Journal Pre-proof

Neurosymbolic learning and domain knowledge-driven explainable AI for enhanced IoT network attack detection and response

Chathuranga Sampath Kalutharage, Xiaodong Liu, Christos Chrysoulas

PII: S0167-4048(25)00007-0
DOI: https://doi.org/10.1016/j.cose.2025.104318
Reference: COSE 104318

To appear in: *Computers & Security*

Received date : 1 October 2024
Revised date : 29 December 2024
Accepted date : 6 January 2025

Please cite this article as: C.S. Kalutharage, X. Liu and C. Chrysoulas, Neurosymbolic learning and domain knowledge-driven explainable AI for enhanced IoT network attack detection and response. *Computers & Security* (2025), doi: https://doi.org/10.1016/j.cose.2025.104318.

Neurosymbolic Learning and Domain Knowledge-Driven Explainable AI for Enhanced IoT network Attack Detection and Response

Chathuranga Sampath Kalutharage[a], Xiaodong Liu[a], Christos Chrysoulas[b]

[a]*Edinburgh Napier University, Scotland, UK*
[b]*Heriot-Watt University, Scotland, UK*

Abstract

In the dynamic landscape of network security, where cyberattacks continuously evolve, robust and adaptive detection mechanisms are essential, particularly for safeguarding Internet of Things (IoT) networks. This paper introduces an advanced anomaly detection model that utilizes Artificial Intelligence (AI) to identify network anomalies based on traffic features, explaining the most influential factors behind each detected anomaly. The model integrates domain knowledge stored in a knowledge graph to verify whether the detected anomaly constitutes a legitimate attack. Upon validation, the model identifies which core cybersecurity principles—Confidentiality, Integrity, or Availability (CIA)—are violated by mapping influential feature values. This is followed by an alignment with the MITRE ATT&CK framework to provide insights into potential attack tactics, techniques, and intelligence-driven countermeasures.

By leveraging explainable AI (XAI) and incorporating expert domain knowledge, our approach bridges the gap between complex AI predictions and human-understandable decision-making, thereby enhancing both detection accuracy and result interpretability. This transparency facilitates faster responses and real-time decision-making while improving adaptability to new, unseen cyber threats. Our evaluation on network traffic datasets demonstrates that the model not only excels in detecting and explaining anomalies but also achieves an overall detection accuracy of 0.97 with the integration of domain knowledge for attack legitimacy. Furthermore, it provides 100% accuracy for threat intelligence based on the MITRE ATT&CK framework, ensuring that security measures are verifiable, actionable, and ultimately strengthen IoT environment defenses by delivering real-time threat intelligence and responses, thus minimizing human response time.

Keywords: Neurosymbolic learning, Attack detection, Explainable artificial intelligence, Expert knowledge, Thraet intelligence.

1. Introduction

In the constantly evolving landscape of cybersecurity, the detection and mitigation of network-based attacks, especially in the context of Internet of Things (IoT) networks, has become a critical challenge. Traditional security mechanisms, while effective for known threats, often fall short against sophisticated and adaptive cyberattacks. As IoT networks expand, the sheer volume of connected devices, coupled with limited computational resources and infrequent security updates, increases the risk of malicious activities. Moreover, the diversity of attack vectors—from Denial of Service (DoS) to Command and Control (C2)—necessitates adaptive, intelligent systems that can not only detect but also explain the underlying causes of anomalies in real time.

To address these challenges, we propose an enhanced anomaly detection model built on Neurosymbolic Learning within the Explainable Artificial Intelligence (XAI) framework, further extended with feature mapping to cybersecurity components (CIA) and the MITRE ATT&CK framework. Neurosymbolic Learning combines the strengths of neural networks and symbolic reasoning, offering both the data-driven pattern recognition capabilities of deep learning and the interpretability of symbolic AI. This integration ensures that the model remains transparent and explainable, a crucial factor in building trust for security operations and facilitating quick, informed responses. Our model leverages SHAP (SHapley Additive exPlanations) values to explain the most influential features responsible for detected anomalies. These feature values are then mapped to Confidentiality, Integrity, and Availability (CIA) violations, ensuring that the model accurately identifies which core cybersecurity principles are at risk. Subsequently, we extend this approach by integrating Large Language Models (LLMs) for feature mapping to the MITRE ATT&CK framework, enabling automatic identification of attack tactics, techniques, and corresponding mitigations. This innovative use of LLMs allows for the real-time correlation of detected anomalies with established attack vectors, significantly enhancing the detection process. By combining expert knowledge embedded in a cybersecurity knowledge graph with the LLM's capacity to map complex anomaly behaviors to the ATT&CK framework, our approach provides a robust defense mechanism that not only identifies attacks but also delivers actionable intelligence for response. This dual-layered system—combining data-driven anomaly detection with symbolic reasoning—ensures that the detection process is both accurate and interpretable, offering a significant advancement over existing black-box models. Our model's ability to deliver clear, context-driven explanations and map detected anomalies to CIA violations and MITRE ATT&CK tactics establishes a comprehensive system for defending IoT environments against increasingly sophisticated cyber threats. Through rigorous evaluation using benchmark datasets and real-time IoT network traffic, our method demonstrates superior performance in both detecting and explaining network attacks, significantly reducing the rate of false positives. The integration of LLM-generated threat intelligence and expert-augmented knowledge graphs ensures that the model is adaptable to evolving threats, making it a powerful tool in the dynamic field of IoT security.

Neurosymbolic artificial intelligence combines neural network-based techniques with symbolic knowledge-based methods, leveraging the strengths of both. Neural networks excel at processing large datasets and identifying complex patterns from raw inputs, while symbolic approaches are known for their proficiency in logical reasoning and structured decision-making. By integrating these two paradigms, neurosymbolic AI not only benefits from the data-driven insights of neural networks but also overcomes their traditional limitations by offering more transparent and interpretable explanations for decision-making processes [1]. Despite the significant advancements in neural networks since the mid-1980s, their adoption beyond academic and commercial settings has been constrained by inherent challenges. On the other hand, symbolic knowledge-based approaches, such as expert systems and rule-based models, are grounded in logical reasoning and structured representation of knowledge. These approaches excel at gathering domain-specific expertise and delivering clear, interpretable explanations for their outcomes [1], [2]. These methods frequently encounter difficulties when dealing with ambiguous or incomplete information and are generally not well-suited for extracting insights from large-scale datasets [1]. In recent years, there has been growing interest in NeuroSymbolic AI, which combines neural and symbolic AI techniques. Although this integration is gaining traction now, the concept of 'Neural-Symbolic' AI actually dates back to the early 2000s [2]. During the 1990s, numerous attempts were made to combine fuzzy rule systems with connectionist methods [3]. The concept of combining the intuitive and logical components of AI was first suggested in the seminal work by McCulloch and Pitts, titled "A Logical Calculus of the Ideas Immanent in Nervous Activity." [4]. The renewed interest in this method can be linked to various reasons, which we will examine within the scope of cybersecurity. In this study, we incorporate neurosymbolic artificial intelligence with our previously established explainable artificial intelligence (XAI) model [5, 6], enhancing the process by extracting attack responses from the MITRE ATT&CK framework as threat intelligence, thereby improving human-speed decision-making with more sophisticated insights. This combination incorporates expert knowledge to improve the detection of cyberattacks while ensuring a clear explanation of the decision-making process and detected attack. The main contributions of this paper are as follows:

- Develop a data-driven cybersecurity knowledge graph to identify legitimate attacks from detected anomalous network behaviours.
- Develop a method for integrating expert knowledge into the existing knowledge graph, thereby bridging the gap between data-driven models and human expertise.
- Develop a main neurosymbolic model with integration of our previous XAI model to enhance cyberattack detection.
- Define security rules based on traffic features (Threshold values for each traffic feature for attack detection).
- Find the violated cyber-security components (CIA) using feature influence.
- Extract the threat intelligence and response from MITRE&CK using AI for reduce the human response time.
- Evaluate the model's performance by comparing it with existing research in the field.

