Neurosymbolic learning in the XAI framework for enhanced cyberattack detection with expert knowledge integration

Chathuranga Sampath Kalutharage[1], Xiaodong Liu[1][0000−0002−7612−9981], Christos Chrysoulas[1][0000−0001−9817−003X], and Oluwaseun Bamgboye[1]

Edinburgh Napier University, Scotland, UK
c.kalutharage,x.liu,c.chrysoulas,O.Bamgboye@napier.ac.uk
https://www.napier.ac.uk/

Abstract. The perpetual evolution of cyberattacks, especially in the realm of Internet of Things (IoT) networks, necessitates advanced, adaptive, and intelligent defence mechanisms. The integration of expert knowledge can drastically enhance the efficacy of IoT network attack detection systems by enabling them to leverage domain-specific insights. This paper introduces a novel approach by applying Neurosymbolic Learning within the Explainable Artificial Intelligence (XAI) framework to enhance the detection of IoT network attacks while ensuring interpretability and transparency in decision-making. Neurosymbolic Learning synergizes symbolic AI, which excels in handling structured knowledge and providing explainability, with neural networks, known for their prowess in learning from data. Our proposed model utilizes expert knowledge in the form of rules and heuristics, integrating them into a learning mechanism to enhance its predictive capabilities and facilitate the incorporation of domain-specific insights into the learning process. The XAI framework is deployed to ensure that the predictive model is not a "black box," providing clear, understandable explanations for its predictions, thereby augmenting trust and facilitating further enhancement by domain experts. Through rigorous evaluation against benchmark IoT network attack datasets, our model demonstrates superior detection performance compared to prevailing models, along with enhanced explainability and the successful incorporation of expert knowledge into the adaptive learning process. The proposed approach not only fortifies the security mechanisms against network attacks in IoT environments but also ensures that the knowledge discovery and decision-making processes are transparent, interpretable, and verifiable by human experts.

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

1 Introduction

In the constantly changing world of cyber security, identifying and preventing cyber threats is a critical challenge. Traditional security methods are often not sufficient against complex and evolving cyber-attacks. This situation calls for a new strategy that not only improves threat detection but also makes the process behind these detections more transparent and understandable. This paper introduces a novel approach that blends advanced learning techniques with human expert knowledge for better detection of cyber threats. This innovative method combines the best of two worlds: the pattern recognition ability of advanced learning systems and the logical problem-solving of human-like reasoning. By integrating insights from cybersecurity experts, this approach ensures that threat detection is not only based on extensive data analysis but also enriched with deep, field-specific knowledge. Our research focuses on how this combined method can improve the detection of cyber threats. We examine its ability to handle large amounts of data, identify complex patterns, and explain its findings in a clear and understandable manner. The importance of including expert knowledge to refine and ensure the accuracy of the detection process is also highlighted. We start by discussing the current issues in detecting cyber threats and then introduce the concepts behind our advanced, integrated approach. Our paper presents a new model that applies this method, showing its effectiveness in identifying a variety of cyber threats. We also discuss the challenges of implementing such a system and offer potential solutions. The primary goal of our study is to show that this approach is not just a theoretical idea but a practical and powerful tool for defending against cyber threats. By offering improved detection capabilities along with clear explanations, this method represents a significant advancement in cyber security, providing stronger protection against an increasingly diverse range of cyber threats. This model is based on neurosymbolic Artificial intelligence and eXplainable Artificial Intelligence (XAI).

