

NETWORK INTRUSION DETECTION USING
EXPLAINABLE AI(XAI)

JIANG SHANWEI

UNIVERSITI KEBANGSAAN MALAYSIA

NETWORK INTRUSION DETECTION USING EXPLAINABLE AI(XAI)

JIANG SHANWEI

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF CYBER SECURITY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

PENGESANAN PENCEROBOHAN RANGKAIAN SECARA EXPLAINABLE
AI(XAI)

JIANG SHANWEI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT UNTUK MEMPEROLEH IJAZAH SARJANA SIBER
KESELAMATAN

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2025

DECLARATION

I hereby declare that the work in this project is my own except for quotations and summaries which have been duly acknowledged.

I have not used any AI tools or technologies to prepare this report.

06 February 2025

JIANG SHANWEI
P137321

ACKNOWLEDGEMENT

Praise the Almighty Allah. First and foremost, I would like to express deepest gratitude to Professor NORUL.HUDA for her guidance, advice and support for my research project.

I would like to thank the college of Information Science and Technology at the National University of Malaysia for providing me with the opportunity to conduct this research.

In addition, I would like to thank all the graduate students majoring in cybersecurity for their help, friendship, and creating a pleasant learning environment during my years at UKM.

And finally, to my family especially my parents who supported and encouraged me throughout my endeavours.

LIBRARY FETSM

ABSTRAK

Ujian ini bertujuan untuk mengatasi cabaran dalam sistem pengesanan gangguan rangkaian dengan menggunakan kecerdasan buatan yang boleh dijelaskan (XAI) dan teknik pembelajaran mesin. Masalah utama yang dihadapi bidang ini ialah kerumitan model kecerdasan buatan tradisional, yang sering gagal memberikan penjelasan yang memuaskan kepada pengguna tentang cara membuat keputusan. Dengan menggunakan teknologi XAI, tujuan kita adalah untuk memperkenalkan kaedah yang tidak hanya meningkatkan kepercayaan sistem, tetapi juga memudahkan bagi pengguna untuk memahami logik di belakang keputusan. Ini penting kerana ia membolehkan pengguna, terutama ahli keselamatan, untuk meninjau secara proaktif dan menyesuaikan sistem berdasarkan pemahaman dalam operasinya. Dalam kajian ini, kami menggunakan set data simulasi yang direka untuk simulasi skenario pengesanan gangguan sebenar. Kami akan mengintegrasikan algoritma XAI dengan model pembelajaran mesin yang ada untuk menguji bagaimana penjelasan yang dijana mempengaruhi interaksi sistem pengguna. Hasil kajian menunjukkan bahawa jelas penjelasan yang dihasilkan oleh sistem diperbaiki secara signifikan, membolehkan pengguna memahami lebih baik mengapa ancaman tertentu dianggap berbahaya. Selain itu, penemuan kajian kami menunjukkan bahawa menggunakan XAI boleh memberikan pandangan yang lebih baik terhadap kelemahan dan kekuatan model, yang boleh membantu optimasi hasil pengesanan. Ini secara langsung mempengaruhi pembangunan dan koordinasi sistem-sistem ini untuk maksimumkan efektiviti mereka dalam mengesan dan menjawab ancaman rangkaian. Ujian ini menyediakan dasar yang kuat untuk peningkatan terus menerus teknologi pengesanan gangguan. Kajian ini telah memberikan kontribusi penting untuk menguatkan sistem pengesanan gangguan, menyediakan alat yang lebih baik untuk menjelaskan dan membenarkan tindakan yang dilakukan, dengan demikian meningkatkan keselamatan rangkaian keseluruhan.

ABSTRACT

This study aims to overcome the challenges in network intrusion detection systems by utilizing explainable artificial intelligence (XAI) and machine learning techniques. The main problem facing this field is the complexity of traditional artificial intelligence models, which often fail to provide satisfactory explanations to users on how to make decisions. By using XAI technology, our goal is to introduce a method that not only improves system reliability, but also makes it easier for users to understand the logic behind decisions. This is important because it enables users, especially security experts, to proactively review and adjust the system based on a deep understanding of its operations. In this study, we used a simulated dataset designed to simulate real intrusion detection scenarios. We will integrate the XAI algorithm with existing machine learning models to test how the generated explanations affect user system interaction. The research results indicate that the clarity of the explanations generated by the system is significantly improved, enabling users to better understand why certain threats are considered risky. In addition, our research findings indicate that using XAI can provide better insights into the weaknesses and strengths of the model, which can help optimize detection results. This directly affects the development and coordination of these systems to maximize their effectiveness in detecting and responding to network threats. This study provides a solid foundation for the continuous improvement of intrusion detection technology. This study has made an important contribution to strengthening intrusion detection systems, providing better tools to explain and justify the actions taken, thereby improving overall network security.