The remainder of the paper is structured as follows: Section 2 provides an overview of the background and related work. Section 3 outlines the proposed algorithm. Section 4 covers the experimental setup, followed by Section 5, which discusses the evaluation process and any necessary modifications. Lastly, Section 6 concludes this work.

2. Background and Related work

2.1. IoT Network Attacks

The Internet of Things (IoT) encompasses a wide range of interconnected devices, from simple sensors to complex industrial tools. This connectivity, while beneficial, exposes networks to various cyber threats. IoT network attacks can be particularly insidious due to the diverse nature and widespread deployment of these devices. Common types of attacks include [7]:

I. DDoS Attacks (Distributed Denial of Service): In these attacks, IoT devices are hijacked to form a botnet that floods a target with overwhelming traffic, causing service disruption.

II. Man-in-the-Middle (MitM) Attacks: Attackers intercept communications between IoT devices and the network to steal or manipulate data.

III. Ransomware and Malware Attacks: Malicious software is used to infect IoT devices, leading to data theft, device malfunction, or ransom demands.

IV. Data and Identity Theft: Attackers target sensitive personal information stored or transmitted by IoT devices.

V. Device Hijacking: Unauthorized access to IoT devices allows attackers to manipulate device functionality, often without the owner's knowledge.

VI. Side-channel Attacks: These exploit information gained from the physical implementation of a system, such as power consumption or electromagnetic leaks.

Detecting network attacks in the realm of the Internet of Things (IoT) is fraught with various distinct challenges [7]. The sheer diversity and volume of IoT devices, each with its own set of protocols and standards, make it hard to establish uniform security across the board. Many of these devices are limited in terms of processing power and memory, hindering the implementation of advanced security algorithms [6]. As the IoT landscape continues to expand rapidly, developing scalable security solutions that can keep pace with this growth is becoming increasingly crucial. Another significant concern is the privacy of data; there's a delicate balance to be maintained between effective security monitoring and the privacy of data collected from IoT devices. A notable issue is that many IoT devices do not receive regular security updates, leaving them vulnerable to known threats. The complexity of IoT ecosystems also presents a challenge, as the interconnected nature of these devices and systems adds difficulty in identifying the source and nature of attacks. Modern threat vectors, such as Distributed Denial-of-Service (DDoS) attacks, exploit the distributed nature of IoT networks, making them increasingly powerful and harder to mitigate effectively. Cirillo et al. [8] introduced a botnet identification algorithm that leverages the concept of message innovation rates (MIR) to distinguish malicious bots from legitimate users, addressing challenges posed by botnets using multiple emulation dictionaries to mimic legitimate traffic patterns. Their proposed cluster expurgation rule ensures high accuracy in isolating malicious traffic, even in complex scenarios. Building on this, Matta et al. [9] extended the approach to tackle multi-clustered botnets, where distinct clusters operate with different portions of emulation dictionaries. They proposed algorithms based on cluster expurgation and union rules to effectively identify diverse botnet clusters, demonstrating robust performance in real-world scenarios and showcasing the scalability of their method. In addressing stealthier threats, Xiang et al. [10] proposed new information-theoretic metrics, including generalized entropy and information distance, to detect low-rate DDoS attacks. These metrics enable earlier detection and reduce false positives, effectively addressing the stealthy nature of such attacks. Additionally, their IP traceback scheme enhances the ability to locate and mitigate attack sources. Tang et al. [11] further contributed to mitigating low-rate DDoS attacks in SDN environments with LtRFT, a Learning-to-Rank-enabled framework that prioritizes malicious flows for eviction. Achieving over 96% accuracy, LtRFT significantly reduces attack durations while maintaining minimal latency, demonstrating its effectiveness and practicality for SDN deployments. However these techniques does not provide a realtime response while they are providing slow datarate DDos detection accurately. Moreover, the necessity for real-time detection and response mechanisms is paramount to maintaining the operational integrity of IoT networks. Unlike traditional cyber systems, many IoT devices are located in public or easily accessible areas, which elevates the risk of physical tampering. This unique set of challenges underscores the need for innovative approaches in securing IoT networks against potential threats.

2.2. MITRE ATT&CK framework

The MITRE ATT&CK framework [1] is a widely recognized cybersecurity knowledge base developed by the MITRE Corporation that categorizes adversarial tactics, techniques, and procedures (TTPs) used in cyberattacks. It provides a comprehensive structure to understand and defend against sophisticated threats by breaking down the various stages of an attack. The framework is organized around three key elements: tactics, which represent the adversary's goals at different stages of the attack, such as initial access or data exfiltration; techniques, which describe the specific methods used by attackers to achieve their goals, such as phishing or credential dumping; and procedures, which detail how these tactics and techniques are implemented in real-world scenarios.

The MITRE ATT&CK framework plays a crucial role in enhancing cybersecurity by offering a standardized language for describing and understanding attacks, making it easier for organizations to share threat intelligence. It also supports security teams in detecting, analyzing, and responding to threats by mapping observed behaviors to known attack methods. Additionally, the framework is a key tool in threat modeling and adversary emulation, allowing organizations to simulate real-world attacks to evaluate and improve their defenses. As a result, the MITRE ATT&CK framework is an invaluable resource for cybersecurity professionals aiming to stay ahead of ever-evolving cyber threats.

2.3. Neurosymbolic AI in Cybersecurity

Neurosymbolic AI seeks to combine the strengths of two approaches: the ability of neural networks to learn and recognize patterns and the interpretability and logical reasoning of symbolic AI. By integrating data-driven techniques with symbolic reasoning, this approach allows for the tracing of the steps or decisions that lead to a model's conclusions. This combination makes a strong argument for the use of neurosymbolic methods in enhancing cybersecurity and privacy efforts [12]. These methods are especially useful for tackling challenges such as threat detection and analysis, where it is important to understand and contextualize patterns across various systems over time, rather than simply identifying them in isolation [13]. Neurosymbolic approaches can address these challenges while preserving privacy by

[1] https://attack.mitre.org/mitigations/ics/

incorporating policies, regulations, and compliance measures. For example, a neurosymbolic model can use logical reasoning to regulate the handling of sensitive network flow data by the neural network detector, ensuring it follows defined privacy guidelines. Furthermore, compliance is maintained through the use of privacy-preserving methods such as differential privacy or secure multi-party computation [14]. Ensuring the security and safety of AI systems is essential. Relying solely on data-driven models for automated vulnerability assessments can be restrictive, as these models are limited to the vulnerabilities they have been trained on. By utilizing a neurosymbolic approach, safety can be improved. In this method, experts simulate adversarial roles during the training of AI-based systems, allowing the model to continuously learn and adapt by applying dynamic rules and policies, rather than depending exclusively on pre-existing vulnerabilities [1]. Additionally, an AI system's reliability and security can be greatly improved by explicitly encoding knowledge from security specification documents using symbolic techniques and enforcing them as behavioral constraints. This approach is particularly relevant to legislators and regulators in many countries. Without the integration of human expertise, advanced AI systems are at considerable risk of producing potentially harmful or dangerous information.

One key advantage of combining rule-based and data-driven approaches is their ability to address the lack of high-quality data, which is often required for drawing reliable conclusions. This issue frequently arises in areas where sensitive data is either limited or difficult to share for experimental purposes. However, alternative sources, such as textual descriptions of sensitive information, may be available. These can be leveraged to create general rules. When the data itself is insufficient for making strong conclusions, these established rules can help support and validate the insights derived from the data [2]. Throughout the learning process, these rules can also be incorporated as input for data-driven models. Additionally, certain fields are highly dynamic, with data accurately reflecting conditions for only a limited time. As a result, conclusions derived from such data may have a short lifespan. This is especially true in areas like fraud detection and cybersecurity. Patterns detected in the current dataset might be effective against present cyber threats but could lose relevance over time. In these cases, it can be beneficial to combine deep network-based detection systems with explicit rules that account for evolving data trends and the temporary applicability of models [15].

