

Course Title: Research Methodology in Computer Science

Course Code: AI800

Weighting: 50%

Lecturer: Dr. Mohammed Anbar

Hand-out Date: 18 July 2026

Hand-in Date: 15 Aug 2026

Hand-in Method: Online Submission via E-Learning.

Report: Explainable Artificial Intelligence for Intrusion Detection Systems: A Comparative Evaluation of Detection Performance and Explanation Quality

**Explainable Artificial Intelligence for
Intrusion Detection Systems: A
Comparative Evaluation of Detection
Performance and Explanation Quality**

Contents

ABSTRACT.....	5
1. INTRODUCTION	6
2. CHAPTER 1.....	7
2.1) CHAPTER 1	7
2.1.1 Intrusion Detection Systems	7
2.1.2 Machine Learning and Deep Learning in IDS	8
2.1.3 The Black-Box Problem	8
2.1.4 Paper4: Explainable Artificial Intelligence	9
2.1.5 XAI in Intrusion Detection Systems	9
2.1.6 Current Challenges in Explainable IDS	10
2.1.7 Problem Statement	10
2.1.8 Research Scope	12
2.2) CHAPTER 2 – LITERATURE REVIEW	13
2.2.1 Explainable Artificial Intelligence in Cybersecurity	13
2.2.2 Development of XAI Evaluation in Cybersecurity	15
2.2.3 Explainable Artificial Intelligence in Intrusion Detection Systems	16
2.2.4 Reliability and Consistency of XAI Explanations	17
2.2.5 Synthesis of Previous Studies and Research Gap	18
2.3) CHAPTER 3 – RESEARCH METHODOLOGY.....	18
2.3.1 Research Design.....	18
2.3.2 Dataset	19
2.3.3 Data Preprocessing	19
2.3.4 Intrusion Detection Models.....	20
2.3.5 Application of XAI Techniques.....	21
2.3.6 Evaluation of Intrusion Detection Performance	21
2.3.7 Evaluation of Explanation Quality.....	22
2.3.8 Comparative Analysis	22
2.3.9 Implementation Tools.....	23
2.3.10 Proposed XAI-IDS Evaluation Framework	23
2.3.11 Framework Validation	25
2.3.12 Ethical Considerations.....	26
2.4) CHAPTER 4 – RESEARCH LIMITATIONS.....	27
2.5) CHAPTER 5 – EXPECTED OUTCOMES	29
2.5) CHAPTER 6 – GANTT CHART	31
3. APPENDIX.....	32
3.1.1 Table 1 Synthesis Matrix for Literature Review	32
3.1.2 Tables Critical Analysis.....	32

3.2) ATTACHMENTS.	32
3.2.1 Cybersecurity Revolution via Large Language Models and Explainable AI.	32
3.2.2 A Survey on Explainable Artificial Intelligence for Cybersecurity.	32
3.2.3 Explainable Artificial Intelligence in Cybersecurity: A Survey.	32
3.2.4 Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research.	32
3.2.5 An assessment framework for explainable AI with applications to cybersecurity.	32
3.2.6 Explainable artificial intelligence for cybersecurity: a literature survey.	32
3.2.7 An Excel file containing a Gantt chart.....	32
3.2) REFERENCES. 32	

Abstract

The growing complexity of cyber threats has increased the use of machine learning and deep learning techniques in Intrusion Detection Systems (IDS). Although these models can provide strong detection performance, many operate as black-box systems, making it difficult for cybersecurity analysts to understand the reasons behind their predictions. Explainable Artificial Intelligence (XAI) techniques have been introduced to improve the transparency and interpretability of such systems. However, existing studies often focus more on detection accuracy than on evaluating the quality and reliability of the explanations produced by different XAI methods. In addition, different explanation techniques may provide inconsistent interpretations for the same IDS model.

This research aims to comparatively evaluate XAI techniques for intrusion detection by considering both detection performance and explanation quality. Selected machine learning and deep learning models will be evaluated using common cybersecurity performance measures such as accuracy, precision, recall, F1-score, and false positive rate. XAI techniques will also be assessed using criteria related to explanation consistency, stability, fidelity, and computational efficiency. The expected outcome is to identify an effective combination of intrusion detection and explainability techniques that provides accurate threat detection while producing clear, reliable, and practical explanations for cybersecurity analysts.