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		vii
LIST OF ILLUSTRATIONS		ix
CHAPTER I	INTRODUCTION	
1.1	Research Background	1
	1.1.1 Introduction to Honeypot and XAI	1
	1.1.2 Importance of Honeypot and XAI	2
	1.1.3 Conclusion	3
1.2	Problem Statement	4
	1.2.1 Technical Issues of Honeypot Technology	4
	1.2.2 Technical Issues of XAI Technology	5
	1.2.3 Technical Issues of Combining Honeypot with XAI	5
1.3	Objective of Research	6
1.4	The Structure of Dissertation	6
CHAPTER II	LITERATURE REVIEW	
2.1	Introduction	8
2.2	Overview and Development of Honeypot	8
2.3	Overview and Development of XAI	16
CHAPTER III	METHODOLOGY	
3.1	Introduction	36
3.2	Dataset	37
3.3	Design and Development	40
	3.3.1 Proposed Framework	40
	3.3.2 Machine Learning Algorithms	44
3.4	Evaluation Method	46

CHAPTER IV	RESULTS AND DISCUSSION	
4.1	Data Preprocessing and Decision Tree Training	49
4.2	Explainable AI with Multi-Layer Perceptron and XGBoost	57
4.3	Using LIME Explaining Predictions	65
4.4	Discussion	67
CHAPTER V	CONCLUSION AND FUTURE WORKS	
5.1	Key Findings	70
5.2	Contributions	70
5.3	Limitations and Challenges	71
5.4	Future Directions	71
REFERENCES		72
APPENDICES		
Appendix A	Program Code for Analysis	76

LIST OF ILLUSTRATIONS

Figure No.		Page
Figure 3.1	Research method flow	37
Figure 3.2	Records distribution and normality test results of the attacks distribution	38
Figure 3.3	List of selected packet' s fields for feature extraction	38
Figure 3.4	List of selected packet' s fields for feature extraction	39
Figure 3.5	List of selected packet' s fields for feature extraction	39
Figure 3.6	Comparison of the percentages of normal and attack records	40
Figure 3.7	Research framework	43
Figure 4.1	Data distribution diagram	50
Figure 4.2	Feature distribution diagram	51
Figure 4.3	Feature importance analysis	52
Figure 4.4	Performance indicators of decision tree classifier	53
Figure 4.5	Decision tree visualization analysis diagram	53
Figure 4.6	Visualization of decision tree	54
Figure 4.7	Feature importance ranking analysis	55
Figure 4.8	Confusion matrix and classification report analysis	57
Figure 4.9	Bar chart of mean importance	58
Figure 4.10	SHAP Summary Plot Analysis	60
Figure 4.11	SHAP INTERACTION VALUE	62
Figure 4.12	avg_t_sent SHAP dependency plot	62
Figure 4.13	srvs SHAP dependency plot	63
Figure 4.14	src_pkts SHAP dependency plot	63
Figure 4.15	dur SHAP dependency plot	64

Figure 4.16	dst_host_count SHAP dependency plot	64
Figure 4.17	dst_avg_byts SHAP dependency plot	65
Figure 4.18	src_byts SHAP dependency plot	65
Figure 4.19	LIME Single Sample Interpretation Analysis	67
Figure 4.20	LIME Single Sample Interpretation Analysis	67

LIBRARY FTSM

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

1.1.1 Introduction to Honeypot and XAI

With the rapid advancement of information technology, network attacks have become increasingly complex and diverse, and network security issues have become more severe, especially threats from malicious software. Therefore, analyzing network activity and detecting malicious behavior has become a key task in the field of network security.

Network honeypot is a security technology that attracts and monitors attackers' behavior by simulating fake systems or services. This technology provides a controlled environment for cybersecurity experts to record the activities, techniques, and attack methods of attackers. Honeypots are designed to look like ordinary network systems or services, such as SSH or Telnet, but in reality, they do not handle any real user requests. Their purpose is to attract malicious attackers and enable researchers to obtain useful intelligence by observing their behavior. The true value of honeypots lies in the fact that during the process of being attacked and destroyed, researchers can monitor and analyze these attack behaviors. Although honeypots are highly effective in analyzing attacker behavior, their main limitation is that they can only capture specific types of attacker behavior, especially those targeting known attack patterns, and may not provide comprehensive intelligence. In addition, to ensure that the honeypot can effectively attract attackers and prevent being identified as a fake target, it is necessary to regularly update and maintain the honeypot. (Hongxia Li et al.,2011)

Explainable Artificial Intelligence (XAI) aims to improve the transparency of machine learning models, making their decision-making processes more understandable, especially for complex "black-box" models like deep neural networks. While traditional machine learning models perform excellently, their "black-box" nature often makes it difficult to explain the rationale behind their decisions, leading to a lack of trust from experts and users. The core goal of XAI is to enhance model interpretability, ensuring that security experts, decision-makers, and other users can understand and trust the reasons behind a model's decisions. The existing interpretable AI (Xai) technology, such as lime and shap, improves the interpretability of the model while maintaining high accuracy. The "black box" problem of the model is particularly prominent in the areas of high reliability and audibility, such as network security. Decision-making decisions can cause compliance problems that make it difficult to correct model bias. With the continued development of machine learning and Xai technology, future research can further enhance the transparency of the model, provide a more flexible and practical interpretation method, and ensure that the model can not only effectively detect attacks, but also provide reliable professional confidence and understanding. (Upadhyay, Kumar et al., 2023)

1.1.2 Importance of Honeypot and XAI

Honeypot technology plays an important role in invasion testing as an aggressive defense. By simulating a vulnerable system, honeypot can capture various types of attacks while recording attacker's actions, tools, techniques and strategies (TTPS), such as scan attacks, authentication attacks, and denial of service attacks. These data will provide valuable information to the security team, to better understand the attacker's intrusion methods and goals, to detect new attack techniques, and to strengthen existing defense systems. Honeypot also protected outside the existing safety infrastructure and blocked many attacks. By collecting actual attack data, honeypot support more accurate intrusion detection models and improve the overall network efficiency. The honeypot also provide detailed data for the response of the security event, helping the team understand the intention of the attacker and increase the accuracy and efficiency of the response.