Neurosymbolic AI, which integrates symbolic AI with neural networks, is becoming increasingly important in cybersecurity. It strengthens key areas like threat intelligence, malware analysis, intrusion detection, and vulnerability assessment, ultimately improving the overall efficiency and effectiveness of security systems [13]. This approach is crucial for transforming Security Operations Centers (SoCs) into next-generation facilities. By integrating AI techniques with human oversight, a more sophisticated and efficient system for managing and responding to security threats is created. For instance, security analysts in SoCs play a key role in safeguarding an organization, relying heavily on their expertise and knowledge of emerging and novel threats. This knowledge becomes particularly valuable when interpreting results from deep neural networks or machine learning systems that analyze incoming data streams. Analysts' familiarity with new attack patterns is essential for accurately identifying potential security breaches. To support them, information from publicly available threat intelligence sources, such as threat feeds or detailed cyberattack reports, can be gathered and organized into a Cybersecurity Knowledge Graph (CKG). We propose two methods for utilizing the structured data within CKGs: the first focuses on explainability through reasoning and inference by creating complex rules using a knowledge engine and real data, forming a rule-based framework. The second method involves developing new cybersecurity strategies (knowledge-guided models) by incorporating these rules into data-driven AI models.

The main objective of a rule-based framework is to create highly effective and resilient rules to safeguard target systems from various threats and malicious activities. These rules, ranging from simple to complex, can be applied across any system or subsystem requiring protection. The emphasis on knowledge-guided models is to address emerging or evolving cyber threats that are not captured in existing datasets for data-driven research. To detect new adversaries and develop corresponding defense mechanisms, techniques such as Reinforcement Learning (RL) and other exploratory modeling approaches are essential. Our experiments demonstrate that Cybersecurity Knowledge Graphs (CKGs) can effectively guide these exploratory methods, improving their efficiency, speed, and overall clarity.

2.4. Explainable Artificial Intelligence (XAI)

Research in Explainable Artificial Intelligence (XAI) is experiencing a resurgence, building upon the earlier contributions of Chandrasekaran, Tanner, and Josephson (1989) [16]. Earlier research primarily concentrated on explaining the decision-making process of knowledge-based and expert systems. The classical learning paradigm, Explanation-Based Learning (EBL), introduced in the early 1980s, is often considered a forerunner of explainability. EBL involves learning a problem-solving method by examining and analyzing the solutions to specific problems [17]. The resurgence of interest in XAI research is largely driven by recent advancements in AI and machine learning (ML), which have been applied across a variety of fields. Additionally, growing concerns about unethical practices and unintended biases in AI models have further contributed to this renewed focus on explainability. Yang and Shafto [18] employed Bayesian Teaching, where a smaller, carefully selected subset of examples is used to train the model, rather than utilizing the entire dataset. These examples are chosen by domain experts for their relevance to the specific problem at hand. However, selecting the appropriate subset of examples in real-world scenarios presents a significant challenge. The convergence of IoT networks and AI technologies poses unique security and interpretability challenges, as explored in [19, 20]. These works highlight the interplay between the physical and cyber domains in IoT environments, emphasizing the critical role of XAI for maintaining trust and security in such systems. Li et al. [19] discuss how ethical AI principles and secure digital twin technologies can enhance trustworthiness in IoT networks. Similarly, Li et al. [20] address the integration of spatiotemporal data with semantic technologies, underscoring the importance of context-aware decision-making in enhancing the interpretability and security of IoT systems through XAI.

AI-based Intrusion Detection Systems (IDSs) have consistently demonstrated strong performance Hodo et al [21]; Shone et al [22]; Kim et al [23]. Shone et al. [22] introduced a hybrid approach combining shallow learning techniques like Random Forest with deep learning models such as Autoencoders. This method is capable of analyzing diverse network traffic and outperforms traditional Deep Belief Networks (DBN). A survey by Dong and Wang (2016) comparing traditional IDS with deep learning-based IDS highlighted that deep learning methods generally offer better accuracy across a wide range of sample sizes and different types of network traffic or attacks. Despite these advancements, challenges such as long training times and the need for human oversight remain prevalent in existing approaches [22]. Offering explanations for outliers can greatly reduce the need for security analysts to manually investigate false alarms. In the system developed by Goodal et al. [24], designed for identifying and interpreting irregularities in computer network traffic and logs, the visualization of contextual information surrounding these outliers serves as the foundation for explanation. Liu et al. [25] introduced the Contextual Outlier Interpretation (COIN) technique, which provides explanations for the outlier anomalies identified by detection systems. Collaris at al. [26] utilized various cutting-edge explanation methods to develop two dashboards, helping domain experts better understand the predictions. These explanations are derived from established techniques, such as partial dependency plots, instance-level feature importance analysis, and local rule mining, which is a modified version of the LIME method. Other studies have proposed an SVM-based approach for malware detection and explanation, focusing on identifying the features that most significantly contribute to detection. This method also verifies whether the identified influential features align with commonly recognized vulnerabilities [27]. Valerio La Gatta et al. [28] introduced a local explanation method called CASTLE (Cluster-Aided Space Transformation for Local Explanations), which generates decision rules for applying model predictions to novel situations while also providing localized insights into the importance of specific features. Kalutharage et al. [29] propose an ensemble-based approach combining an Autoencoder and XGBoost to enhance IoT network attack detection. The study demonstrates how XAI can be used to identify influential features, refine datasets, and reduce computational overhead, enabling lightweight, efficient detection models for resource-constrained IoT environments. Their approach achieves 99.92% accuracy on the CICIDS2017 dataset, showcasing significant advancements over traditional intrusion detection systems while maintaining interpretability and scalability. To the best of our knowledge, no existing models combine domain knowledge with a focus on improving explainability and interpretability while integrating with neurosymbolic learning. Our proposed conceptual model offers enhanced explainability, interpretability, and scalability for large-scale data problems. It reduces false positives by providing legitimate results through domain knowledge, enabling more contextual scenarios and enhancing the model's generalization capability.

3. Proposed Model

3.1. Overview

This research presents an innovative neurosymbolic approach for detecting anomalies in network data. The methodology integrates neural network-based anomaly detection, utilizing autoencoders, with symbolic reasoning through a knowledge graph. By combining the strengths of both neural and symbolic AI, the approach delivers robust anomaly detection while improving interpretability and decision-making. A data-driven method is employed for developing the knowledge graph, with expert knowledge incorporated to enhance it. The model also identifies violated cybersecurity components (CIA) using a knowledge extractor and provides threat intelligence and recommended responses based on the MITRE ATT&CK framework, thereby reducing human intervention and accelerating the process by leveraging AI. Figure 1 illustrates the model's architecture, with each component described in detail.

I. IoT Network Traffic: This represents the data flow within an IoT network, which includes both normal operations and potential security threats.
II. Anomaly Detection: A system or model that processes the IoT network traffic to identify unusual patterns or activities that deviate from the established norm, which could indicate potential security incidents.
III. Benign Traffic: This is the subset of network traffic that has been identified as normal and safe by the anomaly detection system.
IV. Explanation XAI (Explainable Artificial Intelligence): A component that provides insights into the decision-making process of AI models, making the outcomes understandable to humans. In the context of anomaly detection, this would explain why certain traffic was flagged as anomalous.
V. Security Knowledge Graph: A structured representation of cybersecurity knowledge, including concepts, relationships, and rules that define and describe the security aspects of the IoT network.
VI. Security Knowledge Graph Constructor: This is the process or the tool that builds the security knowledge graph, possibly by integrating various data sources and expert input to form a comprehensive security model.
VII. Security Expert: A human expert who provides additional insights and validation to the reasoning model, ensuring that the system's outputs align with real-world cybersecurity knowledge and practices.
VIII. Knowledge Extractor: A tool or process that extracts relevant information from the security knowledge graph to support the reasoning model, providing context and detailed explanations for detected anomalies, aligned with the MITRE ATT&CK framework.

