IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COMPREHENSIVE SURVEY AND IMPLEMENTATION FRAMEWORK FOR DYNAMIC ATTACK DETECTION IN IOT USING EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Ms. Ankita G. Bhagat

Dr. Sushil V. Kulkarni

College of Engineering, Ambajogai

Abstract:

The Internet of Things (IoT) has become a fundamental component of modern digital infrastructures, enabling interconnected devices to exchange information and support intelligent services across diverse sectors, including healthcare, smart cities, industrial automation, and transportation. While the widespread adoption of IoT technologies has created significant opportunities for innovation and efficiency, it has also introduced serious security concerns. The heterogeneous nature of IoT devices, limited computational resources, and continuous connectivity make these environments attractive targets for counterattacks such as Distributed Denial-of-Service (DDoS), botnet activities, spoofing, malware infections, and data tampering. To address these threats, researchers have increasingly employed Machine Learning (ML) and Deep Learning (DL) techniques within Intrusion Detection Systems (IDS). Although these approaches have demonstrated promising detection capabilities, many of the developed models operate as complex black-box systems, making it difficult for security analysts to understand the reasoning behind their predictions. This review paper examines recent advancements in dynamic attack detection for IoT environments and explores the integration of explainable methods such as Shapely Additive explanations (SHAP) and Local Interpreter Model-Agnostic Explanations (LIME) into intrusion detection frameworks. The paper further reviews widely used datasets, machine learning and deep learning models, performance evaluation metrics, and key research challenges. Finally, it highlights emerging trends and future research opportunities aimed at developing adaptive, accurate, and trustworthy IoT security systems capable of addressing evolving cyber threats.

Index Terms - Internet of Things (IoT), Intrusion Detection System (IDS), Cybersecurity, Explainable Artificial Intelligence (XAI), Machine Learning, Deep Learning, SHAP, LIME, Dynamic Attack Detection.

I. INTRODUCTION

The rapid advancement of digital technologies has led to the widespread adoption of the Internet of Things (IoT), a network paradigm that enables physical devices to communicate, exchange data, and perform intelligent operations with minimal human intervention. By integrating sensors, actuators, communication technologies, and cloud-based services, IoT has transformed the way information is collected and utilized across various application domains. Today, IoT plays a vital role in smart homes, healthcare systems, industrial automation, transportation networks, environmental monitoring, and smart city infrastructures, contributing significantly to operational efficiency, automation, and real-time decision-making. Despite its numerous advantages, the growing deployment of IoT devices has introduced substantial security challenges. Most IoT devices are designed with limited processing power, memory capacity, and energy resources, often resulting in inadequate security mechanisms. Furthermore, the heterogeneous nature of IoT environments, combined with the use of diverse communication protocols and continuous internet connectivity, creates multiple entry points for malicious actors. As a result, IoT networks have become frequent targets of cyber threats, including Distributed Denial-of-Service (DDoS) attacks,

botnet-based intrusions, spoofing, ransomware, malware propagation, and unauthorized data manipulation. These attacks can compromise system availability, data integrity, user privacy, and overall network reliability.

To mitigate such threats, Intrusion Detection Systems (IDS) have been widely employed as a critical layer of defense in IoT environments. Conventional IDS approaches typically utilize signature-based or anomaly-based detection mechanisms to identify malicious activities. While signature-based methods are effective against known threats, they often fail to recognize newly emerging attack patterns. Similarly, anomaly-based approaches can detect unknown behaviors but frequently suffer from high false-positive rates. In recent years, Machine Learning (ML) and Deep Learning (DL) techniques have gained significant attention due to their ability to automatically learn complex traffic patterns and improve attack detection accuracy. Models such as Random Forests, Support Vector Machines, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have demonstrated promising performance in identifying both known and previously unseen cyberattacks.

However, despite their effectiveness, many AI-driven intrusion detection models function as black-box systems, providing little insight into how predictions are generated. The lack of interpretability presents a significant challenge, particularly in security-critical environments where understanding the reasoning behind a detection decision is essential for incident response, forensic analysis, regulatory compliance, and stakeholder trust. To address these concerns, Explainable Artificial Intelligence (XAI) has emerged as a promising research direction that enhances the transparency and interpretability of machine learning models. XAI techniques aim to provide meaningful explanations for model predictions, enabling users to better understand and validate automated decisions. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have gained considerable attention for their ability to explain both individual predictions and overall model behavior.

II. LITERATURE REVIEW

2.1 Machine Learning-Based IoT Attack Detection

Machine learning techniques have been widely adopted for intrusion detection in IoT networks. Algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), Naïve Bayes, and K-Nearest Neighbors have demonstrated promising results in identifying malicious network behavior. Al-Garadi et al. (2020) conducted a comprehensive survey on machine learning applications in IoT security and highlighted the effectiveness of supervised learning algorithms for detecting known attack patterns. However, these approaches often struggle to identify previously unseen attacks and require extensive feature engineering. Random Forest classifiers have gained popularity due to their robustness and ability to handle high-dimensional datasets. Nevertheless, despite their strong classification performance, understanding feature contributions remains challenging without additional interpretability techniques.