neurosymbolic artificial intelligence is a blend of neural network-based methods and symbolic knowledge-based approaches. This hybrid technique capitalizes on the strengths of both: neural networks are adept at processing vast amounts of data and discerning intricate patterns from raw input, while symbolic approaches excel in logical reasoning. The integration of these methods not only harnesses the representational abilities of neural networks but also addresses their common challenge of providing transparent explanations for their decisions [12]. The limited application of neural networks beyond academic and commercial research settings, despite a decade of promising development beginning in the mid-1980s, is partly due to certain limitations. In contrast, symbolic knowledge-based methods, such as rule-based systems or expert systems, utilize logical reasoning and clear knowledge representations. These approaches excel in accumulating domain-specific knowledge and providing transparent explanations for their conclusions [12], [5]. However, these techniques often struggle with handling ambiguous or incomplete data and typically lack the capacity to derive insights from vast datasets [12]. Over the past five years, there has been a surge of interest in NeuroSymbolic AI, an approach that integrates neural and symbolic AI methodologies. The fusion of these two paradigms is not a new concept; the term 'Neural-Symbolic' was first introduced as early as the early 2000s [5]. The 1990s witnessed several efforts to integrate fuzzy rules and connectionist approaches [2]. The idea of merging intuitive and logical aspects of AI was hinted in the groundbreaking paper ”A Logical Calculus of the Ideas Immanent in Nervous Activity” by McCulloch and Pitts [7]. The renewed interest in this method can be linked to various reasons, which we will examine within the scope of cybersecurity. In this paper, we integrate neurosymbolic artificial intelligence with our previously developed explainable artificial intelligence model [3], [4]. 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).
- Evaluate the model's performance by comparing it with existing research in the field.

The structure of the rest of the paper is organized as follows: Section 2 presents background and related work. Section 3 details the proposed algorithm. Section 4 discusses the experimental setup, while Section 5 describes the evaluation process and any adjustments made. Finally, the paper concludes in Section 6.

2 Background and Related work

2.1 IoT Network Attacks

The Internet of Things (IoT) is an expanding network of interconnected devices ranging from simple sensors to sophisticated industrial machinery. This interconnectivity, while advantageous, also increases vulnerability to cyber threats, such as DDoS, Man-in-the-Middle, ransomware, data theft, device hijacking, and side-channel attacks.

Security in IoT faces multiple challenges due to device diversity, varying protocols, and the sheer volume of devices, many of which have limited computational resources that impede the use of advanced security measures [4]. The fast-paced growth of the IoT sector demands scalable security solutions. Ensuring data privacy while maintaining effective security is a critical concern. A further complication is the infrequency of security updates for many IoT devices, leaving them open to exploitation. The complexity of IoT ecosystems makes it challenging to pinpoint attack origins and types. Real-time detection and response are essential for the integrity of IoT operations, especially since many devices are physically accessible, increasing the risk of tampering. These unique challenges highlight the importance of developing innovative security strategies for IoT networks [6].

2.2 neurosymbolic AI in cybersecurity

Neurosymbolic AI aims to bring together the best of both worlds: the learning and pattern recognition capabilities of neural networks and the interpretability and logical reasoning of symbolic AI. Firstly, by employing a blend of data-driven methods and symbolic reasoning, it is possible to trace the sequence of events or actions leading to the model's conclusion. This forms a compelling case for adopting neurosymbolic approaches in cybersecurity and privacy [8]. Such approaches are particularly valuable in addressing challenges like threat detection and analysis, where it is crucial to contextualize patterns observed across different systems over time, rather than merely detecting those [11]. neurosymbolic methods are capable of achieving this while maintaining privacy, such as by integrating privacy policies, regulations, and compliance measures. For instance, a neurosymbolic model can apply logical reasoning to manage the use of sensitive network flow data by the neural network detector, adhering to explicit privacy policies. Additionally, it can ensure compliance through the use of privacy-protecting techniques like differential privacy or secure multi-party computation [9]. Secondly, ensuring the safety and security of AI systems is crucial. The reliance on data-driven models for automated vulnerability assessment can be limiting, as these systems only learn from the vulnerabilities they are trained on. Adopting a neurosymbolic approach can enhance safety. In this approach, experts act as simulated adversaries during the training of AI-based software systems. This enables the AI model to dynamically learn and apply rules and policies, rather than relying solely on pre-trained vulnerabilities [12]. Moreover, an AI system's reliability and security can be significantly enhanced if knowledge from security specification documents is explicitly encoded using symbolic methods and applied as behavioural constraints. This aspect is of immediate interest to legislators and regulators in many countries. Without human knowledge or expertise, advanced AI systems run a substantial risk of generating information that could be hazardous or harmful.