Machine learning (ML) is a powerful tool for intrusion detection systems (IDS) and can identify complex network threats by learning and adapting constantly changing data, especially in an environment where attackers continue to adjust strategies to avoid detection. However, in the machine learning model such as the depth neural network, there is insufficient transparency in the process of the decision making, and the safety expert becomes difficult to explain the result, and the problem of correction of the compliance and data deviation is increasing. Therefore, it is the key to solve these problems by developing more interpretable machine learning models and maintaining higher detection performance and providing greater transparency and criminality.

XAI plays an important role in network security, especially intrusion detection, and helps experts understand how the model is categorized into attacks or normal actions. Techniques such as shielding sensitivity can reveal which input features will most affect the prediction, which are important for the identification of potential attack patterns and the coordination of defence strategies. XAI Not only does it improve the transparency of the model, but also increases the operational effect, and the safety expert can better understand the behaviour of the attacker and establish more effective defence measures in confronting complex and dynamic threats. In the XAI field, there are several ways to increase the interpretability of the ML model. Decision Tree provides a clear and interpretable structure by splitting data based on feature values, making it easy to trace the decision path from input features to the final prediction.

LIME (Local Interpretable Model-agnostic Explanations) is interpreted to understand the local response behavior of a particular instance and to understand how the model responds to some inputs.

SHAP (Shapley Additive explanations) assigns an importance value to each feature for a specific prediction, ensuring consistency and accuracy in feature attribution.

1.1.3 Conclusion

Honey_pot is a powerful network security tool that helps the security experts identify and record the behavior of the attacker and provide valuable information to optimize the intrusion detection system. Combined with XAI technology, honey_pot not only

capture the attacker's behavior but also help the safety experts understand the decision-making process of the machine learning model, and can increase the transparency and defense capabilities of the system. XAI In the network security field, it plays an important role in enhancing the transparency and reliability of machine learning models. XAI Security experts can better understand the underlying principles behind the model decision and improve the effectiveness and reliability of the intrusion detection system. LIME、SHAP、Grad-CAM、Occlusion Sensitivity XAI approach provides intuitive interpretation to allow models to understand how to classify attacks and optimize defense strategies. Combining Honeypot data and Xai technology can further enhance the transparency and reliability of the system and help cyber security experts improve defense strategies and increase overall defense efficiency.

1.2 PROBLEM STATEMENT

1.2.1 Technical issues of honeypot technology

LIMITATIONS:Honeypots can only capture the behavior of specific types of attacks. Especially for known attack patterns or when attackers circumvent honeypots, it may not be effective in capturing all potential threats. For example, some attackers may use bypass techniques to avoid identification by honeypots, thus reducing their effectiveness. The design and maintenance of honeypot environments are also challenging, and as attackers advance in technology, honeypots need to be constantly adjusted to ensure that they can still effectively attract and record attacker behavior.

REAL-TIME ISSUES:Honeypots need to provide efficient real-time data collection capabilities to cope with rapidly changing attack patterns and large-scale attack traffic. How to ensure that attack activities can be recorded and analyzed in a timely manner when the amount of data is huge is a major technical challenge.

COMPLEXITY:The data generated by honeypots is typically high-dimensional and complex. Extracting valuable features from this large volume of data and conducting effective analysis is a major challenge for honeypot technology. This is particularly challenging when dealing with complex network attack behaviors, as it is difficult to extract actionable intelligence and develop actionable security strategies.

1.2.2 Technical Issues of XAI Technology

"Black Box" Problem and Transparency: Despite the progress made by XAI techniques (such as LIME, SHAP, etc.) in improving the interpretability of machine learning models, complex models like deep learning still face the "black box" problem, making it difficult to provide users with sufficiently clear decision-making explanations. This issue is particularly critical in the field of cybersecurity, where the transparency of model decisions directly impacts the trust and decision-making quality of security experts.

Balancing Interpretability and Accuracy: While XAI techniques improve the interpretability of models, maintaining high accuracy without sacrificing interpretability remains a long-term challenge. In applications like intrusion detection, where high accuracy is critical, providing better interpretability while maintaining high performance is still a key difficulty in XAI research.

Generalization and Adaptability: Existing XAI methods, such as LIME and SHAP, may perform differently across various datasets and models. Improving the adaptability and generalization capabilities of XAI techniques in the dynamic and evolving environment of cybersecurity is an urgent issue that needs to be addressed.

1.2.3 Technical Issues of Combining Honeypot with XAI

Data Integration and Interpretability: The data generated by honeypots differs from traditional network traffic data. One of the technical challenges in combining these two is how to effectively integrate this data into machine learning models and provide clear explanations through XAI.

