Finally, we will expand scenario diversity - such as incorporating Linux-based LOTL attacks and Command-and-Control (C2) infrastructure - and collaborate with the research community to establish benchmarking standards for cyber-physical attack detection datasets.

CRediT authorship contribution statement

Carol Lo: Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization; **Thu Yein Win:** Writing – review & editing, Supervision, Conceptualization; **Zeinab Rezaeifar:** Writing – review & editing, Supervision, Conceptualization; **Zaheer Khan:** Writing – review & editing, Supervision, Conceptualization; **Phil Legg:** Writing – review & editing, Supervision, Conceptualization.

Data availability

Dataset and code used in this study are available at https://github.com/carolsworld/LOTL-Hunter.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded through PhD studentship scheme at the University of the West of England.

References

[1] D. Ding, Q.-L. Han, X. Ge, J. Wang, Secure state estimation and control of cyber-physical systems: a survey, IEEE Transact. Syst. Man Cybernet.: Syst. 51 (1) (2021) 176–190. https://doi.org/10.1109/TSMC.2020.3041121.

[2] G. Hulme, The Purdue Model's Risky Blindspot, (Nexus Claroty). (2023) https://nexusconnect.io/articles/the-purdue-models-risky-blindspot, 2025 (accessed 28 April 2025).

[3] A. Volkova, M. Niedermeier, R. Basmadjian, H. de Meer, Security challenges in control network protocols: a survey, IEEE Commun. Surv. Tutor. 21 (1) (2019) 619–639. https://doi.org/10.1109/COMST.2018.2872114.

[4] E.C. Balta, M. Pease, J. Moyne, K. Barton, D.M. Tilbury, Digital twin-based cyber-attack detection framework for cyber-physical manufacturing systems, IEEE Trans. Autom. Sci. Eng. 21 (2) (2024) 1695–1712. https://doi.org/10.1109/TASE.2023.3243147.

[5] A. Nelson, S. Rekhi, M. Souppaya, K. Scarfone, NIST SP 800-61 Rev. 3 Incident Response Recommendations and Considerations for Cybersecurity Risk Management: a CSF 2.0 Community Profile, 2025. https://doi.org/10.6028/NIST.SP.800-61r3.

[6] R. Canonico, G. Esposito, A. Navarro, S.P. Romano, G. SperlÃ¬, A. Vignali, Empowered cyber-physical systems security using both network and physical data, Comput. Secur. 152 (2025) 104382. https://doi.org/10.1016/j.cose.2025.104382.

[7] A.A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, S. Sastry, Attacks against process control systems: risk assessment, detection, and response, in: Proc. 6th ACM Symp. Inf. Comput. Commun. Secur. (ASIACCS), ACM, New York, USA, 2011, p. 355–366. https://doi.org/10.1145/1966913.1966959.

[8] J. Slowik, Evolution of ICS Attacks and the Prospects for Future Disruptive Events, 2019. Dragos, 2019, https://api.semanticscholar.org/CorpusID:201756877 (accessed 30 May 2025).

[9] B. Stojanović, K. Hofer-Schmitz, U. Kleb, APT datasets and attack modelling for automated detection methods: a review, Comput. Secur. 92 (2020) 101734. https://doi.org/10.1016/j.cose.2020.101734.

[10] N.I. Che Mat, N. Jamil, Y. Yusoff, M.L. Mat Kiah, A systematic literature review on advanced persistent threat behaviors and its detection strategy, J. Cybersecur. 10 (1) 2023 https://doi.org/10.1093/cybsec/tyad023.

[11] MITRE Corporation, MITRE ATT&CK, 2025, https://attack.mitre.org (accessed 30 May 2025).

[12] M. Conti, D. Donadel, F. Turrin, A survey on industrial control system testbeds and datasets for security research, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2248–2294. https://doi.org/10.1109/COMST.2021.3094360.

[13] M. Fujimoto, Y. Itani, T. Mitsunaga, Lessons learned for practical penetration test against industrial control systems, in: 2023 IEEE Intl. Conf. on Computing (ICOCO), 2023, pp. 59–64. https://doi.org/10.1109/ICOCO59262.2023.10397899.