1. Introduction

The rapid growth of digital networks, cloud services, Internet of Things (IoT) devices, and connected systems has increased both the volume of network traffic and the complexity of cyber threats. Traditional intrusion detection approaches, particularly signature-based systems, remain useful for identifying known attacks but are limited when dealing with new, evolving, or previously unseen threats. As a result, machine learning (ML) and deep learning (DL) techniques have increasingly been integrated into Intrusion Detection Systems (IDS) because of their ability to identify complex patterns and anomalies within large amounts of network traffic data.

Despite their strong detection capabilities, many ML- and DL-based IDS models operate as black-box systems. Although these models can classify network activities as normal or malicious, security analysts may not clearly understand which features or patterns contributed to a particular prediction. This lack of transparency can reduce trust in automated IDS decisions, especially when analysts must determine whether an alert represents a genuine attack or a false positive. The problem becomes particularly significant in cybersecurity because incorrect or unexplained decisions may affect the speed and quality of incident response.

Explainable Artificial Intelligence (XAI) has emerged as an approach for addressing the interpretability limitations of complex AI models. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) can provide information about the features that influence individual or overall model predictions. Recent studies have demonstrated that these techniques can improve the transparency of IDS and assist cybersecurity analysts in understanding how AI-based detection models reach their decisions. For example, recent experimental work has applied SHAP and LIME to machine learning and deep learning intrusion detection models using datasets such as UNSW-NB15 and TON_IoT, demonstrating that explainability can be incorporated without necessarily sacrificing high detection performance.

However, integrating XAI into IDS introduces additional challenges. Existing research identifies issues including computational overhead, scalability, the trade-off between predictive performance and interpretability, and the absence of standardized criteria for evaluating explanation quality. Furthermore, recent experimental evidence indicates that different XAI techniques may produce conflicting explanations for the same intrusion detection model. This raises an important concern: an explanation may be available, but its reliability, stability, and consistency cannot automatically be assumed.

Therefore, this research focuses on the comparative evaluation of Explainable Artificial Intelligence techniques for Intrusion Detection Systems. Rather than evaluating IDS models solely on conventional metrics such as accuracy, precision, recall, F1-score, and false positive rate, the study will also consider the quality of the explanations generated by selected XAI techniques. Factors such as explanation fidelity, stability, consistency, and computational efficiency will be examined alongside intrusion detection performance. The purpose is to identify approaches that provide an appropriate balance between accurate threat detection and explanations that are sufficiently reliable and practical for cybersecurity analysts.

2. Chapter 1

2.1) Chapter 1

2.1.1 Intrusion Detection Systems

Intrusion Detection Systems (IDS) are an important component of cybersecurity because they monitor network traffic or system activities to identify suspicious or unauthorized behavior. Traditional IDS approaches generally rely on predefined rules, signatures, or known attack patterns. Although these approaches can be effective for detecting previously identified threats, they have limitations when dealing with new or evolving attacks that do not match existing signatures. Modern IDS therefore increasingly use data-driven techniques to identify abnormal behavior and previously unseen attack patterns.

IDS can generally be divided into Network Intrusion Detection Systems (NIDS) and Host Intrusion Detection Systems (HIDS). NIDS monitor network traffic and communication patterns, while HIDS focus on activities occurring within individual devices or hosts. Both approaches contribute to cybersecurity protection, but the growing volume and complexity of network data have increased the need for more adaptive detection techniques.

2.1.2 Machine Learning and Deep Learning in IDS

Machine Learning (ML) and Deep Learning (DL) have become increasingly important in intrusion detection because they can learn patterns from historical network traffic and distinguish between legitimate and malicious activities. ML techniques such as Decision Trees, Random Forest, XGBoost, CatBoost, Logistic Regression, and Support Vector Machines have been applied to intrusion detection, while DL approaches such as Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), autoencoders, and recurrent architectures are able to capture more complex relationships within network data.