2.2 Deep Learning for IoT Security

Deep learning models have significantly advanced intrusion detection capabilities by automatically extracting complex feature representations from network traffic data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Autoencoders have been successfully applied to cybersecurity tasks. Lopez-Martin et al. (2019) proposed a deep learning framework capable of detecting IoT attacks with high accuracy. Their study demonstrated that deep neural networks outperform traditional machine learning algorithms in identifying sophisticated attack patterns. LSTM networks are particularly effective in analyzing sequential network traffic data and detecting temporal attack behaviors. Autoencoders have also been employed for anomaly detection by learning normal network behavior and identifying deviations indicative of cyberattacks. Despite their superior performance, deep learning models often suffer from a lack of transparency, making them difficult to deploy in security-critical environments where explainability is essential.

2.3 Explainable Artificial Intelligence in Cybersecurity

The increasing adoption of AI in cybersecurity has led to growing concerns regarding transparency and accountability. Explainable Artificial Intelligence seeks to address these concerns by providing human-understandable explanations for AI predictions. Ribeiro et al. (2016) introduced Local Interpretable Model-Agnostic Explanations (LIME), which explains individual predictions by approximating complex models with simpler interpretable models. LIME enables security analysts to identify the factors contributing to attack classifications. Similarly, Lundberg and Lee (2017) developed SHAP (SHapley Additive exPlanations), a game-theoretic approach that quantifies the contribution of each feature to a model's prediction. SHAP provides both local and global explanations, making it highly suitable for cybersecurity applications. Recent studies indicate that integrating SHAP and LIME with machine learning-based IDS frameworks significantly improves trust and usability by offering detailed insights into attack detection decisions.

III. DYNAMIC ATTACK DETECTION FRAMEWORK FOR IOT

Conventional Intrusion Detection Systems (IDS) employed in Internet of Things (IoT) environments primarily depend on predefined attack signatures and static learning mechanisms. Although these approaches are effective in identifying known threats, they often struggle to recognize newly emerging and sophisticated attack patterns. The dynamic and continuously evolving nature of cyber threats necessitates the development of adaptive security frameworks capable of learning from changing network behaviors and responding to previously unseen attacks. To address these challenges, a Dynamic Attack Detection Framework for IoT integrates advanced machine learning techniques, explainable artificial intelligence methods, and adaptive learning mechanisms to provide accurate, transparent, and real-time threat detection.

The framework begins with the Data Collection Layer, which is responsible for gathering network traffic generated by various IoT devices and communication infrastructures. Data can be collected through packet capture tools such as Wireshark, network monitoring systems, and traffic analyzers deployed within IoT environments. In addition to real-time traffic acquisition, benchmark datasets such as CICIoT2023 and CICIDS2017 are frequently utilized for model training, validation, and performance evaluation. Following data acquisition, the information is forwarded to the Data Preprocessing Layer, where raw network traffic is transformed into a structured format suitable for machine learning and deep learning models. This stage involves several important operations, including handling missing values, eliminating redundant information, extracting meaningful network features, and selecting the most relevant attributes for classification. Furthermore, normalization techniques are applied to standardize feature values and improve model convergence. Since cybersecurity datasets often contain imbalanced distributions between normal and malicious traffic, class balancing methods are also employed to ensure fair and effective model training.

IV. EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES FOR IOT ATTACK DETECTION

4.1 Shapely Additive explanation (SHAP)

SHapley Additive exPlanations (SHAP) is a widely used Explainable Artificial Intelligence (XAI) technique that interprets machine learning predictions by quantifying the contribution of individual features to the final decision. Based on cooperative game theory, SHAP assigns an importance value to each feature, enabling both local and global interpretation of model behavior. In IoT intrusion detection systems, SHAP helps identify critical network characteristics such as packet size, protocol type, flow duration, and traffic statistics that significantly influence attack classification. The primary advantage of SHAP is its ability to provide consistent and reliable explanations across different machine learning and deep learning models. By highlighting feature importance, SHAP enhances transparency, supports security analysis, and improves trust in AI-driven attack detection systems. However, the computational complexity of SHAP may increase when applied to large-scale IoT datasets and complex deep learning architectures.

4.2 Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) is an explainability technique designed to explain individual predictions generated by complex machine learning models. It works by creating a simple interpretable model around a specific prediction and analyzing the influence of nearby data samples. Since LIME is model-agnostic, it can be applied to a wide range of machine learning and deep learning algorithms used in IoT security applications. In intrusion detection systems, LIME assists security analysts in understanding why a particular network flow has been classified as malicious by identifying the most influential features contributing to the decision. The technique provides intuitive visual explanations and supports forensic investigations by making attack detection outcomes easier to interpret. Although LIME is computationally efficient and easy to implement, its explanations are primarily local in nature and may not fully represent the overall behavior of the detection model.

V. RESEARCH GAP ANALYSIS

Despite significant advancements in machine learning-based intrusion detection systems for Internet of Things (IoT) environments, several research challenges remain unresolved. Existing studies primarily focus on improving detection accuracy while giving limited attention to model interpretability and adaptability. Most traditional Intrusion Detection Systems (IDS) rely on static learning mechanisms and predefined attack signatures, making them ineffective against evolving and previously unseen cyber threats. Recent machine learning and deep learning approaches have demonstrated superior performance in detecting complex attacks; however, many of these models function as black-box systems. As a result, security analysts often struggle to understand the rationale behind attack classification decisions, which reduces trust and limits deployment in critical environments such as healthcare, industrial automation, and smart city infrastructures.

Although Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME have been increasingly adopted in cybersecurity research, their integration into dynamic IoT intrusion detection systems remains limited. Most existing studies evaluate explainability and detection performance separately rather than combining them into

a unified adaptive security framework. Furthermore, only a few studies address real-time learning, continuous model updating, and detection of zero-day attacks while simultaneously providing transparent explanations. Another significant gap exists in the availability of explainable and adaptive frameworks suitable for large-scale IoT environments. Current solutions often face challenges related to computational complexity, scalability, and deployment on resource-constrained IoT devices. Therefore, there is a need for a comprehensive framework that integrates dynamic attack detection, adaptive learning mechanisms, and explainable decision-making capabilities to improve both security effectiveness and user trust.

5.1 Identified Research Gaps

1. Limited integration of XAI techniques within real-time IoT intrusion detection systems.
2. Lack of adaptive learning mechanisms for detecting evolving and zero-day attacks.
3. Insufficient focus on transparency and interpretability in deep learning-based IDS models.
4. High false-positive rates in anomaly-based intrusion detection systems.
5. Limited scalability of existing explainable security frameworks for large IoT networks.
6. Lack of unified frameworks combining detection accuracy, explainability, and continuous learning.

VI. PROPOSED ARCHITECTURE

The proposed architecture presents an integrated framework for detecting cyberattacks in IoT environments using Machine Learning (ML), Deep Learning (DL), and Explainable Artificial Intelligence (XAI) techniques. The framework begins with the Data Collection Layer, where network traffic is gathered from IoT devices, gateways, and benchmark datasets such as CICIoT2023 and CICIDS2017. The collected data is then processed through the Data Preprocessing Layer, which performs data cleaning, feature extraction, feature selection, normalization, and class balancing to improve data quality and model performance.

The preprocessed data is forwarded to the Detection Engine, where machine learning and deep learning algorithms, including Random Forest, XGBoost, CNN, and LSTM, are employed to classify network traffic as normal or malicious. To enhance transparency and trustworthiness, the Explainability Layer incorporates SHAP and LIME techniques, which provide interpretable insights by identifying the features that contribute most significantly to attack detection decisions. The generated explanations are presented to security analysts through dashboards, alerts, and reports, enabling better understanding and validation of model predictions. Furthermore, the framework includes an Adaptive Learning and Feedback Loop, which continuously updates and retrains the detection models using new traffic patterns and feedback. This capability enables the system to adapt to evolving cyber threats, improve detection accuracy, and reduce false-positive rates, making it suitable for real-time IoT security applications.

VII. FUTURE SCOPE

The integration of Explainable Artificial Intelligence with IoT intrusion detection systems offers several promising research opportunities. Future work can focus on developing lightweight XAI algorithms suitable for deployment on resource-constrained edge devices. The incorporation of federated learning and edge intelligence can further enhance privacy and reduce communication overhead in distributed IoT environments. Another important research direction involves the development of real-time adaptive intrusion detection systems capable of identifying sophisticated zero-day attacks without requiring extensive retraining. Hybrid explainability frameworks combining SHAP, LIME, and attention-based neural explanations may provide deeper insights into model behavior and improve decision transparency. Future studies can also investigate the integration of blockchain technology with explainable intrusion detection systems to enhance data integrity, trust management, and secure information sharing among IoT devices. Furthermore, large-scale validation using real-world industrial and healthcare IoT datasets will be essential for assessing the practical applicability of proposed frameworks. Overall, the convergence of Explainable Artificial Intelligence, adaptive machine learning, edge computing, and intelligent cybersecurity mechanisms is expected to play a crucial role in building trustworthy and resilient IoT ecosystems capable of addressing future cyber threats.

REFERENCES

[1] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[3] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A Survey of Machine Learning in Internet of Things Security," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1686–1721, 2020.

[4] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things," IEEE Access, vol. 5, pp. 18042–18050, 2019.

[5] Y. Zhang, Y. Weng, and J. Lund, "Explainable Artificial Intelligence for Cybersecurity Applications," IEEE Security & Privacy, vol. 20, no. 2, pp. 52–61, 2022.

[6] H. Hindy et al., "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems," IEEE Access, vol. 8, pp. 104650–104675, 2020.

[7] CICIoT2023 Dataset, "A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," IEEE Data Repository, 2023.

[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[9] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in Proc. Int. Conf. Learning Representations (ICLR), 2017.

[10] J. Kim, H. Kim, and K. Kim, "Explainable Deep Learning-Based Intrusion Detection for IoT Networks," IEEE Internet of Things Journal, vol. 10, no. 5, pp. 4120–4132, 2023.