[14] A. Bécue, E. Maia, L. Feeken, P. Borchers, I. Praça, A new concept of digital twin supporting optimization and resilience of factories of the future, Appl. Sci. 10 (13) . (2020) https://doi.org/10.3390/app10134482.

[15] M. Eckhart, A. Ekelhart, Towards security-aware virtual environments for digital twins, in: Proceedings of the 4th ACM Workshop on CPS Security, CPSS '18, ACM, New York, NY, USA, 2018, p. 61–72. https://doi.org/10.1145/3198458.3198464.

[16] F. Akbarian, E. Fitzgerald, M. Kihl, Intrusion detection in digital twins for industrial control systems, in: 2020 Intl. Conf. on Software, Telecom. and Computer Networks (SoftCOM), 2020, pp. 1–6. https://doi.org/10.23919/SoftCOM50211.2020.9238162.

[17] B.E. Alliance, Zeek OPC UA Binary Parser, 2023, https://github.com/cisagov/icsnpp-opcua-binary (accessed 30 May 2025).

[18] Wazuh, Wazuh SIEM, 2025, https://wazuh.com (accessed 30 May 2025).

[19] P.G.D. Group, PostgreSQL Database, 2025, www.postgresql.org (accessed 30 May 2025).

[20] A.P. Mathur, N.O. Tippenhauer, SWaT: a water treatment testbed for research and training on ICS security, in: 2016 Intl. Workshop on CPS for Smart Water Networks (CySWater), IEEE, 2016, pp. 31–36. https://doi.org/10.1109/CySWater.2016.7469060.

[21] C.M. Ahmed, V.R. Palleti, A.P. Mathur, WADI: a water distribution testbed for research in the design of secure cyber physical systems, in: Proceedings of the 3rd Intl. Workshop on CPS for Smart Water Networks, CySWATER '17, ACM, New York, NY, USA, 2017, p. 25–28. https://doi.org/10.1145/3055366.3055375.

[22] B. Green, A. Le, R. Antrobus, U. Roedig, D. Hutchison, A. Rashid, Pains, gains and PLCs: ten lessons from building an industrial control systems testbed for security research, in: Proceedings of the 10th USENIX Conference on Cyber Security Experimentation and Test, CSET'17, USENIX Association, USA, 2017. https://dl.acm.org/doi/10.5555/3241074.3241078.

[23] D.I. Urbina, J.A. Giraldo, A.A. Cardenas, N.O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, H. Sandberg, Limiting the impact of stealthy attacks on industrial control systems, in: Proceedings of the 2016 ACM SIGSAC Conf. on Computer and Comm. Security, CCS '16, ACM, New York, NY, USA, 2016, p. 1092–1105. https://doi.org/10.1145/2976749.2978388.

[24] S. Abaimov, J. Gardiner, E. Samanis, J. Williams, M. Samanis, F. Shahbi, A. Rashid, Capture the industrial flag: lessons from hosting an ICS cybersecurity exercise, in: Proceedings of the 10th ACM CPS Security Workshop, CPSS '24, ACM, New York, NY, USA, 2024, p. 98–106. https://doi.org/10.1145/3626205.3659148.

[25] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, M. Thompson, NIST SP 800-82 Rev. 3 Guide to Operational Technology (OT) Security, (2023). https://doi.org/10.6028/NIST.SP.800-82r3.

[26] W. Danilczyk, Y. Sun, H. He, ANGEL: an intelligent digital twin framework for microgrid security, in: 2019 North American Power Symposium (NAPS), 2019, pp. 1–6. https://doi.org/10.1109/NAPS46351.2019.9000371.

[27] A. Castellani, S. Schmitt, S. Squartini, Real-world anomaly detection by using digital twin systems and weakly supervised learning, IEEE Trans. Ind. Inf. 17 (7) (2021) 4733–4742. https://doi.org/10.1109/TII.2020.3019788.

[28] M. Eckhart, A. Ekelhart, E. Weippl, Enhancing cyber situational awareness for cyber-physical systems through digital twins, in: 2019 24th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA), 2019, pp. 1222–1225. https://doi.org/10.1109/ETFA.2019.8869197.