Recent studies demonstrate the potential of these models. Mohale and Obagbuwa (2025) evaluated several ML models using the UNSW-NB15 dataset and reported that XGBoost and CatBoost achieved the highest accuracy of approximately 87% among the evaluated models. Alabbadi and Bajaber (2025), using seven TON_IoT datasets, reported a test accuracy of 99.24% for a CNN model on the network dataset and very high accuracy across several IoT datasets. These findings demonstrate that ML and DL can provide strong intrusion detection capabilities.

2.1.3 The Black-Box Problem

Despite their detection capabilities, many advanced ML and DL models are difficult to interpret. Their internal decision-making mechanisms may involve complex relationships between a large number of parameters and features. As a result, cybersecurity analysts may receive an alert indicating that network activity is malicious without receiving a clear explanation of why the model reached that conclusion.

This black-box problem is particularly important in intrusion detection. Security analysts need to determine whether an alert is valid, identify the factors responsible for the detection, and decide what response should be taken. A lack of transparency can reduce confidence in automated decisions and make it more difficult to investigate false positives, model errors, biases, or unexpected behavior.

2.1.4 Paper4: Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) aims to make AI and machine learning decisions more understandable to human users. Instead of presenting only a prediction, XAI techniques provide additional information about the features or factors that influenced that prediction. In cybersecurity, this can help analysts understand why particular network activity was classified as normal or malicious.

Two commonly used XAI techniques are SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). SHAP estimates the contribution of individual features to model predictions using concepts derived from Shapley values, while LIME creates a simpler local approximation of a complex model around a particular prediction. These techniques can therefore provide different perspectives on how intrusion detection models make their decisions.

Other approaches include rule-based explanations, interpretable decision trees, ELI5, and surrogate-model techniques such as TRUSTEE. Different XAI techniques vary in their ability to provide local or global explanations, their computational requirements, and their suitability for different intrusion detection models.

2.1.5 XAI in Intrusion Detection Systems

The integration of XAI with IDS aims to combine strong detection performance with understandable model decisions. Recent research has applied XAI techniques to both machine-learning and deep-learning intrusion detection models. Mohale and Obagbuwa (2025), for example, applied SHAP, LIME, and ELI5 to models including XGBoost, CatBoost, Decision Trees, and MLP using UNSW-NB15. Their analysis showed that XAI can identify influential network features such as Source-to-Destination Time-to-Live (sttl) and Destination Service Count (ct_srv_dst).

Similarly, Alabbadi and Bajaber (2025) integrated SHAP and LIME with CNN, DNN, and TabNet models using TON_IoT datasets. Their framework was designed not only to classify network and IoT attacks but also to identify the features responsible for model predictions.

More recent work has moved beyond simply generating explanations and has begun examining whether those explanations are reliable. Kumar and Thing evaluated state-of-the-art DL-based NIDS models using SHAP and TRUSTEE and found that explanations generated by different XAI methods can conflict for the same intrusion detection model. They also examined explanation stability and adversarial robustness, showing that the quality of an explanation cannot be assumed simply because an XAI method has been applied.

2.1.6 Current Challenges in Explainable IDS

Although XAI improves transparency, several challenges remain. A systematic review by Mohale and Obagbuwa (2025) identifies computational overhead, scalability, privacy risks, real-time performance, and the trade-off between interpretability and detection accuracy as important unresolved problems. Techniques such as SHAP and LIME can require additional processing, which may be problematic in high-speed network environments where intrusion detection decisions must be generated rapidly.

Another important challenge is the evaluation of explanation quality. Conventional IDS metrics such as accuracy, precision, recall, F1-score, and false positive rate measure detection effectiveness but do not determine whether an explanation is faithful, stable, consistent, or computationally practical. The reviewed literature therefore emphasizes the need for standardized interpretability metrics that allow meaningful comparison between XAI techniques.

Furthermore, the finding that different XAI techniques may produce conflicting explanations creates an additional reliability concern. For cybersecurity analysts, receiving different explanations for the same IDS prediction may make it difficult to determine which explanation should be trusted. Therefore, evaluating intrusion detection performance together with explanation fidelity, stability, consistency, and computational efficiency is an important direction for the development of trustworthy and practical XAI-based intrusion detection systems.