Real-time Issues and Computational Complexity: When combining honeypots and XAI, ensuring that decision explanations can be provided quickly and efficiently during real-time intrusion detection and attack analysis is another key issue. Especially in cases of large-scale data streams and complex models, the computational complexity of XAI may lead to a decrease in real-time performance.

Attack Patterns and Model Bias: Honeypot data itself may contain biases, and machine learning models may produce biases due to imbalanced training data. Addressing these bias issues when integrating honeypots with XAI, and ensuring the fairness and accuracy of the model, remains a significant challenge.

1.3 OBJECTIVE OF RESEARCH

1. To classify UKM-IDS network intrusion attacks using XAI techniques to provide stronger interpretability, enabling defenders to better understand attack patterns and model decisions, thereby enhancing the effectiveness and reliability of the defense system.

2. To optimize feature selection based on SHAP and LIME feature importance and dependency for an effective XAI decision model.

1.4 THE STRUCTURE OF DISSERTATION

This paper is divided into five chapters, and the contents of each chapter are summarized as follows:

Chapter 1: Introduction. Introduce the research background, problem statement, research purpose and problem, and paper structure.

Chapter 2: Literature review. Review the development history and technical points of honeypot technology, analyze the advantages and disadvantages of existing malware analysis methods, explore malware behavior analysis technology in honeypot environment, and combine explainable artificial intelligence (XAI) technology to explore how to improve the interpretability of honeypot data to optimize malware detection and defense strategies.

Chapter 3: Methodology. Explain the research design, data collection methods, analysis methods and experimental settings, as well as the expected experimental results.

Chapter 4: Experimental results and analysis. Show the experimental results, analyze the behavioral characteristics of malware in the sandbox environment, and verify the effectiveness and accuracy of the proposed method.

Chapter 5: Conclusion and Prospect. Summarize the research results, discuss the limitations of the research and possible future research directions.

LIBRARY FTSM

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This section is a literature review, which comprehensively outlines the literature related to this study to help readers understand the research background and experimental environment. The literature materials mainly come from academic journals, blog posts and online articles, and the search tools include databases such as Google Scholar, Web of Science and IEEE Xplore.

Firstly, the background and development history of honeypot technology were introduced, from its initial concept to its widespread application in the field of network security. As a secure resource, the value of honeypots lies in being detected, attacked, or invaded, and monitoring, detecting, and analysing these attack activities.

Next, we discussed various applications of XAI technology in network intrusion detection. Meanwhile, we introduced the latest research achievements and technological progress of XAI technology in network intrusion detection in recent years.

Overall, this chapter presents the important role and development prospects of honeypot technology and XAI technology in network intrusion detection through literature review, providing a solid foundation for subsequent research.

2.2 OVERVIEW AND DEVELOPMENT OF HONEYPOT

In recent years, as the network threat becomes complicated, the conventional signature-based malware detection method shows an obvious limit. Mahajan and Singh proposed a framework that combines network learning algorithms with modern honey network

to capture the attacker's behavior and classify malware to enhance network security. Framework MHN A low interacting honey tank, such as a dionea sensor, is used to effectively collect malicious flow samples and generate detailed attack reports. Data analysis phase, malware sample random forest K-Nearest Neighbor Classify using decision tree such as decision tree. Among them, the expression of random forest is the best, the accuracy reaches 94.65%, the F1 is the highest and the better the traditional method. Compared to existing studies, the novelty point of this job is that it combines the modern honeypot and machine learning algorithms to increase the efficiency and accuracy of data collection and analysis. This framework provides a prospective way to detect and classify malware. Future improvements include integrating highly interactive honey tanks and larger datasets to further optimize detection performance and provide protection from complex network threats. (V. Mahajan & J. Singh,2023)

Honeypots are already an important tool in network security and provide a controlled environment to observe and analyze the behavior of attackers. Lee and Kim wangkin (2011) have made a full view of network honeypots technology and discussed their operational mechanisms, advantages and challenges. The study underlined the mechanism of honeypots, such as tempting the attacker to capture their actions, detecting malicious activity, and recording detailed logs. These logs are important for understanding attack patterns and creating defense strategies. The authors also evaluated the benefits of honeypots, such as the ability to collect valuable information about network threats, and at the same time it was easy to detect resources and be easily detected by skilled attackers. This paper also examines the dynamic interaction between attack and defense techniques in a honeypot system and considers it as a sustainable adaptation game. The authors predicted the future development of honeypot technology with emphasis on enhanced detection methods and more complex interaction models. This study provides valuable knowledge for the development and implementation of honeypots in the modern network security framework. However, it also emphasizes the need to address issues such as integration with extensibility and more extensive security systems. Future tasks in this field can explore the application of machine learning in the enhancement of honeypots and the potential of highly interactive honeypot in improving threat detection and relaxation. (Li, Chen, & Jin, 2011)