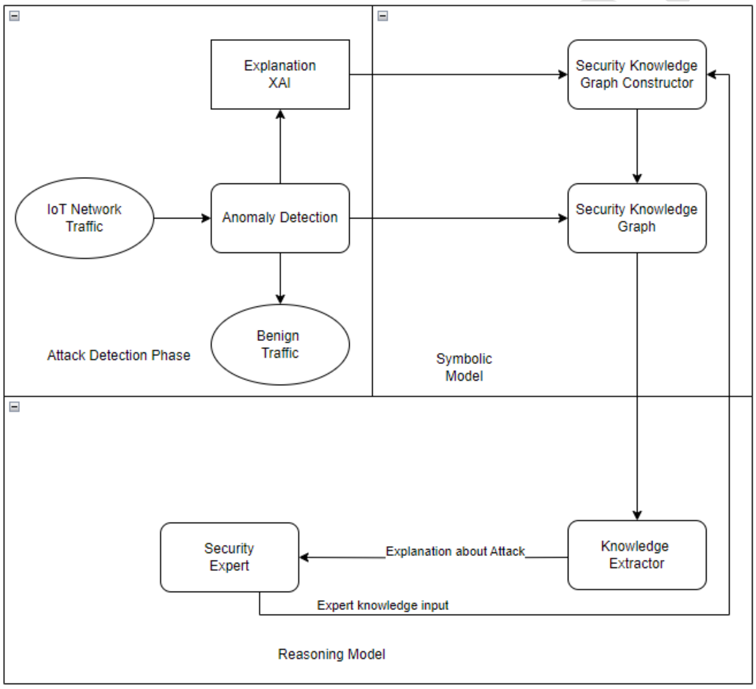

Figure 1: Proposed Neurosymbolic learning in the XAI framework architecture for IoT attack detection.

3.2. Neural Network-Based Anomaly Detection

The methodology relies on an autoencoder, a type of neural network known for its ability to generate compact data representations. The autoencoder functions through two key stages: encoding and decoding. During the encoding phase, it reduces the network data to a lower-dimensional form, preserving the most important features. In the subsequent decoding phase, the compressed data is reconstructed back to its original size. The effectiveness of the autoencoder is measured by the reconstruction error, which calculates the difference between the original input and the reconstructed output. A frequently used metric for this evaluation is the Mean Absolute Error (MAE). In the context of anomaly detection, MAE plays a crucial role in determining whether the reconstruction error surpasses a predefined threshold, signaling a potential anomaly. This threshold is typically based on the error distribution observed in normal data. The underlying assumption is that normal data will produce smaller reconstruction errors, while anomalous data will result in larger errors due to significant deviations from the patterns learned by the model.

3.3. Symbolic Reasoning with SHAP and Knowledge Graphs

To improve the model's interpretability and decision-making capabilities, we incorporate SHAP (SHapley Additive exPlanations) values, rooted in game theory, to assign importance to individual features in anomaly detection. SHAP values are crucial for identifying the contribution of each feature to the anomalies detected, thereby providing insights into the model's decision-making process. For each anomaly identified by the model, SHAP values reveal which features play the most significant role in signaling the anomaly, allowing for a detailed analysis of the model's behavior. Alongside this, we construct a domain-specific knowledge graph using real-world attack data to map anomalous behaviors that indicate legitimate cybersecurity threats, as outlined in Algorithm 1. This knowledge graph, designed for network security, serves as a structured representation of expert knowledge and heuristic rules. Each node represents individual network data features, while the edges reflect the complex relationships and constraints between them. The graph effectively captures the intricate network dynamics that may indicate potential security breaches.

In the context of detected anomalies, the knowledge graph plays a critical role by leveraging the Maximum Mean Absolute Error (Max MAE)—a metric that reflects the model's highest deviation in reconstruction error when it encounters an anomalous pattern. This metric helps distinguish between normal and abnormal behavior. By linking Max MAE with actual feature values corresponding to known attack types stored in the knowledge graph, it becomes possible to determine whether a detected anomaly signifies a legitimate attack or merely unusual but harmless network activity. The integration of SHAP values and the knowledge graph serves two key purposes: first, SHAP values offer a detailed explanation of why certain instances are classified as anomalies by highlighting the contributions of specific features. Second, the knowledge graph cross-references these anomalies with real-world attack patterns to differentiate genuine threats from false positives. This dual approach enhances the model's accuracy in detecting attacks while providing a clear understanding of each anomaly, ensuring a more robust and reliable network security system.

3.4. Neurosymbolic Integration

Our methodology embodies the integration of neural network outputs and symbolic reasoning, creating a unified framework for detecting anomalies in IoT networks. The process begins by evaluating each data instance using an autoencoder, which computes both the reconstruction error and SHAP values. These SHAP values are essential as they highlight the influence of individual features on the model's predictions. In this framework, SHAP values play a critical role by being assessed against predefined thresholds and rules within a custom-built knowledge graph. Initially developed from data-driven insights, the knowledge graph encapsulates normal network behavior and recognized anomaly patterns. When a SHAP value identifies a feature as highly influential, the model cross-references the corresponding original feature value with the maximum Mean Absolute Error (MAE). If this value exceeds the feature's threshold in the knowledge graph, the instance is classified as an attack.

Given the context-specific nature of IoT networks, generalizing models poses a challenge. To address this, we augment the data-driven knowledge graph with expert knowledge, which brings a deeper and more nuanced understanding of network behavior and threat landscapes—elements that may not be fully captured by data alone. The integration of expert insights significantly enhances the model's ability to detect and validate anomalies. When an instance is flagged based on influential SHAP values, the model employs symbolic reasoning, grounded not only in data-driven thresholds but also in expert-derived rules. This comprehensive approach ensures more accurate, contextually relevant interpretations of anomalies and offers actionable recommendations for responding to potential threats. In summary, our approach seamlessly combines data-driven analysis with expert knowledge. SHAP values highlight the most critical features for identifying anomalies, while the enhanced knowledge graph, infused with expert insights, validates these findings. This integration ensures that the model's interpretations and responses are accurately tailored to the complex and evolving landscape of IoT network security.

Algorithm 1 Neurosymbolic Anomaly Detection with SHAP and Knowledge Graph Integration

---

Require: X — Anomaly instance that needs to be explained, $X_{1..i}$ — instances used by kernel SHAP, `autoencoder_model` — trained autoencoder model for anomaly detection, `expert_knowledge` — expert knowledge integrated into the knowledge graph, `Feature_thresholds` — thresholds for Feature values derived from the knowledge graph.

Ensure: `shap_top_features` — SHAP values for each feature within the top R features, `detected_anomalies` — list of detected anomalies with decision reasoning.

$top\_R\_features \leftarrow$ top value from Error List derived from reconstruction errors
for each i in $top\_R\_features$ **do**
    $explainer \leftarrow$ `shap.KernelExplainer(autoencoder_model.predict, X_1..i)`
    $shap\_values[i] \leftarrow$ `explainer.shap_values(X, i)`
end for
$knowledge\_graph \leftarrow$ `construct_knowledge_graph(expert_knowledge)`
for each $feature, Original\_value$ in `shap_top_features` **do**
    if $knowledge\_graph.nodes[feature]['threshold'] < Original\_value$ **then**
        `detected_anomalies.append(feature)`
        `symbolic_reasoning(feature, Original_value, knowledge_graph)`
    end if
end for
return `detected_anomalies`

Journal Pre-proof

3.5. Mapping Features to Violated Cybersecurity Components

To pinpoint the violated cybersecurity components, we apply the CIA principles—confidentiality, integrity, and availability—as domain knowledge. By analyzing different types of attacks within the dataset, we assess how each one affects the individual components of the CIA triad as shown Table 1. DoS and DDoS attacks primarily target the availability of services or data, aiming to overwhelm systems and make them inaccessible to legitimate users. Similarly, Port Scan attacks are associated with a compromise in confidentiality, as attackers send probes to various ports to gather information about available services and operating systems. SSH Patator and FTP Patator are brute-force attacks that typically lead to a breach of confidentiality by attempting to guess login credentials. Additionally, Heartbleed vulnerability is linked to a breach of confidentiality, as it allows attackers to access sensitive information stored in the memory of systems running a vulnerable version of OpenSSL. In the case of Infiltration attacks, they usually exploit software vulnerabilities, such as those in Adobe Acrobat Reader, to create backdoors and exfiltrate confidential information like IP addresses, thus compromising confidentiality. Web attacks, such as SQL injection, can affect all three components of the CIA triad. They compromise confidentiality and integrity by allowing unauthorized access to read and modify data, while also jeopardizing availability by overwhelming databases with complex queries. Lastly, Botnets—networks of compromised devices—pose a multifaceted threat, as they can allow attackers to perform actions like remote shell access, file manipulation, screenshot capture, and keylogging. Consequently, botnets have the potential to compromise confidentiality, integrity, and availability.