2.1.7 Problem Statement

Intrusion Detection Systems (IDS) are essential for identifying malicious activities and protecting modern networks from increasingly complex cyber threats. Machine learning and deep learning techniques have improved the ability of IDS to detect both known and previously unseen attacks. However, many high-performing models operate as black-box systems, meaning that cybersecurity analysts may receive an intrusion alert without clearly understanding the reasons behind the model's decision. This lack of transparency can reduce trust in automated IDS and make it more difficult for analysts to validate alerts, investigate false positives, and respond appropriately to security incidents.

Explainable Artificial Intelligence (XAI) has been introduced to address this problem by providing explanations for machine learning and deep learning predictions. Existing studies have successfully integrated techniques such as SHAP, LIME, ELI5, and TRUSTEE into intrusion detection models. For example, Alabbadi and Bajaber (2025) applied SHAP and LIME to CNN, DNN, and TabNet-based IDS models, while Mohale and Obagbuwa (2025) used SHAP, LIME, and ELI5 with several machine learning classifiers. These studies demonstrate that XAI can identify the network features that contribute to intrusion detection decisions and therefore improve model transparency.

Nevertheless, providing an explanation does not necessarily mean that the explanation is reliable or suitable for operational cybersecurity use. Current research still faces challenges related to computational overhead, scalability, real-time explainability, and the trade-off between detection accuracy and interpretability. In addition, there is no widely standardized approach for evaluating the quality of explanations generated by XAI techniques in IDS. Traditional performance measures such as accuracy, precision, recall, F1-score, and false positive rate evaluate detection effectiveness, but they do not determine whether an explanation is faithful, stable, consistent, or computationally efficient.

A further concern is that different XAI methods may provide different explanations for the same intrusion detection model. Kumar and Thing found that explanations generated by SHAP and TRUSTEE were in conflict for several evaluated deep-learning-based network intrusion detection systems. Their results also showed differences between explanation techniques in terms of fidelity, stability, and robustness. This creates an important practical problem for cybersecurity analysts because conflicting or unstable explanations may reduce confidence in XAI-assisted intrusion detection decisions.

Therefore, there is a need for a comparative evaluation of XAI techniques in Intrusion Detection Systems that considers both ****detection performance and explanation quality****. Such an evaluation should examine conventional IDS performance metrics alongside factors such as explanation fidelity, stability, consistency, and computational efficiency. Addressing this gap can help identify XAI approaches that not only maintain effective intrusion detection performance but also provide reliable, understandable, and practical explanations that can support cybersecurity analysts in making informed security decisions.

2.1.8 Research Scope

This research focuses on the application and comparative evaluation of Explainable Artificial Intelligence (XAI) techniques within ****Network Intrusion Detection Systems (NIDS)****. The study will examine how selected machine learning and deep learning models can detect malicious network activities while also providing understandable explanations for their predictions. The scope is limited to network-based intrusion detection and does not cover host-based intrusion detection systems or other cybersecurity applications such as malware, phishing, or fraud detection.

The experimental study will use the ****UNSW-NB15 dataset**** as the main benchmark dataset. This dataset has been widely used in intrusion detection research and contains normal network traffic together with multiple categories of cyberattacks. It has also been used in recent XAI-based IDS research, making it suitable for comparing both intrusion detection performance and model interpretability.

A selected group of machine learning and deep learning models will be considered, such as ****XGBoost, Random Forest, and a Deep Neural Network (DNN)****. These models represent different levels of complexity and interpretability and therefore provide a suitable basis for examining the relationship between predictive performance and explainability. Recent studies have demonstrated the use of both traditional machine learning and deep learning approaches in explainable IDS environments.

For explainability, the study will mainly focus on ****SHAP and LIME****, as these techniques are among the most commonly applied model-agnostic XAI approaches in intrusion detection research. SHAP will be used to analyze feature contributions at both global and

local levels, while LIME will provide local explanations for selected intrusion predictions. Previous studies have shown the usefulness of both techniques in identifying important features that influence IDS decisions.