Another reason for the usefulness of combining rule-based and data-driven methods is the scarcity of high-quality data needed for reliable conclusions. This challenge is common in fields where sensitive data is scarce or challenging to distribute for experiments. Nevertheless, alternative sources, such as textual descriptions of the sensitive data, might be accessible. These alternatives can be utilized to establish common rules. When the available data alone is not enough to make reliable conclusions, these derived rules can be used to reinforce the conclusions drawn from the data [5]. During the learning process, they can also be supplied as an input to the data-driven model. Furthermore, certain areas are highly dynamic, with data that accurately represents conditions only for a brief period. Consequently, conclusions drawn from such data may also be short-lived.

This situation is particularly relevant in fields like fraud detection and cybersecurity. Patterns identified from our existing dataset may be effective against current cyberattacks but may not remain relevant in the future. In scenarios like these, combining deep network-based detection systems with explicit rules that account for changes in data trends or the time-limited applicability of a model can be advantageous [10].

Neurosymbolic AI, which combines symbolic AI with neural networks, is increasingly relevant in cybersecurity. It enhances areas such as threat intelligence, malware analysis, intrusion detection, and vulnerability assessment, thereby bolstering overall security system effectiveness [11]. This approach is pivotal in evolving Security Operations Centres (SoCs) into next-generation facilities. Here, the integration of AI methods with human monitoring creates a more sophisticated and efficient system for managing and responding to security threats. Consider a scenario where security analysts, working in a Security Operations Centre, play a vital role in upholding an organization's security. Their effectiveness in detecting attacks relies significantly on their experience and knowledge of emerging and novel threats. This expertise is especially important when interpreting outputs from deep neural networks or machine learning (ML)-based systems that analyze incoming data streams. Their prior understanding of new attack patterns is crucial in effectively identifying potential security breaches. To assist analysts, we can collect information from publicly available threat intelligence sources, such as threat feeds or detailed accounts of cyberattacks. This data is then organized and stored in a Cybersecurity Knowledge Graph (CKG). We propose two methods that utilize the structured data within CKGs for subsequent tasks, focusing particularly on explainability through reasoning and inference. The first method involves creating complex rules using an existing knowledge engine and actual data, forming a rule-based framework. The second method revolves around developing new cybersecurity strategies (knowledge-guided models) by incorporating established rules into subsequent data-driven AI models.

The primary aim of a rule-based framework is to develop the most effective and robust rules possible for protecting target machines from all forms of threats and hostile activities. These rules, varying from simple to complex, are to be applied to any system or subsystem needing defence. The focus on knowledge-guided models is to tackle emerging or evolving cyberthreats that aren't covered in existing datasets for data-driven research. To identify novel adversaries and consequently new defence mechanisms, techniques such as Reinforcement Learning (RL) and other exploratory modelling methods are crucial. Our experiments have shown that Cybersecurity Knowledge Graphs (CKGs) can guide these exploratory learning methods, enhancing their efficiency, speed, and clarity

3 Proposed Model

3.1 Overview

This research introduces a novel neurosymbolic approach for anomaly detection in network data. The methodology synergistically combines neural network-based anomaly detection using autoencoders with symbolic reasoning via a knowledge graph. This integration leverages the strengths of both neural and symbolic AI, providing robust anomaly detection while enhancing interpretability and decision-making. Finally, we use a data-driven approach for the knowledge graph development and expert knowledge integration for the enhanced knowledge graph. Figure 1 illustrates the architecture of the model, with each components describe as follows.

1. IoT Network Traffic: This represents the data flow within an IoT network, which includes both normal operations and potential security threats.
2. 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.
3. Benign Traffic: This is the subset of network traffic that has been identified as normal and safe by the anomaly detection system.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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 about detected anomalies.