Wireless networks are an important component of modern connections, but their vulnerabilities lead to serious security challenges. An 802.11 nectar study has introduced an accurate way to understand the WLAN risk by attracting attackers and placing honeypots to analyze their behavior. The study WLAN Emphasizing two important factors affecting security: vulnerability vulnerable to attacks and network attacks. By simulating a mobile WLAN environment, you can capture the behavior of attackers, such as a privilege upgrade attempt, denial of service attacks, and conversation fraud. This technology design includes a monitoring system for detecting data link level detectors, a target server for recording login attempts, and an access point associated with the log. Snort This research provided a structured method for evaluating 802.11 network vulnerabilities, through traffic classification and an extension index to specific activities of WLAN based on. Note that this study highlights the key weaknesses in WLAN encryption, authentication and session processing and emphasizes the risk of employing wireless technology without strong security measures. This study is consistent with previous studies using honeypots as a deception based safety tool, emphasizing the use in wireless environments. It provides a practical view of the attacker's methodology; WLAN Provide a basis for improving security practices and implement, for example, a high-level monitoring and inspection mechanism. Future research can extend this work by combining machine learning techniques to enhance attack pattern identification, and adjust the design of honeypots to accommodate emerging wireless technologies. These advances will further enhance the role of honeypots in modern network infrastructure protection. (www.loud-fat-bloke.co.uk,n.d.)

The zero day attack is a serious challenge to cyber security, and innovative intrusion detection and prevention methods are necessary. Anagnostakis The shadow honeypot architecture proposed by (2005) proposed a mixture model combining an anomaly detection system (ads) and honeypots to improve the accuracy and reliability of detecting malicious flow. This architecture utilizes an anomaly detection sensor to mark suspicious traffic and is processed by a shadowy tank (an instrumental copy of the original application). This integration will resolve the critical limits of independent honeypots and ads. Although honeypot be used to detect the scan worm, it is difficult to respond to target attacks, but ads can detect a broader threat, but misinformation

occurs. A shadow honeypot ADS Fill this difference by validating the rollback by predicting the state change by malicious activity. Apache The experimental results using the web server and Mozilla Firefox browser verified the effectiveness of this method. This architecture reduces misinformation while maintaining detection accuracy even under high flow conditions. Note that this system is in high performance environments (such as web server fields) with scalable overhead. Compared with traditional honeypots, shadow honeypot extend the detection ability to client attacks such as malicious content in web browsers. The feedback mechanism also enables self correction of the anomaly detection system to improve the long-term system efficiency. The limitations of this method include the dependence of an accurate anomaly detector configuration and the challenge of dynamic attack scenes in pine coupling settings. In the future work, we can integrate machine learning techniques to improve detection accuracy and to expand the shadowed nectar tank to respond to advanced threats such as polymorphisms and deformation attacks. This study is useful for the development of intrusion detection and prevention systems, and provides powerful solutions to the risk of zero day attacks and network security. (Anagnostakis et al., 2005)

Intrusion detection system (IDS) is important for network security, but it is often troubled by high false reporting rates and reduces reliability. Khosravifar and bantahar (2008) proposed a mixture system combining IDs and honeypots, and solved this problem. This system integrates Honeyd with distributed agents with IDS components, honey is a honeypot designed for detailed analysis of suspicious flow. The proposed system performs initial detection using IDs and verifies the alarm using honeypots. By this method IDS The legitimate traffic marked as malicious by the error can be redirected to the original destination after verification and can reduce interruptions. In addition, honeypot create a new signature and create detailed malicious activity logs to improve detection rules and improve the long-term performance of IDS. This architecture has been tested under different network load conditions. This system improves the ability to filter the flow rate before reaching the nectar, and also to improve the performance under high loads. However, the authors point out that further studies are needed to test the system in a more radical flow mode and improve the coordination model between agents and honeypots. IDS It contributes to the development of the intrusion detection field by providing practical solutions to solve one of the most

persistent issues in the introduction. Future developments can combine machine learning techniques to further enhance the accuracy and scalability of the test and provide more adaptive and robust protection against complex network threats. (Khosravifar & Bentahar, 2008)

In modern network security, honeypot systems are widely employed for intrusion detection and malicious activity analysis. However, traditional honeypot systems face challenges in effectively handling complex network attacks. Tang et al. proposed an enhanced honeypot system, named HonIDS, which integrates the TFRPP model and the Bayesian model to improve detection capabilities. HonIDS adopts a five-layer architecture consisting of the service layer, host layer, event layer, data modeling layer, and detection layer. This modular design simplifies implementation and extension.

The experimental results show that HonIDS demonstrated high detection rates in experiments and effectively reduced false positives and negatives. Additionally, by combining the two models, the system significantly improved the quality of automatically generated intrusion detection signatures. The design and implementation of HonIDS highlight the potential of honeypot systems in intrusion detection. However, the study also identifies limitations, such as the inability to detect unforeseen attack behaviors and the need for parameter optimization to adapt to different network scales. (Tang, Hu, Lu, & Wang, 2006)

In the field of modern network security, honeypot-based intrusion detection systems (IDS) are widely applied to defend against cyberattacks and enhance overall system security. Veena et al. proposed an advanced intrusion detection solution using honeypot servers, combining IP validation and voice recognition techniques to improve detection accuracy through honeypot technology and machine learning models.

Methodology and Architecture: The system employs a two-stage validation process (IP validation and voice recognition) to distinguish between legitimate and unauthorized users. Packets from validated users are forwarded to the server, while unauthorized packets are routed to honeypot servers for further analysis. By integrating

Naive Bayes classifiers and other algorithms, the system effectively detects and records attack behaviors for subsequent analysis.

Performance and Results: Experimental results indicate a detection rate exceeding 95% and a false alarm rate of less than 1%, demonstrating significant improvements in accuracy and real-time response capabilities. The proposed solution offers effective protection for network systems through real-time responses and low complexity.

Contributions and Future Directions: Honeypot technology not only captures attack behaviors but also generates new detection rules to enhance IDS performance. The authors suggest future research directions, including the adoption of advanced machine learning algorithms to further improve detection efficiency and the integration of additional security technologies, such as AI-driven real-time threat responses.

This study highlights the vital role of honeypot technology in modern network security and provides new directions for the development of advanced intrusion detection systems. (Veena et al., 2023)

As cyber threats grow increasingly sophisticated, traditional intrusion detection methods exhibit significant limitations. Raghul et al. propose an integrated solution combining advanced honeypot systems with the Suricata intrusion detection system (IDS) to enhance threat detection and response capabilities.

Research Objectives and Methodology: The study aims to leverage honeypot technologies to attract and capture attack behaviors while utilizing Suricata's deep packet inspection and rule-based analysis to improve detection accuracy. The system adopts a modular design, integrating multiple honeypots (e.g., Cowrie and Honeytrap) with Suricata IDS to enable multi-dimensional threat detection.

Performance and Experimental Results: In simulated attack scenarios, the system demonstrated strong performance in terms of detection rates and false-positive

reduction, particularly excelling in identifying zero-day and unknown threats. The T-Pot honeypot platform's visualization capabilities allow real-time monitoring of network activities, generating high-quality logs for subsequent analysis.

Innovations and Limitations: By combining the deception capabilities of honeypots with Suricata's rule-based detection, the study provides a proactive approach to network security. The authors note that deploying and maintaining a multi-component system may increase resource overhead, requiring robust hardware and continuous oversight.

Future Directions: The study suggests incorporating machine learning techniques to further improve detection efficiency for emerging threats. Integrating real-time log analysis and dynamic rule adjustment mechanisms into the system is recommended to enhance adaptability and responsiveness.

This research offers an effective network security solution, laying a solid foundation for detecting and mitigating advanced cyber threats through the synergy of honeypots and IDS. (Raghul et al., 2024)

Blockchain-based IoT (BIoT) offers decentralized and transparent features that support various industries but remains vulnerable to sophisticated cyber threats. Daniel Commey et al. proposed a dynamic honeypot deployment model powered by artificial intelligence (AI) and game theory, aiming to optimize defense strategies against advanced threats in BIoT systems. The model integrates AI-driven intrusion detection systems (IDS) with smart contract functionalities to enable dynamic honeypot deployment. Suspicious network traffic is redirected to honeypots for isolated analysis, while legitimate traffic flows uninterrupted. Bayesian game theory is employed to analyze interactions between defenders and attackers, optimizing honeypot deployment to balance detection efficiency and operational costs. Simulations demonstrated the model's effectiveness in detecting sophisticated attacks, such as those disguised as legitimate traffic, while reducing false positives. The dynamic strategy showed strong adaptability across various attack rates, improving detection accuracy and minimizing resource expenditure. By combining game theory and AI, this study provides an

innovative solution for proactive defense in BIoT networks. It addresses current threats while offering scalability and adaptability for evolving challenges, laying a foundation for robust and future-proof BIoT security systems. Despite its promising results, the model's high computational requirements and reliance on accurate parameter optimization present challenges. Future research should focus on validating the model in real-world BIoT scenarios, incorporating machine learning techniques, and enhancing dynamic capabilities to improve robustness and scalability. (Commey et al., 2024)

The study by Sujata Yeldi et al. integrates intrusion detection systems (IDS) with honeypot technology, proposing an enhanced security solution called “Mirage.” This system aims to strengthen network defenses by leveraging honeypot deception and traffic redirection while collecting attacker behavior data to refine security strategies. Core Contributions includes:

Role of Honeypots: Honeypots serve as decoys to attract attackers, recording their activities and preventing access to actual resources. The study distinguishes between low-interaction and high-interaction honeypots, combined with multi-layered logging mechanisms for thorough data capture.

Integration with IDS: Mirage integrates Snort IDS with the honeypot system. When suspicious activity is detected, traffic is redirected to the honeypot for further analysis, enabling dynamic threat response.

Multi-Layer Logging: The system employs a five-layer logging approach (e.g., Ethereal packet capture, Snort rule-based detection, Tripwire file integrity monitoring) to collect comprehensive data on attacker behavior.

Mirage effectively isolates malicious traffic, protecting critical network resources while generating high-quality log data for subsequent analysis. Results demonstrate superior responsiveness to unknown threats compared to traditional static defense methods.

The system's high demands on hardware resources and real-time processing may limit its applicability in certain environments. Future research could focus on developing more efficient log analysis tools and automated response mechanisms to enhance scalability and adaptability. (Yeldi et al., 2003)

2.3 OVERVIEW AND DEVELOPMENT OF XAI

Network Intrusion Detection System (NIDS) is an important tool for modern network security. It ensures the integrity and reliability of the system by identifying and preventing potential network threats. However, many NIDS models based on machine learning (ML) are difficult to explain their decision logic due to their "black box" nature, which limits their application in security-sensitive fields. Pap M. Corea et al. proposed a method based on explainable artificial intelligence (XAI) to conduct an in-depth analysis of the feature sensitivity and robustness of various ML models in intrusion detection tasks. Machine learning models (such as random forests, multi-layer perceptrons MLP and decision trees) perform well in intrusion detection, but their dependence on high-weight features lacks detailed research. The authors analyzed the performance changes of the model in the absence of key features through the XAI tool "occlusion sensitivity" to explore how to optimize feature engineering and model selection.