The evaluation will cover two main dimensions. First, intrusion detection performance will be assessed using metrics such as **accuracy, precision, recall, F1-score, and false positive rate**. Second, explanation quality will be examined using criteria such as **fidelity, stability, consistency, and computational efficiency**. This combined evaluation is intended to address the limitation of assessing IDS models only through predictive performance while giving insufficient attention to the reliability and practicality of the explanations generated by XAI techniques.

The study will not attempt to develop a completely new XAI algorithm or deploy the proposed approach in a live production network. It will also not focus on privacy-preserving XAI, adversarial attacks against explanation methods, or direct human-user evaluation involving cybersecurity analysts. These areas are recognized as important future research directions but are outside the scope of the present study. The primary aim is to conduct a controlled comparative evaluation that identifies which combination of intrusion detection model and XAI technique provides an appropriate balance between effective threat detection and reliable explanation quality.

2.2) Chapter 2 – Literature Review

2.2.1 Explainable Artificial Intelligence in Cybersecurity

- Artificial Intelligence has become an important component of modern cybersecurity systems due to its ability to process large amounts of security data and identify complex patterns associated with malicious activities. Machine learning and deep learning models have been applied to different cybersecurity tasks, including intrusion detection, malware analysis, phishing detection, authentication, vulnerability assessment, and fraud detection. However, as these models become more complex, understanding how they reach their decisions becomes increasingly difficult. This limitation has motivated growing interest in Explainable Artificial Intelligence (XAI), which attempts to make AI-based security decisions more transparent and understandable.

- Rjoub et al. (2023) conducted a systematic survey of XAI applications in cybersecurity and showed that explainability can improve transparency, trust, and analysts' understanding of AI-based security systems. Their study categorized existing XAI approaches and examined their use in different cybersecurity domains, including intrusion detection. The authors identified techniques such as SHAP and LIME as commonly used approaches for interpreting complex machine learning decisions. However, the study also noted that practical deployment of explainable cybersecurity systems remains limited and that further research is required to improve robustness, reduce false positives, and develop more reliable AI-based security systems.
- Similarly, Capuano et al. (2022) reviewed more than 300 studies related to explainable AI in cybersecurity. Their work provided a broad classification of XAI techniques, including intrinsic and post-hoc approaches, local and global explanations, and model-specific and model-agnostic methods. The study showed that explainability can support several cybersecurity applications, including intrusion detection, malware detection, phishing, botnet detection, fraud detection, and digital forensics. An important observation from this study is that explainability can also represent a security risk. While explanations can increase trust among security analysts, excessive information about the internal behavior of a security model may expose information that an attacker could exploit. Therefore, XAI in cybersecurity should consider not only transparency but also the security of the explanations themselves.
- Zhang et al. (2022) also examined the state of XAI applications in cybersecurity and emphasized the contribution of explainability to trust, accountability, and transparency. Their review considered datasets, XAI techniques, cybersecurity applications, and existing challenges. Similar to Capuano et al. (2022), they identified intrusion detection as an important application of explainable AI. However, their research highlighted several unresolved problems, including privacy, adversarial attacks, legal considerations, and the trade-off between predictive performance and interpretability. Another significant issue identified by the study was the absence of an accepted evaluation system for determining whether one explanation technique is better than another.
- Charmet et al. (2022) expanded this discussion by examining XAI from two related perspectives: the application of XAI to cybersecurity and the security of XAI itself. Their literature survey considered issues such as fairness, integrity, privacy, confidentiality, and robustness in addition to conventional explainability. The study showed that explanations can help cybersecurity operators understand alerts and reduce the difficulties associated with interpreting complex security models. Nevertheless, the authors also observed that much of the existing research was based on benchmark

datasets rather than operational environments. This limits the ability to determine whether current XAI approaches will behave in the same way when deployed in real-world cybersecurity systems.

Together, these early studies established the importance of XAI in cybersecurity but also revealed recurring limitations. Although SHAP, LIME, rule-based explanations, and other approaches can provide greater transparency, research still faces challenges related to evaluation, robustness, privacy, security, computational requirements, and practical deployment.