3.2 Neural Network-Based Anomaly Detection

The methodology is based on the use of an autoencoder, which is a type of neural network with a proficiency for creating compact representations of data. The autoencoder operates through two main processes: encoding and decoding. During encoding, it compresses network data into a lower-dimensional space, retaining the essential features. Subsequently, in the decoding process, the compressed

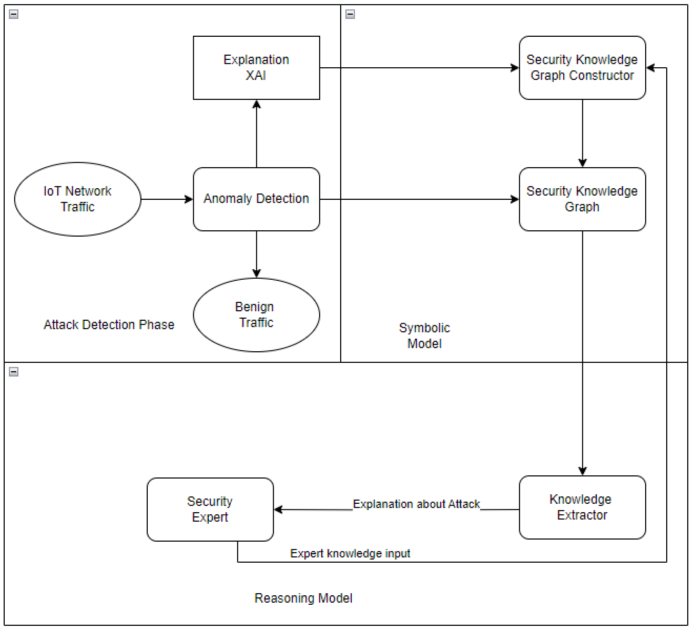

Fig. 1. Proposed Neurosymbolic learning in the XAI framework architecture for IoT attack detection.

data is expanded back to its original dimension. The critical metric used to evaluate the performance of an autoencoder in this setting is the reconstruction error, which assesses the discrepancy between the original data and the reconstructed output. A commonly employed measure for this error is the Mean Absolute Error (MAE). In the realm of anomaly detection with autoencoders, the MAE is especially significant. It provides insight into whether the reconstruction error exceeds a certain threshold, which would indicate an anomaly. This threshold is typically set based on the error distribution of normal data instances. The fundamental assumption is that normal data will have lower reconstruction errors, while anomalous data will exhibit higher errors due to significant deviations from the model's learned patterns.

3.3 Symbolic Reasoning with SHAP and Knowledge Graphs

To enhance the interpretability and decision-making capabilities of the model, we integrate SHAP (SHapley Additive exPlanations) values, grounded in game theory, to attribute significance to individual features in anomaly detection. SHAP values are instrumental in pinpointing the contribution of each feature to the detected anomalies, thereby unravelling the rationale behind the model's decisions. Specifically, for each anomalous instance detected by the model, SHAP values elucidate which features are most influential in signalling the anomaly, enabling a granular analysis of the model's behaviour. Parallel to this, we develop a domain-specific knowledge graph, leveraging real attack data to map anomalous behaviours indicative of genuine cybersecurity threats as shown in Algorithm 1. This knowledge graph, tailored to network security, constitutes a structured representation of expert insights and heuristic rules. Its nodes symbolize individual features of network data, while its edges represent the intricate relationships and constraints among these features. The graph effectively encapsulates the complex interplay of network characteristics that signify potential security breaches.

Crucially, in the context of detected anomalies, the knowledge graph utilizes the Maximum Mean Absolute Error (Max MAE) – a metric derived from the model's performance – to delineate normal from abnormal behaviour. The Max MAE reflects the highest deviation in reconstruction error when the model encounters an anomalous pattern. By associating the Max MAE with real feature values corresponding to known attack classes in the knowledge graph, we can ascertain whether a detected anomaly constitutes a legitimate attack or merely an unusual but benign network behaviour. In essence, the integration of SHAP values and the knowledge graph achieves a two-fold objective: Firstly, SHAP values provide an in-depth explanatory analysis of why certain instances are flagged as anomalies, based on feature contributions. Secondly, the knowledge graph validates these anomalies against real-world attack patterns, discerning genuine threats from false alarms. This dual approach not only bolsters the model's accuracy in detecting attacks but also offers a comprehensive understanding of the nature of each detected anomaly, ensuring robust and reliable network security.