From the original dataset's feature ranking, we identified the top three most important features for each type of attack based on their significance using autoencoder and SHAP (Shapley Additive Explanations). These features were then mapped to their associated compromises under the CIA principles (as shown in Table 2)). For instance, the feature Average Packet Size is denoted as Avg Packet Size - A, where A signifies its relevance to a compromise in availability (refer to Table 2). To establish this mapping between features and associated compromises, we first determine the relationship between each attack and the related compromises (derived from Table 1) and formulated in Equation 2). Essentially, Formula 1 identifies the attack for which the feature ranks in the top three in terms of importance, while Formula 2 links the attack to the relevant compromises under confidentiality, integrity, or availability. Using domain knowledge, we narrowed down the

Table 1: Mapping of network attack with related component of CIA principles

Attack	Related component of CIA
Heartbleed	C
SSH-Patator	C
FTP-Patator	C
Infiltration	C
PortScan	C
Web Attack	C, I, A
Bot	C, I, A
DoS GoldenEye	A
DoS Hulk	A
DoS Slowhttp	A
DoS Slowloris	A
DDoS	A

features to 22 (as displayed in Table 2) from an initial set of features, which we now refer to as the domain features for CIA triads. Table 2) provides detailed descriptions of these features.

$$f(\text{feature}) \rightarrow \text{attack} \tag{1}$$

$$f(\text{attack}) \rightarrow C, I, \text{or} A \tag{2}$$

Table 2: Mapping of feature with related component of CIA principles

Feature	**Description**	**Top features of attack**	**Domain Knowledge feature**
Average Packet Size	Average size of packet	DDoS	Avg Packet Size - A
Flow Duration	Duration of the flow in Microseconds	DDoS, DoS Slowloris, DoS Hulk, DoS Slowhttp, Infiltration, Heartbleed	Flow Duration - AC
Bwd IAT Mean	Mean time between two packets sent in backward direction	DoS Hulk, DoS GoldenEye, DDoS, Heartbleed, DoS Hulk	Bwd IAT Mean - A
Fwd IAT Mean	Mean time between two packets in forward direction	DoS Slowloris	Fwd IAT Mean - A
Active Mean	Mean time a flow was active before idle	DoS Slowhttp	Active Mean - AC
Bwd Packet Length Std	Standard deviation of packet length in backward direction	DoS Slowloris, DoS GoldenEye	Bwd Packet Length Std - AC
Flow IAT Std	Standard deviation of inter-arrival time	DDoS, DoS Slowhttp, DoS Hulk	Flow IAT Std - A
Flow IAT Mean	Mean inter-arrival time of packet	DoS GoldenEye	Flow IAT Mean - A
Flow IAT Min	Minimum inter-arrival time of packet	DoS GoldenEye	Flow IAT Min - A
Active Min	Minimum time a flow was active before idle	DoS Slowhttp	Active Min - A
Init Win Bytes Forward	Total bytes sent in initial window in forward direction	Web Attack	Init Win Bytes Fwd - C
SYN Flag Count	Number of packets with SYN	FTP-Patator	SYN Flag Count - C
Fwd Packet Length Mean	Mean size of packet in forward direction	Benign, Bot, FTP-Patator	Fwd Packet Length Mean - CIA
Fwd Packets/s	Number of forward packets per second	FTP-Patator	Fwd Packets/s - C
Fwd PSH Flags	Number of times PSH flag was set in forward packets	FTP-Patator	Fwd PSH Flags - C
ACK Flag Count	Number of packets with ACK	SSH-Patator, DoS Slowhttp, Infiltration	ACK Flag Count - C
Bwd Packets/s	Number of backward packets per second	Bot, PortScan	Bwd Packets/s - CIA
PSH Flag Count	Number of packets with PSH	PortScan	PSH Flag Count - C
Subflow Fwd Bytes	Average number of packets in subflow in forward direction	Benign, SSH-Patator, Web Attack, Bot, Heartbleed, Infiltration	Subflow Fwd Bytes - CIA
Total Length of Fwd Packets	Total size of forward packets	FTP-Patator, Benign, SSH-Patator, Web Attack, Bot, Heartbleed, Infiltration	Total Length of Fwd Packets - CIA

4. Experimental Setup

4.1. Dataset

The USBIDS dataset was not only chosen for its comprehensive feature explanations but also served as the foundational data for model training in our study. Comprising seventeen labelled CSV files, this dataset encapsulates a breadth of network traffic information. It includes sixteen files that detail a range of non-standard network conditions, with one file exclusively documenting benign traffic flows that have not been subjected to attacks, alongside records of combined defence modules and Denial of Service (DoS) attack data. These network flows were meticulously measured using the CIC FlowMeter2, ensuring precise data for analysis. Each of the sixteen non-normative CSV files is named to provide immediate insight into the data collection context. For instance, 'HULK-NoDefense.csv' denotes network flows captured during the HULK attack, conducted without the deployment of defensive strategies. This dataset, with its explicit annotations and diverse traffic scenarios, provided a robust platform for training our model, enabling it to learn and adapt to a wide spectrum of network behaviours and potential security threats.

4.2. Experimental Environment

Our experimental setup was designed to evaluate the model's ability to distinguish between normal and anomalous network traffic. The model was trained solely on benign data, allowing it to learn the patterns of typical network behavior. For testing, we used a combination of benign data and two separate attack datasets, challenging the model to detect deviations indicative of network intrusions. The model's architecture was a fully connected autoencoder with a Rectified Linear Unit (RELU) activation function. The structure was intentionally kept simple, consisting of two hidden layers with 10 and 32 neurons, respectively, to capture essential data patterns while maintaining a lightweight design. An anomaly detection threshold was established by calculating the maximum Mean Absolute Error (MAE) during the training phase with benign data. This threshold was key in differentiating between normal traffic and potential threats during testing. The implementation of our proposed algorithm was carried out in Python, utilizing TensorFlow Lite and the Keras library for their efficiency and ease of use. The Adam optimizer was employed for model optimization due to its strong performance across a variety of conditions. The training and testing processes were conducted over 40 epochs, with a learning rate of 0.01 to balance speed and accuracy.

The hardware used for our experiments included an ASUS ZenBook with a 2.30 GHz Intel Core i7 processor and 16 GB of RAM, ensuring fast computation and high efficiency. Additionally, a Raspberry Pi Model B with 4 GB of RAM was utilized, demonstrating the model's adaptability and its potential for deployment in resource-constrained IoT environments. The experiment utilized a comprehensive dataset that included both benign and malicious network traffic. The dataset was normalized before being processed by the trained autoencoder. Anomaly thresholds were derived from the reconstruction error distribution of the benign samples. Concurrently, the knowledge graph was populated with feature-specific thresholds and rules informed by network security expertise.

5. Evaluation and Adjustment

5.1. Case 1 Experiment with Data-driven Knowledge Graph

In the first case study, we conducted an evaluation of our model using the USBIDS dataset, complemented by a data-driven knowledge graph. The initial phase involved training the model with the dataset and subsequently testing it to validate its performance. During testing, we determined the most influential features for each anomalous instance, which served as a critical step in understanding the anomalies. Subsequently, we constructed a knowledge graph. This construction process was based on identifying the maximum Mean Absolute Error (MAE) from the benign data during the reconstruction error analysis. For each feature corresponding to this maximum MAE, we recorded its original values.