2.2.2 Development of XAI Evaluation in Cybersecurity

- More recent research has attempted to move beyond describing the benefits of XAI and address how explanations can be evaluated. Ghazal et al. (2024) explored the integration of Large Language Models (LLMs), machine learning, and XAI for cybersecurity applications. Their proposed approach considered techniques such as SHAP, LIME, and Grad-CAM together with language models for applications including intrusion detection, phishing detection, and insider threat monitoring. The study suggested that combining LLMs and XAI could improve transparency and potentially assist security analysts in understanding AI-generated decisions. However, the proposed framework remains mainly conceptual, and the authors identified scalability, privacy, integration complexity, and the absence of standardized datasets and evaluation metrics as unresolved challenges.
- Calzarossa et al. (2025) addressed the evaluation problem more directly by proposing an assessment framework for explainable AI in cybersecurity. Instead of evaluating a model only by its predictive accuracy, their framework considered properties such as explanation complexity and robustness. The approach was demonstrated using a phishing detection dataset containing 10,000 observations and 48 URL and HTML features, with models including Logistic Regression, Random Forest, SVM, KNN, and XGBoost. Their findings showed that useful explanations can be produced with lower computational cost than SHAP while maintaining similar important-feature rankings. However, the framework was evaluated only in phishing detection rather than intrusion detection and used a single dataset. Therefore, its findings cannot automatically be generalized to network-based IDS.

- These studies show an important development in XAI research. Earlier work focused mainly on whether AI decisions could be explained, while newer research increasingly focuses on whether explanations are robust, efficient, and suitable for practical use. However, the evaluation of explanation quality remains less standardized than the evaluation of predictive performance.

2.2.3 Explainable Artificial Intelligence in Intrusion Detection Systems

- The growing use of machine learning and deep learning in Intrusion Detection Systems has made explainability particularly important. Security analysts often need to determine why network activity was classified as malicious before deciding how to respond. Consequently, an IDS that provides high detection accuracy but no understandable reasoning may still be difficult to trust or use effectively.
- Mohale and Obagbuwa (2025) conducted a systematic review specifically focused on integrating XAI into IDS. The researchers initially identified 78 studies, narrowed them to 35 after screening, and ultimately selected 20 studies for detailed analysis. Their review examined SHAP, LIME, decision trees, rule-based methods, and hybrid approaches. The findings showed that SHAP and LIME are among the most widely adopted post-hoc methods in explainable IDS, while decision-tree and rule-based approaches remain attractive because of their natural interpretability.
- However, the review also identified several significant challenges. SHAP can be computationally expensive, which may limit its use in high-speed network environments, while LIME can produce localized explanations that may not generalize effectively across different instances. The authors also highlighted interpretability-accuracy trade-offs, privacy concerns, scalability limitations, and real-time processing requirements. Most importantly, they identified a continuing need for standardized interpretability metrics that allow different XAI-enhanced IDS approaches to be compared consistently.
- While the systematic review provides an overview of the field, Mohale and Obagbuwa (2025) also conducted an experimental study evaluating machine-learning-based IDS with XAI. Using the UNSW-NB15 dataset, which contains more than 2.5 million network records, the researchers evaluated models including Decision Tree, XGBoost, CatBoost, Multilayer Perceptron, Random Forest, Logistic Regression, and Gaussian Naive Bayes. SHAP, LIME, and ELI5 were incorporated to improve transparency and identify features that influenced intrusion predictions.
- XGBoost and CatBoost achieved the highest reported accuracy of approximately 87%, and the XAI analysis identified features such as Source-to-Destination Time-to-Live (sttl) and Destination Service Count (ct_srv_dst) as important indicators of malicious network behavior. The study demonstrates that XAI can provide meaningful information about IDS decisions while maintaining useful predictive performance. Nevertheless, its experimental evaluation was based primarily on UNSW-NB15, and although multiple XAI techniques were applied, the explanation methods were not compared using a comprehensive standardized set of quantitative explanation-quality measures.

