applied sciences

MDPI

Article

A CTI-Enriched GCN-LSTM Architecture for Multiclass Cyberattack Classification in Critical Infrastructure

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

[1] Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, Bogotá 111711, Colombia; ydonoso@uniandes.edu.co
[2] Faculty of Fundamental Sciences, Department of Informatics Engineering, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; luis.herrera@vilniustech.lt
[3] Department of Computer and Information Sciences, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand; jairo.gutierrez@aut.ac.nz
* Correspondence: ya.pinto10@uniandes.edu.co

Abstract

Critical infrastructures (CI) are essential to modern society, providing vital services such as energy, water, and transportation. However, these systems are increasingly targeted by sophisticated cyberattacks, exploiting vulnerabilities in both IT (Information Technology) and OT (Operational Technology) environments, posing significant risks to safety, economic stability, and national security. Despite advancements, current anomaly detection models for CI often cannot effectively integrate diverse data sources or provide detailed attack classifications. To address these challenges, we propose a novel Graph Convolutional Network (GCN) model integrated with Long Short-Term Memory (LSTM) layers for effective anomaly detection and attack classification in CI. The model leverages Cyber Threat Intelligence (CTI) and MITRE ATT&CK techniques, integrating network traffic and physical device data to enhance detection of sophisticated threats. Unlike approaches using binary classification, our model performs multiclass classification to recognize specific attack types, bridging the gap in understanding complex attack patterns within CI. By incorporating Indicators of Compromise (IoCs) from MISP (Malware Information Sharing Platform) with the SWAT (Secure Water Treatment) dataset, we developed a graph-based data structure where nodes represent entities like SCADA tags and IP addresses. The model processes this dynamic graph using convolutional layers for spatial feature extraction and LSTM layers for temporal dependencies. Results indicate a significant improvement over existing solutions, achieving a test accuracy of 99.04% and a macro F1-score of 0.9151. The integration of multiple data sources enhances the model's capacity to handle evolving cyber threats, making it well-suited for protecting CI.

Keywords: cyber threat intelligence; cybersecurity; critical infrastructures; graph neural networks; MITRE ATT&CK framework

Academic Editor: Christos Bouras

Received: 8 May 2026
Revised: 25 May 2026
Accepted: 26 May 2026
Published: 3 June 2026

1. Introduction

In recent years, cyberattacks targeting CI have become increasingly persistent and sophisticated. Following the COVID-19 pandemic, cyber threats targeting CI escalated significantly; the FBI's Internet Crime Complaint Center (IC3) reported that ransomware incidents affected at least 14 of the 16 recognized critical infrastructure sectors in the U.S. during a single year [1]. Real-world manifestations of these systemic vulnerabilities include the 2021 Colonial Pipeline ransomware attack, which caused widespread fuel disruptions

along the U.S. East Coast [2], and the same year's breach of the Oldsmar water treatment plant in Florida involved attempts to alter chemical levels in the water supply [3]. These attacks have far-reaching consequences, not only resulting in significant financial losses but also threatening the availability of essential services such as electricity and water supply.

These CI are uniquely vulnerable due to a convergence of factors, including the prevalence of legacy OT systems with long lifecycles, the use of insecure-by-design industrial protocols (e.g., Modbus, DNP3), and insufficient network segmentation between IT and OT environments. Attackers exploit these weaknesses to move laterally from enterprise networks into control systems, where they can directly manipulate physical processes, making advanced, context-aware detection models essential.

Cyber-Physical Systems (CPS), which play a pivotal role in CI by integrating physical devices such as sensors and actuators, enable direct interaction with the real world. Despite the numerous advantages of integrating CPS with digital systems, they remain vulnerable to a variety of physical and cyber threats [4]. While there is some overlap between IT systems and OT systems, a comprehensive understanding of the physical environment is often the key factor in mounting an adequate defense against potential cyberattacks, thereby preventing successful breaches [5].

In this context, CTI becomes an essential component in enhancing CI security. CTI plays a critical role by collecting, processing, analyzing, and applying information about cyber threats to improve an organization's security posture [6]. Including CTI in security systems significantly enhances their effectiveness by providing contextual awareness, improving threat attribution, and enabling proactive threat detection. CTI supplies essential context that helps distinguish between false positives and genuine threats, allowing security systems to respond more precisely. Additionally, CTI facilitates better threat attribution by mapping detected anomalies to known Indicators of Compromise (IoCs) and Tactics, Techniques, and Procedures (TTPs) used by specific threat actors. This enhanced attribution supports a deeper understanding of adversary motives and helps prioritize mitigation strategies, ultimately strengthening the defense of CI.

The Cyber Kill Chain and the MITRE ATT&CK Matrix are two of the most widely adopted cyber-attack representation models in the industry [7]. The Cyber Kill Chain is effective for depicting malware attacks; however, it is not intended to cover all key aspects of cyber-attacks and lacks flexibility when dealing with human-centric attack vectors, such as social engineering or insider threats. This limitation is compounded by the disparity in information between the attackers and the forensics and response teams attempting to counter these threats. These shortcomings have led some to mistakenly assume that the MITRE ATT&CK framework is simply an extension of the Cyber Kill Chain. However, MITRE ATT&CK offers a more detailed and adaptable representation of adversary tactics and techniques. Unlike the Cyber Kill Chain's linear approach, MITRE ATT&CK is not meant to follow a strict chronological sequence of events; instead, it aims to represent an adversary's entire attack lifecycle, making it better suited to evolving and sophisticated threats, particularly those involving human elements [8]. By offering a more comprehensive and nuanced view of the various stages and techniques employed by adversaries, MITRE ATT&CK provides an invaluable resource for understanding the complex tactics and behaviors of attackers [9]. Its flexible structure allows cybersecurity teams to adapt their defenses more effectively, improving detection and response strategies against a broader spectrum of threats that go beyond traditional malware-based attacks.

MITRE ATT&CK provides a common language and structured approach for understanding and categorizing the actions taken by threat actors during cyberattacks [10]. For Industrial Control Systems (ICS), the MITRE ATT&CK framework helps identify potential attack vectors specific to OT environments, thereby enabling the development of more

targeted defenses. Classifying attacks based on MITRE ATT&CK enhances the ability to detect, analyze, and mitigate threats by providing insights into the techniques used by attackers, making it possible to build detection systems that are better aligned with real-world threat scenarios. By leveraging the MITRE ATT&CK framework, organizations can improve their understanding of adversary behaviors and implement security measures that are tailored to the unique challenges faced by CI. For instance, mapping the tactics used in the Stuxnet attack to the framework can help in developing defenses against similar sophisticated threats. The increasing adoption of MITRE ATT&CK in industry stands in contrast to the limited body of academic research exploring its benefits, a gap that has also been observed in the literature [8].

Incorporating CTI into datasets used for building detection systems is vital in addressing existing security gaps. Despite advancements in security technologies, current detection systems often lack the ability to fully leverage CTI within a dynamic threat landscape, leading to suboptimal detection and mitigation strategies. By including intelligence information, detection models are better equipped to identify and correlate malicious activity with known threats, thereby improving detection rates and reducing false positives. This integration represents an innovative advancement by bridging the gap between static data analysis and dynamic threat intelligence, providing detection systems with real-time context and adaptability.

By applying Graph Convolutional Networks (GCNs), our model can more effectively integrate CTI data, capturing complex relationships between threat indicators and network entities. GCNs are particularly well-suited for CI security because they can model complex relationships and dependencies within network traffic, capturing both spatial and temporal features of cyber-physical environments [11]. The ability of GCNs to effectively represent interactions between entities such as devices, protocols, and threat indicators makes them a powerful tool for enhancing the detection and prevention of sophisticated attacks on CI [12].

Given the limitations of current detection systems in fully leveraging CTI and the challenges in integrating the MITRE ATT&CK framework, this research seeks to address the following question: How can integrating CTI and the MITRE ATT&CK framework with GCNs enhance the detection and classification of cyberattacks on CI?

While prior works have applied GNNs for binary anomaly detection or used CTI for threat modeling, a critical gap exists in unifying these approaches for context-rich, multiclass threat classification. Our research addresses this gap by proposing a novel model that performs a synergistic integration of CTI-enriched graph data structures with a GCN-LSTM architecture. This allows our model to move beyond simple anomaly flagging to provide actionable, tactic-level classification of attacks based on the MITRE ATT&CK framework, representing a key advancement in developing adaptive defense mechanisms for CI. The key contributions of this study are:

- Development of a CTI-Enhanced Detection Model: We introduce a detection model that integrates CTI, enhancing the system's ability to recognize and respond to known threats with greater accuracy.
- Attack Classification Using MITRE ATT&CK Framework: Our model classifies detected attacks based on the MITRE ATT&CK framework, providing a structured and standardized approach to understanding adversary tactics and techniques.
- Application of GCNs: We leverage GCNs to model complex relationships within network traffic data, enhancing the detection of sophisticated cyber threats while reducing false positives. Additionally, we incorporate interpretability capabilities to identify potential attack sources.

By combining intelligence-driven insights with the advanced modeling capabilities of GCNs and leveraging the structured knowledge provided by the MITRE ATT&CK framework, our approach offers a comprehensive defense mechanism. This integration is poised to enhance the resilience of CI security systems, enabling them to proactively detect and mitigate threats, thereby ensuring the reliability and availability of essential services.

This study employs Design Science Research Methodology (DSRM) as the foundational framework to systematically develop and evaluate the proposed detection model. DSRM offers a rigorous and structured approach for creating and assessing artifacts intended to solve identified problems within a specific domain [13]. As illustrated in Figure 1, our study adhered to the DSRM process. We began with *Problem Identification*, recognizing that existing IDS struggle to leverage CTI and classify attacks in CI. This led to *Defining the Objectives* for a solution: a model that integrates structured threat intelligence for enhanced multiclass detection. The *Design and Development* phase involved the core technical work of creating the CTI-enriched graph structure and the GCN-LSTM architecture. The model was then *Demonstrated* by applying it to the SWAT dataset, and its performance was rigorously *Evaluated* using metrics like F1-score to assess its effectiveness against imbalanced, real-world data. Finally, the findings are communicated to contribute to both academic knowledge and practical applications in cybersecurity defense. This methodological approach not only advances theoretical integration but also provides tangible solutions for strengthening cybersecurity in real-world CI environments.

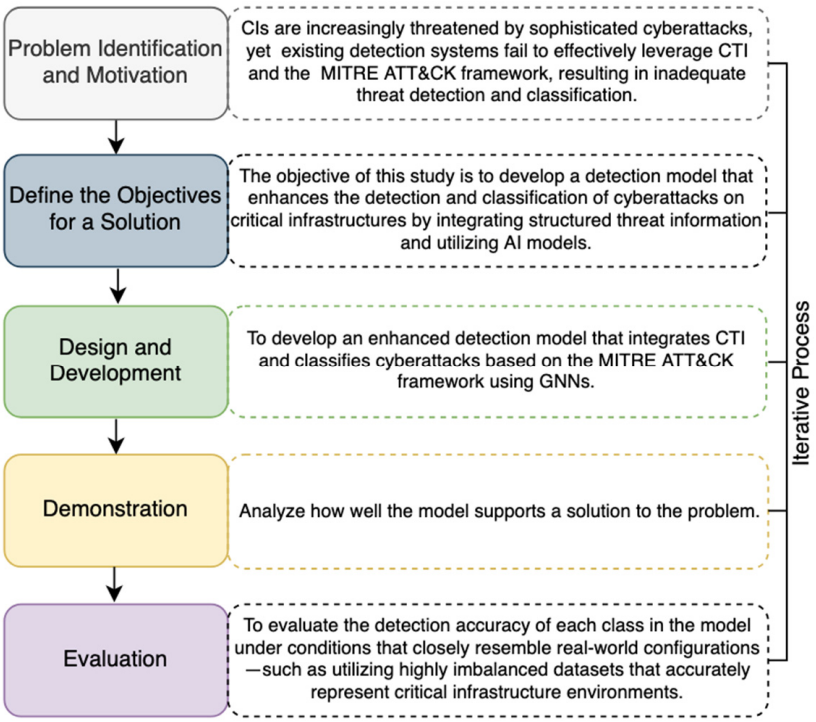


Figure 1. DRSM Methodology.

The remainder of this paper is organized as follows: Section 2, Background, provides essential information on CTI, the MITRE ATT&CK framework, GNNs, and the SWAT dataset. Section 3 reviews related work and existing literature in the field. Section 4, Materials and Methods, details the preprocessing steps, the integration of CTI with the MITRE ATT&CK framework, and the architectural design of the proposed GCN-LSTM model. Section 5, Results and Discussion, presents the experimental findings, the ablation study, and model interpretability. Finally, Section 6, Conclusions and Future Work, summarizes our contributions and suggests future research directions.

2. Background

The increasing sophistication of cyberattacks targeting CI has necessitated the development of advanced methodologies for effective threat detection and mitigation. This background section provides a comprehensive overview of the fundamental frameworks and technologies underpinning this research to enhance cybersecurity in CI. The MITRE ATT&CK framework serves as an extensive knowledge base for adversary tactics and techniques, supporting the systematic identification and categorization of cyber threats. CTI plays a pivotal role in gathering, analyzing, and utilizing threat-related data, thereby enhancing situational awareness and improving defense mechanisms. To analyze and model complex relationships within network data, GCNs are introduced as an advanced AI approach capable of capturing the intricate dependencies and dynamics within CI systems.

Furthermore, the SWAT (Secure Water Treatment) dataset is employed as a benchmark to evaluate the performance of cybersecurity models within an ICS environment, representing realistic operational conditions and potential threat scenarios. Together, these concepts establish a robust foundation for developing comprehensive, adaptive, and intelligence-driven cybersecurity solutions for CI environments.

2.1. MITRE ATT&CK for ICS

The MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework is an internationally recognized, comprehensive knowledge base that systematically catalogs adversary tactics and techniques derived from real-world cyber incidents [14]. It offers a structured methodology for understanding, analyzing, and categorizing cyber threats, thereby enabling organizations to anticipate, mitigate, and respond effectively to potential adversarial actions.

MITRE ATT&CK for ICS is a specialized extension of the standard framework, meticulously tailored to address the unique challenges associated with securing ICS environments [9]. These environments are critical components of essential infrastructure sectors, including energy, water treatment, manufacturing, and transportation. These systems differ fundamentally from traditional IT environments due to their stringent real-time operational requirements, safety-critical functions, and the utilization of specialized hardware and communication protocols. The framework's components—tactics and techniques—are essential for dissecting and analyzing cyber threats: (1) Tactics, represent the adversary's tactical objectives during an attack, answering the question of "why" an adversary performs a certain action. They are the high-level goals that an attacker aims to achieve at each stage of the intrusion, and (2) Techniques, describe the specific methods adversaries use to achieve their tactical objectives, addressing the "how" of an attack [15], as illustrated in Table 1. Detailed information on each technique, including descriptions, examples, detection recommendations, and potential mitigations, can be found online. The list of techniques and tactics is continuously updated; at the time this study was conducted, a total of 12 tactics and 94 techniques constituted the matrix for ICS.

The ATT&CK for ICS framework is of paramount importance for organizations operating CI for several reasons: (1) *Enhanced Threat Modeling*, the framework enables systematic identification and categorization of vulnerabilities and potential attack vectors specific to ICS environments, facilitating comprehensive threat modeling and risk assessment; (2) *Improved Detection and Response Capabilities*, by providing detailed insights into adversary behaviors and methodologies, the framework supports the development of targeted detection mechanisms and incident response strategies aligned with documented adversarial tactics and techniques; (3) *Standardization of Cybersecurity Practices*, the framework establishes a common taxonomy and language for discussing cyber threats, enhancing collaboration and information sharing among cybersecurity professionals, organizations, and

industry sectors, and (4) *Alignment with Regulatory and Compliance Requirements*, adoption of the framework can assist organizations in meeting industry-specific cybersecurity standards and regulatory obligations by demonstrating a proactive and structured approach to threat management.

Table 1. MITRE ATT&CK for ICS Summary.

Tactics	Adversary Objectives	Techniques (Example)
Initial Access	The adversary is attempting to gain unauthorized entry into ICS environments.	• Drive-by Compromise. • Exploit Public-Facing Application. • Exploitation of Remote Services.
Execution	The adversary aims to execute code or manipulate system functions, parameters, and data without authorization.	• Autorun Image. • Change Operating Mode. • Command-Line Interface.
Persistence	The adversary is attempting to maintain persistent access within the ICS environment.	• Hardcoded Credentials. • Modify Program. • Module Firmware. • Project File Infection.
Privilege Escalation	The adversary is attempting to obtain elevated privileges.	• Exploitation for Privilege Escalation. • Hooking.
Evasion	The adversary is attempting to evade security defenses.	• Change Operating Mode. • Exploitation for Evasion.
Discovery	The adversary is gathering information to evaluate and identify potential targets within the environment.	• Network Connection—Enumeration. • Network Sniffing. • Remote System Discovery.
Lateral Movement	The adversary is attempting to move within the ICS environment.	• Default Credentials. • Exploitation of Remote Service. • Hardcoded Credentials.
Collection	The adversary is attempting to collect data and domain knowledge from the ICS environment to support their objectives.	• Adversary-in-the-Middle. • Automated Collection. • Data from Information Repositories.
Command and Control	The adversary is attempting to establish communication with and control compromised systems, controllers, and platforms within the ICS environment.	• Commonly Used Port. • Connection Proxy. • Standard Application Layer Protocol.
Inhibit Response Function	The adversary is attempting to inhibit safety, protection, quality assurance, and operator intervention functions from responding to failures, hazards, or unsafe conditions.	• Activate Firmware Update Mode. • Alarm Suppression. • Block Command Message. • Block Reporting Message. • Block Serial COM.
Impair Process Control	The adversary aims to manipulate, disable, or damage physical control processes.	• Brute Force I/O. • Modify Parameter. • Module Firmware. • Spoof Reporting Message.
Impact	The adversary is attempting to manipulate, disrupt, or destroy ICS systems, data, and their surrounding environment.	• Damage to Property. • Denial of Control. • Denial of View. • Loss of Availability. • Loss of Control.

2.2. Cyber Threat Intelligence (CTI)

CTI constitutes a systematic methodology for the collection, analysis, and utilization of information concerning both existing and emerging cyber threats, enabling a deeper understanding of adversary behaviors, tactics, techniques, and procedures. By converting raw data into actionable intelligence, CTI offers a comprehensive perspective on the dynamic threat landscape, thereby empowering organizations to proactively anticipate, detect, and respond to cyber threats with heightened effectiveness [16]. For CI, which encompass essential services such as energy, water, and transportation, integrating CTI into cybersecurity strategies is imperative. These infrastructures are increasingly targeted by sophisticated cyberattacks that exploit vulnerabilities across both information technology IT and OT, rendering traditional reactive defenses inadequate. The incorporation of CTI fosters a proactive defense posture by providing real-time situational awareness, facilitating the early identification of IoCs, and ensuring that cybersecurity measures are continuously aligned with the latest threat intelligence. This intelligence-driven approach significantly enhances the resilience of CI, mitigating the impact of cyber incidents through reduced response times and improved adaptability to evolving threats [17].

IoCs, as an essential element of CTI, are critical pieces of forensic data that indicate evidence of potential breaches or ongoing cyberattacks. These indicators, including IP addresses, domain names, file hashes, and unusual network behaviors, offer valuable insights into adversary tactics and provide a traceable footprint of malicious activities [18]. IoCs are derived from past cyber incidents and play a crucial role in enhancing an organization's detection and response capabilities by allowing security teams to compare current system data with known malicious patterns. Given their ability to identify and signal potential threats, IoCs are invaluable in improving situational awareness and strengthening the overall cybersecurity posture of critical infrastructures.

Incorporating IoCs into Artificial Intelligence (AI)-driven cybersecurity models is a novel and powerful approach to threat detection and mitigation. Feeding AI systems with IoC data, such as known malicious IP addresses or ports associated with previous attacks, enables these models to identify patterns indicative of a potential threat, even in its nascent stages. Unlike static rule-based systems, AI models utilizing IoCs can learn and generalize, recognizing threats that share similarities with known attacks while adapting to altered adversarial tactics. This intelligence-driven integration not only strengthens predictive capabilities but also enhances the accuracy and efficiency of intrusion detection systems. By leveraging well-established IoCs, AI-based security solutions for critical infrastructures can develop a more comprehensive and proactive strategy, ultimately improving resilience and adaptability against an increasingly sophisticated threat landscape.

2.3. Graph Neural Networks (GNNs)

GNNs are a class of deep learning models specifically tailored for data represented as graphs [12], making them particularly effective for modeling intricate dependencies in network traffic and physical sensor data within cybersecurity contexts. Unlike traditional neural network architectures that are optimized for structured data formats such as images or sequential data, GNNs are designed to effectively capture the relationships inherent in non-Euclidean structures [19], such as the complex interdependencies in network topologies and sensor networks within ICS and CI.

Mathematically, GNNs employ an iterative message-passing mechanism to propagate node features throughout the graph, enabling nodes to learn representations that encapsulate not only their inherent features but also contextual information from their neighbors. For a given node v, the embedding at the k-th layer is updated according to:

$$h_v^{(k)} = \sigma(\sum_{u \in \mathcal{N}(v)} W^{(k)} h_u^{(k-1)} + b^{(k)}),$$

where $h_v^{(k)}$ is the node embedding at layer k, $\mathcal{N}(v)$ denotes the set of neighboring nodes of v, $W^{(k)}$ and $b^{(k)}$ are learnable parameters, and σ represents an activation function. This approach allows for the aggregation of local information, resulting in higher-order structural learning that captures the nuanced relationships between nodes [20].

In cybersecurity applications, GNNs have proven highly effective in modeling network traffic data by representing devices or hosts as nodes and network interactions as edges. This graph-based representation enables the identification of anomalous behaviors that may indicate sophisticated cyber threats, thereby enhancing detection capabilities beyond traditional signature-based methods. Furthermore, when applied to physical sensor data in ICS, GNNs can model the complex relationships between interconnected sensors, allowing for precise detection of anomalies indicative of system failures or malicious activities. The capacity of GNNs to effectively aggregate both local and global graph information provides a robust framework for anomaly detection, improved situational awareness, and a strengthened security posture, positioning them as a critical tool in safeguarding CI systems against evolving cyber threats.

The capacity of GNNs to reduce false positives stems from their ability to learn contextual relationships within the graph structure. Unlike methods that assess data points in isolation, a GNN's message-passing mechanism evaluates a node's features in the context of its neighbors. An anomalous sensor reading that might trigger an alarm in a traditional system can be correctly identified as benign by a GNN if the states of connected components are consistent with that reading, thus averting a false positive.

2.4. Dataset

The SWAT dataset was selected as the foundational evaluative benchmark due to its rigorous emulation of a real-world ICS environment. While newer iterations of the SWAT testbed data exist, they predominantly isolate physical process logs. The original dataset remains uniquely suited for this research because it comprehensively integrates both network traffic and physical process data, providing a rich environment for testing multi-layer intrusion detection systems. This synchronized dual-modality is a strict architectural prerequisite for generating the spatio-temporal graph inputs required by the proposed GCN-LSTM framework, justifying its use over more recent, yet single-modality, repositories. Developed by the iTrust Center for Research in Cyber Security at the Singapore University of Technology and Design, SWAT is universally recognized in the literature as a gold-standard academic benchmark for cyber-physical security [21]. Recent high-impact studies published in top-tier journals have consistently relied on the SWAT dataset to validate advanced anomaly detection architectures, further underscoring its credibility as a robust evaluative standard [22–24]. The decision to validate the proposed architecture primarily on this benchmark is driven by strict architectural prerequisites rather than arbitrary selection. The GCN-LSTM model requires synchronized, dual-modality inputs encompassing both continuous physical sensor telemetry and discrete industrial network traffic (CIP/EtherNet/IP). Most conventional cybersecurity benchmarks (e.g., NSL-KDD, UNSW-NB15) are strictly IT-centric and lack physical process data, rendering them structurally incompatible with this spatio-temporal framework. Consequently, SWAT remains one of the few rigorously documented, multi-layered environments capable of validating the proposed integration of CTI, network topology, and physical state deviations [21]. The SWAT system comprises a six-stage water treatment process, integrating 25 sensors and 26 actuators that monitor and control physical processes. The dataset encompasses detailed time-series data capturing both normal operational states and 41 series of orchestrated cy-

berattacks. These attacks span a wide range of adversarial tactics and techniques, targeting different components and stages of the water treatment process.

In addition to the physical process data, the SWAT dataset includes comprehensive network traffic data collected from a star network topology at Level 1 of the SWAT network architecture. This level represents the communication channels between the Supervisory Control and Data Acquisition (SCADA) system and the six Programmable Logic Controllers (PLCs). The network utilized the Common Industrial Protocol (CIP) over EtherNet/IP, a standard protocol in industrial communications. The dataset comprises 18 distinct features, including Date, Time, Origin, Type, Interface, Direction, Source IP, Destination IP, Protocol, Proxy Source IP, Application Name, Modbus Function Code, Modbus Function Description, Modbus Transaction ID, SCADA Tag, Modbus Value, Destination Port, and Source Port.

To manage the substantial volume of data, the network traffic was segmented into multiple CSV files, each containing up to 500,000 packets. Due to data capture occurring at one-second intervals, there are instances where multiple entries share the same timestamp but represent different network activities, reflecting the concurrent processes typical in ICS environments. The network data spans a 11-day period, which includes the final four days dedicated to various attack scenarios meticulously crafted to simulate real cyber threats.

Although Table 2 lists all the cyberattacks conducted, certain limitations affected their inclusion in our analysis. Specifically, five attacks—numbered 5, 9, 12, 15, and 18—did not produce any physical impact on the system. While reconnaissance and network-probing activities represent critical early-stage adversarial behaviors, they were excluded from our classification targets.

The primary objective of this proposed GCN-LSTM architecture is the detection of sophisticated attacks that manifest as measurable cyber-physical state deviations. Consequently, our analysis is strictly scoped to anomalies that produce a tangible disruption to the physical processes, ensuring the model's focus remains on high-impact operational threats. Attack number 4, which targeted a motorized valve designated as MV-504, could not be analyzed due to the absence of data related to this valve in the official dataset. Furthermore, attacks 24 and 34 exhibited minimal or no discernible effect on system performance. Attacks numbered 13, 14, and 29 were unsuccessful in their execution. Additionally, attacks 6, 19, 20, and 38 were aimed at chemical sensors that were not operational at the time. Although these attacks were intended to influence other sensors, various malfunctions prevented this from occurring, thereby impeding the detection of these cyberattacks. Consequently, our analysis focuses on detecting a total of 26 attacks that had a tangible impact on the system.

Table 2. Attack Scenarios in SWAT Dataset, adapted from [21] with proposed MITRE ATT&CK mapping.

No.	# SWAT Dataset	Attack Point	Attack	Expected Impact or Attacker Intent	Mapping to MITRE ATT&CK	
1	1	MV-101	Open MV-101	Tank overflow	T0831	Manipulation of Control—Opening/Closing valves or other actuators
2	10	FIT-401	Set value of FIT-401 as <0.7	UV shutdown; P-501 turns off		
3	11	FIT-401	Set value of FIT-401 as 0	UV shutdown; P-501 turns off		
4	17	MV-303	Do not let MV-303 open	Tank overflow		
5	23	P-602, DIT-301, MV-302	Value of DPIT-301 set to >0.4 bar; Keep MV-302 open; Keep P-602 closed	Change in water quality		
6	27	P-302, LIT-401	Keep P-302 on continuously; Value of LIT401 set as 600 mm till 1:26:01	Tank overflow		
7	39	FIT-401, AIT-502	Set value of FIT-401 as 0.5; Set value of AIT-502 as 140 mV	UV will shut down and water will go to RO		
8	40	FIT-401	Set value of FIT-401 as 0	UV will shut down and water will go to RO		

Table 2. *Cont.*

No.	# SWAT Dataset	Attack Point	Attack	Expected Impact or Attacker Intent	Mapping to MITRE ATT&CK	
9	3	LIT-101	Increase by 1 mm every second	Tank Underflow; Damage P-101	T0836	Modify Parameter—Changing device parameters to alter behavior
10	7	LIT-301	Water level increased above HH	Stop of inflow; Tank underflow; Damage P-301		
11	16	LIT-301	Decrease water level by 1 mm each second	Tank Overflow		
12	25	LIT-401, P-401	Set value of LIT-401 as 1000; P402 is kept on	Tank underflow		
13	26	P-101, LIT-301	P-101 is turned on continuously; Set value of LIT-301 as 801 mm	Tank 101 underflow; Tank 301 overflow		
14	30	LIT-101, P-101, MV-201	Turn P-101 on continuously; Turn MV-101 on continuously; Set value of LIT-101 as 700 mm; P-102 started itself because LIT301 level became low	Tank 101 underflow; Tank 301 overflow		
15	31	LIT-401	Set LIT-401 to less than L	Tank overflow		
16	36	LIT-101	Set LIT-101 to less than LL	Tank overflow		
17	41	LIT-301	decrease value by 0.5 mm per second	Tank overflow		
18	2	P-102	Turn on P-102	Pipe bursts	T0879	Damage to Property—Causing physical damage to equipment or infrastructure
19	21	MV-101, LIT-101	Keep MV-101 on continuously; Value of LIT-101 set as 700 mm	Possible damage to RO		
20	32	LIT-301	Set LIT-301 to above HH	Tank underflow; Damage P-302		
21	33	LIT-101	Set LIT-101 to above H	Tank underflow; Damage P-101		
22	8	DPIT-301	Set value of DPIT as >40 kpa	Backwash process is started again; Normal operation stops; Decrease in water level of tank 401. Increase in water level of tank 301	T0881	Service Stop—Normal operation stop.
23	22	UV-401, AIT-502, P-501	Stop UV-401; Value of AIT502 set as 150; Force P-501 to remain on	System freeze		
24	28	P-302	Close P-302	Stop inflow of tank T-401		
25	35	P-101; P-102	Turn P-101 off; Keep P-102 off	Stops outflow		
26	37	P-501, FIT-502	Close P-501; Set value of FIT-502 to 1.29 at 11:18:36	Reduced output		

Note: The # symbol indicates the specific attack number within the SWAT dataset.

3. Related Work

Research into the application of frameworks such as MITRE ATT&CK to enhance cybersecurity in CI has been previously conducted. For example, in [9], the authors explore the use of the MITRE ATT&CK and D3FEND frameworks to enhance maritime cybersecurity. The study aims to illustrate how these frameworks can be used to model both cyberattacks and corresponding defense mechanisms, offering proactive and reactive cybersecurity strategies for the maritime sector. However, the paper is limited by the absence of a maritime-specific adaptation of the MITRE ATT&CK matrix, highlighting a gap in fully capturing threats unique to maritime environments. Additionally, the study focuses solely on modeling attack-defense scenarios without incorporating risk management aspects, which limits the overall applicability of the findings. This paper is relevant to the present study, as it demonstrates the utility of MITRE ATT&CK in modeling attack tactics and MITRE D3FEND for countering threats, which aligns with the research's focus on enhancing threat detection and classification through CTI. Integrating such frameworks within AI-based approaches, such as GNNs, can significantly enhance the detection of sophisticated cyberattacks on critical infrastructure systems, thereby contributing to the development of a robust cybersecurity posture.

An additional study that highlights the importance of incorporating CTI into the cybersecurity posture is presented in [14]. The authors aim to address the growing complexities of enterprise IT systems by providing a structured method for modeling and simulating cyberattacks, thereby supporting decision-making to improve cybersecurity resilience. They emphasize the importance of integrating multiple CTI sources for a more comprehensive understanding of potential threats. Although the study utilizes the MITRE Enterprise ATT&CK Matrix, both case studies presented involve CI, including OT. This underscores the need to also employ the MITRE ATT&CK Matrix to more accurately represent threats in these hybrid environments. Nonetheless, this research aligns with the current study by underscoring the value of comprehensive threat modeling and leveraging diverse CTI sources to strengthen cybersecurity defenses. Another proposal to integrate threat modeling and structured security frameworks, such as MITRE ATT&CK, to enhance CI security is presented in [4]. The authors propose a threat modeling framework specifically for smart firefighting systems, which are a form of CPS. The study uses the MITRE ATT&CK Matrix in combination with System Requirement Collection (SRC) to generate a threat list for smart firefighting systems. This threat list is then mapped to NIST security and privacy controls to develop a mitigation framework that aims to improve the security posture of smart firefighting CPS. The primary purpose of this research is to enhance the understanding of potential security risks in CPS by integrating threat-informed methodologies and aligning them with widely recognized security standards.

Previous research has explored incorporating CTI and MITRE ATT&CK information to enhance AI models, though most studies have focused on enterprise infrastructures rather than critical CI. For example, in [25], the authors present a hybrid model for intrusion detection in IoT environments connected to traditional IT systems. The approach combines a basic autoencoder (bAE), one-class support vector machine (OCSVM), deep autoencoder (dAE), and DBSCAN clustering to address challenges like dimensionality reduction and anomaly detection. The model, tested on CIC-IDS2017 and CSECIC-IDS2018, achieved over 98% accuracy. However, the authors do not display class-specific detection accuracy, which may indicate bias toward majority classes in highly imbalanced scenarios. The MITRE ATT&CK framework is used to guide feature selection and ensure a representative variety of attacks, but the model output remains a binary classification (attack vs. normal). Despite promising accuracy, the model's limitations include high computational requirements, reliance on high-quality data, and interpretability issues that may hinder real-time deployment in resource-constrained IoT environments.

In [17], authors leverage the MITRE ATT&CK framework to validate security measures in ICS by modeling specific attack scenarios such as Denial of Service (DoS) and ARP poisoning.

The study employs open-source intrusion detection systems like Suricata and Snort, utilizing IoCs to enhance detection capabilities and situational awareness. Despite its contributions, the work is limited by its focus on a narrow set of attack scenarios and the use of controlled simulations, which may not fully capture the complexities of real-world ICS environments. In contrast, our study incorporates IoCs information into the training of an AI-driven model using GNNs to enhance anomaly detection capabilities. This approach enables the model to learn from IoCs, providing a more dynamic and generalized response to evolving threats beyond predefined scenarios. By integrating IoCs into the AI training process, our model offers a more robust and adaptive cybersecurity solution for ICS, addressing the evolving threat landscape more effectively compared to traditional rule-based methods.

Researchers have recently shifted their focus to GNNs due to their powerful ability to analyze graph-structured data derived from network topologies, providing new

perspectives and insights. Over the past five years, numerous variations of GNNs have emerged, including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), Graph-SAGE, among others. For example, a GRU-based anomaly detector for ICS was proposed in [22], integrating GNNs to enhance anomaly detection capabilities. The model aims to improve detection accuracy, prevent accidents, and provide interpretability by capturing complex dependencies among sensor signals. Evaluation against nine state-of-the-art algorithms shows that the proposed model achieves superior performance, particularly in precision and recall. By leveraging GRU for temporal analysis and GNNs for modeling interdependencies, the model effectively addresses challenges such as gradient vanishing and exploding. A similar approach is presented in [23], where the authors introduce the Global-Local Integration Network (GLIN) to enhance anomaly detection in ICS by integrating local and global information. The methodology includes graph construction, node embedding through an encoder, and pooling to generate global representations. GLIN was trained and evaluated against baseline models, demonstrating superior performance in anomaly detection, with accuracy improvements of up to 96.86% and F1-scores reaching 0.95. Recent advancements further validate this trajectory; for instance, a hybrid GNN-LSTM framework has recently been proposed specifically for attack vector reconstruction. Similar to our approach, this contemporary research utilizes GNNs to analyze structural relationships mapped to the MITRE ATT&CK framework, while leveraging LSTM networks to model the temporal dynamics of attack sequences [26].

In their study on Dual-Process Dynamic Graph-Based Anomaly Detection for Multivariate Time Series, the authors present a novel method called DPDGAD, aimed at enhancing anomaly detection in CPSs by classifying spatiotemporal features into steady and dynamic components. The approach involves using a dual-process structure: a Dynamic Graph Construct Block (DGCB) for representing the steady and dynamic features as graphs and a Trend-Guided Predictor (TGP) for forecasting time series values to detect anomalies. The model was tested on six real-world datasets, achieving F1 scores of 0.9009 on the SWAT dataset [24]. Similarly, the recent FST-AD framework utilizes spatio-temporal Graph Neural Networks to capture dynamic dependencies among SWAT variables, highlighting that deep fusion of temporal patterns and graph structures yields significant gains in robustness against CPS cyber threats [27]. However, like the aforementioned studies, it remains focused on binary anomaly detection rather than multi-class tactic classification.

To emphasize the importance of modeling both spatial and temporal aspects of sensor data, a study presented in [28] introduces a model that integrates GATs to improve anomaly detection in multivariate time series from industrial systems. The model employs a time-series encoder and a sustainable graph structure learning method to enhance the learning of sensor relationships over time. The methodology involves continuous updates to the graph structure, combining local sensor data with temporal analysis to accurately detect anomalies. The model achieved an F1 score of 92.98% for the SWAT dataset.

While most of these advancements prioritize binary anomaly detection, recent literature from 2024 and 2025 has increasingly focused on multiclass threat classification within industrial systems. For instance, recent studies have employed advanced deep learning architectures, such as Transformer-based models and ensemble networks, to categorize distinct cyberattacks across ICS testbeds [29,30]. However, a significant limitation of these contemporary approaches is their reliance on arbitrary, dataset-specific labels (e.g., ‘Attack Type 1’ or ‘Data Injection’). They do not map these anomalies to a standardized operational taxonomy. Furthermore, these multiclass models often process network traffic in isolation, lacking the spatial-temporal graph mechanisms required to correlate network events with physical sensor deviations. The GCN-LSTM architecture proposed in this study directly addresses these limitations. By fusing spatial topology with temporal tracking and explicitly

mapping the multiclass outputs to standardized MITRE ATT&CK tactics, our approach ensures that the classifications are both mathematically rigorous and operationally actionable for incident response.

In summary, while the reviewed literature shows progress in applying AI to ICS security, key gaps remain. Methodologies often focus on either threat modeling with CTI or GNN-based anomaly detection but rarely integrate both. Furthermore, most detection models are limited to binary classification ("Normal" vs. "Anomaly"), which lacks the contextual detail needed for effective incident response. Our work directly addresses these gaps. As summarized in Table 3, our primary contribution is the synergistic integration of CTI-enriched graph data with a spatio-temporal GCN-LSTM model to perform multiclass classification of threats into specific MITRE ATT&CK tactics.

Table 3. Comparison of Proposed Model with State-of-the-Art Approaches.

Ref.	Methodology	CTI Integration	MITRE ATT&CK Use	Classification	Key Limitation Addressed by Our Work
[17]	Rule-based IDS (Suricata, Snort)	Uses IoCs for rule creation	Validates specific scenarios	Binary (Alert/No Alert)	Lacks learning/generalization; our GNN model learns from IoCs directly.
[25]	Unsupervised (AE, OCSVM)	Indirectly, via feature selection	Guides attack variety in dataset	Binary	Does not provide attack type; our model offers multiclass classification into tactics.
[22,23]	GNN-based (GRU, GLIN)	None	None	Binary Anomaly	Lack threat context; our model enriches graph data with CTI for context-aware detection.
Our Model	GCN-LSTM	Direct (IoCs enrich graph features)	Defines output classes (Tactics)	Multiclass	Provides actionable, tactic-level classification based on enriched spatio-temporal data.

4. Materials and Methods

To address the challenges of detecting sophisticated cyber-physical threats, this section outlines the systematic approach used to develop and evaluate the proposed detection model. Following the DSRM, the process encompasses data preparation, feature enrichment with external intelligence, and the implementation of a hybrid deep learning architecture. The methodology is divided into three primary phases: the rigorous preprocessing of the SWAT dataset to ensure high-fidelity inputs, the synergistic integration of CTI with the MITRE ATT&CK framework, and the design of the GCN integrated with LSTM layers.

4.1. Preprocessing the SWAT Dataset

The preprocessing of the network data followed a structured, multi-step process to ensure quality and optimize model training. Initially, invalid entries, such as NaN values, were systematically removed to maintain data integrity. Subsequently, features with low or zero variance—such as Origin, Type, Interface Name, Interface Direction, Protocol, Application Name, Modbus Function Code, and Service—were excluded, as they offered little informational value, thereby enhancing model performance.

Measurements from physical sensors, initially encoded under the Modbus Value feature in a little-endian, single-precision floating-point format, were parsed and converted into integer format to facilitate analysis.

Using detailed documentation from the SWAT developers, the Modbus Value feature was disaggregated into multiple columns, each corresponding to specific physical devices in the scaled-down hydroelectric plant, resulting in 24 distinct features. To handle missing values, linear interpolation was applied. This method was specifically selected because the missing data intervals within the dataset were transient (typically spanning less than three seconds). Given the physical inertia inherent to the water treatment testbed (e.g., gradual changes in tank levels or flow rates), linear interpolation accurately estimates these continuous physical variables without introducing artificial statistical trends or skewing

the temporal dependencies learned by the subsequent LSTM layers. Finally, categorical data—including source and destination IP addresses, Modbus Function Description, Modbus Transaction ID, SCADA Tag, and Source Port—was encoded using the $pd.factorize()$ method to enable effective integration into the machine learning model. This comprehensive preprocessing approach ensured that the dataset was feature-rich, well-balanced, and optimized for anomaly detection and network analysis tasks, ultimately supporting a more robust model performance.

4.2. Combining CTI with MITRE ATT&CK for Advanced Threat Detection in ICS

The integration CTI and the application of the MITRE ATT&CK for ICS framework are pivotal in enhancing the security capabilities of ICS. CTI provides valuable contextual insights by leveraging IoCs, threat actor profiles, and Tactics, Techniques, and Procedures (TTPs), enabling the dataset to be enriched with real-world threat data. This study incorporates CTI into the SWAT dataset, creating a rich contextual environment for better anomaly detection. In parallel, the MITRE ATT&CK techniques were systematically selected and aligned with the characteristics of the SWAT dataset to accurately reflect common adversary behaviors in ICS environments. By combining CTI and MITRE ATT&CK, this section demonstrates how structured threat intelligence and adversarial knowledge can be seamlessly integrated into ICS data to enhance threat detection, resilience, and adaptability to evolving cyber threats.

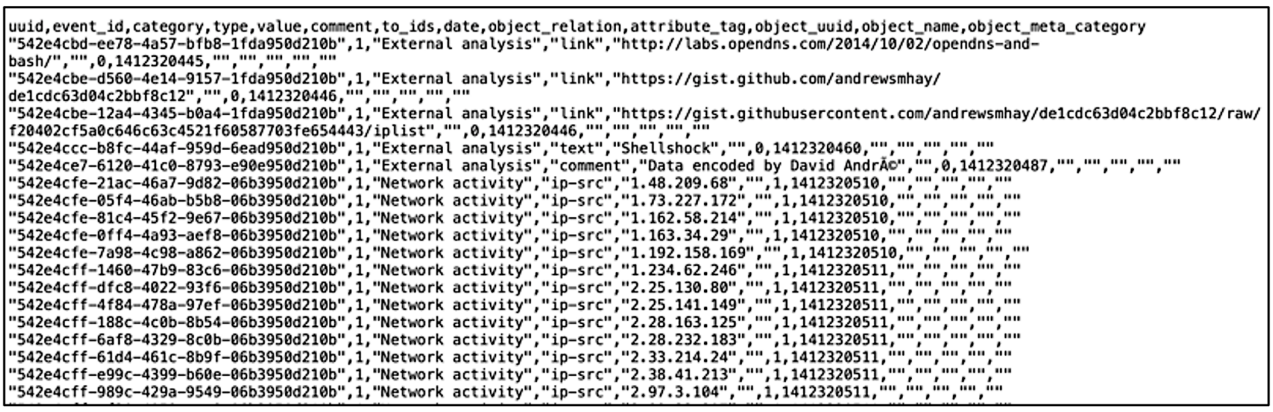
4.2.1. CTI-Driven Feature Engineering

The Malware Information Sharing Platform (MISP) is an open-source threat intelligence platform that facilitates the collection, storage, sharing, and analysis of CTI data. It provides a collaborative environment for disseminating information on IoCs, threat actors, TTPs, and other threat-related artifacts to enhance situational awareness and strengthen organizational defense capabilities.

MISP can be leveraged to extract valuable information that serves as input features for graph nodes and to capture temporal relationships within network data. Specifically, IoCs—such as IP addresses, domains, and network ports—can be represented as graph nodes, while their interconnections form graph edges. The efficacy of transforming unstructured threat intelligence into structural graph components is supported by recent literature. For instance, recent frameworks like Warning-Graph have successfully demonstrated that modeling easily accessible IoCs (such as IP addresses and ports) as nodes in a heterogeneous information network allows Graph Neural Networks to capture rich threat attributes and significantly reduce dependency on labeled samples [31]. Additionally, TTPs provide insights into adversary behaviors, which are instrumental in capturing temporal dependencies through LSTM networks. By harnessing the structured threat data from MISP, models can be trained to recognize real-world attack patterns, making them well-suited for anomaly detection and the enhancement of the security posture of CI systems. An example of information extracted in CSV format from MISP is depicted in Figure 2.

The threat intelligence data extracted from MISP was subsequently integrated into the SWAT dataset to enhance its network traffic data with contextual threat insights. Specifically, information from MISP—including source and destination IP addresses, as well as port numbers that matched existing columns in the SWAT dataset—was incorporated to improve the feature richness of the dataset. The identified IoCs, such as known malicious IPs and flagged ports, were used to label or augment data points in the SWAT dataset, adding a contextual layer indicating potential threat activity. The process is illustrated in Figure 3. However, the amount of anomalous traffic remained low at approximately 9.10%, resulting in an imbalanced scenario that is typical for models used to detect cyberattacks through

network traffic data. Because the SWAT testbed operates as a closed-loop system, its native traffic inherently lacks external malicious IP addresses. To bridge this gap without compromising the dataset's natural statistical distribution, we performed a volumetrically bounded synthetic data augmentation. By strictly mapping external IoCs from our MISP CTI database to the temporal windows of existing physical anomalies, the overall ratio of anomalous to benign traffic was artificially preserved at its native baseline of approximately 9.10%. This controlled injection ensures that the original network packet distribution remains unaltered, effectively simulating unrepresented attack configurations (e.g., data exfiltration) while allowing the model to learn the contextual relationship between external threat actor infrastructure and localized cyber–physical anomalies. This integration adds a crucial dimension of threat awareness to the dataset, thereby allowing the model to learn from both sensor and network traffic patterns, enriched with insights derived from historical cyberattacks. Such enrichment significantly strengthens the model's capability to detect sophisticated cyber threats by incorporating both real-time operational data and strategic context provided by threat intelligence. Consequently, this integration facilitates a more adaptive and comprehensive anomaly detection system tailored to meet the security needs of CI.



```
uuid,event_id,category,type,value,comment,to_ids,date,object_relation,attribute_tag,object_uuid,object_name,object_meta_category
"542e4cbd-ee78-4a57-bfb8-1fda950d210b",1,"External analysis","link","http://labs.opendns.com/2014/10/02/opendns-and-
bash/","",0,1412320445,"","","","",""
"542e4cbe-d560-4e14-9157-1fda950d210b",1,"External analysis","link","https://gist.github.com/andrewsmhay/
de1cdc63d04c2bbf8c12","",0,1412320446,"","","","",""
"542e4cbe-12a4-4345-b0a4-1fda950d210b",1,"External analysis","link","https://gist.githubusercontent.com/andrewsmhay/de1cdc63d04c2bbf8c12/raw/
f20402cf5a0c646c63c4521f60587703fe654443/iplist","",0,1412320446,"","","","",""
"542e4ccc-b8fc-44af-959d-6ead950d210b",1,"External analysis","text","Shellshock","",0,1412320460,"","","","",""
"542e4ce7-6120-41c0-8793-e90e950d210b",1,"External analysis","comment","Data encoded by David AndrÃ©","",0,1412320487,"","","","",""
"542e4cfe-21ac-46a7-9d82-06b3950d210b",1,"Network activity","ip-src","1.48.209.68","",1,1412320510,"","","","",""
"542e4cfe-05f4-46ab-b5b8-06b3950d210b",1,"Network activity","ip-src","1.73.227.172","",1,1412320510,"","","","",""
"542e4cfe-81c4-45f2-9e67-06b3950d210b",1,"Network activity","ip-src","1.162.58.214","",1,1412320510,"","","","",""
"542e4cfe-0ff4-4a93-aef8-06b3950d210b",1,"Network activity","ip-src","1.163.34.29","",1,1412320510,"","","","",""
"542e4cfe-7a98-4c98-a862-06b3950d210b",1,"Network activity","ip-src","1.192.158.169","",1,1412320510,"","","","",""
"542e4cff-1460-47b9-83c6-06b3950d210b",1,"Network activity","ip-src","1.234.62.246","",1,1412320511,"","","","",""
"542e4cff-dfc8-4022-93f6-06b3950d210b",1,"Network activity","ip-src","2.25.130.80","",1,1412320511,"","","","",""
"542e4cff-4f84-478a-97ef-06b3950d210b",1,"Network activity","ip-src","2.25.141.149","",1,1412320511,"","","","",""
"542e4cff-188c-4c0b-8b54-06b3950d210b",1,"Network activity","ip-src","2.28.163.125","",1,1412320511,"","","","",""
"542e4cff-6af8-4329-8c0b-06b3950d210b",1,"Network activity","ip-src","2.28.232.183","",1,1412320511,"","","","",""
"542e4cff-61d4-461c-8b9f-06b3950d210b",1,"Network activity","ip-src","2.33.214.24","",1,1412320511,"","","","",""
"542e4cff-e99c-4399-b60e-06b3950d210b",1,"Network activity","ip-src","2.38.41.213","",1,1412320511,"","","","",""
"542e4cff-989c-429a-9549-06b3950d210b",1,"Network activity","ip-src","2.97.3.104","",1,1412320511,"","","","",""
```

Figure 2. MISP information.

Figure 3. Process to integrate CTI data into SWAT dataset.

After incorporating CTI data into the SWAT dataset to enrich its network traffic information, an additional step was undertaken to further enhance the dataset by introducing authorization features. This process was informed by a combination of threat intelligence and domain-specific knowledge of the ICS. To strictly prevent data leakage, the dataset was partitioned using a chronological split prior to feature engineering. Lists of authorized source IP addresses, destination IP addresses, and network ports were curated a priori, relying exclusively on CTI insights and legitimate network behaviors observed solely within the training partition. Using these isolated lists, three new binary features—*is_src_authorized*, *is_dst_authorized*, and *is_port_authorized*—were added to indicate authorization status, guaranteeing that the testing phase remained entirely blinded and performance metrics were not artificially inflated. These features serve as binary indicators of the legitimacy of

each network interaction, with a value of 1 denoting an authorized entity and 0 representing unauthorized activity. The addition of these features results in a dataset that is not only enriched with historical threat intelligence but also contains critical information regarding the legitimacy of network entities, derived from operational knowledge. This augmentation facilitates a more precise and context-aware analysis of network traffic, enabling the model to more effectively distinguish between benign and malicious activities. By integrating both historical threat context and operational domain knowledge, this approach enhances the model's capability to detect sophisticated and evolving cyber threats, thereby contributing to a more robust and adaptive anomaly detection system for safeguarding critical infrastructure.

4.2.2. Mapping Dataset Scenarios to MITRE ATT&CK Tactics

Following the integration of CTI into the SWAT dataset and the development of authorization features to further contextualize legitimate network behavior, the next critical phase involved leveraging the MITRE ATT&CK for ICS framework to systematically classify attack scenarios. This classification process provides a structured approach to understanding potential attack vectors within an ICS environment, thereby informing the development of effective defense strategies.

To determine the appropriate classifications, we began with an extensive analysis of the 94 techniques outlined in the MITRE ATT&CK for ICS framework. The final selection was based on a direct mapping to the ground truth of the 26 successful attacks simulated in the SWAT dataset. We analyzed the documented intent and impact of each scenario to categorize them under the most appropriate high-level tactic. The chosen tactics—Manipulation of Control (T0831), Modify Parameter (T0836), Service Stop (T0881), and Damage to Property (T0879)—were selected because they strictly encompass the entirety of the 26 physically impactful adversarial goals demonstrated within the dataset. Tactics restricted to reconnaissance or lateral movement, which produced no measurable physical state deviation, were deliberately excluded. Consequently, the current GCN-LSTM architecture operates under a closed-world assumption. When confronted with unseen or zero-day tactics outside this training distribution, the model successfully detects the subsequent cyber-physical deviations as anomalies but will mathematically map the novel attack to the closest learned spatio-temporal disruption pattern. Adapting the framework with open-set recognition capabilities to isolate and label entirely novel tactics represents a necessary vector for future research. The classification serves multiple purposes, including providing insights into the most vulnerable areas within the ICS, identifying potential attack vectors, and creating the foundation for enhancing detection capabilities.

A detailed description of all 41 attack scenarios present in the original SWAT dataset is provided in Table 2. From this comprehensive list, our analysis focused on the 26 attacks that produced a discernible impact on the system. We systematically mapped each of these 26 scenarios to one of four overarching MITRE ATT&CK tactics, which serve as the final class labels for our model. Table 4 summarizes these four tactics and provides illustrative examples of the attack scenarios that were mapped to each category.

The integration of CTI and the use of selected MITRE ATT&CK for ICS techniques were tailored to align with the information contained in the SWAT dataset, ensuring that the process effectively captures the unique characteristics of the system under study. This approach enhances the model's ability to identify and respond to complex cyber threats, leveraging real-world threat intelligence to improve anomaly detection accuracy and resilience.

While the current study is specific to the SWAT dataset, the methodology—including CTI enrichment and MITRE ATT&CK-based classification—provides a flexible and adaptive

framework that can be applied to other CI environments. By adjusting the selection of techniques to fit the specific characteristics of different CI systems, this approach has the potential to effectively secure a variety of critical sectors, offering a comprehensive solution adaptable to diverse operational contexts.

Table 4. Mapping of MITRE ATT&CK for ICS Categories to SWAT Attack Scenarios.

Tactic ID	Tactic Name	Adversarial Goal and Description	Representative SWAT Scenarios
T0831	Manipulation of Control	Adversaries gain unauthorized control over physical actuators (valves, pumps) to disrupt the intended process flow.	• Attack No. 1: Opening valve MV-101, leading to a tank overflow. • Attack No. 17: Preventing valve MV-303 from opening, causing an overflow.
T0836	Modify Parameter	Adversaries alter system setpoints, thresholds, or device parameters to force components to operate beyond safe limits.	• Attack No. 3: Abnormally increasing the LIT-101 water level reading, leading to underflow. • Attack No. 6: Modifying the LIT-301 setpoint to induce overflow.
T0881	Service Stop	Adversaries force the shutdown of critical components, halting normal operations and causing service interruptions.	• Attack No. 28: Closing pump P-302 to stop inflow to a tank. • Attack No. 22: Stopping the UV-401 process, leading to a system freeze.
T0879	Damage to Property	Adversaries execute actions specifically designed to cause physical damage to ICS components, leading to costly repairs and downtime.	• Attack No. 2: Turning on P-102 while P-101 is active, causing a pipe burst. • Attack No. 32: Setting LIT-301 to exceed safe limits, damaging pump P-302.

5. Model Design and Results

Our model's architecture is founded on a hybrid design that synergistically combines Graph Convolutional Networks (GCNs) for spatial analysis with Long Short-Term Memory (LSTM) networks for temporal analysis. This approach is motivated by the dual nature of CI data, which possesses both spatial structure (the static and dynamic relationships between devices) and temporal structure (the evolution of device states over time). For each time step t, the system's state is captured as a graph, $G\_t = (V, E)$, where nodes V represent entities like sensors and IP addresses, and edges E signify interactions between them. The GCN layers process each graph snapshot to extract high-level spatial features, learning the complex patterns of inter-device relationships. The output is a compact vector embedding for the entire graph at time t. This sequence of graph embeddings is then fed into the LSTM layers, which excel at identifying temporal patterns and dependencies across the sequence. This dual process allows the model to learn not just what a system's state is, but how it is evolving, enabling a more nuanced and accurate detection of sophisticated, multi-stage attacks.

The GCN component is responsible for learning relational features from each graph snapshot. While other advanced architectures, such as Graph Attention Networks (GATs), employ dynamic attention mechanisms to prioritize neighbor interactions, a GCN was explicitly selected for this framework due to the static physical topology of the SWAT testbed. In industrial control systems with fixed sensor-actuator configurations, the unweighted, symmetric message-passing mechanism of a GCN captures permanent spatial dependencies more efficiently, avoiding the computational overhead and overfitting risks associated with GATs in rigid physical environments. This structural advantage allows our model to scale into multiclass tactic classification, outperforming the operational utility of recent binary GAT baselines (e.g., [32]). The GCN operates through this iterative message-passing mechanism, where each node aggregates information from its neighbors. This process is formally defined as follows. For each node $v \in V$, its initial feature vector at layer $k = 0$ is

given as $h_v^{(0)}$. During each layer k, GCN performs the following update to compute new node features:

$$h_v^{(k)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \frac{1}{c_{vu}} W^{(k)} \boldsymbol{h_u^{(k-1)}} + b^{(k)}\right)$$

where:

- $h_v^{(k)}$ is the hidden representation of node v at layer k.
- $\mathcal{N}(v)$ is the set of neighbors of node v.
- $W^{(k)}$ is the weight matrix of the k-th layer.
- $b^{(k)}$ is the bias term.
- c_{vu} is a normalization constant.
- σ is the activation function [11].

After applying multiple GCN layers, global mean pooling is used to summarize the graph-level information: $z_t = \text{GlobalMeanPool}\left(\left\{h_v^{(K)} \mid v \in V\right\}\right)$, where z_t is the pooled feature vector representing the entire graph at time t.

Then, the sequence of graph-level embeddings $\{z_1, z_2, \ldots, z_T\}$ is passed through the LSTM to capture temporal relationships. The LSTM update equations are as follows:

$$f_t = \sigma\left(W_f z_t + U_f h_{t-1} + b_f\right)$$

$$i_t = \sigma(W_i z_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma(W_o z_t + U_o h_{t-1} + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c z_t + U_c h_{t-1} + b_c)$$

$$h_t = o_t \odot \tanh(c_t)$$

where:

- f_t, i_t, o_t are the forget, input, and output gates, respectively, at time step t.
- c_t is the cell state, and h_t is the hidden state at time step t.
- W_*, U_*, and b_* are learnable weight matrices and bias vectors.
- σ is the sigmoid activation function, and tanh is the hyperbolic tangent activation function.
- $\odot$ denotes element-wise multiplication [30].

After processing the entire sequence, we extract the hidden state from the last time step (h_T) and pass it through a fully connected layer for the final prediction: $y = W_{\text{out}} h_T + b_{\text{out}}$, where:

- y is the output (e.g., class scores for classification).
- W_{out} and b_{out} are the weights and bias of the fully connected layer.

The complete architecture of our proposed model, from data enrichment to final classification, is illustrated in Figure 4. To feed the proposed model, an additional processing step involves creating a graph-based data structure. The process begins by identifying unique nodes, such as source IPs, destination IPs, SCADA tags, and ports, and mapping these nodes to maintain consistency across graphs. The data is then grouped by second-level timestamps to create a graph for each time step, capturing the state of the network at that moment. Edges are formed between nodes to represent interactions, preserving temporal relationships. This methodology of transforming sequential network snapshots into structured graphs is strongly supported by the latest research in the field.

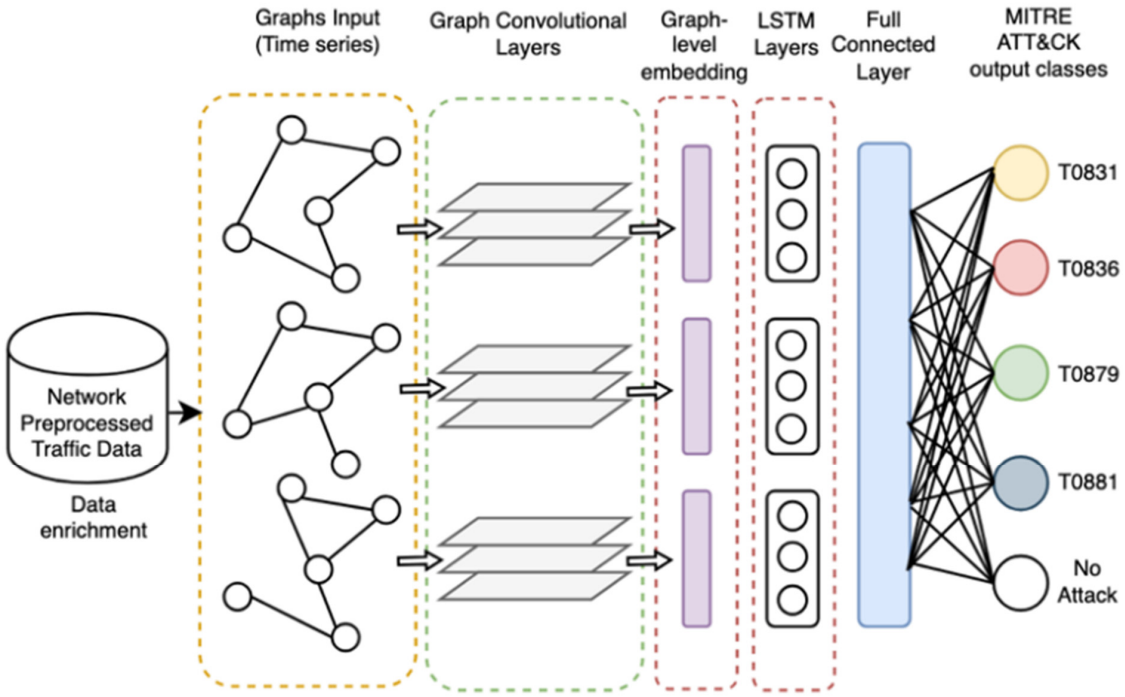


Figure 4. GCN-LSTM architecture.

Recent applications of hybrid GNN-LSTM architectures have demonstrated that capturing network states as discrete temporal graphs allows the GNN to model behavioral similarity at any given second, while the LSTM effectively tracks how these malicious patterns evolve across multi-stage attacks [33]. Node features, such as Modbus transaction IDs and authorization attributes, are assigned based on operational data, with Modbus values specifically allocated to destination nodes. Authorization features are enriched using CTI to provide additional context for network activity. Each graph is labeled with the most frequent MITRE ATT&CK technique observed, making it suitable for model training. The resulting PyTorch Geometric (v2.5.2) Data objects represent a time series of graphs, which are subsequently fed into the GCN-LSTM model.

The pseudocode for constructing the graph-based data structures is detailed in Algorithm 1:

Algorithm 1: Creation of Spatio-Temporal Graph Data Structure

Input: Preprocessed network traffic and sensor data, CTI authorization attributes
Output: graph_list (Sequence of PyTorch Geometric Data objects)

```
Extract unique nodes from src, dst, SCADA_Tag, and s_port features.
Generate a consistent mapping assigning a unique ID to each node v ∈ V.
Group dataset records by discrete second-level timestamps t.
for each timestamp group t do
    Initialize edge list E_t ← ∅ and edge timestamps T_t ← ∅.
    for each record in group t do
        Construct edges connecting src, dst, SCADA_Tag, and s_port nodes.
        E_t ← E_t ∪ {edge}, T_t ← T_t ∪ {timestamp}.
    end for
  Initialize node feature matrix X_t for all v ∈ V.
  Assign categorical features and CTI attributes (is_src_auth, etc.) to X_t.
  Assign Modbus sensor values to corresponding destination nodes in X_t.
  y_t ← majority MITRE ATT&CK technique label for group t.
  Construct graph object G_t = (X_t, E_t, T_t, y_t).
  Append G_t to graph_list.
end for
return graph_list
```

The forward pass and spatio-temporal execution sequence of the proposed GCN-LSTM network are detailed in Algorithm 2:

Algorithm 2: GCN-LSTM Spatio-Temporal Forward Pass

Input: graph_list = {G_1, G_2, . . ., G_T}, initialized hyperparameters
Output: Predicted MITRE ATT&CK tactic $\hat{y}$

```
Initialize GCN layers, LSTM layers, and Fully Connected (FC) layer.
Initialize empty sequence tensor S ← ∅.
for each graph G_t ∈ graph_list do
    h_t^(0) ← extract_node_features(G_t)
    for k = 1 to num_layers do
        h_t^(k) ← GCNConv(h_t^(k−1), G_t.E_t)
        h_t^(k) ← Dropout (LeakyReLU(h_t^(k)))
    end for
    z_t ← GlobalMeanPool(h_t^(K))
  Append z_t to S
end for
H ← LSTM(S)
h_T ← Extract final hidden state from H
ŷ ← FC(h_T)
return ŷ
```

A significant challenge in training the architecture was the inherent class imbalance within the SWAT dataset, where normal operational traffic heavily outweighs anomalous traffic, and certain attack tactics (such as T0831) are underrepresented. To mitigate this during the training phase, we implemented a class-weighted Cross-Entropy loss function. By assigning higher penalty weights to misclassifications of minority attack classes, the gradient updates are prevented from being disproportionately dominated by the 'No_Attack' class, encouraging the network to better learn the subtle feature representations of less frequent cyber-physical manipulations.

To optimize the model and guarantee reproducibility, we conducted a comprehensive hyperparameter search using Bayesian Optimization with the Tree-structured Parzen Estimator (TPE) sampler. A summary of the explored ranges and optimal selected values is presented in Table 5. Based on this optimization, the final deployed architecture consists of a single GCN layer (123 hidden units) utilizing a Leaky ReLU activation (negative slope = 0.1055) and a dropout rate of 0.2475, followed by a single-layer LSTM (69 hidden units). The model was trained using the Adam optimizer (learning rate = 1.384×10^{-4}) and a class-weighted Cross-Entropy loss function over a chronological 80/20 train-test split, ensuring the reported performance reflects genuine generalization to unseen temporal sequences rather than training-set memorization. An analysis of architectural complexity established strict operational bounds. Reducing the hidden representations below minimum thresholds degraded the macro F1-score by failing to capture multi-stage attack nuances. Conversely, increasing the depth beyond a single GCN or LSTM layer yielded diminishing returns; deeper GCNs induced over-smoothing—rendering distinct physical sensor nodes mathematically indistinguishable—while deeper LSTMs merely inflated computational latency without corresponding accuracy gains. To validate the robustness of this minimal-complexity configuration, a sensitivity analysis demonstrated that variations within $\pm 15\%$ of the optimal learning rate resulted in a marginal macro F1-score variance of less than 0.02. Similarly, scaling the LSTM hidden dimension between 64 and 128 units-maintained baseline accuracy above 98.5%, confirming a robust optimal region. To ensure absolute independent reproducibility, the complete Python implementation and trained network weights are publicly archived via the provided Zenodo repository."

Table 5. Hyperparameter space search.

Hyperparameter	Range	Optimal Selected Value
Hidden Dimension (hidden_dim)	16–128	123
LSTM Hidden Dimension (lstm_hidden_dim)	32–256	69
Number of Layers (num_layers)	1–3	1
Dropout Rate (dropout_rate)	0.0–0.5	0.2475
Activation Function (activation)	ReLU, Leaky ReLU, ELU	Leaky ReLU
Learning Rate (lr)	1×10^{-4}–1×10^{-2} (log scale)	0.0001384
Batch Size (batch_size)	128–1024	808
Early Stopping Patience (early_stopping_patience)	5–15	13
Number of LSTM Layers (num_lstm_layers)	1–3	1
Leaky ReLU Negative Slope (negative_slope)	0.01–0.2	0.1055

5.1. Experimental Results

The machine learning procedures were carried out on a server equipped with an Intel Xeon Silver 4310 CPU @ 2.10 GHz, running Ubuntu 22.04.3 LTS. The system utilized NVIDIA-SMI 525.147.05 with Driver Version 525.147.05 and CUDA Version 12.0 for GPU acceleration. Deep neural network models were implemented using Torch 1.13.1, with hyperparameter optimization performed via Optuna 3.2.0, in conjunction with Python 3.10.12.

The evaluation of the proposed model was conducted using a variety of metrics, assessed both at the per-class level and globally to provide a comprehensive understanding of model performance. Specifically, precision, recall, and F1-score were computed for each class to determine the model's effectiveness in detecting different attack types and normal activities. This per-class evaluation is crucial for understanding the model's sensitivity and specificity across diverse, often imbalanced, classes—particularly important for critical infrastructure security where specific attack scenarios may be rare but highly impactful. Additionally, overall accuracy, weighted F1, and macro F1-score were computed to provide a global perspective on model performance. The weighted F1-scores accounts for the frequency of each class, providing insights into the model's general effectiveness even when certain classes are underrepresented. In contrast, the macro F1-score treats all classes equally, highlighting the model's balanced capability to handle various attack types. The combination of these metrics ensures a detailed and robust evaluation, emphasizing the model's reliability in distinguishing between multiple types of anomalous and normal behaviors.

The evaluation results of the model demonstrate strong overall performance in detecting various types of network activities, achieving an accuracy of 99.04%. The per-class precision shows a high ability to correctly identify positive instances for most classes. Specifically, for the 'No_Attack' class, the model achieved a precision of 1.00, indicating perfect precision in identifying benign activities. The model's high precision for the 'No_Attack' class is partially attributed to the CTI-derived authorization features, which provided a strong signal for distinguishing legitimate from unauthorized network traffic. However, for the 'T0831' technique (Manipulation of Control), the precision drops to 0.532, which suggests potential challenges in correctly identifying this attack type, likely due to class imbalance. Moreover, this lower precision on the T0831 class suggests that some normal operational changes to valves or actuators may share features with manipulative attacks, presenting a challenge for the model. The precision for 'T0879' (Damage to Property) and

'T0836' (Modify Parameter) techniques are 0.939 and 0.992, respectively, indicating robust detection of these attack types, while 'T0881' (Service Stop) achieved 0.998.

The recall scores indicate that the model effectively detects true positive cases across most classes, with 'No_Attack', 'T0836', and 'T0881' all achieving recall values close to 1.0, suggesting nearly perfect detection rates. However, 'T0879' achieved a recall of 0.982, and 'T0881' showed a lower recall at 0.867, suggesting that the model may occasionally miss instances of these attacks. As illustrated in Table 6.

Table 6. Results of the GCN-LSTM per class.

Mitre Technique ID	Class Name	Precision	Recall	F1-Score
No_Attack	Normal Activity	1.0000	0.9999	0.9999
T0831	Manipulation of Control	0.5320	0.9854	0.6910
T0879	Damage to Property	0.9394	0.9819	0.9602
T0836	Modify Parameter	0.9924	1.0000	0.9962
T0881	Service Stop	0.9978	0.8675	0.9281

A per-class comparison with the models cited in Table 7 is not feasible, as those works frame the problem as binary classification ('Normal' vs. 'Anomaly') and thus do not report performance against specific attack types. This limitation in existing literature highlights a significant contribution of our methodology. By performing multiclass classification mapped to MITRE ATT&CK tactics, we provide a more granular and operationally relevant evaluation, demonstrating the model's effectiveness in distinguishing the nature of different threats, not just their presence.

Table 7. Comparison results among different studies.

Ref.	Accuracy	Recall	Precision	F1-Score
[22]	NR	0.5909	0.9986	0.7496
[23]	0.9686	0.9463	0.9668	0.9564
[24]	NR	0.8485	0.9602	0.9009
[28]	NR	0.8992	0.9627	0.9298
[34]	NR	0.7040	0.9819	0.8200
[35]	NR	0.7593	0.9761	0.8541
GCN-LSTM (OURS)	0.9904	0.9669	0.8923	0.9151

NR: Not Reported in the source paper. Our model performs multiclass classification; other models are binary. Metrics are macro-averaged for fair comparison.

The F1-score, which balances precision and recall, highlights consistent performance across most classes. For 'No_Attack', the F1-score is 0.999, reflecting the model's ability to accurately classify normal activities. Our experimental results provide strong support for the claim that GNNs reduce false positives. As shown in Table 6, the model achieved a precision of 1.0000 for the 'No_Attack' class. This is a critical finding, as it indicates that zero instances of normal activity were misclassified as an attack during testing. For the attack classes, 'T0831' achieved an F1-score of 0.691, indicating the need for improvement in distinguishing this attack type, whereas 'T0879', 'T0836', and 'T0881' all have F1-scores above 0.92, showcasing the model's robustness in classifying these attack types.

While the overall accuracy of 99.04% and the weighted F1-score of 0.9915 reflect highly accurate general classification, these globally weighted metrics are inherently influenced by the overwhelming volume of benign operational traffic. Although the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is frequently utilized to offset such

class imbalance, it can yield overly optimistic assessments in extreme scenarios; the massive number of true negatives (the 'No_Attack' class) artificially depresses the false positive rate. Consequently, this study deliberately bypasses AUC-ROC to evaluate discriminative robustness using the stricter Macro F1-score (0.9151) and class-specific F1 metrics. By computing the harmonic mean of precision and recall independently for each class and unweighting the average, the Macro F1-score strictly prevents the majority class from masking minority performance, providing a highly rigorous, transparent, and penalizing assessment of the model's true capability to detect rare cyber-physical manipulations. This high performance is directly attributable to the model's architecture, where the GCN layers effectively model the complex spatial relationships between ICS components and the LSTM layers capture the temporal evolution of attack sequences, enabling a nuanced distinction between benign and malicious patterns. These results indicate that the model is highly suitable for detecting both normal and anomalous behaviors in critical infrastructure.

To demonstrate the model's practical detection capability, we analyze its response to Attack No. 3 from the SWAT dataset. In this scenario, an adversary manipulates the sensor readings for 'LIT-101', causing the value to increase by a constant 1 mm every second. This manipulation is intended to deceive the control system, potentially leading to a tank underflow and physical damage to pump P-101. As shown in Figure 5, this attack creates a distinct and unnatural linear trend (in red) that stands out from the normal operational fluctuations (in blue).

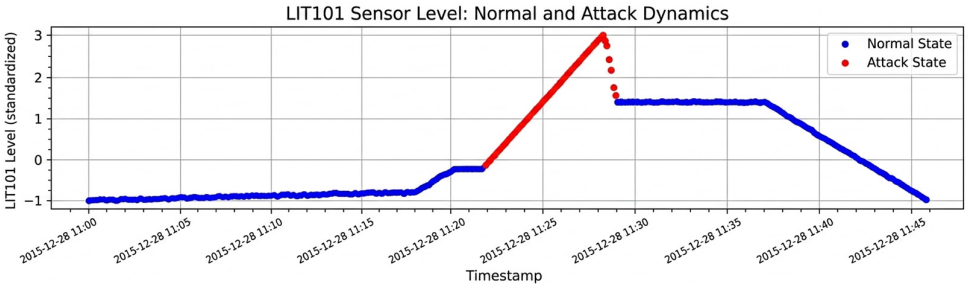


Figure 5. Time-series plot of the LIT-101 sensor during Attack No. 3. The red segment highlights the malicious, constant increase in the sensor's reported value, characteristic of the 'Modify Parameter' attack.

Our GCN-LSTM model successfully identifies this attack. The GCN layers process the graph at each time step and detect a significant deviation in the feature values associated with the LIT-101 node. Crucially, the model also processes the features of neighboring nodes—such as the status of inflow valve MV-101 and pump P-101—and recognizes that their states do not correlate with the rapidly rising water level. The LSTM layers then identify this sustained, uncorrelated increase as a malicious temporal pattern, distinct from normal fluctuations. This spatio-temporal analysis leads to a confident classification of the activity as T0836 (Modify Parameter), providing the operator with specific, actionable insight into the nature of the attack.

In comparison to prior works, as illustrated in Table 7, our GCN-LSTM model demonstrates superior performance across several key metrics, highlighting its effectiveness in anomaly detection within critical infrastructure environments. Specifically, our model achieved an accuracy of 99.04%, surpassing the best accuracy reported by previous models, such as [23] with 96.86%, demonstrating a significantly higher capability to correctly classify network traffic and physical device data. Our recall score of 96.69% highlights the model's robustness in detecting true positive cases, a substantial improvement over studies like [22] (59.09%) and [24] (84.85%). This high recall is particularly crucial for critical infrastructure, where missing an attack could have severe consequences. Furthermore, our model

achieved a precision of 89.23%, reflecting its effectiveness in minimizing false positives. While our model's precision is competitive, the work in [22] reports an exceptionally high precision of 99.86%, though as noted above, it suffers from a significantly lower recall. Our model achieves a more balanced performance, which is vital for operational efficiency. Additionally, the F1-score of our model (91.51%) indicates a balanced performance between precision and recall, outperforming studies like [34] with an F1-score of just 0.82. Moreover, unlike most previous studies, which relied solely on physical device data, our model incorporates both network traffic and physical device data, providing a more comprehensive view of potential threats and leading to more accurate anomaly detection. This integrated approach, combined with the capability to perform multiclassification of anomalies aligned with MITRE ATT&CK techniques, makes our model not only highly accurate but also contextually aware, thus offering a more adaptive and effective solution for protecting critical infrastructure.

Furthermore, to systematically isolate the contributions of both the external feature enrichment and the internal architectural modules, a comprehensive multi-step ablation study was conducted. First, to assess reliance on the engineered CTI authorization features (is_src_authorized, is_dst_authorized, is_port_authorized), a baseline model trained strictly on raw physical and network features yielded an accuracy of 95.28% and a macro F1-score of 0.8645. Second, to evaluate the structural components, the spatial and temporal modules were independently ablated. A GCN-only architecture (omitting the LSTM layers) achieved a macro F1-score of 0.8420, demonstrating a degraded ability to track prolonged, multi-stage attack sequences. Conversely, an LSTM-only configuration (omitting the GCN layers) resulted in a macro F1-score of 0.8115, struggling to map simultaneous topological deviations across the ICS testbed. These stepwise deltas mathematically confirm that while CTI enrichment is vital for minimizing false positives, the synergistic spatio-temporal fusion of the GCN and LSTM layers is the indispensable core driver of the model's peak 0.9151 macro F1-score.

5.2. Model Interpretability

A key contribution of the proposed GNN-based architecture is its systemic interpretability. To quantitatively establish transparency across the entire classification space, GNN explainer attribution weights were aggregated across all instances of the four targeted MITRE ATT&CK tactics. This global analysis confirmed that the CTI-derived is_src_authorized feature and the specifically targeted physical sensor nodes consistently ranked within the top 7th percentile of attribution importance regardless of the attack class, mathematically proving a systemic, logical feature hierarchy. To operationalize this global finding, a localized case study of Attack No. 3 (T0836: Modify Parameter) was analyzed, as shown in Figure 5. The objective was to identify the specific nodes and causal relationships driving this individual classification. The analysis revealed that the top 3 most influential factors in the model's decision were:

- The LIT-101 sensor node, which was the direct target of data manipulation.
- The is_src_authorized feature, specifically its value indicating that the traffic originated from an unauthorized source.
- The P-101 pump node, whose operation is directly dependent on the LIT-101 sensor readings.

This result is significant for two reasons. First, it experimentally validates our CTI-driven feature engineering, confirming that the authorization features provide a powerful signal for the model. Second, and more importantly, the high influence of the P-101 pump node demonstrates that the model is not merely correlating network features; it is learning the physical topology and causal relationships of the critical infrastructure itself. It correctly

identified that a primary consequence of manipulating the LIT-101 sensor is the operational state of the P-101 pump. This provides a transparent outcome that links a cyber event to its potential physical impact, which is crucial for security analysts and system operators who need to understand the "why" behind an alert for effective incident response.

6. Conclusions and Future Work

This study presents a foundational proof-of-concept for a novel detection model that integrates Cyber Threat Intelligence (CTI) and classifies cyberattacks using the MITRE ATT&CK framework, leveraging GCNs to enhance anomaly detection in Critical Infrastructure (CI). Rather than asserting a universal defense mechanism, our research provides a robust architectural blueprint that addresses the evolving cyber threat landscape within specific dual-modality environments.

The main contributions of this work are as follows: First, we developed a CTI-enhanced detection model that demonstrates the potential to improve contextual awareness, aiding in the detection of known threats and allowing for better response actions. Second, our use of the MITRE ATT&CK framework ensures a structured and standardized method for attack classification, providing a deeper understanding of adversary tactics. The operational importance of this multiclass approach is significant: for a security analyst, a binary 'anomaly' alert is of limited value. In contrast, an alert from our model specifying, for instance, 'T0879—Damage to Property' provides immediate, actionable intelligence about the adversary's intent, enabling a prioritized incident response. Third, the application of GNNs allows the model to capture complex relationships within network traffic data, improving its ability to isolate sophisticated cyber threats while reducing false positives. Furthermore, we integrated post hoc interpretability features to help identify potential attack sources, adding a crucial layer of transparency to the detection process.

The experimental results, demonstrated on the SWaT testbed, indicate the efficacy of our approach within this specific context. The model's robustness is highlighted by its high macro F1-score of 0.9151 on a highly imbalanced dataset. This metric gives equal weight to all classes, proving the model's capability to detect underrepresented attack types rather than overfitting to the majority 'No_Attack' class.

However, the reliance on a single benchmark dataset—necessitated by the model's strict requirement for synchronized IT-network traffic and OT-physical sensor telemetry—constitutes a primary limitation of this study. Consequently, broad claims regarding the model's universal effectiveness across diverse CI sectors cannot yet be fully substantiated. To ensure these findings are transparent and independently reproducible, the exact network topology, optimal hyperparameters, and training pipelines have been rigorously documented and open-sourced.

Future research will focus on several critical pathways. First, while the current architecture successfully classifies anomalies into high-level MITRE ATT&CK tactics, future iterations will aim to achieve sub-tactic, technique-level granularity. Advancing the model to automatically distinguish between specific techniques—such as differentiating a generalized 'Modify Parameter' (T0836) attack from a specific 'Spoof Reporting Message' manipulation—would provide even deeper intelligence for security orchestration. Second, validating this open-source model against emerging multi-modality CI datasets is an essential requirement to prove its generalizability across a broader range of industrial environments. Finally, while the proposed architecture achieves high detection efficacy, the computational complexity inherent to fusing spatial GCNs with temporal LSTM layers presents a potential bottleneck for real-time industrial deployment. Future research must rigorously profile the model's inference latency, memory footprint, and overall resource consumption. Exploring model optimization techniques—such as graph pruning, quantiza-

tion, or knowledge distillation—will be a critical priority to ensure the framework strictly adheres to the low-latency constraints required for active threat mitigation in time-sensitive cyber-physical environments.

Author Contributions: A.P. was responsible for the study conceptualization, methodology, formal analysis, investigation, data curation, visualization, and writing the original draft. Y.D. contributed to the conceptualization, formal analysis, investigation, validation, resource provision, manuscript review and editing, and overall supervision. L.-C.H. was involved in validation, investigation, and the review and editing process. J.G. contributed to validation, investigation, manuscript review and editing, and provided supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Universidad de los Andes (research support awarded to corresponding author A.P.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study, along with the graph neural network implementation code, are available in the Zenodo repository at https://zenodo.org/records/18498474 (accessed on 5 May 2026). In the course of drafting this manuscript, the Gemini large language model was employed specifically to enhance linguistic clarity and overall readability. Following the application of this tool, the authors meticulously reviewed, refined, and validated all text, thereby assuming complete responsibility for the final published content.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ATT&CK	Adversarial Tactics, Techniques, and Common Knowledge
bAE	Basic Autoencoder
CI	Critical Infrastructures
CIP	Common Industrial Protocol
CPS	Cyber-Physical Systems
CPU	Central Processing Unit
CTI	Cyber Threat Intelligence
dAE	Deep Autoencoder
DGCB	Dynamic Graph Construct Block
DoS	Denial of Service
DPDGAD	Dual-Process Dynamic Graph-Based Anomaly Detection
DSRM	Design Science Research Methodology
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GLIN	Global-Local Integration Network
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
ICS	Industrial Control Systems
IDS	Intrusion Detection Systems
IoC	Indicator of Compromise

IoT	Internet of Things
IT	Information Technology
LSTM	Long Short-Term Memory
MISP	Malware Information Sharing Platform
NIST	National Institute of Standards and Technology
OCSVM	One-Class Support Vector Machine
OT	Operational Technology
PLC	Programmable Logic Controller
SCADA	Supervisory Control and Data Acquisition
SRC	System Requirement Collection
SWAT	Secure Water Treatment
TGP	Trend-Guided Predictor
TPE	Tree-structured Parzen Estimator
TTP	Tactics, Techniques, and Procedures

References

1. Federal Bureau of Investigation. "Internet Crime Report 2021" Washington, DC, USA, 2022. Available online: https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf (accessed on 21 May 2026).
2. U.S. Government Accountability Office. "Critical Infrastructure Protection: TSA Is Taking Steps to Address Some Pipeline Security Program Weaknesses," Washington, DC, USA, GAO-21-105263, July 2021. Available online: https://www.gao.gov/products/gao-21-105263 (accessed on 25 May 2026).
3. Cybersecurity and Infrastructure Security Agency (CISA). Compromise of U.S. Water Treatment Facility. Available online: https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-042a (accessed on 17 October 2024).
4. Zahid, S.; Mazhar, M.S.; Abbas, S.G.; Hanif, Z.; Hina, S.; Shah, G.A. Threat modeling in smart firefighting systems: Aligning MITRE ATT&CK matrix and NIST security controls. *Internet Things* **2023**, *22*, 100766. [CrossRef]
5. Imran, M.; Siddiqui, H.U.R.; Raza, A.; Raza, M.A.; Rustam, F.; Ashraf, I. A performance overview of machine learning-based defense strategies for advanced persistent threats in industrial control systems. *Comput. Secur.* **2023**, *134*, 103445. [CrossRef]
6. Preuveneers, D.; Joosen, W. Sharing Machine Learning Models as Indicators of Compromise for Cyber Threat Intelligence. *J. Cybersecur. Priv.* **2021**, *1*, 140–163. [CrossRef]
7. Mashima, D. MITRE ATT&CK Based Evaluation on In-Network Deception Technology for Modernized Electrical Substation Systems. *Sustainability* **2022**, *14*, 1256. [CrossRef]
8. Pirca, A.M.; Lallie, H.S. An empirical evaluation of the effectiveness of attack graphs and MITRE ATT&CK matrices in aiding cyber attack perception amongst decision-makers. *Comput. Secur.* **2023**, *130*, 103254. [CrossRef]
9. Yousaf, A.; Zhou, J. From sinking to saving: MITRE ATT &CK and D3FEND frameworks for maritime cybersecurity. *Int. J. Inf. Secur.* **2024**, *23*, 1603–1618. [CrossRef]
10. Kim, Y.; Lee, I.; Kwon, H.; Lee, K.; Yoon, J. BAN: Predicting APT Attack Based on Bayesian Network with MITRE ATT&CK Framework. *IEEE Access* **2023**, *11*, 91949–91968. [CrossRef]
11. Wang, Y.; Liu, J.; Qian, G. Hierarchical FFT-LSTM-GCN based model for nuclear power plant fault diagnosis considering spatio-temporal features fusion. *Prog. Nucl. Energy* **2024**, *171*, 105178. [CrossRef]
12. Tran, D.H.; Park, M. FN-GNN: A Novel Graph Embedding Approach for Enhancing Graph Neural Networks in Network Intrusion Detection Systems. *Appl. Sci.* **2024**, *14*, 6932. [CrossRef]
13. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [CrossRef]
14. Xiong, W.; Legrand, E.; Åberg, O.; Lagerström, R. Cyber security threat modeling based on the MITRE Enterprise ATT&CK Matrix. *Softw. Syst. Model.* **2022**, *21*, 157–177. [CrossRef]
15. Ahn, G.; Jang, J.; Choi, S.; Shin, D. Research on Improving Cyber Resilience by Integrating the Zero Trust Security Model with the MITRE ATT&CK Matrix. *IEEE Access* **2024**, *12*, 89291–89309. [CrossRef]
16. Alaeifar, P.; Pal, S.; Jadidi, Z.; Hussain, M.; Foo, E. Current approaches and future directions for Cyber Threat Intelligence sharing: A survey. *J. Inf. Secur. Appl.* **2024**, *83*, 103786. [CrossRef]
17. Afenu, D.S.; Asiri, M.; Saxena, N. Industrial Control Systems Security Validation Based on MITRE Adversarial Tactics, Techniques, and Common Knowledge Framework. *Electronics* **2024**, *13*, 917. [CrossRef]
18. Zhang, S.; Chen, P.; Bai, G.; Wang, S.; Zhang, M.; Li, S.; Zhao, C. An Automatic Assessment Method of Cyber Threat Intelligence Combined with ATT&CK Matrix. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7875910. [CrossRef]

19. Ji, Y.; Wang, J.; Li, S.; Li, Y.; Lin, S.; Li, X. An Anomaly Event Detection Method Based on GNN Algorithm for Multi-data Sources. In *Proceedings of the BSCI 2021—Proceedings of the 3rd ACM International Symposium on Blockchain and Secure Critical Infrastructure, Hong Kong, 7 June 2021*; Association for Computing Machinery Inc.: New York, NY, USA, 2021; pp. 91–96. [CrossRef]
20. Alkahtani, H.K.; Mahmood, K.; Khalid, M.; Othman, M.; Al Duhayyim, M.; Osman, A.E.; Alneil, A.A.; Zamani, A.S. Optimal Graph Convolutional Neural Network-Based Ransomware Detection for Cybersecurity in IoT Environment. *Appl. Sci.* **2023**, *13*, 5167. [CrossRef]
21. Singapore University of Technology and Design. Secure Water Treatment (SWAT) Dataset. Available online: https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/ (accessed on 15 November 2025).
22. Tang, C.; Xu, L.; Yang, B.; Tang, Y.; Zhao, D. GRU-Based Interpretable Multivariate Time Series Anomaly Detection in Industrial Control System. *Comput. Secur.* **2023**, *127*, 103094. [CrossRef]
23. L(y)u, S.; Wang, K.; Zhang, L.; Wang, B. Global-local integration for GNN-based anomalous device state detection in industrial control systems. *Expert Syst. Appl.* **2022**, *209*, 118345. [CrossRef]
24. Liao, J.; Li, J.; Chen, Y.; Gu, R.; Zhu, Y.; Peng, W. DPDGAD: A Dual-Process Dynamic Graph-based Anomaly Detection for multivariate time series analysis in cyber-physical systems. *Adv. Eng. Inform.* **2024**, *61*, 102547. [CrossRef]
25. Kaliyaperumal, P.; Periyasamy, S.; Thirumalaisamy, M.; Balusamy, B.; Benedetto, F. A Novel Hybrid Unsupervised Learning Approach for Enhanced Cybersecurity in the IoT. *Future Internet* **2024**, *16*, 253. [CrossRef]
26. Vitulyova, Y.; Babenko, T.; Kolesnikova, K.; Kiktev, N.; Abramkina, O. A Hybrid Approach Using Graph Neural Networks and LSTM for Attack Vector Reconstruction. *Computers* **2025**, *14*, 301. [CrossRef]
27. Chen, Z.; Bian, X.; Yang, M.; Liu, C.; Wu, X.; Lu, S. FST-AD: Anomaly Detection for Cyber-Physical Systems via Frequency-Spatio-Temporal GNNs. In *Proceedings of the 2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guiyang, China, 14–17 November 2025*; IEEE Computer Society: Los Alamitos, CA, USA, 2025; pp. 2319–2326. [CrossRef]
28. Zhang, W.; He, P.; Qin, C.; Yang, F.; Liu, Y. A graph attention network-based model for anomaly detection in multivariate time series. *J. Supercomput.* **2024**, *80*, 8529–8549. [CrossRef]
29. Ahmed, A.A.; Abdullah, T.A.A. An Enhanced XGBoost-Based Framework for Efficient Multi-Class Cyber Threat Detection in Industrial IoT Networks. *Technologies* **2026**, *14*, 274. [CrossRef]
30. Wahab, S.A.; Sultana, S.; Tariq, N.; Mujahid, M.; Khan, J.A.; Mylonas, A. A Multi-Class Intrusion Detection System for DDoS Attacks in IoT Networks Using Deep Learning and Transformers. *Sensors* **2025**, *25*, 4845. [CrossRef]
31. Zhang, S.; Wang, Y.; Zhang, Z.; Hao, Q.; Hou, Y.; Yang, W.; Liu, L. Warning-Graph: An Early Warning Framework for APT Attacks Based on Threat Intelligence Modeling. *IEEE Trans. Dependable Secur. Comput.* **2026**, *23*, 3880–3897. [CrossRef]
32. Zhao, M.; Peng, H.; Li, L.; Ren, Y. Graph Attention Network and Informer for Multivariate Time Series Anomaly Detection. *Sensors* **2024**, *24*, 1522. [CrossRef]
33. Zhang, Y.; Xu, S.; Zhang, L.; Jiang, W.; Alam, S.; Xue, D. Short-term multi-step-ahead sector-based traffic flow prediction based on the attention-enhanced graph convolutional LSTM network (AGC-LSTM). *Neural Comput. Appl.* **2025**, *37*, 14869–14888. [CrossRef]
34. Babenko, T.; Kolesnikova, K.; Bakhtiyarova, Y.; Yeskendirova, D.; Sansyzbay, K.; Sysoyev, A.; Kruchinin, O. Hybrid GNN–LSTM Architecture for Probabilistic IoT Botnet Detection with Calibrated Risk Assessment. *Computers* **2026**, *15*, 26. [CrossRef]
35. Jo, H.; Lee, S.W. Edge conditional node update graph neural network for multivariate time series anomaly detection. *Inf. Sci.* **2024**, *679*, 121062. [CrossRef]