3.4 Neurosymbolic Integration

Our methodology epitomizes the synergy of neural network outputs and symbolic reasoning, forming an integrated framework for anomaly detection in IoT networks. This begins with each data instance being evaluated by the autoencoder, which calculates the reconstruction error and SHAP values. These SHAP values are crucial, as they indicate the influence of individual features on the model's predictions. In this setup, SHAP values are instrumental. They are carefully assessed against predetermined thresholds and rules within a custom-built knowledge graph. Initially formed from data-driven insights, this graph encapsulates typical network behaviour and recognized anomaly patterns. Crucially, when the SHAP value identifies a feature as highly influential, the model checks the corresponding original feature value against the maximum Mean Absolute Error (MAE). If this value surpasses the feature's threshold in the knowledge graph, the instance is identified as an attack. IoT networks, known for their context-specific characteristics, present challenges in generalizing models. To address this, we enhance our initially data-driven knowledge graph with expert knowledge. This addition is vital, as it incorporates a deeper, more nuanced understanding of network behaviours and threat landscapes – aspects that might not be completely apparent from data alone. This integration of expert knowledge into the knowledge graph substantially improves the model's ability to detect and confirm anomalies. When an instance is flagged based on influential SHAP values, the model undertakes symbolic reasoning. This reasoning is grounded not only in data-driven thresholds but also in expert insights and rules. Such a comprehensive approach ensures more accurate and contextually relevant interpretations of anomalies and provides potential recommendations and actions. In essence, our methodology marries data-driven analysis with expert insights. While SHAP values direct us to the most significant features in detecting anomalies, the enhanced knowledge graph, enriched with expert understanding, corroborates these detections. This ensures that the model's interpretations and responses are precisely aligned with the complex and dynamic nature of IoT network security.

Algorithm 1 Neurosymbolic Anomaly Detection with SHAP and Knowledge Graph Integration

```
Require: $X$ — Anomaly instance that needs to explain, $X_{1..i}$ — 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 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 $detected\_anomalies$
```

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 environment was established with the objective of assessing the model's capability to discern between normal and anomalous network traffic. The training phase exclusively utilized benign data, fostering a model attuned to recognizing typical network behaviour. For testing, we integrated benign data with two distinct sets of attack data, challenging the model to identify deviations indicative of network attacks. The model's architecture was a fully connected autoencoder with a Rectified Linear Unit (RELU) activation function. It featured a concise network structure with just two hidden layers, deliberately designed to minimize the model's complexity. These layers comprised 10 and 32 neurons respectively, sufficient for capturing essential data patterns without overburdening the system. To define the threshold for anomaly detection, we computed the highest mean absolute error (MAE) during the training phase using benign data. This threshold was crucial for distinguishing between normal traffic flows and potential threats during the evaluation phase. The implementation of our proposed algorithm was executed using Python, leveraging TensorFlow lite and the Keras library for their efficiency and ease of use. The optimization of the model was facilitated by the Adam optimizer, selected for its robust performance in various conditions. Our training and testing processes spanned over 40 epochs, with a learning rate set at 0.01 to balance speed and accuracy. The hardware employed for our experiments included an ASUS ZenBook, equipped with a 2.30 GHz Intel Core i7 processor and 16 GB of RAM, ensuring swift computation and high efficiency. Additionally, a Raspberry Pi Model B with 4 GB of RAM was used, showcasing the model's adaptability and potential for deployment in resource-constrained environments typically found in IoT networks. Our experiment involves a comprehensive dataset comprising both benign and malicious network traffic. The dataset is first normalized and then fed into the trained autoencoder. The reconstruction error thresholds are determined based on the distribution of errors in benign samples. Meanwhile, the knowledge graph is 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 US-BIDS 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 1. 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 1. Proposed model comparison with the current state of the art [1]

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 [1] and the CICFlowMeter [2] tools. Wireshark facilitated manual experiments, capturing live data traffic, whereas the CICFlowMeter was instrumental in extracting features from the PCAP files. 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 2

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 2. 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.