After establishing the knowledge graph, we conducted tests on the model using a distinct set of attack data. This step was crucial for assessing the model's practical effectiveness and its ability to differentiate between normal network operations and potential security threats. In our evaluations of various models, the one described earlier stood out due to its exceptional performance in diverse attack scenarios. Specifically, it achieved a 0.98 detection rate for the 'Attack Hulk No Defense', and it successfully identified both the 'Attack Hulk Evasive' and the 'Attack Hulk Reqtimeout' scenarios with perfect scores of 1.0. Notably, when tested against the combined dataset comprising all 16 attack types, the model maintained an overall accuracy of

96.8%Post detection, each instance marked anomalous undergoes a reasoning phase where decisions are assessed against the knowledge graph. This phase aims not only to validate the anomalies but also to iteratively refine the model by incorporating new insights and patterns observed in the data as Table 3. This model significantly reduces the rate of false positives compared to current state-of-the-art approaches by validating identified anomalies with the knowledge graph. It distinguishes whether each anomaly represents a legitimate attack or just normal, anomalous behaviour.

Table 3: Proposed model comparison with the current state of the art [30]

Detection Method	Hulk No Defense	Hulk Evasive	Hulk Reqtimeout	Overall
DT	0.97	0.06	0.97	-
RF	0.98	0.00	0.98	-
DNN	0.67	0.05	0.66	-
Proposed model	0.98	1.0	1.0	0.96

5.2. Case 2 Nurosymbolic integration

In the second experimental scenario, we utilized a dataset uniquely compiled by our team, which was gathered from various IoT environments, each with its distinct context. In our experiment, we utilized a real-time IoT network to gather network traffic data, focusing on the impact of various types of attacks on a target device. The experiment spanned five days within a smart home network environment, consisting of eight IoT devices and three non-IoT devices.The IoT devices, procured from local stores, varied in types and functions. This diversity was crucial to understanding how different devices generate traffic and interact within the network. All IoT devices were connected via Wi-Fi, while the router was categorized as a non-IoT device. For network traffic capture, we employed Wireshark [2] and the CICFlowMeter [3] tools. When addressing the complexity of implementation, we leveraged a distributed architecture tailored for scalability and practical deployment. The anomaly detection component was deployed on a Raspberry Pi 4 Model B within the smart home network, functioning as an edge device, while computationally intensive tasks such as threat intelligence processing, validation with a knowledge-graph-driven framework, and explain-

[2] https://www.wireshark.org/

[3] https://github.com/ahlashkari/CICFlowMeter

able AI reasoning were handled on an edge server. To optimize the anomaly detection model for resource-constrained edge devices, we applied pruning and quantization techniques to the autoencoder, significantly reducing its memory and computational footprint without compromising detection accuracy. This optimization enables real-time anomaly detection on lightweight devices, such as the Raspberry Pi, ensuring efficient performance in small-scale IoT networks. This framework dynamically retrieves data from the MITRE ATT&CK framework and other threat intelligence sources to contextualize detected anomalies. Pre-trained Large Language Models (LLMs) were accessed via APIs to process the retrieved data and generate explanations without requiring local hosting or fine-tuning. This approach significantly reduces resource requirements, enabling broader adoption in resource-constrained IoT environments.

Wireshark facilitated manual experiments, capturing live data traffic, whereas the CICFlowMeter was instrumental in extracting features from the PCAP files.To validate the robustness of our proposed model against real-world attack scenarios, we employed modern and actively maintained open-source tools. SlowHTTPTest was used to simulate low-rate and application-layer DDoS attacks, testing the model's ability to detect stealthy, low-traffic threats. Hping3 was utilized to craft custom packets and simulate both low-rate and volumetric DDoS attacks, providing comprehensive coverage of network-based attack vectors. A specific device was designated to simulate attack traffic towards the victim device, replicating several scenarios and conditions akin to those in the USBIDS dataset. The generated attack data was meticulously recorded and saved in CSV format for subsequent experimental analysis. Then we experimented with the above model without changing knowledge graph values. It reduces the accuracy of the model significantly and increases the false positives as shown in Table 3.

Then we consulted a few cybersecurity experts from academia and industry and asked them to update the knowledge graph values based on their expertise. They closely monitored the network traffic, and they updated the values of the knowledge graph based on their expertise as shown in Algorithm 2. For this, we gave another function to update features of the existing data-driven knowledge graph as shown in algorithm. after updating all the corresponding most influential features respective to detect legitimate attacks and again we have done the experiment with this dataset with an updated knowledge graph and model. It achieves higher accuracy for the overall model as shown in comparison in Table 3. Our model's accuracy is determined through a systematic process. Firstly, we establish ground truth by selecting a labelled dataset distinct from our training data and categorizing instances as 'normal' or 'anomalous.' Next, we deploy our trained autoencoder on this dataset to detect anomalies. During this phase, SHAP values are calculated for each instance to pinpoint the most influential features. We then consult our knowledge graph, which uses Max MAE values, to assess whether the detected anomalies signify actual attacks. Finally, we compare our model's predictions against the dataset's ground truth, identifying true positives, false negatives, false positives, and true negatives. This method provides a thorough evaluation of our model's ability to accurately detect anomalies.

Algorithm 2 Update Node Attributes in a Graph

function UPDATE_NODE_ATTRIBUTES($graph$, $feature$, new_value)
  if $graph$ has a node with the given $feature$ **then**
    $graph.nodes[feature]['original\_value'] \leftarrow new\_value$
  else
    print "Feature '$feature$' not found in the graph."
  end if
end function

Manually updating the graph with new values:
UPDATE_NODE_ATTRIBUTES(G, 'Flow Packets/s', 21830)
UPDATE_NODE_ATTRIBUTES(G, 'PSH Flags', 15)

Table 4 showcases the accuracy of our model, which integrates expert knowledge, compared to the performance of a purely data-driven knowledge graph in our IoT network setup. This comparison highlights that IoT networks are highly context-sensitive systems, making it challenging for data-driven approaches to generalize across diverse IoT infrastructures effectively. In such scenarios, our neuro symbolic approach demonstrates a higher attack detection rate with a minimal false positive rate. This is primarily due to our model's ability to adapt system features by integrating expert knowledge pertinent to each specific context. In addition to enhancing detection accuracy, the model also elucidates the underlying factors of each identified attack by pinpointing the most influential features. This level of detailed explanation proves invaluable for cybersecurity professionals, empowering them to make informed decisions and take appropriate actions in response to the detected threats.

Table 4: Comparison of Model Accuracy: Data-Driven (DDKG) vs. Expert Knowledge Integrated Knowledge Graph (EKIKG) on the real-time IoT data

Detection Method	No Defense	Evasive	Reqtimeout	Overall
DDKG	0.91	0.94	0.93	0.91
EKIKG	0.98	0.99	0.98	0.97

We acknowledge that reliance on a static knowledge graph may limit the model's ability to adapt to entirely novel threats that do not align with predefined patterns. To address this limitation, we have implemented mechanisms for continuous updating of the knowledge graph, as detailed in Algorithm 2. By integrating expert feedback and real-time threat intelligence, the graph evolves dynamically to include emerging attack patterns and novel vulnerabilities. Furthermore, our approach combines the knowledge graph with SHAP-based feature importance ranking and anomaly detection. This hybrid methodology enables the model to identify and highlight unknown threats based on data-driven anomalies, even when the knowledge graph lacks corresponding patterns. In future work, we plan to automate the updating process of the knowledge graph by leveraging reinforcement learning techniques and incorporating insights from network traffic features mapped to the MITRE ATT&CK framework and open threat intelligence data. This will enable the system to adapt dynamically to the evolving threat landscape, reducing dependence on manual updates and ensuring its robustness against novel threats.