[29] M. Dietz, M. Vielberth, G. Pernul, Integrating digital twin security simulations in the security operations center, in: Proceedings of the 15th Intl. Conf. on Availability, Reliability and Security, ARES '20, ACM, New York, NY, USA, 2020, p. 9. https://doi.org/10.1145/3407023.3407039.

[30] P. Empl, G. Pernul, Digital-twin-based security analytics for the Internet of Things, Information 14 (2). (2023) https://doi.org/10.3390/info14020095.

[31] S. Liu, G. Peng, H. Zeng, J. Fu, A survey on the evolution of fileless attacks and detection techniques, Comput. Secur. 137 (2024) 103653. https://doi.org/10.1016/j.cose.2023.103653.

[32] R. Ning, W. Bu, J. Yang, S. Duan, A survey of detection methods research on living-off-the-land techniques, in: 2023 IEEE Intl. Conf. on Sensors, Electronics and Computer Engineering (ICSECE), 2023, pp. 159–164. https://doi.org/10.1109/ICSECE58870.2023.10263445.

[33] C. Alcaraz, J. Lopez, Digital twin-assisted anomaly detection for industrial scenarios, Intl. J. Criti. Infrastruct. Protect. 47 (2024) 100721. https://doi.org/10.1016/j.ijcip.2024.100721.

[34] M. Dahlmanns, J. Lohmöller, I.B. Fink, J. Pennekamp, K. Wehrle, M. Henze, Easing the conscience with OPC UA: an internet-wide study on insecure deployments, in: Proceedings of the ACM Internet Measurement Conference, IMC '20, ACM, New York, NY, USA, 2020, p. 101–110. https://doi.org/10.1145/3419394.3423666.

[35] M.A. Umer, K.N. Junejo, M.T. Jilani, A.P. Mathur, Machine learning for intrusion detection in industrial control systems: applications, challenges, and recommendations, Intl. J. Criti. Infrastruct. Protect. 38 (2022) 100516. https://doi.org/10.1016/j.ijcip.2022.100516.

[36] Cisco, Snort, 2025, www.snort.org.

[37] O.I.S. Foundation, Suricata, 2025, https://suricata.io.

[38] R. Mitchell, I.-R. Chen, A survey of intrusion detection techniques for cyber-physical systems, ACM Comput. Surv. 46 (4) (2014). https://doi.org/10.1145/2542049.

[39] J. Suaboot, A. Fahad, Z. Tari, J. Grundy, A.N. Mahmood, A. Almalawi, A.Y. Zomaya, K. Drira, A taxonomy of supervised learning for IDSs in SCADA environments, ACM Comput. Surv. 53 (2). (2020) https://doi.org/10.1145/3379499.

[40] S. Ghosh, S. Sampalli, A survey of security in SCADA networks: current issues and future challenges, IEEE Access 7 (2019) 135812–135831. https://doi.org/10.1109/ACCESS.2019.2926441.

[41] D. Garcia-Retuerta, R. Casado-Vara, J. Prieto, Enhanced cybersecurity in smart cities: integration methods of OPC UA and suricata, in: J.M. Corchado, S. Trabelsi (Eds.), Sustainable Smart Cities and Territories, Springer Intl Publishing, Cham, 2022, pp. 61–67. https://doi.org/10.1007/978-3-030-78901-5_6.

[42] O.P. Team, OSSEC, 2025, www.ossec.net.

[43] M. Russinovich, T. Garnier, Sysmon, 2025, https://download.sysinternals.com/files/Sysmon.zip.

[44] Y. Liu, P. Ning, M.K. Reiter, False data injection attacks against state estimation in electric power grids, in: Proceedings of the 16th ACM Con. on Computer and Comm. Security, CCS '09, ACM, New York, NY, USA, 2009, p. 21–32. https://doi.org/10.1145/1653662.1653666.

[45] 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 Comput. Surv. 51 (4). (2018) https://doi.org/10.1145/3203245.

[46] C. Lo, J. Christie, T.Y. Win, Z. Rezaeifar, Z. Khan, P. Legg, TRIST: towards a container-based ICS testbed for cyber threat simulation and anomaly detection, in: Proceedings of the Intl. Conf. on Cybersecurity, Situational Awareness and Social Media, Springer Nature Singapore, 2025, pp. 225–239. https://doi.org/10.1007/978-981-96-0401-2_13.

[47] C.M. Ahmed, M.R. Gauthama Raman, A.P. Mathur, Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems, in: Proceedings of the 6th ACM on CPS Security Workshop, CPSS '20, ACM, New York, NY, USA, 2020, p. 23–29. https://doi.org/10.1145/3384941.3409588.

[48] D. Umsonst, H. Sandberg, A game-theoretic approach for choosing a detector tuning under stealthy sensor data attacks, in: 2018 IEEE Conference on Decision and Control (CDC), 2018, pp. 5975–5981. https://doi.org/10.1109/CDC.2018.8619338.

[49] C. Lo, T.Y. Win, Z. Rezaeifar, Z. Khan, P. Legg, Digital twins of cyber physical systems in smart manufacturing for threat simulation and detection with deep learning for time series classification, in: 2024 29th Intl. Conf. on Automation and Computing (ICAC), 2024, pp. 1–6. https://doi.org/10.1109/ICAC61394.2024.10718749.

[50] L. Faramondi, F. Flammini, S. Guarino, R. Setola, Evaluating machine learning approaches for cyber and physical anomalies in SCADA systems, in: 2023 IEEE Intl. Conf. on Cyber Security and Resilience (CSR), 2023, pp. 412–417. https://doi.org/10.1109/CSR57506.2023.10224915.

[51] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477. https://doi.org/{https://doi.org/10.1109/JPROC.2015.2460697}.

[52] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.

[53] L. Li, X. Zhao, J. Fan, F. Liu, N. Liu, H. Zhao, A trust worthy security model for IIoT attacks on industrial robots, Futur. Generat. Comput. Syst., 153 (2024) 340–349. https://doi.org/10.1016/j.future.2023.11.027.

[54] A. Abid, F. Jemili, O. Korbaa, Real-time data fusion for intrusion detection in industrial control systems based on cloud computing and big data techniques, Clust. Comput. 27 (2) (2023) 2217–2238. https://doi.org/10.1007/s10586-023-04087-7.

[55] T.J. Williams, The Purdue enterprise reference architecture, Comput. Ind. 24 (2) (1994) 141–158. https://doi.org/10.1016/0166-3615(94)90017-5.

[56] UK DSIT, Digital Twin Definition, 2025, https://www.gov.uk/government/publications/digital-twin-definition. gov.uk (accessed 2025-11-27).

[57] Siemens, Siemens NX Mechatronic Concept Designer, 2025, https://plm.sw.siemens.com/en-US/nx/.

[58] Siemens, Siemens S7-PLCSIM Advanced, 2025, https://www.siemens.com/global/en/products/automation/systems/industrial/plc/s7-plcsim-advanced.html.

[59] I. Automation, Ignition SCADA/ HMI, 2025, https://inductiveautomation.com. (accessed 30 May 2025).

[60] Siemens, Siemens TIA Portal, 2025, https://www.siemens.com/global/en/products/automation/industry-software/automation-software/tia-portal.html.

[61] U. Automation, UA Expert, 2025, https://www.unified-automation.com/, (accessed 30 May 2025).

[62] Wireshark Foundation, Wireshark, 2025, www.wireshark.org.

[63] Department of Homeland Security, USA, Identifying and Mitigating Living Off the Land Techniques, 2024, https://www.cisa.gov/resources-tools/resources/identifying-and-mitigating-living-land-techniques.

[64] L. OffSec Services, Kali Linux, 2025, www.kali.org.

[65] Python Software Foundation, Python, 2025, www.python.org.

[66] T.J. Project, Jupyter Notebook, 2025, www.jupyter.org.

[67] M. Löning, sktime/sktime: v0.37.0, 2025, https://doi.org/10.5281/zenodo.15203127.

[68] F. Pedregosa, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[69] Google, Gemini 2.5 Flash, 2025, https://gemini.google.com. (accessed 30 May 2025).