Research methods: Data processing and feature selection, using the UNSW-NB15 dataset, combined with real network behavior and synthetic attack behavior, feature cleaning and correlation screening are performed, and finally high-correlation features are retained for modeling. Model training and task classification, binary classification task: distinguish normal traffic from intrusion traffic. Multi-classification task: Identify different attack types (such as DoS, vulnerability exploits, etc.). The models used include random forest, decision tree, linear SVM, KNN and MLP, and analyze the feature sensitivity and performance degradation of each model. Performance evaluation, use the occlusion sensitivity method to evaluate the model performance after occluding key features. Compare the accuracy, robustness and feature dependency of the models in binary and multi-classification tasks.

Main findings: Feature sensitivity, most models (such as decision trees, linear SVM) are highly dependent on a small number of key features (usually less than 3), especially time-related features (such as TTL). Random forests show minimal performance degradation when features are occluded, showing strong robustness. Model robustness, random forests and KNN are stable in feature distribution changes and have better adaptability than other models. In multi-classification tasks, the impact of feature selection on accuracy is more significant, indicating that reasonable feature engineering is the core of complex tasks. The key role of feature engineering, studies have shown that model performance improvement depends more on feature engineering rather than complex model design, which emphasizes the importance of feature optimization.

Research Contribution and Inspiration: Model Optimization Guidance, XAI technology reveals the feature utilization pattern of the model, and random forest is recommended as an ideal intrusion detection model due to its robustness and balance. Practical Applicability, the importance of time-related features to model performance highlights the necessity of feature selection, but it may also affect the generalization ability of the model in other scenarios. Future Research Direction, explore cross-domain feature engineering technology to reduce the model's dependence on specific features. Combine more XAI tools (such as SHAP and LIME) to improve the transparency and interpretability of the model.

This study provides a new perspective on the application of machine learning in intrusion detection through XAI technology. Random forest has been proven to be the best choice for intrusion detection due to its high robustness and wide applicability, and the priority of feature engineering lays the foundation for building a transparent and efficient detection system. (Corea et al., 2024)

With the widespread application of cloud computing, its dynamics and complexity have brought new challenges to traditional intrusion detection systems (IDS), including high false alarm rates, underreporting problems, and lack of transparency in the decision-making process. In response to these problems, Pap M. Corea et al. proposed a solution combining explanatory artificial intelligence (XAI) to

enhance the application performance of IDS in dynamic cloud environments by improving its explanatory power and adaptability. In the study, XAI technologies such as SHAP and LIME were applied to analyze the contribution of key features in AI model decisions, and the main indicators that trigger alarms were identified through attention mechanisms and integrated gradient techniques, thereby helping security analysts quickly distinguish between normal traffic and malicious activities. Experiments show that XAI-based IDS significantly reduces the false alarm rate and exhibits stronger detection capabilities when dealing with complex attack patterns (such as zero-day attacks). In addition, the study pointed out that in the future, the computational efficiency of XAI can be further optimized, balancing model complexity and explanatory power, while combining dynamic characteristics and deep learning to improve the robustness and scalability of the system. Overall, this study provides an important direction for the development of IDS in cloud environments. By introducing XAI technology, it not only improves the detection performance but also provides strong support for security decision-making. (Upadhyay et al.,2023)

In recent years, the rapid development of edge computing and artificial intelligence (AI) technologies has brought great changes to intelligent surveillance systems. However, while complex deep learning models improve the performance of edge camera systems, they also introduce issues of explainability and fairness. In the literature, Nguyen et al. proposed a diagnostic framework based on explainable artificial intelligence (XAI) to improve the fairness and performance of human detection models in edge cameras. The study focuses on the Bytetrack model, analyzes the model's insufficient performance in detecting individuals with occluded or missing body parts through XAI technology, and identifies training data bias as the main source of problems. The study reviews existing XAI techniques, such as the perturbation-based D-RISE method and the back-propagation-based Grad-CAM, and explores the applicability of these techniques in explaining complex detection models. Subsequently, the authors proposed a multi-stage framework that includes data selection, statistical analysis, problem identification, and model improvement. Specifically, the framework uses XAI tools to classify prediction errors, such as "under-detection", "over-detection", and "mislocalization", and generates improvement solutions through these explanations, including data augmentation, re-labeling, and model parameter adjustment.