5.3. Expert knowledge based Treat Intelligence and Response

After the model identifies an anomaly, it validates the detected attack using expert knowledge. As demonstrated in Table 4, the expert knowledge-integrated model outperforms traditional models. Following this, the model's knowledge extractor identifies the domain-specific features and maps them to the most influential features of the detected attack. It then determines the violated cybersecurity components, such as confidentiality, integrity, or availability, providing a detailed explanation of the compromised aspects of the networks shown in Figure 2. In the next step of the model, we integrate a Large Language Model (LLM) alongside the MITRE ATT&CK API to generate natural language explanations for detected anomalies based on the mapped feature values and corresponding MITRE ATT&CK techniques. This integration enhances the model's interpretability by delivering human-readable explanations (as shown Figure 2) that network security analysts can easily understand. We use OpenAI's API to generate these explanations, where anomalous feature values—such as Flow packets per second (PPS), SYN flags, and port activity—are fed into the GPT model, alongside relevant MITRE ATT&CK techniques retrieved via the MITRE ATT&CK API. The GPT model then generates a natural language explanation, detailing the potential implications of the anomaly and its impact on network security. For instance, detecting a high PPS rate and an abnormal number of SYN flags may indicate a Denial of Service (DoS) attack, while unusual port activity could point to Network Scanning, a common precursor to more advanced attacks.After obtaining the results, we validated them against the MITRE ATT&CK framework by manually (as shown in Figure 3 and Figure 4) verifying the findings as part of the experimental process and proof of concept. The results were 100% accurate in identifying threats, as confirmed through this manual validation process. However, further experimentation and automated validation are necessary to ensure the model's consistent performance and scalability. Ongoing work will focus on refining the validation process and improving overall accuracy.

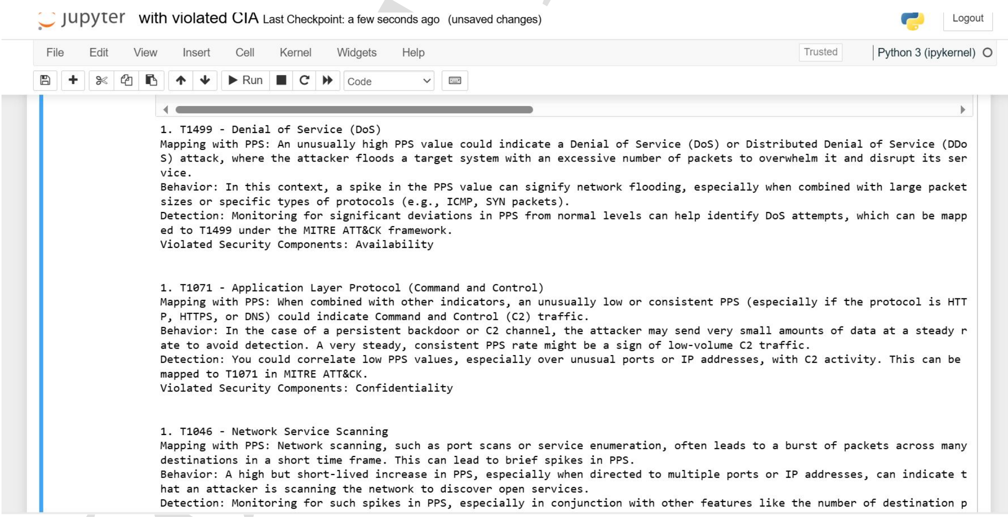

Figure 2: Automated Threat Inteligence and Response.

This step significantly improves the system by enabling it not only to detect and map anomalies but also to explain them in a manner accessible to non-experts. The AI-driven explanations, combined with the MITRE ATT&CK framework, help reduce the workload on security analysts by providing immediate, context-aware insights, allowing them to better understand potential threats and respond more efficiently. By incorporating GPT-generated explanations and leveraging the MITRE ATT&CK API, the system bridges the gap between machine-driven anomaly detection and human interpretation, enhancing its capability to provide actionable intelligence in dynamic cybersecurity environments.

5.4. Results Discussion

Our evaluation of the proposed neurosymbolic learning model clearly demonstrates its superiority over traditional models in both accuracy and interpretability. Notably, the model achieved an overall detection accuracy of 0.97 by integrating domain knowledge, which significantly enhanced its ability to verify the legitimacy of detected attacks. Furthermore, the integration with the MITRE ATT&CK framework enabled 100% accuracy for threat intelligence, validating each detected anomaly with corresponding tactics, techniques, and mitigations.

Compared to state-of-the-art detection models such as deep learning-based IDS (e.g., Autoencoder, Random Forest, and Decision Tree classifiers), our model not only excelled in detection rates but also substantially reduced false positives. The combined use of SHAP values for feature importance ranking and the knowledge graph for attack legitimacy validation ensures that security measures are both verifiable and actionable. This approach bridges the gap between anomaly detection and real-time threat intelligence by providing contextual explanations that are aligned with cybersecurity standards like CIA (Confidentiality, Integrity, and Availability).

Moreover, the capability to map detected anomalies to the MITRE ATT&CK framework provides deeper insights into potential attack patterns, allowing for faster and more accurate threat responses. This dual-layered system—comprising neural anomaly detection with symbolic reasoning—ensures that IoT environments are protected against evolving threats while minimizing human intervention.

To address the scalability challenges inherent in IoT networks, we deployed the detection component of our model on a Raspberry Pi 4 B model as an edge device using TensorFlow Lite. Experimental results demonstrated

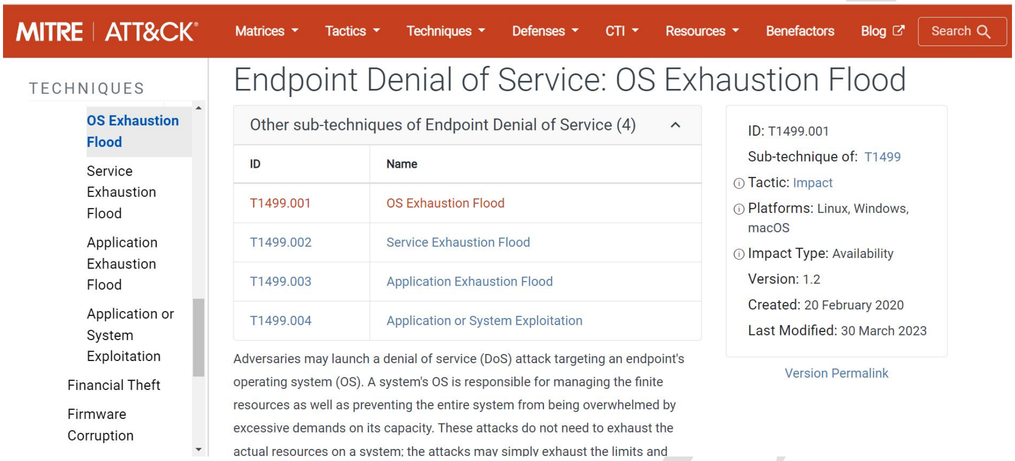

Figure 3: This image shows the MITRE ATT&CK T1499 details, which match the response generated by our model.

that the model could effectively operate on resource-constrained environments, including Microcontroller Processor Units (MCUs), while maintaining real-time detection capabilities. The intelligence components, such as explainable AI processing, knowledge graph mapping, and MITRE ATT&CK framework integration, were hosted on an edge server with higher computational resources. This distributed architecture reduced communication overhead and computational bottlenecks while ensuring low-latency threat detection and response.

In future work, we aim to deploy the detection model on ESP32 devices as edge IoT devices, leveraging Real-Time Operating System (RTOS)-enabled machine learning techniques. This step will extend the applicability of our approach to ultra-resource-constrained environments, further enhancing its scalability and practicality in diverse IoT scenarios.