Table 2 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

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

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

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)
```

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 2. 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

6 Conclusion

This study introduced a cutting-edge neurosymbolic method for detecting attacks in IoT networks, combining neural network-based autoencoders with SHAP explanations and expert-augmented knowledge graphs. This approach significantly outperformed traditional models by accurately identifying and explaining attacks, leveraging SHAP values and expert insights to effectively differentiate between actual attacks and benign activities. The focus on key features for anomaly detection enabled detailed, context-sensitive explanations, crucial in the diverse and interconnected environment of IoT networks.

Experimental validation using the USBIDS dataset and real IoT network data demonstrated the model's superior accuracy and lower false positives, highlighting its adaptability and deep insight into network security. This neurosymbolic model's success in a real-world setting points to a promising future for cybersecurity, emphasizing the role of neurosymbolic AI in improving anomaly detection systems' interpretability and reliability. With the ongoing expansion of IoT networks, such innovative approaches are vital for defending against complex cyber threats. Future work will incorporate Large Language Models for enhanced attack explanation and publish the collected IoT network data for research. This study not only marks a significant advancement in IoT security but also paves the way for further neurosymbolic AI research and applications in complex, dynamic domains.

References

1. Catillo, M., Del Vecchio, A., Pecchia, A., Villano, U.: Transferability of machine learning models learned from public intrusion detection datasets: the cicids2017 case study. Software Quality Journal **30**(4), 955–981 (2022)
2. Joshi, A., Ramakrishman, N., Houstis, E.N., Rice, J.R.: On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques. IEEE Transactions on Neural Networks **8**(1), 18–31 (1997)
3. Kalutharage, C.S., Liu, X., Chrysoulas, C.: Explainable ai and deep autoencoders based security framework for iot network attack certainty. In: International Workshop on Attacks and Defenses for Internet-of-Things. pp. 41–50. Springer (2022)
4. Kalutharage, C.S., Liu, X., Chrysoulas, C., Pitropakis, N., Papadopoulos, P.: Explainable ai-based ddos attack identification method for iot networks. Computers **12**(2), 32 (2023)
5. Kambhampati, S.: Polanyi's revenge and ai's new romance with tacit knowledge. Communications of the ACM **64**(2), 31–32 (2021)
6. Kaur, B., Dadkhah, S., Shoeleh, F., Neto, E.C.P., Xiong, P., Iqbal, S., Lamontagne, P., Ray, S., Ghorbani, A.A.: Internet of things (iot) security dataset evolution: Challenges and future directions. Internet of Things p. 100780 (2023)
7. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics **5**, 115–133 (1943)
8. Piplai, A., Joshi, A., Finin, T.: Offline rl+ ckg: A hybrid ai model for cybersecurity tasks. UMBC Faculty Collection (2023)
9. Piplai, A., Kotal, A., Mohseni, S., Gaur, M., Mittal, S., Joshi, A.: Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy. IEEE Internet Computing **27**(5), 43–48 (2023)
10. Piplai, A., Mittal, S., Joshi, A., Finin, T., Holt, J., Zak, R.: Creating cybersecurity knowledge graphs from malware after action reports. IEEE Access **8**, 211691–211703 (2020)
11. Piplai, A., Ranade, P., Kotal, A., Mittal, S., Narayanan, S.N., Joshi, A.: Using knowledge graphs and reinforcement learning for malware analysis. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 2626–2633. IEEE (2020)
12. Sheth, A., Roy, K., Gaur, M.: Neurosymbolic artificial intelligence (why, what, and how). IEEE Intelligent Systems **38**(3), 56–62 (2023)