Experimental results show that re-labeling the bounding boxes in the dataset significantly improves the Bytetrack model's ability to detect partially occluded humans, especially in real-world surveillance scenarios. In addition, the study also highlights the improvement in the detection of individuals with partially occluded or hidden bodies after model tuning, laying the foundation for future intelligent edge computing applications. Future research can further combine more advanced XAI methods to continuously improve the fairness and transparency of detection models. (Nguyen et al.,2024)

With the rapid popularization of artificial intelligence (AI) technology in many fields, its "black box" characteristics have aroused people's concerns about transparency, fairness and trust. In their study, Kamate et al. explored the importance of explainable artificial intelligence (XAI) and focused on the advantages of the SHAP method based on Shapley values in improving the interpretability of models. In the study, the actual effect of SHAP technology in explaining the prediction results of the model was evaluated by applying voice call quality data. As a post-processing technique, the SHAP method can provide a detailed feature contribution decomposition for the prediction of each specific instance, thereby revealing the degree of influence of each feature on the model output. The study shows that SHAP has significant advantages, including meeting the principles of efficiency, symmetry, zero feature contribution and additivity, which makes it widely applicable in many machine learning models. At the same time, the influence of features on the prediction results is further intuitively presented through visualization tools such as waterfall charts, force diagrams and decision diagrams, providing strong support for analysts. However, the study also pointed out that the SHAP method has the problem of high computational complexity when processing large-scale data, especially in high-dimensional data and complex models. To this end, the authors explored the Tree SHAP technology based on tree model optimization. By using conditional expectations instead of marginal expectations, the computational overhead was greatly reduced, making it more suitable for practical application scenarios. Through experiments on voice call quality data, the study demonstrated the potential of SHAP in analyzing customer experience and optimizing network performance, while revealing the positive and negative effects of specific features on prediction results. This study emphasizes the important role of XAI technology in

improving model transparency and user trust, and suggests that in the future, while developing more efficient XAI methods, we should continue to explore its applicability and performance improvements in different application areas. (Kamate et al.,2024)

With the popularity of smart computing devices and the increase in the use of Android devices, the frequency of malware attacks has also increased significantly. To address this problem, Patel et al. proposed an Android malware detection framework AMD-XAI-ML based on explainable artificial intelligence (XAI), which improves the accuracy and transparency of malware detection by combining machine learning (ML) models and XAI technology. The study used the CICAndMal2019 dataset and evaluated multiple ML models, including random forest (RF), extreme gradient boosting (XGB), and extra tree classifier (Extra Tree). The results show that XGB has the highest detection accuracy of 98.54%, followed by Extra Tree and RF, which are 98.52% and 98.42%, respectively. The framework achieves effective detection through three stages: data preprocessing, feature optimization, and model interpretation. XAI techniques (such as SHAP) are used to reveal the importance of each feature in the model decision, thereby helping analysts better understand and optimize the detection system. Experimental results show that feature importance analysis based on SHAP can significantly improve the system's detection ability for specific types of malware, especially when facing complex or obfuscated attacks. In addition, the study also emphasized the important role of XAI technology in reducing false positive rates and improving model robustness. Future work directions include developing more advanced ML models for multi-category malware detection and combining cross-platform datasets to improve the model's detection capabilities for zero-day attacks. This study provides an efficient and explainable solution for malware detection on Android devices, demonstrating the great potential of XAI technology in intelligent computing environments. (Patel & Ghosh, 2024)

In recent years, as artificial intelligence (AI) technology has been increasingly applied across various fields, its transparency and interpretability have become critical areas of focus, especially in high-stakes domains such as healthcare, finance, and autonomous driving. Traditional AI models often operate as "black boxes," where their internal decision-making processes remain opaque to both users and developers. This

lack of transparency raises trust and ethical concerns, further compounded by the growing demands of regulatory compliance. To address these challenges, Jaibir Singh et al. proposed a novel framework that seeks to balance high performance with high interpretability, advancing the development of Explainable Artificial Intelligence (XAI). The study reviews progress in the XAI field, covering techniques such as feature importance analysis, model-agnostic methods (e.g., LIME and SHAP), and interactive visualization tools. These methods have proven effective in enhancing model transparency, identifying biases, and improving fairness. However, the research highlights that traditional approaches often compromise performance to achieve interpretability. The proposed framework demonstrates that, with the adoption of advanced techniques, it is possible to maintain or even enhance model performance while improving its interpretability. Experimental results indicate that the framework significantly improves transparency and user trust in complex datasets while achieving performance levels comparable to, or better than, traditional "black-box" models. This finding challenges the conventional belief that interpretability necessarily comes at the cost of performance, offering a new pathway for developing AI systems that combine high performance with high transparency. The study further emphasizes the importance of embedding interpretability into the design of AI systems from the early stages rather than treating it as an afterthought. By introducing this framework, the research provides empirical support for advancing XAI technologies in complex decision-making scenarios, showcasing their potential for building trustworthy, efficient, and ethically sound AI systems. Future work is encouraged to explore the cross-domain applicability of the framework and develop more intuitive tools to address the diverse interpretability needs of stakeholders. (Singh et al., 2024)

Internet of Things (IoT) networks have developed rapidly in recent years, but their security issues have become increasingly prominent, especially in critical service areas. Due to the limitations of traditional network security mechanisms in IoT environments, intrusion detection systems (IDS) have gradually become an important direction of network security research. In 2021, Harshil Patel wrote about focusing on intrusion detection systems (IDS) in IoT network security, exploring the application of machine learning (ML) technology and its interpretability issues. The author pointed out that traditional security mechanisms are difficult to meet the complex and diverse