By achieving real-time threat intelligence and response, our model outperforms existing solutions by enabling quicker, more efficient decision-making processes, as well as better adaptability to new and unseen attack vectors. This advance significantly strengthens IoT network defenses, as demonstrated through rigorous experimental validation

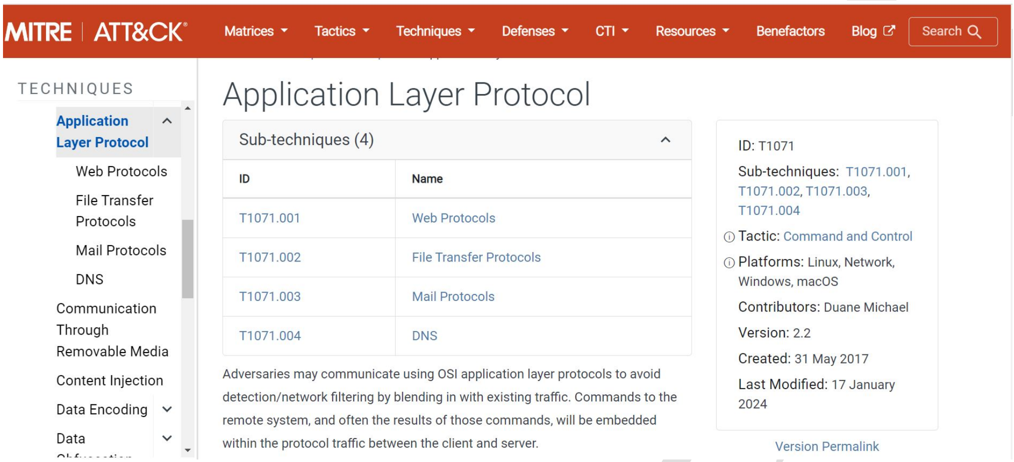

Figure 4: This image shows the MITRE ATT&CK T1071 details, which match the response generated by our model.

6. Conclusion

This study presents an innovative neurosymbolic approach for detecting attacks in IoT networks by integrating neural network-based autoencoders with SHAP explanations and expert-enhanced knowledge graphs. This method outperformed traditional models by accurately identifying and explaining attacks, leveraging SHAP values and expert knowledge to effectively distinguish between genuine threats and benign activities. By focusing on key features for anomaly detection, the model delivered detailed, context-aware explanations, essential for navigating the complexity and diversity of IoT networks.

The experimental validation, conducted using the USBIDS dataset and real IoT network data, showcased the model's superior accuracy and reduced false positive rate, highlighting its adaptability and deep understanding of network security. The success of this neurosymbolic model in real-world applications underscores its potential for advancing cybersecurity, especially in improving the interpretability and reliability of anomaly detection systems. As IoT networks continue to expand, such innovative solutions are crucial for defending against increasingly sophisticated cyber threats. Future work will apply this model to various IoT environments, including critical infrastructure, to further enhance its applicability. This research marks a significant advancement in IoT security and sets the stage for continued exploration of neurosymbolic AI, offering promising prospects for reducing human involvement and accelerating threat intelligence and response processes.

References

[1] A. Sheth, K. Roy, M. Gaur, Neurosymbolic artificial intelligence (why, what, and how), IEEE Intelligent Systems 38 (3) (2023) 56–62.

[2] S. Kambhampati, Polanyi's revenge and ai's new romance with tacit knowledge, Communications of the ACM 64 (2) (2021) 31–32.

[3] A. Joshi, N. Ramakrishman, E. N. Houstis, J. R. Rice, On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques, IEEE Transactions on Neural Networks 8 (1) (1997) 18–31.

[4] W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics 5 (1943) 115–133.

[5] C. S. Kalutharage, X. Liu, C. Chrysoulas, Explainable ai and deep autoencoders based security framework for iot network attack certainty, in: International Workshop on Attacks and Defenses for Internet-of-Things, Springer, 2022, pp. 41–50.

[6] C. S. Kalutharage, X. Liu, C. Chrysoulas, N. Pitropakis, P. Papadopoulos, Explainable ai-based ddos attack identification method for iot networks, Computers 12 (2) (2023) 32.

[7] B. Kaur, S. Dadkhah, F. Shoeleh, E. C. P. Neto, P. Xiong, S. Iqbal, P. Lamontagne, S. Ray, A. A. Ghorbani, Internet of things (iot) security dataset evolution: Challenges and future directions, Internet of Things (2023) 100780.

[8] M. Cirillo, M. Di Mauro, V. Matta, M. Tambasco, Botnet identification in ddos attacks with multiple emulation dictionaries, IEEE Transactions on Information Forensics and Security 16 (2021) 3554–3569.

[9] V. Matta, M. Di Mauro, M. Longo, Botnet identification in multi-clustered ddos attacks, in: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2017, pp. 2171–2175.

[10] Y. Xiang, K. Li, W. Zhou, Low-rate ddos attacks detection and traceback by using new information metrics, IEEE transactions on information forensics and security 6 (2) (2011) 426–437.

[11] D. Tang, Y. Yan, C. Gao, W. Liang, W. Jin, Ltrft: Mitigate the low-rate data plane ddos attack with learning-to-rank enabled flow tables, IEEE Transactions on Information Forensics and Security 18 (2023) 3143–3157.

[12] A. Piplai, A. Joshi, T. Finin, Offline rl+ ckg: A hybrid ai model for cybersecurity tasks, UMBC Faculty Collection (2023).

[13] A. Piplai, P. Ranade, A. Kotal, S. Mittal, S. N. Narayanan, A. Joshi, Using knowledge graphs and reinforcement learning for malware analysis, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 2626–2633.

[14] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, A. Joshi, Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy, IEEE Internet Computing 27 (5) (2023) 43–48.

[15] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, R. Zak, Creating cybersecurity knowledge graphs from malware after action reports, IEEE Access 8 (2020) 211691–211703.

[16] B. Chandrasekaran, M. C. Tanner, J. R. Josephson, Explaining control strategies in problem solving, IEEE Intelligent Systems 4 (01) (1989) 9–15.

[17] G. DeJong, Generalizations based on explanations., in: IJCAI, Vol. 81, 1981, pp. 67–69.

[18] S. C.-H. Yang, P. Shafto, Explainable artificial intelligence via bayesian teaching, in: NIPS 2017 workshop on teaching machines, robots, and humans, Vol. 2, 2017.

[19] K. Li, Y. Cui, W. Li, T. Lv, X. Yuan, S. Li, W. Ni, M. Simsek, F. Dressler, When internet of things meets metaverse: Convergence of physical and cyber worlds, IEEE Internet of Things Journal 10 (5) (2022) 4148–4173.

[20] K. Li, B. P. L. Lau, X. Yuan, W. Ni, M. Guizani, C. Yuen, Towards ubiquitous semantic metaverse: Challenges, approaches, and opportunities, IEEE Internet of Things Journal (2023).

[21] E. Hodo, X. Bellekens, E. Iorkyase, A. Hamilton, C. Tachtatzis, R. Atkinson, 2 intrusion detection system.

[22] N. Shone, T. N. Ngoc, V. D. Phai, Q. Shi, A deep learning approach to network intrusion detection, IEEE transactions on emerging topics in computational intelligence 2 (1) (2018) 41–50.

[23] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.

[24] J. R. Goodall, E. D. Ragan, C. A. Steed, J. W. Reed, G. D. Richardson, K. M. Huffer, R. A. Bridges, J. A. Laska, Situ: Identifying and explaining suspicious behavior in networks, IEEE transactions on visualization and computer graphics 25 (1) (2018) 204–214.

[25] N. Liu, D. Shin, X. Hu, Contextual outlier interpretation, arXiv preprint arXiv:1711.10589 (2017).

[26] D. Collaris, L. M. Vink, J. J. van Wijk, Instance-level explanations for fraud detection: A case study, arXiv preprint arXiv:1806.07129 (2018).

[27] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, C. Siemens, Drebin: Effective and explainable detection of android malware in your pocket., in: Ndss, Vol. 14, 2014, pp. 23–26.

[28] V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, Castle: Cluster-aided space transformation for local explanations, Expert Systems with Applications 179 (2021) 115045.

[29] C. S. Kalutharage, X. Liu, C. Chrysoulas, O. Bamgboye, Utilizing the ensemble learning and xai for performance improvements in iot network attack detection, in: European Symposium on Research in Computer Security, Springer, 2023, pp. 125–139.

[30] M. Catillo, A. Del Vecchio, A. Pecchia, U. Villano, Transferability of machine learning models learned from public intrusion detection datasets: the cicids2017 case study, Software Quality Journal 30 (4) (2022) 955–981.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: