Turkish Journal of Electrical Engineering and Computer Sciences

Volume 32 | Number 3 Article 2

5-20-2024

Unveiling anomalies: a survey on XAI-based anomaly detection for IoT

Esin EREN

Feyza YILDIRIM OKAY

Suat ÖZDEMİR

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2024) 32: 358 – 381
© TÜBİTAK
doi:10.55730/1300-0632.4075

Unveiling anomalies: a survey on XAI-based anomaly detection for IoT

Esin EREN[1*], **Feyza YILDIRIM OKAY**[2], **Suat ÖZDEMİR**[1]
[1]Department of Computer Engineering, Hacettepe University, Ankara, Turkiye
[2]Department of Computer Engineering, Gazi University, Ankara, Turkiye

Received: 14.11.2023 • **Accepted/Published Online:** 12.02.2024 • **Final Version:** 20.05.2024

Abstract: In recent years, the rapid growth of the Internet of Things (IoT) has raised concerns about the security and reliability of IoT systems. Anomaly detection is vital for recognizing potential risks and ensuring the optimal functionality of IoT networks. However, traditional anomaly detection methods often lack transparency and interpretability, hindering the understanding of their decisions. As a solution, Explainable Artificial Intelligence (XAI) techniques have emerged to provide human-understandable explanations for the decisions made by anomaly detection models. In this study, we present a comprehensive survey of XAI-based anomaly detection methods for IoT. We review and analyze various XAI techniques, including feature-based approaches, model-agnostic methods, and post-hoc explainability techniques, and discuss their applicability and limitations in the context of IoT. We also discuss the challenges and future research directions in XAI-based anomaly detection for IoT. This survey aims to provide researchers and practitioners in the field of IoT security with a better understanding of the current state of XAI techniques and their potential for enhancing anomaly detection in IoT systems.

Key words: Internet of Things (IoT), anomaly detection, explainable artificial intelligence (XAI), interpretable machine learning (IML)

1. Introduction

The rapid proliferation of IoT technologies has revolutionized numerous industries, enabling seamless connection and communication between physical devices and the digital world. However, this interconnectedness has also introduced new challenges, particularly in terms of ensuring the security and reliability of IoT systems. The identification of anomalous patterns or behaviors that differ from typical operations is a critical function of anomaly detection in the mitigation of these difficulties. With the increasing complexity and scale of IoT networks, traditional anomaly detection techniques often struggle to provide accurate and interpretable results [1]. Traditional anomaly detection techniques cannot provide adequate results in the face of the increasing complexity and scale of IoT networks. This highlights the need for more sophisticated and adaptive approaches in IoT systems.

XAI has emerged as a promising approach to enhance the transparency and interpretability of machine learning models. XAI techniques aim to provide human-understandable explanations for the decisions made by AI systems, thereby enabling users to trust and comprehend the underlying reasoning [2]. As technological landscapes advance, the need for clear and understandable explanations in AI systems becomes increasingly crucial. The integration of XAI into anomaly detection for IoT marks a significant paradigm shift, addressing

[*] Correspondence: erenogluesin@gmail.com

the inherent black-box character that often hides conventional methods. This revolutionary integration not only improves interpretability but is essential for accelerating high-stakes decision-making in the complex IoT environment. Furthermore, the interpretability offered by XAI extends its impact beyond elucidation. It serves as a proactive tool in identifying potential vulnerabilities, thereby contributing significantly to the development of robust security measures tailored for the complex deployments of IoT systems [3].

An in-depth examination of XAI-based anomaly detection techniques for IoT is essential due to the increasing reliance on IoT systems in critical applications such as smart cities, healthcare, transportation, and industrial automation. These systems generate vast amounts of data, making it challenging to detect anomalies accurately and promptly. Additionally, the consequences of undetected anomalies can be severe, ranging from financial losses to safety hazards. Therefore, the development of robust and explainable anomaly detection methods is important to ensure the integrity, reliability, and security of IoT networks [4]. These methods are necessary to support the widespread use of IoT technology and elevate industry standards. As a result, a detailed examination of developments in this field stands out as a step towards encouraging the widespread adoption of secure and effective IoT applications.

This paper contributes to the literature by presenting a systematic and comprehensive overview of the recent studies focusing on XAI-based anomaly detection mechanisms in IoT networks. The study explores the existing literature and research efforts in this domain, analyzing the strengths and limitations of different approaches. It also highlights the importance of interpretability in anomaly detection and discusses the potential benefits and challenges associated with integrating XAI techniques into IoT systems. This survey intends to serve as a useful resource for researchers, practitioners, and decision-makers involved in IoT security and anomaly detection by understanding the state-of-the-art methods and advancements in XAI-based anomaly detection. By providing a systematic overview of the current state-of-the-art methods and their potential applications, this survey aims to encourage further research and development in XAI techniques to enhance the security and reliability of IoT systems. Additionally, by highlighting the strengths and weaknesses of existing methods, the survey provides readers interested in understanding the current landscape of XAI-based anomaly detection in IoT systems with a detailed perspective. In this context, it thoroughly addresses the role of interpretability in the anomaly detection process, examining the potential benefits and challenges associated with the integration of XAI techniques into IoT systems. This comprehensive survey seeks to be a valuable resource for researchers, practitioners, and decision-makers involved in IoT security and anomaly detection by offering a deep insight into the current state of affairs. It aims to shape and encourage efforts to enhance the reliability of IoT systems both in academia and industrial applications.

The paper is organized as follows: Section 2 provides detailed information about IoT, attacks in IoT, and anomaly detection in IoT. Section 3 introduces XAI, its terminology, the taxonomy of XAI, and explainability methods. Section 4 presents a comprehensive review of existing XAI-based anomaly detection approaches, categorizing them based on their underlying methodologies. Section 5 discusses the challenges and potential future works. Finally, Section 6 concludes the survey and provides insights into the significance of XAI in addressing the limitations of conventional anomaly detection techniques for IoT applications.

1.1. Research methodology

Once the motivation for the study is identified, a comprehensive research methodology is developed. This precisely designed methodology not only explains the overall approach but also details the systematic procedures employed throughout the paper selection process. This strategic approach aims to uphold the integrity of the research process and contribute to the overall coherence and reliability of the study.

1. Literature search phase: The first step is selecting search phrases that are specific to the topic. These search phrases include "XAI", "Anomaly Detection", "Anomaly Detection in IoT", "Interpretable ML(IML)", "XAI in IoT", and "Intrusion Detection in IoT". Related studies have been researched from digital databases such as ScienceDirect [1], IEEE Xplore [2], Springer [3], ACM Digital Library [4]."
2. Paper selection criteria: We used the following criteria to determine the papers to be excluded from this study:
   - papers not directly related to IoT,
   - white papers,
   - papers published before 2018,
3. Paper classification: Due to page limitations, we selected 24 papers that meet our selection criteria. We classified these papers based on the explanation methods they used as follows: 16 papers employing feature-based explanations, 4 papers employing perturbation-based explanations, one paper employing rule-based explanations, one paper employing example-based explanations, and others are studies with more than one explanation type.

2. Anomaly detection in IoT

2.1. IoT

IoT is a technology in which more and more smart devices are located, and these devices communicate with each other. It has been included in many parts of our lives, from smart home systems we use in our daily lives to smart devices used in industrial areas. IoT supports diverse applications, including smart homes, smart cities, healthcare, agriculture, logistics, transportation, and energy. It basically creates a network where data is transmitted, collected, and processed through devices. There is a transition from certain layers in the transfer of data. These layers are called the perception layer, network layer, and application layer [5]. The perception layer consists of physical devices and sensors that collect data from the surroundings. Data transmission from the perception layer to the application layer is facilitated by the network layer. The application layer then processes this data to give end users useful information. IoT architecture is designed to be scalable, flexible, and interoperable, enabling effective communication between distinct devices and applications [6].

The increasing number of devices in IoT requires data transmission over long distances. It has been observed that wireless connection is suitable for long distances. A protocol is needed to ensure the connection between devices. Wi-Fi, Bluetooth, ZigBee, and LoRaWAN are some of the common protocols used in IoT [7]. There are sensors where data is generated or collected at one end of the established wireless connection. These sensors can be GPS units that receive location information or cameras that collect images. These received data can be transmitted to the cloud with the help of the specified protocol or stored locally. The data sent to the cloud is processed here, and the necessary action is taken.

2.2. IoT attacks

IoT technology has a profound impact on our daily lives since it allows devices to connect to the Internet and communicate with each other. This connectivity makes our world smarter and more interconnected, but it

[1] ScienceDirect [online], https://www.sciencedirect.com/, accessed [05/01/2024].
[2] IEEE Xplore [online], https://ieeexplore.ieee.org/Xplore/home.jsp, accessed [05/01/2024].
[3] Springer Link [online], https://link.springer.com/, accessed [05/01/2024].
[4] ACM Digital Library [online], https://dl.acm.org/, accessed [05/01/2024].

also introduces new opportunities and risks for cyber attackers. IoT devices often have security weaknesses, allowing hackers to expand their targets and launch attacks through IoT networks. These vulnerabilities in IoT devices allow attackers to conduct hacks, data breaches, delays in service, and other harmful actions. Issues such as poor password management, missing updates, or inherent vulnerabilities can grant attackers access to and control devices [8]. This may compromise both individual and organizational security. It is possible to classify attacks that compromise IoT security according to IoT architecture [9].

1. Perception layer attacks: This layer includes sensors, cameras, and other devices designed to detect and gather data from the physical environment within an IoT system. It collects, analyzes, and distributes contextual data for decision-making. However, this layer is susceptible to cyberattacks since attackers may alter the sensors and insert false information or malicious requests into the system. Hardware attacks are commonly found in this layer [7]. The probable attack types include hardware tampering, fake node injection, malicious code injection, and WSN Node Jamming [9].

2. Network layer attacks: The network layer facilitates communication among all devices in the IoT system. It encompasses network interfaces, communication channels, network management, information repair, and intelligent processing [10]. This layer serves as the central hub for collecting and integrating communications from various devices, enabling the routing of data to specific devices, typically through a gateway. Attacks in this layer target network traffic by taking advantage of weak points in the network. Some common attacks in this layer include RFID Spoofing, Sinkhole attacks, Man-in-the-middle attacks, Denial of Service, Routing Information attacks, and Sybil attacks [9].

3. Application layer attacks: The application layer serves as an interface for users. This layer facilitates actions such as accessing and managing IoT devices, processing data from sensors, and making decisions. Privacy and confidentiality issues are important at this layer, as they pertain to the user and their sensitive information. Consequently, this layer becomes a prime target for various types of cyberattacks. The common attacks in this layer are Phishing attacks, Viruses, Worms, Trojan Horses, Spyware, and Denial of Service [11].

2.3. Anomaly detection techniques in IoT

It has become important to process and transmit data to these devices accurately and securely with the expanding number of devices in the IoT. Detecting anomalies or attacks that may occur in the IoT network is critical to ensure IoT security. It is essential to identify the typical behavioral patterns of users within the system to determine abnormal situations. Anomaly is a term used to describe unusual or unexpected behavior in a system. The process of detecting such behavior is known as anomaly detection [12]. Anomalies can be categorized into three types: point anomalies, contextual anomalies, and collective anomalies [13]. Data points that deviate significantly from the rest of the dataset are known as point anomalies. Data points that are normal in one context but unusual in another are referred to as contextual anomalies. Collective anomalies are a group of data points that are unusual when seen collectively but appear normal when viewed individually [14].

Anomaly detection techniques are important in many industries, including cybersecurity, finance, network monitoring, and industrial processes, since they improve the discovery and mitigation of potential hazards or problems. There are several techniques commonly used for anomaly detection, categorized differently in the literature. One of them is statistical-based anomaly detection, which uses statistical methods to identify anomalies [15]. This technique is appropriate when the system's usual behavior can be represented using statistical methods. It does not require existing security knowledge and can detect new attacks, making it useful for long-term monitoring and recognizing denial-of-service attacks. Another technique is the data mining approach [16]. It is valuable for extracting patterns from big data stores in order to detect known and novel attacks more effectively. By providing information important for anomaly detection, this strategy decreases the storage of vast amounts of data. Various data mining technologies have been employed to detect known and unknown attacks. Another anomaly detection technique is knowledge-based detection [17]. It is based on predefined rules or knowledge to identify anomalies in a dataset. This strategy focuses on recognizing anomalous patterns or data points by utilizing an expert's expertise and experience. Data is evaluated in this method to find probable abnormalities based on predefined rules. Anomalies are identified as data points that violate rules or deviate from the expected norm. The last technique is machine learning-based detection [18], which uses machine learning algorithms to identify anomalies. It includes training the system on a dataset of normal behavior and then using that information to detect deviations from the normal behavior. The advantage of this method is that it can adapt to new types of attacks and detect previously unknown attacks. Supervised and unsupervised learning methods are the most commonly employed techniques in the literature for machine learning-based anomaly detection. Supervised methods tend to yield more accurate results as they involve data labeling during the classification process, thereby providing more reliable outcomes. In contrast, unsupervised methods do not require a labeling process, which enables the use of larger datasets and helps reduce time and resource costs associated with data labeling. Furthermore, unsupervised methods can detect previously undefined or unexpected anomalies. While both approaches have their respective advantages and disadvantages, the choice of method should be guided by specific application scenarios and the characteristics of the data [19].

Network-based and host-based network security approaches represent two fundamental strategies used to ensure information security [20]. Each provides different levels of protection and monitoring, often incorporating complementary elements. The network-based approach focuses on securing the overall network of an organization by monitoring, filtering, or controlling various security measures on network traffic. Additionally, it is used to strengthen the defense mechanisms of the organization against cyber threats. Network-based security enhances security at the network level with the ability to provide comprehensive protection. However, this approach may also have disadvantages, such as a lack of detailed content analysis or limitations in being effective only at the network level. Therefore, it is typically combined with host-based security to create a more comprehensive security strategy [21]. On the other hand, the host-based approach aims to protect each computer or host itself. This approach includes security measures at the level of the host's operating system and applications. Antivirus software, firewalls, and computer-specific security settings are among the elements of host-based security measures. In this way, an environment is created where each computer is individually secure [22]. Combining these two security approaches can create a comprehensive security strategy. The use of both approaches together enhances not only the overall security of the network but also strengthens the protection of individual computers.

2.4. XAI-based anomaly detection

Anomaly detection is critical in environments with continuous data flow, such as IoT. Fast and accurate detection of anomalies that may occur due to security breaches and system errors is a great necessity for security and efficiency. The reasons and consequences of anomaly detection must be explainable to the users. In this context, XAI techniques are widely used in anomaly detection in IoT systems.

The first step in the anomaly detection mechanism involves collecting data from IoT devices, which is gathered into a data pool via devices such as sensors and cameras. These data are generally large in volume and can be of different types such as textual, image, or time-series. The gathered data is preprocessed to become proper for evaluation. During this stage, missing data is imputed, noise is reduced, and data transformations are applied. Effective data cleaning and organization are the basis of a successful anomaly detection mechanism [23]. Following the data preprocessing phase, machine learning algorithms are employed to analyze the dataset. During this stage, learning algorithms begin to model the data for anomaly detection. Traditionally, the models used at this stage are often considered ”Black-box” models, implying that their internal workings or decision-making processes are not fully transparent or explainable. It can be challenging to provide a clear understanding of why a model made a particular prediction or decision. To enhance reliability and interpretability in anomaly detection, XAI techniques are integrated into the mechanism. XAI is employed to elucidate and interpret the decisions made by machine learning models. One of the XAI methods evaluates each feature's impact on the result at the feature extraction stage. Various explanation methods, such as LIME (Local Interpretable Model-Agnostic Explanation), SHAP (Shapley Additive exPlanations), and Decision Trees (DT), are utilized [24].

These processes collectively contribute to a more comprehensive understanding of the origins and implications of abnormal situations. The decisions provided by the anomaly detection model are clarified using explanatory models built with XAI techniques. They offer both operational teams and their end users the opportunity to understand the causes of such abnormal situations better. The use of XAI in IoT anomaly detection is shown in Figure 1.

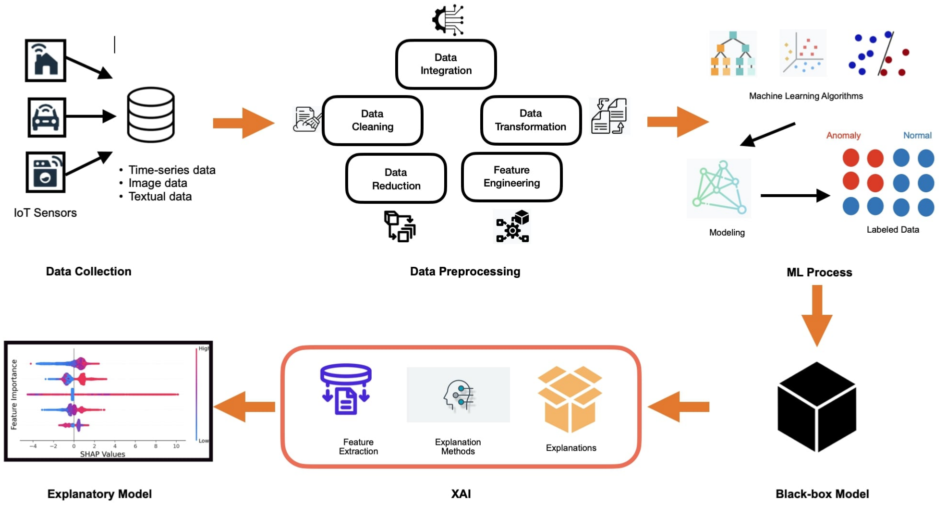

Figure 1. The structure of XAI-based anomaly detection mechanism.

3. Explainable artificial intelligence (XAI)

XAI is the process of artificial intelligence systems explaining their decisions and results in a manner that is understandable to humans. XAI aims to improve transparency, comprehensibility, and reliability of seemingly complex AI models and algorithms [25]. It helps people understand and trust AI technologies, as explainability is critical for understanding, validating, and questioning AI system decisions. Users often desire insights into how AI systems work and what factors influence their outcomes, particularly in areas where significant decisions are involved. XAI simplifies the tasks of tracking and analyzing the system's decisions by providing essential information.

XAI focuses on the transparent understanding of model decisions, a field that has emerged with increasing demand for the complexity of powerful learning models. In this context, human-based evaluations of XAI include important factors such as usability and user experience (UX). Human-centered evaluations focus on evaluating the reliability and acceptability of XAI systems, as well as ensuring that end users effectively understand these systems. Many different methodologies are used for human-based evaluations of XAI systems [26]. Various methods such as usability testing in laboratory environments, participant observation, surveys, and field studies are widely used to evaluate how XAI performs under real-world usage conditions. These methodologies provide a significant range of tools for understanding and improving users' interaction with XAI systems. The impact of XAI on UX may vary depending on application domains and context. For example, the user interactions of an XAI system in the financial sector and a system in the healthcare sector may differ. Therefore, XAI developers must design and optimize their systems taking into account the application context.

Furthermore, addressing UX and understandability is crucial in the context of XAI. UX plays a vital role. Researches [27, 28] show that making XAI more descriptive positively affects the user's interaction with the system. The connection between understandability levels and UX guides developers when designing and optimizing XAI systems [27]. The human factor involved also brings with it ethical and privacy issues. Ethical issues and privacy concerns that may arise during the use of XAI systems may affect the sustainable adoption of this technology. Protection of users' personal information and fair and transparent use of XAI, compliance with ethical standards are important issues that should be emphasized. The study [29] highlights the importance of using XAI to ensure trust and include ethical elements in IoT systems. It identifies the weaknesses of IoT systems and emphasizes the need for tamper detection systems. It also provides an overview of XAI methods for IoT systems, guiding their advantages, disadvantages, and ethical considerations.

Recent studies employing XAI in the realm of IoT anomaly detection have demonstrated significant achievements. These endeavors often focus not only on accuracy rates but also on performance metrics such as precision, recall, F1-score, and AUC-ROC. As a case in point, Alani et al. [30] showcased an XAI-based anomaly detection model surpassing a 99.94% accuracy rate. Additionally, a separate investigation led by Abououf et al. [31] yielded impressive results, particularly in terms of F1 score, indicating a substantial performance improvement compared to traditional methods. These findings underscore the effectiveness of XAI in IoT anomaly detection, providing a foundation for further exploration in this field.

3.1. XAI taxonomy

In the literature, XAI taxonomy has been approached differently in various studies [32, 33]. Taxonomy serves to systematize the process of understanding and explaining AI systems by categorizing and arranging XAI approaches using diverse dimensions. As shown in Figure 2, the taxonomy of XAI is given according to the XAI implementation, scope, and applicability.

1. Ante-hoc & post-hoc: Two approaches are used in the application of explainability to the models: ante-hoc and post-hoc. Ante-hoc approaches incorporate transparency principles and techniques directly into the model creation process. Therefore, ante-hoc models do not require an additional XAI method for producing explainability. Post-hoc approaches are used after the model has been trained to explain the

model's predictions [32]. It analyzes the decision processes of complex or black-box models and applies explainability techniques afterward. These techniques offer the advantage of allowing transparency and explainability aspects into existing models later. In this way, more sophisticated or higher-performance models can be used, and the decision processes of these models can be made understandable.

2. Local & global: Local explanations deal with explaining the prediction of a given instance or input, whereas global explanations are concerned with providing an understanding of the model's overall behavior [34]. Local explanations are useful for understanding why a particular prediction was made, while global explanations help understand how the model works overall. Both approaches are key components of XAI and increase the transparency of AI systems. They differ from each other depending on the scope and target of the explanations. Local explanations are essential in understanding the inner workings of a particular data point and can be used to provide the user with information on a specific instance. Additionally, global explanations explain how the model works and general trends [24].

3. Model-specific & model-agnostic: These approaches can be distinguished based on the specific model they are intended for and their general applicability. Model-specific methods are designed to explain the decisions of a specific model, while model-agnostic methods are designed to explain the decisions of any model. Model-specific methods are tailored to a particular model's design and attributes; however, model-agnostic methods are more general and can be applied to any model. Model-specific approaches are frequently more accurate and efficient, but they take more time to create and may not be transferable to other models [33, 35].

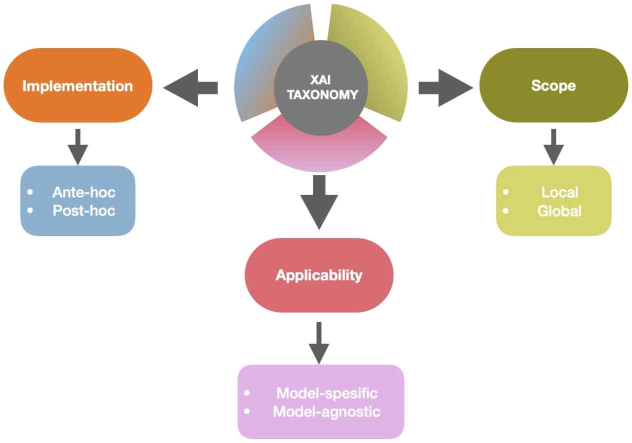

Figure 2. XAI Taxonomy

3.2. Explainability methods

Many explainability methods have emerged with the prominence of the concept of XAI. In this study, feature-based, rule-based, perturbation-based, and example-based methods are discussed.

1. Feature-based explanations: These methods explain what the inputs used for the model contribute to the model output. Many methods have been used to reveal the factors affecting the decision of the model.

These are feature importance, Partial Dependence Plots (PDP), LIME, SHAP, Individual Conditional Expectation (ICE), etc [35]. LIME and SHAP are frequently encountered methods in the literature. LIME modifies the input of data samples and assesses how predictions change, aiming to comprehend the model. This approach is valuable for understanding what human interests are when the output of a model is observed. SHAP relies on game theory to explain the output of a model. It considers all feature combinations to determine their contribution. It can be applied to various types of models.

2. Rule-based explanations: These explanations use logical rules and circumstances to explain the decision-making process of models. These rules can be either manually created or automatically generated [36]. It requires providing accessible and interpretable insights into how the model makes predictions or categorizes data. Rule-based explanations are frequently expressed as if-then statements in which specific circumstances or combinations of features are described, and associated decisions or effects are decided. XAI supports trust, openness, and interpretability by allowing people to understand and reason about the underlying logic and components that influence the model's decisions.

3. Perturbation-based explanations: These explanations are post-hoc methods for explaining model decisions. They involve degrading or altering the input properties of a model to observe their impact on results. By analyzing the changes in the output of the model due to perturbations, it becomes possible to determine which input features are most important to the model's decision. This technique is model-agnostic and applicable to any model, but it can be computationally expensive and less suitable for large datasets. There are different perturbation-based methods such as LIME, RISE (Random Input Sampling for Explanations), and Occlusion [37]. The LIME method is an example of a perturbation-based XAI technique that operates on information or feature superpixels. Visual descriptions of individual superpixels are developed by progressively supplying input patches. RISE is one of the methods for improving the predictability and functionality of AI models. In the Occlusion method, first, a data sample is selected, and a specific region or feature is selected from that sample. The selected area will then be temporarily closed or modified. The model makes predictions again on this changing data. In the last step, the estimated difference between the original and modified data is checked.

4. Example-based explanations: These explanations are used to explain the behavior of models by using data examples. They select instances of certain input-output pairs to explain the model's output and allow users to comprehend the model's decision-making process. It is commonly used in complex tasks like image recognition and natural language processing. It illustrates the model's features and the factors that influence its decisions by showing how the model reaches a particular outcome. It is model-independent and applicable to a wide range of models. Instead of interfering with or altering the model, example-based explanation approaches analyze by choosing cases from the dataset, which makes the method highly generalized and applicable [33].

4. XAI-based anomaly detection in IoT networks

This section provides a thorough overview of XAI-based anomaly detection studies in IoT application domains. There are several survey studies in the literature addressing IoT, XAI, and anomaly detection. Especially in recent years, many researchers have carefully investigated XAI techniques. For instance, some papers [38–41] only discuss the XAI system's functionality and overview and do not include the concepts of IoT and anomaly detection. Tjoa et al. [41] discuss the need for explanations in machine decisions, the challenges posed by the black-box nature of deep learning, and provide a categorization of interpretability approaches in the healthcare domain. On the other hand, some papers [42, 43] have achieved results by not only focusing on anomaly detection but also analyzing studies across all IoT subdomains. Moustafa et al. [44] have focused only on post-hoc and model-agnostic methods in their study although addressing the concepts of IoT, XAI and anomaly detection. Javed et al. [45] evaluated the examined studies in terms of domain and technical aspects. The performance metrics obtained in the studies were not discussed, and there were no examinations regarding XAI taxonomy. In our study, the concepts of IoT, XAI, and anomaly detection have been comprehensively addressed, establishing connections among them. The selected works have not only been evaluated in terms of XAI taxonomy but have also been examined through a general security approach. A comparison of our study with other existing survey works is presented in Table 1.

Different from other survey studies, we conducted a literature review covering XAI-based anomaly detection mechanisms in IoT networks and summarized them according to their taxonomies and methodologies, as presented in Table 2. Additionally, in Table 3, we provided a summary of studies based on their security approach and anomaly type. These help researchers and practitioners better adapt to their needs and application scenarios when choosing a particular explanation method. The most preferred methods in the field of anomaly detection in IoT may vary depending on the complexity of the data and the problem context. However, based on the studies in the literature, it has been observed that the most favored explanation method is the feature-based explanation method. When IoT data emphasizes the importance of specific features like temperature, humidity, and pressure, feature-based methods are often preferred. Thus, the complexity of IoT data is reduced to a simpler and more understandable level, typically enhancing the explainability of the system. Zolanvari et al. [46] discussed various examples of XAI applications in different fields such as industrial environments, smart grids, 5G telecommunications, smart homes, and healthcare. The authors argued that XAI is essential to ensure transparency, accountability, and reliability. The paper proposed a feature-importance-based approach, using a separate model development and statistical techniques to explain the behavior of the AI model. The proposed approach outperformed other methods in terms of accuracy and interpretability. This work can be considered an important step toward increasing the transparency and explainability of AI systems. Khan et al. [47] proposed an Autoencoder-based detection framework using convolutional and recurrent networks for discovering cyber threats in Industrial IoT (IIoT) networks and explaining the model. The framework incorporated an XAI method to provide explanatory reasoning for prediction decisions and underlying data evidence. By employing a two-step Sliding Window (SW) approach, the framework effectively extracted features that capture the contexts of malicious patterns. The empirical results confirmed the framework's robustness in detecting malicious events, surpassing contemporary modern methods and exhibiting its potential for practical application in real-world IIoT-based networks. Dong et al. [48] presented FEDFOREST, a learning-based Network Intrusion Detection System (NIDS) that combined the Gradient Boosting Decision Tree (GBDT) and Federated Learning (FL) framework. FEDFOREST provided high accuracy while maintaining data locality and privacy, as well as providing interpretability. Local attack features were extracted by several clients for server training and detection. FL privacy was further protected by privacy-enhancing technology. FEDFOREST has been tested on a variety of datasets and has proven to be effective, efficient, interpretable, and extendable. The practical implications included real-world NIDS deployment, effective detection for different tasks, enhanced privacy in FL, interpretability for experts, and potential extension to other domains requiring privacy-preserving ML. Alani et al. [49] introduced DeepIIoT, an explainable deep learning-based Intrusion Detection System (IDS) developed for industrial IoT. A Multi-Layer Perceptron (MLP) classifier was used in the proposed system to

Table 1. Comparison between our study and existing studies

Reference	Year	Time	IoT	XAI	AD	Key Findings	Limitations
Capuano et al. [38]	2022	2018 - 2022	X	✓	X	Highlights the extensive analysis of XAI applications in CyberSecurity, and emphasizing the importance of transparency and explainability in enhancing CyberSecurity practices.	It focuses on general security applications and is not specific to IoT.
Li et al. [39]	2023	1998 - 2022	X	✓	✓	Underscores the insufficient attention to explainability in anomaly detection, highlight the deficiencies in existing surveys, and emphasize the study's goal of providing a comprehensive overview of state-of-the-art explainable anomaly detection techniques	Practical implementation and evaluation of explainable anomaly detection techniques are not covered in the study.
Neupane et al. [40]	2022	2018 - 2022	X	✓	✓	Proposes the concept of Explainable Intrusion Detection Systems (X-IDS) for cybersecurity, review XAI methods, present a taxonomy for explainability, and providing design guidelines for X-IDS.	It focuses only on IDS and not IoT.
Tjoa et al. [41]	2020	2000 - 2019	X	✓	X	Medical practitioners are provided with insights through a survey on applications of XAI in the medical field.	The paper only concentrates on exploring the functions of XAI within the healthcare domain.
Kök et al. [42]	2023	2008 - 2023	✓	✓	X	Provides clear explanations of XAI terminology and techniques, present a thorough review of current studies on XAI in the IoT domain, and outlining emerging challenges, open issues, and future research directions in XAI from an IoT perspective.	A specific IoT domain was not addressed and studies related to anomaly detection were not examined.
Jagatheesaperumal et al. [43]	2022	2018 - 2022	✓	✓	X	Underscores the importance of XAI in the IoT domain, offering insights into XAI frameworks, their characteristics, and support for various IoT applications.	The studies are not evaluated in terms of XAI taxonomy and do not focus on anomaly detection.
Moustafa et al. [44]	2023	2017 - 2022	✓	✓	✓	Emphasizes the focus on XAI techniques for anomaly-based intrusion detection in IoT networks, highlighting their effectiveness in ensuring reliability, interpreting security events, and integrating into cyber defense systems.	It focuses only on post-hoc and model-agnostic methods.
Javed et al. [45]	2023	2017 - 2022	✓	✓	X	Provides significant findings by comprehensively surveying the current state and future developments of XAI technologies for smart cities.	Studies were examined only in terms of application and technical aspects.
Our study	2024	2018 - 2024	✓	✓	✓	This study provides a summary of the existing literature by examining XAI-based anomaly detection mechanisms in IoT networks and emphasizes the significance of XAI in this context. While analyzing the strengths and weaknesses of various approaches, the review discusses the potential benefits and challenges associated with integrating XAI techniques into IoT systems.	-

recognize various types of attacks such as backdoors, Denial of Service (DoS), and command injection attacks. This system was trained and tested using WUSTL-IIoT-2021 dataset, achieving an accuracy of more than 99% while preserving low false-positive and false-negative rates. The implementation of SHAP values improved the explainability of the system. DeepIIoT outperformed other IDSs in comparison tests. Improving IoT security, understanding model decisions through SHAP values, and reaching potential improvements for larger implementation were among the practical implications. Overall, the study contributed a high-performance and explainable intrusion detection method for industrial IoT domains. Patil et al. [50] offered a new IDS based on machine learning ensemble methods like DTs, Random Forests (RFs), and Support Vector Machines (SVMs). The proposed model was trained and evaluated on the CICIDS-2017 dataset with the goal of improving classification accuracy and decreasing false positives. The XAI algorithm LIME was used in the study to improve the explainability and comprehensibility of the black-box technique for trustworthy intrusion detection. It also gave a thorough examination and exploration of SVM-based intrusion detection and feature selection methods, discussed different sources of data, and established an IDS taxonomy for different machine learning techniques in this domain. The study highlighted the importance of using machine learning for IDSs and examined available NIDS implementation tools and datasets. Ultimately, it provided useful insights and contributions to the design and implementation of successful IDSs based on machine learning techniques. Huong et al. [51] presented FedeX, a revolutionary Federated Learning-based Explainable Anomaly Detection architecture built for Industrial Control Systems (ICSs) in smart factories. FedeX employed advanced approaches such as Variational Autoencoder (VAE), Federated Learning, Support Vector Data Description (SVDD), and SHAP to provide reliable and interpretable anomaly detection. FedeX ensured interactive training while protecting data privacy and security in a spread setting by employing Federated Learning. VAE was an effective detection model that captured regular behavior patterns in ICSs data. SVDD was used to calculate anomaly detection thresholds automatically. The use of SHAP allows for the interpretation of the black-box learning model, which provides insights into anomaly predictions. Experimental results demonstrated the superior performance of FedeX. It outperformed 14 existing anomaly detection methods across multiple parameters, demonstrating its ability to detect anomalies in ICSs. Particularly, FedeX performed well on the liquid storage and Secure Water Treatment (SWaT) datasets, with a recall of 1 and an F1-score of 0.9857. FedeX was also extremely fast, with a training period of only 7.5 min, and it was low in terms of hardware needs, utilizing only 14% of RAM. FedeX was well-suited for real-time deployment and edge computing architecture due to these properties. Hussain et al. [52] described a technique for detecting explainable anomalies in IoT-based industrial processes, meeting the demand for transparency and interpretability in complex systems. The suggested technique employed dual substitute models to provide explanations for black-box model outputs. The authors used treeSHAP to compute feature importance values, which aided in comprehending the significance of distinct characteristics in discovered anomalies. SHAP force plots and SHAP dependency plots provided insights into the components contributing to the anomalies, providing a local explanation of the black-box model output. The study also included an interactive dashboard that combined the deep learning explanation technique with previous records to provide a thorough perspective of the observed abnormalities. This dashboard was intended to cater to several personas with diverse levels of technical skill, ensuring good communication and comprehension of the anomalies. Notably, the study underlined the need to take into account the social context when providing explanations, emphasizing the importance of establishing trust and confidence in the system. Overall, the suggested technique advanced the field of IoT-based industrial processes by providing a viable solution for real-time anomaly detection and thorough explanations, resulting in increased efficiency, decreased downtime, and informed decision-making. Oseni et al. [53] presented a deep learning-based intrusion detection methodology for IoT-enabled transportation networks. For cybersecurity experts, the framework utilized the SHAP mechanism to analyze judgments produced by deep learning-based IDS. The architecture aimed to improve the transparency and resilience of IDS in IoT networks by providing explainability. The proposed framework was validated using the ToN_IoT dataset, and it achieved high performance with 99.15% accuracy and 98.83% F1 score, demonstrating its effectiveness in securing IoT networks, particularly those related to the Internet of Vehicles (IoV). Improving the security and design of IoT networks, assisting in root cause investigation, and enabling the development of more resilient IDSs were some of the practical implications. The proposed explainable framework, its validation and comparison with other methodologies, and its possible implementation in real-world scenarios such as intrusion detection and threat intelligence in IoT and Industrial IoT networks were the paper's contributions. Overall, the findings highlighted the proposed framework's capacity to protect IoV networks from sophisticated cyberattacks. Djenouri et al. [54] presented a new framework for intrusion detection in the next-generation IoT. Several methods were used in the framework, including MinMax normalization for data collection and preprocessing, the Marine Predator algorithm for feature selection, a sophisticated recurrent neural network for training the selected features, and the Shapley score as an explainability method. The MinMax normalization method was utilized to preprocess the data. This method reduced the data to a defined range of 0 to 1. The Marine Predator algorithm was used to identify significant features. This algorithm selected features depending on their importance to the learning process. The selected features were subsequently used during the training process. The experimental results showed that the suggested system obtained a high detection rate of over 94% for both true negative and true positive detection, outperforming existing methods on the difficult NSL-KDD datasets, where their rates were less than 90%. Alani et al. [30] described a simple and effective method for selecting universal features from IoT intrusion detection datasets. The method aimed to create machine learning-based IDSs for IoT devices that were highly accurate and efficient. The suggested approach was applied to three datasets, giving six generic network-flow features. The approach was tested successfully with a high accuracy of 99.62% and a significant reduction in prediction time. The implementation of SHAP provided insight into the selected features and their compatibility with current attack strategies. There were also implications of the study such as improving IoT device and network security, identifying various forms of attacks, and reducing prediction time. Abououf et al. [31] presented a novel online Event and Anomaly Detection (EAD) method designed for healthcare monitoring systems in the context of the Medical Internet of Things (MIoT). The method combines the XAI approach KernelSHAP using a lightweight AutoEncoder to give easy-to-understand reasons for anomalies detected. The suggested method shows robustness in identifying and categorizing events through extensive simulations using the Medical Information Mart for Intensive Care (MIMIC) dataset, demonstrating consistency across a range of anomaly percentages. The dataset was divided into 70% for training and 30% for testing. The EAD model was trained using the alerts generated during the EAD step. Utilizing waterfall plots, the study improves the interpretability of the model's decision-making process by visually representing each feature's impact to deviating the projected value from the actual value. The proposed method effectively enhances anomaly detection in healthcare monitoring systems dedicated to the MIoT. The combination of these techniques proves successful in providing transparent and reliable insights into the presence of anomalies. Djenouri et al. [55] proposed a framework that combines deep learning, evolutionary computation and XAI to address problems in IoT. They used the Gamian angular field to convert data gathered from various sensors in the IoT ecosystem into an image database, and they used the VGG16 architecture for image training. The integration of XAI technology and hyper-parameter optimization facilitates a thorough examination of the effect of input values and an increased understanding of the weights within the deep learning model. After completing comprehensive testing on two different IoT datasets, IPFlow and N-BaIoT, the framework outperforms baseline methods in terms of accuracy and runtime. The new framework focuses on addressing IoT issues by examining connections between sensor data components and utilizing deep learning for prediction and intrusion detection tasks. Furthermore, XAI was used to improve the comprehension of each feature's impact on the model output, providing insightful information on the significance of each feature. Experiments on multivariate time series datasets with various properties are carried out on IPFlow, which contains flow data from IoT devices, and N-BaIoT, which focuses on IoT botnet anomalies. Houda et al. [56] proposed a framework called FedIoT that combines XAI techniques and Blockchain to secure FL-based IDS in IoT networks. FedIoT detects local model changes and mitigates FL-based attacks by leveraging innovative XAI algorithms. Additionally, it integrates a reputation system based on blockchain to guarantee the dependability and trustworthiness of the FL training method. The efficacy of the framework in identifying threats is evaluated by using the UNSW-NB15 dataset. It is demonstrated that FedIoT can efficiently enable federated learning among different users and detect malicious actions. Rathod et al. [57] proposed a secure data dissemination architecture for IoT-enabled critical infrastructure, combining AI and blockchain technologies to address security and privacy challenges. The architecture included dimensionality reduction using PCA (Principal Component Analysis) and XAI and utilized AI classifiers such as RF, DT, SVM, perceptron, and GaussianNB for data classification. Additionally, an IPFS-driven blockchain network was implemented to ensure the security of nonmalicious data, with an anomaly detection approach to identify and eliminate poisoned data. The performance of the proposed architecture was evaluated using various metrics, with the RF classifier achieving the highest accuracy at 98.46%. The article also discussed the experimental setup, including the use of Google Colab and Remix IDE, highlighting the importance of integrating AI and blockchain technologies for enhancing security in IoT-based critical infrastructure. Hasan et al. [58] proposed an explainable ensemble deep learning approach for intrusion detection in IIoT systems. The method utilized LIME and SHAP techniques to offer insights into the choices made by deep learning-based IDSs. The suggested framework aims to improve the transparency and robustness of IDSs in IIoT networks. The ToN_IoT dataset was used to evaluate the efficacy of the framework, and experiments showed the effectiveness of ensemble learning in improving the results. The paper also implemented the extreme learning machines (ELM) model as a baseline IDS and compared it with other models. The results highlight the importance of explainability in IDSs and how it can aid cybersecurity professionals in assessing system effectiveness and developing more cyber-resilient solutions. Sharma et al. [59] focused on intrusion detection in IoT networks using DL models. It introduces a DL model designed to classify various attacks within the dataset, employing a filter-based approach to emphasize essential aspects and constrain the number of features. Two DL models, Deep neural network (DNN) and Convolution Neural Network (CNN), are built and tested on publicly accessible datasets, NSL-KDD and UNSW-NB 15. The DL model shows better accuracy rates for both datasets. To address the challenge of understanding DL models, the study applies the concept of XAI using LIME and SHAP methods. The study also discusses data preprocessing, feature selection, and feature preprocessing techniques used in the study.

Perturbation-based methods are also preferred, as they can be employed in any AI model, regardless of the specific model type. This method is valuable for assessing how adding noise or variations in the data affects predictions. Huang et al. [60] offered an Energy-efficient and Trustworthy Unsupervised Anomaly Detection Framework (EATU) for the Industrial IoT. The framework employed a two-stage approach for feature extraction: Autoencoder-based feature extraction at the first level and Efficient DeepExplainer-based feature selection at the second level. The Efficient DeepExplainer method is a post-hoc explainable feature selection method that selects the most significant features for anomaly detection and gives local explanations for the binary classifier's decisions. The proposed framework was validated on three real-world IIoT datasets with high-dimensional characteristics, and the experimental findings showed that it outperforms state-of-the-art approaches in terms of accuracy, trustworthiness, and energy efficiency. Almuqren et al. [61] introduced XAIID-SCPS, an XAI

Table 2. Summary of XAI-based anomaly detection in IoT according to their taxonomies.

Explanation Method	Reference	IoT Domain	ML/DL Model	XAI Model	Ante Hoc / Post Hoc	Model Specific / Model Agnostic	Local / Global
Feature-based	Zolanvari et al. [46]	IIoT	-	TRUST	-	Model agnostic	-
	Khan et al. [47]	IIoT	Combination of CNNs and LSTM	LIME	Post-hoc	Model-agnostic	Local
	Dong et al. [48]	IoT Security	GBDT	GBDT	Post-hoc	Model-specific	Local & Global
	Alani et al. [49]	IIoT	CNN	SHAP	Post-hoc	Model-agnostic	Local & Global
	Patil et al. [50]	IoT Security	DT, RF, SVM	LIME	Post-hoc	Model-agnostic	Local
	Huong et al. [51]	Smart Manufacturing, IIoT	SVDD, CNN, and LSTM	SHAP	Post-hoc	Model-agnostic	Local & Global
	Hussain et al. [52]	IIoT	AE consisting of LSTM layers	SHAP	Post-hoc	Model-agnostic	Local
	Oseni et al. [53]	Transportation	CNN	SHAP	Post-hoc	Model-agnostic	Local
	Djenouri et al. [54]	IoT Security	RNN	Shapley values	Post-hoc	Model-agnostic	Global
	Alani et al. [30]	IoT Security	Fuzzy rule-based classifiers (FRBCs)	SHAP	Post-hoc	Model agnostic	-
	Abououf et al. [31]	Healthcare	AutoEncoder	KernelSHAP	Post-hoc	Model-agnostic	Local & Global
	Djenouri et al. [55]	IoT Security	RNN-LF, kNN-TF, LOF-TF	RuleFit, SHAP	Post-hoc	Model-specific	Local
	Houda et al. [56]	IoT Security	VGG16	Shapley value	Post-hoc	Model-agnostic	Local & Global
	Rathod et al. [57]	Critical Infrastructure	RF, DT, SVM, Perceptron, Gaussian Naive Bayes	PCA	Post-hoc	Model-agnostic	Global
	Hasan et al. [58]	IIoT	CNN	LIME, SHAP	Post-hoc	Model-agnostic	Local & Global
	Sharma et al. [59]	Critical Infrastructure	CNN, DNN	LIME, SHAP	Post-hoc	Model-agnostic	Local & Global
Perturbation-based	Huang et al. [60]	IIoT	AE	Efficient Deep-Explainer	Post-hoc	Model-specific	Local
	Almuqren et al. [61]	IoT Security	IENN	LIME	Post-hoc	Model-agnostic	Local
	Sharma et al. [62]	IoT Security	KNN/SVM, DNN/CNN	LIME	Post-hoc	Model-agnostic	Local
	Anello et al. [63]	IIoT	Isolation Forest	AcME	Post-hoc	Model-agnostic	Global
Rule-based	Sivapalan et al. [64]	Healthcare	ANN	Rule mining	Post-hoc	Model-specific	-
Example-based	Guerra-Manzanares et al. [65]	IoT Security	DT, kNN, RF	LIME	Post-hoc	Model-agnostic	Local
Feature-based, Example-based, Perturbation-based	Abou El Houda et al. [66]	IoT Security	DNNs	RuleFit, LIME, SHAP	Post-hoc	Model-agnostic	Local & Global
Feature-based, Perturbation-based	Keshk et al. [67]	IoT Security	LSTM	SPIP	Post-hoc	Model-agnostic	Local & Global

Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems. This technique involved various subprocesses such as data preprocessing, Improved Elman Neural Network (IENN)-based classification, and HESGO-based feature selection. XAIID-SCPS utilized the XAI methodology LIME for increased understanding and explainability of the black-box method employed for accurate intrusion classification. The results showed that XAIID-SCPS performs well when compared to other current techniques, with a high accuracy of 98.87%. Future work on improving detection performance through data clustering, outlier removal techniques, and extending the model with ensemble voting classifiers was suggested in the study. As a whole, XAIID-SCPS provided an effective and explainable solution for intrusion detection in Cyber-Physical Systems. Sharma et al. [62] proposed a deep learning-based model for intrusion detection in IoT networks. The model employed a DNN architecture and a filter-based technique for feature reduction. The model was trained and tested on the NSL_KDD dataset, and its accuracy was compared to that of other machine-learning approaches. The chosen DNN model obtained the best accuracy of 0.993 with certain hyperparameters. The LIME explainability method was used in the study to provide insights into the model's predictions. The study had practical consequences such as increased IoT system security, reduced computational complexity, and improved real-time intrusion detection in IoT systems. The contributions of the study involved the proposed DNN model, a filter-based feature reduction method. The results showed that DNN model outperforms other strategies in terms of accuracy, showing its potential for IoT intrusion detection. Anello et al. [63] presented a thorough method for detecting explainable anomalies in IIoT systems. The suggested method combined Isolation Forest, a machine learning methodology for anomaly identification, with AcME, a rapid and model-agnostic interpretability tool. The system was designed to detect anomalies in real-time and provided interpretable reasons for those anomalies, allowing for root-cause analysis and better decision-making. The AcME technique allowed for local explanations, allowing maintenance personnel to understand the root causes of anomalies and conduct appropriate repair actions. The proposed approach's usefulness was demonstrated through trials performed in real-world industrial instances, demonstrating its effectiveness in detecting abnormalities and presenting interpretable answers. When compared to the modern SHAP approach, AcME provided comparable or greater interpretability while being computationally more efficient. This discovery had major practical consequences since it enables preventative maintenance, lowers downtime, improves operational efficiency, and increases the reliability of IIoT systems. Overall, the suggested method provided an effective and interpretable solution for anomaly detection in IIoT, allowing for better root cause analysis and corrective actions.

Rule-based methods are preferred in cases where IoT data can be explained with certain rules and threshold values. Sivapalan et al. [64] introduced an explainable rule-mining technique for giving importance to anomalous class detection in ECG data. The proposed method implemented a biased-trained Artificial Neural Network (ANN) with input features taken from ECG beat sequences. It generated a set of criteria at all nodes of a tree-like search method, with the rule base built from important features detected in the ANN via gradient analysis. The resulting model was a rule-based system that determines unusual heartbeats using quantitative and morphological ECG data. The system provided great accuracy and sensitivity while being simple to implement, making it ideal for healthcare applications, particularly in IoT-enabled wearable edge sensors. The model achieved an accuracy of 93% with only nine nodes and an evaluation accuracy of 90% and 80%, respectively for VEB and SVEB beat types when tested on previously unknown ECG data from the INCART database.

Example-based methods can be more comprehensible to end users because they explain model predictions using specific examples. Furthermore, if IoT data exhibits differences between certain events or situations, example-based methods can be preferred. Guerra-Manzanares et al. [65] investigated the relation between feature selection and post-hoc interpretation methods in an IoT botnet machine learning workflow. The authors proposed employing Fisher's Score for feature selection and LIME for post-hoc interpretation. The study illustrated that highly accurate and interpretable learning models may be generated with fewer features using both processes. The study also addressed the interpretability gap in machine learning-based IDSs and offered a metric for measuring detection accuracy and interpretability together. The findings showed that Fisher's Score and LIME were effective at selecting features and generating explanations, providing accurate and interpretable models for IoT botnet detection.

Each explanation method provides a unique perspective on the model. For example, a feature-based method can reveal the model's focus on specific features, while a perturbation-based method can indicate its sensitivity to the data. Thus, studies often employ multiple methods. Abou El Houda et al. [66] proposed a two-stage XAI architecture for IDSs in IoT networks. To identify IoT-based threats, the system utilized a DNN architecture and combined XAI techniques such as RuleFit, LIME, and SHAP to provide local and global justifications for the IDS's actions. The goal was to improve communication and trust between the deep IDS system and cybersecurity specialists. The proposed framework was validated using the NSL-KDD and UNSW-NB15 datasets, proving its value for enhancing IDS interpretability against IoT attacks and supporting cybersecurity specialists in understanding decision-making processes. Improving IDS interpretability, promoting better decision comprehension, and developing more efficient IDSs for IoT networks were some of the practical

Table 3. Summary of the studies according to their security approach and anomaly type.

Reference	Dataset	ML Approach	Security Approach	Anomaly Type	Evaluation Metrics	IoT Attack Type
Zolanvari et al. [46]	WUSTL-IIoT, NSL-KDD, UNSW	-	-	Point	WUSTL-IIoT: {Accuracy(Acc): 99.98%, Matthew's Correlation Coefficient(MCC): 99.86%, Undetected Rate(UR): 0.23%}, NSL-KDD: {Acc:99.24&, MMC: 98.47%, UR: 1.18% }, UNSW: {Acc: 97.77%, MCC: 94.78%, UR: 1.27% }	Network layer
Khan et al. [47]	Real-world GPS data	Unsupervised	Network-based	Collective	Recall: 97.17%, Acc: 98.26%, F-measure: 98.21%, Precision: 98.29%	Network Layer & Application layer
Sivapalan et al. [64]	MIT-BIH Arrhythmia Database	Supervised	Host-based	Point	Acc: 93%, Sensitivity: 88%, Specificity: 94%, Positive Predictive Rate: 67%, Fβ Score: 90%	Perception layer
Huang et al. [60]	SECOM, Wafer, APS	Unsupervised	Network-based	Point	SECOM: {AUC-ROC: 82.72%, F1-score: 84.53%}, Wafer: {AUC-ROC: 88.16%, F1-score: 89.37%}, APS: {AUC-ROC: 93.25%, F1-score: 97.76% }	Application layer
Guerra-Manzanares et al. [65]	Custom dataset	Supervised	Host-based	Point	DT: {Acc: $>$ 97.5% }, kNN: {Acc: $>$ 97.5% }, RF: {Acc: $>$ 96.25% }	Network layer
Djenouri et al. [54]	NSL-KDD	Supervised	Network-based	Collective	TPR(True Positive Rate): $>$ 94% TNR(True Negative Rate): $>$ 94%	Network layer
Dong et al. [48]	CIC-DDoS2019, CICMalDroid2020, CIC-Darknet2020, CIRA-CIC-DoHBrw-2020	Supervised	Network-based	Collective	Acc: { DDoS2019: 67.03%, MalDroid2020: 89.63%, Darknet2020: 86.76%, DoHBrw2020: 79.63% }, Miss rate: { DDoS2019: 4.40%, MalDroid2020: 7.72%, DoHBrw2020: 0.71% }, F1-score: { DDoS2019: 494.60%, MalDroid2020: 88.59%, DoHBrw2020: 99.54% }	Network layer
Almuqren et al. [61]	NSL-KDD	Supervised	Network-based	-	Acc: 98.87%, Precision: 98.95%, Recall: 98.87%, F1-score: 98.91%, AUC-Score: 98.87%	Network layer
Alani et al. [30]	TON_IoT, IoT-ID, Aposemat IoT-23	Supervised	Host-based	-	Acc: 99.62%, Precision: 99.55%, Recall: 99.61%, F1-score: 99.58%, Training Time: 0.3339(s), Testing Time: 0.4549 (μs)	Network layer
Sharma et al. [62]	NSL-KDD	Supervised	Network-based	Collective	Acc: 99.3%, Loss: 0.00001	Network layer
Alani et al. [49]	WUSTL-IIoT-2021	Supervised	Host-based	Collective	Acc: 99.94%, Precision: 99.92%, Recall: 99.95%, F1-score: 99.94%	Network layer
Patil et al. [50]	CICIDS-2017	Supervised	Network-based	Point Collective	Acc: 99.25%, Precision: 89%, Recall: 89%, F1-score: 89%	Network layer
Huong et al. [51]	SCADA liquid storage infrastructure dataset, SWaT	Unsupervised	Network-based	Point	Threshold: 0.26, Acc: 90.17%, Precision: 90.59%, Recall: 98.06%, F1-score: 94.18%, AUC: 90%	-
Anello et al. [63]	Roller Coaster dataset, Compacting Machine dataset	Unsupervised	-	Point	-	-
Oseni et al. [53]	ToN_IoT	Supervised	Network-based	-	Acc: 99.15%, Precision: 99.10%, Recall: 99.15%, F1-score: 98.83%	Perception Layer & Network Layer & Application layer
Abou El Houda et al. [66]	NSL-KDD, UNSW-NB15	Supervised	Host-based	-	Acc: 88%, Precision: 96%, Recall: 88%, F1-score: 88%	Perception Layer & Network Layer & Application layer
Hussain et al. [52]	SWaT	Unsupervised	Host-based	Point	-	-
Keshk et al. [67]	NSL-KDD, UNSW-NB15, ToN_IoT	Supervised	Network-based	Collective	NSL-KDD: {Acc: 0.931, Precision: 0.958, Recall: 0.829, F1-score: 0.889}, UNSW-NB15: {Acc: 0.840, Precision: 0.799, Recall: 0.943, F1-score: 0.866}, ToN_IoT: {Acc: 0.987, Precision: 0.822, Recall: 0.902, F1-score: 0.859}	Perception Layer & Network Layer & Application layer
Abououf et al. [31]	MIMIC dataset	Unsupervised	Network-based	Point	Precision: 1.0, Recall: 0.94, F1-score: 0.97	Application layer
Djenouri et al. [55]	PFlow, N-BaIoT	Supervised	Network-based	-	-	Application layer
Houda et al. [56]	UNSW-NB15	Supervised	Network-based	-	Acc: 99%, Precision: 99%, Recall: 99%, F1-core: 99%	Application layer
Rathod et al. [57]	IEC 60870-5-104 IDS dataset	Supervised	Network-based	Point	Acc: 98.46%, Precision: 97.56%, Recall: 96.53%, F1-core: 96.65%	Application layer

Sharma et al. [59]	NSL-KDD Cup, UNSW-NB15	Supervised	Network-based	-	NSL-KDD: { Precision: 1.0, Recall: 1.0, F1-core: 1.0}
					UNSW-NBnew: { Precision: 0.57, Recall: 0.04 F1-core: 0.07}
						Network layer

Hasan et al. [58]	ToN_IoT	Supervised	Network-based	Point	Acc: 99.69%, Precision: 100%, Recall: 100%, F1-core: 100%, Error: 0.0030, Sensitivity: 99.68%, Specificity: 99.63%	Perception Layer & Network Layer & Application layer

consequences. The unique XAI-based framework, the integration of DNN architecture with XAI methodologies, and the validation of the system's performance using real-world datasets are the contributions. Overall, the results showed that the proposed approach had the potential to improve IDS interpretability in IoT networks. Keshk et al. [67] presented a unique explainable IDS for IoT networks using an LSTM model. To extract input features and evaluate the LSTM model, the system integrated a unique SPIP (SHAP, Permutation Feature Importance, ICE, PDP) architecture. The SPIP framework integrated feature-based and perturbation-based methodologies to discover relevant characteristics and examine their effect on model output. The paper discussed related work on intrusion detection and explainable AI-based IDS in IoT networks, and it employed a variety of XAI approaches, including SHAP, PFI, ICE, and PDP, to assess the model's operation and predictions. The practical consequences included supporting administrators and decision-makers in understanding complicated attack behavior, increasing the effectiveness and efficiency of IDSs, and offering local and global explanations. The unique intrusion detection framework, the SPIP framework for feature extraction, and the coupling of XAI approaches with LSTM for intrusion detection and explanation are the contributions. The results showed good detection accuracy and interpretability, as well as the potential to improve IDSs in IoT networks.

5. Challenges and future directions

The challenges and future works in XAI-based anomaly detection for the IoT encompass several key aspects that need to be addressed to enhance the effectiveness and applicability of anomaly detection techniques. In this section, we discuss the challenges faced in analyzing complex and large-volume IoT data, as well as potential future research directions.

- Data complexity: IoT generates vast amounts of unstructured and heterogeneous data from numerous sources, including sensors and devices. This data is multidimensional and lacks a predefined structure, often necessitating sophisticated analysis techniques. Managing the abundance of variables and addressing potential data sparsity pose challenges in accurately detecting anomalies. Although XAI methods and feature selection techniques may have overlapping capabilities, they can complement each other effectively. XAI approaches can be used in conjunction with feature selection techniques to reduce dimensionality, thereby helping mitigate data complexity [68].

- Scalability: The exponential growth of IoT devices causes an increase in data volume and velocity as well. This large influx of data must be handled efficiently by anomaly detection algorithms in real time. Researchers are investigating distributed and parallel processing solutions that leverage cloud and edge computing to distribute the workload and boost scalability. Algorithm optimization, data division, sampling, and approximate approaches all help to improve scalability. In the constantly developing IoT ecosystem, addressing scalability issues is critical to ensuring effective and real-time anomaly detection. XAI techniques can be scalable to handle large datasets and real-time processing, where anomalies can occur rapidly [69].

- Interoperability vs. accuracy: Balancing interpretability and accuracy poses a challenge in XAI-based anomaly detection for IoT. XAI aims to provide explicit explanations for AI model decisions; nevertheless, sophisticated models frequently compromise interpretability for greater accuracy, whereas simpler models are more interpretable but less accurate. It is critical to find the correct balance. Researchers are investigating hybrid models that combine accuracy with interpretability or improve the interpretability of complex models through visualization or rule extraction. It is also critical to include contextual information. Finding the right trade-off remains a difficulty in developing transparent and accurate anomaly detection systems that users can trust and understand.

- Dynamic nature of anomalies: Anomalies can alter over time, and as the system matures, new sorts of anomalies may emerge. Conventional anomaly detection approaches based on static models encounter difficulties in promptly adapting and identifying anomalies. Researchers are addressing this issue by creating algorithms that can learn and update anomaly detection models in real time. These algorithms use online learning and adaptive modeling techniques to dynamically adapt to shifting patterns and identify abnormalities as they occur. Unsupervised learning, clustering, and outlier identification are being investigated as methods for capturing the dynamic characteristics of anomalies without relying on existing labels or training data. XAI-based techniques can promote preemptive reactions and prevent potential damages in IoT systems by offering real-time and adaptive anomaly detection [70].

- Data privacy: The integration of XAI into IoT raises concerns regarding data privacy. In the quest for transparent and interpretable artificial intelligence, XAI developers must carefully consider the privacy implications, given that the analysis often involves sensitive data. Preserving user privacy poses a challenge in XAI, emerging as a fundamental step in obtaining informed consent [71]. It is essential to ensure that users know how their data will be used for XAI purposes and provide consent accordingly. Transparency requires clear communication on how data is collected, processed, and used in XAI systems. Striking a sensitive balance between protecting individual privacy and extracting meaningful insights is a challenge that requires the exploration of advanced anonymization and pseudonymization techniques. Regulatory compliance introduces another complex issue, as data protection laws such as the General Data Protection Regulation (GDPR) demand stringent measures for the ethical use of personal information [72]. XAI systems must be designed and implemented with careful adherence to these legal frameworks, promoting a secure and compliant environment. The integration of XAI into IoT systems not only advances interpretability but also necessitates a conscientious approach to user data privacy. Addressing these challenges requires a multifaceted strategy encompassing user awareness, regulatory compliance, exploration of advanced anonymization techniques, and implementation of robust security measures [73].

- Data breach: The use of XAI presents a complex problem in the effort to improve transparency in black-box systems, especially with the rising risk of data breaches. AI aims to enhance transparency by providing understandable explanations for the decisions made by AI models. However, in the process of achieving transparency, there is a risk of unintended data breaches [74]. XAI often operates on comprehensive datasets to elucidate complex algorithms and decision-making processes. The effort to clarify decision processes can potentially lead to the unintentional revelation of sensitive information within the data. This dilemma highlights the need to strike a balance between the transparency goal of XAI and the imperative to safeguard against inadvertent data leaks. Addressing data breaches in the realm of XAI involves developing strategies and protective measures to mitigate the risk of exposing sensitive information during the explanation generation process [75]. This may include implementing robust anonymization techniques, access controls, and encryption methods to ensure that the benefits of transparency provided by XAI do not compromise data security and confidentiality.

XAI technologies have the potential to provide solutions to important problems in many fields by combining them with different technologies in the future. They can be used to detect abnormal situations and explain these anomalies by analyzing large amounts of data from IoT devices. These technologies can play a key role in IoT applications spanning from factories to healthcare and more. Additionally, these technologies can help develop better tools to explain AI model results that are difficult to understand. In the future, advanced XAI methods that address privacy and security issues may help users and organizations have greater trust in these technologies [76]. Another important issue is that IoT networks generally have a distributed structure, and such networks must combine XAI technologies to account for local and global deployment. Finally, XAI can offer the ability to understand the details of specific application domains and produce detailed descriptions, especially through application-specific interfaces [77]. This could open ways to integrate XAI with different technologies and problems in the future.

6. Conclusion

In this paper, we highlight the importance of XAI-based anomaly detection for the IoT. XAI techniques provide transparent and interpretable explanations for anomaly detection in IoT systems, enabling users to understand model decisions and trust in model decisions. The survey comprehensively examined the existing studies in the literature comparatively by the XAI taxonomy and methodology. As a result of the investigations, it is observed that more studies are carried out in the Industrial IoT domain and mostly the feature-based approach was addressed. Also, post-hoc explainability methods have been widely preferred. It is observed that LIME and SHAP are frequently preferred techniques in studies. In addition, the challenges of XAI techniques in IoT are mentioned. Understanding these challenges will form the basis for developing sustainable and effective solutions for future research. Finally, potential future directions are given based on the challenging issues. These provide a framework for encouraging more in-depth studies in the field of XAI-based anomaly detection and increasing knowledge in this field.

References

[1] Patel K, Patel S. Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies, Application & Future Challenges. International journal of engineering science and computing 2016; 6 (5).

[2] Gunning D, Aha D. Darpa's explainable artificial intelligence (XAI) program. AI Magazine 2019; 40 (2): 44-58. doi: https://doi.org/10.1609/aimag.v40i2.2850

[3] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020; 58: 82-115. doi: https://doi.org/10.1016/j.inffus.2019.12.012

[4] Tritscher J, Krause A, Hotho A. Feature relevance XAI in anomaly detection: Reviewing approaches and challenges. Frontiers in Artificial Intelligence 2023; 6: 1099521. doi: https://doi.org/10.3389/frai.2023.1099521

[5] Burhan M, Rehman RA, Khan B, Kim BS. IoT elements, layered architectures and security issues: A comprehensive survey. Sensors 2018; 18 (9): 2796. doi: https://doi.org/10.3390/s18092796

[6] Mahmoud R, Yousuf T, Aloul F, Zualkernan I. Internet of things (IoT) security: Current status, challenges and prospective measures. In: 10th International Conference for Internet Technology and Secured Transactions (ICITST); London, UK; 2015. pp. 336-341. doi: 10.1109/ICITST.2015.7412116

[7] Ahemd MM, Shah MA, Wahid A. IoT security: A layered approach for attacks & defenses. In: International Conference on Communication Technologies (ComTech); Rawalpindi, Pakistan; 2017. pp. 104-110. doi: 10.1109/COMTECH.2017.8065757

[8] Suo H, Wan J, Zou C. Security in the internet of things: a review. In: International Conference on Computer Science and Electronics Engineering; Hangzhou, China; 2012. pp. 648-651. doi: 10.1109/ICCSEE.2012.373

[9] Aarika K, Bouhlal M, Abdelouahid RA, Elfilali S, Benlahmar E. Perception layer security in the internet of things. Procedia Computer Science 2020; 175: 591-596. doi: https://doi.org/10.1016/j.procs.2020.07.085

[10] Deogirikar J, Vidhate A. Security attacks in IoT: A survey. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud); Palladam, India; 2017. pp. 32-37. doi: 10.1109/I-SMAC.2017.8058363

[11] Swamy SN, Jadhav D, Kulkarni N. Security threats in the application layer in IoT applications. In: International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud); Palladam, India; 2017. pp. 477-480. doi: 10.1109/I-SMAC.2017.8058395

[12] Bezerra F, Wainer J, Aalst WM. Anomaly detection using process mining. In: International Workshop on Business Process Modeling, Development and Support; Berlin, Heidelberg; 2009. pp 149-161. doi: https://doi.org/10.1007/978-3-642-01862-6_13

[13] Araya DB, Grolinger K, ElYamany HF, Capretz MA, Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption. Energy and Buildings 2017; 144: 191-206. doi: https://doi.org/10.1016/j.enbuild.2017.02.058

[14] Chatterjee A, Ahmed BS. IoT anomaly detection methods and applications: A survey. Internet of Things 2022; 19: 100568. doi: https://doi.org/10.1016/j.iot.2022.100568

[15] Lim SY, Jones A. Network anomaly detection system: The state of art of network behaviour analysis. In: International Conference on Convergence and Hybrid Information Technology; Daejeon, Korea (South); 2008. pp. 459-465. doi: 10.1109/ICHIT.2008.249

[16] Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Computer Science 2015; 60: 708-713. doi: https://doi.org/10.1016/j.procs.2015.08.220

[17] Lunt TF, Jagannathan R, Lee R, Whitehurst A, Listgarten S. Knowledge based intrusion detection. In: Proceedings of the Annual AI Systems in Government Conference; Washington, DC; 1989.

[18] Omar S, Ngadi A, Jebur HH. Machine learning techniques for anomaly detection: an overview. International Journal of Computer Applications 2013; 79 (2).

[19] Anton SD, Kanoor S, Fraunholz D, Schotten HD. Evaluation of machine learning-based anomaly detection algorithms on an industrial modbus/tcp data set. In: Proceedings of the 13th international conference on availability, reliability and security; Hamburg, Germany; 2018. pp. 1-9. doi: https://doi.org/10.1145/3230833.3232818

[20] Singh AP, Singh MD. Analysis of host-based and network-based intrusion detection system. International Journal of Computer Network and Information Security 2014; 6 (8): 41-47. doi: 10.5815/ijcnis.2014.08.06

[21] Ariyapala K, Do Hoang G, Huynh NA, Wee KN, Conti M. A host and network based intrusion detection for android smartphones. In: 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA); Crans-Montana, Switzerland; 2016. pp. 849-854. doi: 10.1109/WAINA.2016.35

[22] Hu J. Host-based anomaly intrusion detection. In: Handbook of information and communication security. Berlin, Heidelberg: Springer, 2010, pp. 235- 255. doi: https://doi.org/10.1007/978-3-642-04117-4_13

[23] Sivamohan S, Sridhar SS. An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework. Neural Computing and Applications 2023; 35 (15): 11459-11475. doi: https://doi.org/10.1007/s00521-023-08319-0

[24] Zhang Y, Xu F, Zou J, Petrosian OL, Krinkin KV. XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction. In: II International Conference on Neural Networks and Neurotechnologies (NeuroNT); Saint Petersburg, Russia; 2021. pp. 13-16. doi: 10.1109/NeuroNT53022.2021.9472817

[25] Gunning D, Stefik M, Choi J, Miller T, Stumpf S et al. XAI—Explainable artificial intelligence. Science Robotics 2019; 4 (37). doi: 10.1126/scirobotics.aay712

[26] Mukherjee S, Rupe J, Zhu J. XAI for Communication Networks. In: IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW); Charlotte, NC, USA; 2022. pp. 359-364. doi: 10.1109/ISSREW55968.2022.00093

[27] Nguyen TN, Choo R. Human-in-the-loop xai-enabled vulnerability detection, investigation, and mitigation. In: 36th IEEE/ACM International Conference on Automated Software Engineering (ASE); Melbourne, Australia; 2021. pp. 1210-1212. doi: 10.1109/ASE51524.2021.9678840

[28] Gentile D. Jamieson G. Donmez B. Evaluating human understanding in XAI systems. In: ACM CHI XCXAI Workshop; 2021.

[29] Mahalle PN, Patil RV, Dey N, Crespo RG, Sherratt RS et al. Explainable AI for Human-Centric Ethical IoT Systems. IEEE Transactions on Computational Social Systems 2023; 1-13. doi: 10.1109/TCSS.2023.3330738

[30] Alani MM, Miri A. Towards an explainable universal feature set for IoT intrusion detection. Sensors 2022; 22 (15): 5690. doi: https://doi.org/10.3390/s22155690

[31] Abououf M, Singh S, Mizouni R, Otrok H. Explainable AI for Event and Anomaly Detection and Classification in Healthcare Monitoring Systems. IEEE Internet of Things Journal 2023; 11 (2): 3446 - 3457. doi: 10.1109/JIOT.2023.3296809

[32] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv 2020; 11371. doi: https://doi.org/10.48550/arXiv.2006.11371

[33] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access 2018; 6: 52138-52160. doi: 10.1109/ACCESS.2018.2870052

[34] Hariharan S, Rejimol Robinson RR, Prasad RR, Thomas C, Balakrishnan N. XAI for intrusion detection system: comparing explanations based on global and local scope. Journal of Computer Virology and Hacking Techniques 2023; 19 (2): 217-239. doi: https://doi.org/10.1007/s11416-022-00441-2

[35] Neves I, Folgado D, Santos S, Barandas M, Campagner A et al. Interpretable heartbeat classification using local model-agnostic explanations on ECGs. Computers in Biology and Medicine 2021; 133: 104393. doi: https://doi.org/10.1016/j.compbiomed.2021.104393

[36] Macha D, Kozielski M. Wróbel Ł, Sikora M. RuleXAI—A package for rule-based explanations of machine learning model. SoftwareX 2022; 20: 101209. doi: https://doi.org/10.1016/j.softx.2022.101209

[37] Qiu L, Yang Y, Cao CC, Liu J, Zheng Y et al. Resisting Out-of-Distribution Data Problem in Perturbation of XAI. arXiv preprint arXiv:2107.14000 2021; doi: https://doi.org/10.48550/arXiv.2107.14000

[38] Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in cybersecurity: A survey. IEEE Access 2022; 10: 93575-93600. doi: 10.1109/ACCESS.2022.3204171

[39] Li Z, Zhu Y, Van Leeuwen M. A survey on explainable anomaly detection. ACM Transactions on Knowledge Discovery from Data 2023; 18 (1): 1-54. doi: https://doi.org/10.1145/3609333

[40] Neupane S, Ables J, Anderson W, Mittal S, Rahimi S et al. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access 2022; 10: 112392-112415. doi: 10.1109/ACCESS.2022.3216617

[41] Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems 2020; 32 (11): 4793-4813. doi: 10.1109/TNNLS.2020.3027314

[42] Kök İ, Okay FY, Muyanlı Ö, Özdemir S. Explainable artificial intelligence (xai) for internet of things: a survey. IEEE Internet of Things Journal 2023; 10 (16): 14764 - 14779. doi: 10.1109/JIOT.2023.3287678

[43] Jagatheesaperumal SK, Pham QV, Ruby R, Yang Z, Xu C et al. Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions. IEEE Open Journal of the Communications Society 2022; 3: 2106 - 2136. doi: 10.1109/OJCOMS.2022.3215676

[44] Moustafa N, Koroniotis N, Keshk M, Zomaya AY, Tari Z. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. IEEE Communications Surveys & Tutorials 2023; 25 (3): 1775 - 1807. doi: 10.1109/COMST.2023.3280465

[45] Javed AR, Ahmed W, Pandya S, Maddikunta PKR, Alazab M et al. A survey of explainable artificial intelligence for smart cities. Electronics 2023; 12 (4): 1020. doi: https://doi.org/10.3390/electronics12041020

[46] Zolanvari M, Yang Z, Khan K, Jain R, Meskin N. TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security. IEEE Internet of Things Journal 2021; 10 (4): 2967 - 2978. doi: 10.1109/JIOT.2021.3122019

[47] Khan IA, Moustafa N, Pi D, Sallam KM, Zomaya AY et al. A New Explainable Deep Learning Framework for Cyber Threat Discovery in Industrial IoT Networks. IEEE Internet of Things Journal 2021; 9 (13): 11604-11613. doi: 10.1109/JIOT.2021.3130156

[48] Dong T, Li S, Qiu H, Lu J. An interpretable federated learning-based network intrusion detection framework. arXiv preprint arXiv:2201.03134 2022; doi: https://doi.org/10.48550/arXiv.2201.03134

[49] Alani MM, Damiani E, Ghosh U. DeepIIoT: An explainable deep learning based intrusion detection system for industrial IOT. In: IEEE 42nd International Conference on Distributed Computing Systems Workshops (ICDCSW); Bologna, Italy; 2022. pp. 169-174. doi: 10.1109/ICDCSW56584.2022.00040

[50] Patil S, Varadarajan V, Mazhar SM, Sahibzada A, Ahmed N et al. Explainable artificial intelligence for intrusion detection system. Electronics 2022; 11 (19): 3079. doi: https://doi.org/10.3390/electronics11193079

[51] Huong TT, Bac TP, Ha KN, Hoang NV, Hoang NX et al. Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access 2022; 10: 53854-53872. doi: 10.1109/ACCESS.2022.3173288

[52] Hussain MT, Perera C. Explainable sensor data-driven anomaly detection in Internet of Things systems. In: IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI); Milano, Italy; 2022. pp. 80-81. doi: 10.1109/IoTDI54339.2022.00021

[53] Oseni A, Moustafa N, Creech G, Sohrabi N, Strelzoff A et al. An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. IEEE Transactions on Intelligent Transportation Systems 2022; 24 (1): 1000-1014. doi: 10.1109/TITS.2022.3188671

[54] Djenouri Y, Belhadi A, Srivastava G, Lin JCW, Yazidi A. Interpretable intrusion detection for next generation of internet of things. Computer Communications 2023; 203: 192-198. doi: https://doi.org/10.1016/j.comcom.2023.03.005

[55] Djenouri Y, Belhadi A, Srivastava G, Lin JCW. When explainable AI meets IoT applications for supervised learning. Cluster Computing 2023; 26 (4): 2313-2323. doi: https://doi.org/10.1007/s10586-022-03659-3

[56] Abou El Houda Z, Moudoud H, Khoukhi L. Securing Federated Learning through Blockchain and Explainable AI for Robust Intrusion Detection in IoT Networks. In: IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); Hoboken, NJ, USA; 2023. pp. 1-6. doi: 10.1109/INFOCOMWKSHPS57453.2023.10225769

[57] Rathod T, Jadav NK, Tanwar S, Polkowski Z, Yamsani N et al. AI and Blockchain-Based Secure Data Dissemination Architecture for IoT-Enabled Critical Infrastructure. Sensors 2023; 23 (21): 8928. doi: https://doi.org/10.3390/s23218928

[58] Hasan MK, Sulaiman R, Islam S, Rehman AU, Khan AR. An Explainable Ensemble Deep Learning Approach for Intrusion Detection in Industrial Internet of Things. IEEE Access 2023; 11: 115047 - 115061. doi: 10.1109/ACCESS.2023.3323573

[59] Sharma B, Sharma L, Lal C, Roy S. Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. Expert Systems with Applications 2024; 238: 121751. doi: https://doi.org/10.1016/j.eswa.2023.121751

[60] Huang Z, Wu Y, Tempini N, Lin H, Yin H. An energy-efficient and trustworthy unsupervised anomaly detection framework (eatu) for iIoT. ACM Transactions on Sensor Networks 2022; 18 (4): 1-18. doi: https://doi.org/10.1145/3543855

[61] Almuqren L, Maashi MS, Alamgeer M, Mohsen H, Hamza MA et al. Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems. Applied Sciences 2023; 13 (5): 3081. doi: https://doi.org/10.3390/app13053081

[62] Sharma B, Sharma L, Lal C. Anomaly-Based DNN Model for Intrusion Detection in IoT and Model Explanation: Explainable Artificial Intelligence. In: Proceedings of Second International Conference on Computational Electronics for Wireless Communications: ICCWC 2022; Singapore; 2023. pp. 315-324. doi: https://doi.org/10.1007/978-981-19-6661-3_28

[63] Anello E, Masiero C, Ferro F, Ferrari F, Mukaj B et al. Anomaly Detection for the Industrial Internet of Things: an Unsupervised Approach for Fast Root Cause Analysis. In: IEEE Conference on Control Technology and Applications (CCTA); Trieste, Italy; 2022. pp. 1366-1371. doi: 10.1109/CCTA49430.2022.9966158

[64] Sivapalan G, Nundy KK, James A, Cardiff B, John D. Interpretable rule mining for real-time ecg anomaly detection in IoT edge sensors. IEEE Internet of Things Journal 2023; 10 (15): 13095 - 13108. doi: 10.1109/JIOT.2023.3260722

[65] Guerra-Manzanares A, Nõmm S, Bahsi H. Towards the integration of a post-hoc interpretation step into the machine learning workflow for IoT botnet detection. In: 18th IEEE International Conference On Machine Learning And Applications (ICMLA); Boca Raton, FL, USA; 2019. pp. 1162-1169. doi: 10.1109/ICMLA.2019.00193

[66] Abou El Houda Z, Brik B, Khoukhi L. “Why Should I Trust Your Ids?”: An explainable deep learning framework for intrusion detection systems in Internet of Things networks. IEEE Open Journal of the Communications Society 2022; 3: 1164-1176. doi: 10.1109/OJCOMS.2022.3188750

[67] Keshk M, KoronIoTis N, Pham N, Moustafa N, Turnbull B et al. An explainable deep learning-enabled intrusion detection framework in IoT networks. Information Sciences 2023; 639: 119000. doi: https://doi.org/10.1016/j.ins.2023.119000

[68] Ribeiro J, Silva R, Cardoso L, Alves R. Does Dataset Complexity Matters for Model Explainers?. In: IEEE International Conference on Big Data (Big Data); Orlando, FL, USA; 2021. pp. 5257-5265. doi: 10.1109/BigData52589.2021.9671630

[69] Botana ILR, Eiras-Franco C, Alonso-Betanzos A. Regression tree based explanation for anomaly detection algorithm. Proceedings 2020; 54 (1): 7. doi: https://doi.org/10.3390/proceedings2020054007

[70] Wawrowski Ł, Michalak M, Białas A, Kurianowicz R, Sikora M et al. Detecting anomalies and attacks in network traffic monitoring with classification methods and XAI-based explainability. Procedia Computer Science 2021; 192: 2259-2268. doi: https://doi.org/10.1016/j.procs.2021.08.239

[71] Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems 2023; 263: 110273. doi: https://doi.org/10.1016/j.knosys.2023.110273

[72] Colaner N. Is explainable artificial intelligence intrinsically valuable?. AI & SOCIETY 2022; 37: 231–238. doi:https://doi.org/10.1007/s00146-021-01184-2

[73] Majumdar S. Fairness, explainability, privacy, and robustness for trustworthy algorithmic decision-making. Big Data Analytics in Chemoinformatics and Bioinformatics 2023; 61-95. doi: https://doi.org/10.1016/B978-0-323-85713-0.00017-7

[74] Kuppa A, Le-Khac NA. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In: International Joint Conference on neural networks (IJCNN); Glasgow, UK; 2020. pp. 1-8. doi: 10.1109/IJCNN48605.2020.9206780

[75] Srivastava G, Jhaveri RH, Bhattacharya S, Pandya S, Maddikunta PKR et al. XAI for cybersecurity: state of the art, challenges, open issues and future directions. arXiv preprint arXiv:2206.03585 2022; doi: https://doi.org/10.48550/arXiv.2206.03585

[76] Ha DT, Bac TP, Tran KD, Tran KP. Efficient and Trustworthy Federated Learning-Based Explainable Anomaly Detection: Challenges, Methods, and Future Directions. In: Artificial Intelligence for Smart Manufacturing: Methods, Applications, and Challenges. Cham: Springer International Publishing, 2023, pp. 145-166.

[77] Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Transactions on Industrial Informatics 2022; 18 (8): 5031-5042. doi: 10.1109/TII.2022.3146552