- Alabbadi and Bajaber (2025) investigated explainability from a deep-learning perspective. They proposed an IDS architecture using customized 1-D CNN and DNN models together with TabNet. The experiments used seven datasets from the TON_IoT collection, consisting of one network dataset and six IoT/IIoT datasets. SHAP and LIME were applied to explain model predictions and identify the features that contributed to different intrusion classes.
- The results showed strong predictive performance. The CNN achieved a test accuracy of 99.24% on the network dataset, while CNN and DNN models achieved very high accuracy across several IoT datasets. The study therefore demonstrates that explainability can be integrated with high-performing deep learning IDS models. However, the emphasis of the XAI component was mainly on identifying influential features. It did not provide a comprehensive comparative analysis of explanation properties such as fidelity, stability, consistency, or explanation-generation cost. In addition, all seven datasets belonged to the TON_IoT environment, which limits conclusions regarding performance across different network environments.

2.2.4 Reliability and Consistency of XAI Explanations

- An important development in recent research is the recognition that generating an explanation is not sufficient by itself. The explanation must also be reliable. Kumar and Thing investigated this issue by evaluating XAI techniques on state-of-the-art deep-learning-based Network Intrusion Detection Systems, including Kitsune, HorusEye, ENIDrift, and HELAD. Their work applied SHAP and TRUSTEE and examined explanation properties including fidelity, stability, adversarial robustness, and agreement between local and global explanations.
- The study produced an important finding: different XAI methods frequently generated conflicting explanations for the same NIDS models. The researchers proposed a TRUSTEE-SHAP agreement score to examine the level of agreement between global explanations generated by TRUSTEE and local feature explanations generated by SHAP. Their experiments also showed that the explainability of an IDS depends partly on the underlying model. For example, some evaluated models produced useful TRUSTEE explanations, while others resulted in very low or negative fidelity values.
- The researchers further evaluated stability and found differences between XAI techniques and NIDS architecture. These findings raise an important issue for cybersecurity practitioners. If two accepted XAI methods produce significantly different explanations for the same intrusion detector, simply presenting an explanation to an analyst does not guarantee that the explanation accurately represents the behavior of the underlying model. Therefore, explanation quality should be examined independently rather than being assumed from the use of an XAI technique.

2.2.5 Synthesis of Previous Studies and Research Gap

- The reviewed literature demonstrates a clear progression in research on explainable AI for cybersecurity and intrusion detection. Early studies such as Capuano et al. (2022), Zhang et al. (2022), Charmet et al. (2022), and Rjoub et al. (2023) established the need for transparency in AI-based cybersecurity systems. These studies consistently identified the black-box nature of complex ML and DL models as a limitation and presented XAI techniques such as SHAP and LIME as potential solutions. At the same time, they highlighted unresolved concerns related to privacy, robustness, security, computational requirements, and standardized evaluation.
- Recent studies have moved toward more specialized and experimental IDS applications. Mohale and Obagbuwa (2025) demonstrated how SHAP, LIME, and ELI5 can explain several ML-based IDS models, while Alabbadi and Bajaber (2025) integrated SHAP and LIME with CNN, DNN, and TabNet architectures. These studies provide evidence that XAI can improve the transparency of intrusion detection without necessarily preventing models from achieving useful predictive performance.
- However, a common limitation remains. Traditional IDS performance metrics, including accuracy, precision, recall, F1-score, and false positive rate, are well established, whereas the quality of XAI explanations is evaluated less consistently. The systematic review by Mohale and Obagbuwa identifies standardized interpretability metrics as an important research need, while the experimental findings of Kumar and Thing show that different XAI methods can generate conflicting explanations for the same detection system.
- Therefore, the main research gap is no longer simply the absence of explainability in intrusion detection systems. XAI techniques are already being applied to both machine-learning and deep-learning IDS. Instead, there is a need for a more systematic **comparative evaluation of XAI techniques that considers explanation quality together with intrusion detection performance**. In particular, explanation fidelity, stability, consistency, and computational efficiency should be considered alongside accuracy, precision, recall, F1-score, and false positive rate.
- Addressing this gap can provide a clearer understanding of which combinations of intrusion detection models and XAI techniques offer an appropriate balance between detection effectiveness and trustworthy explanation. This forms the basis of the present research, which focuses on comparing selected XAI approaches in Network Intrusion Detection Systems rather than assuming that the availability of an explanation automatically makes an IDS transparent or trustworthy

2.3) Chapter 3 – Research Methodology

2.3.1 Research Design

This study will adopt a **quantitative experimental and comparative research design** to evaluate Explainable Artificial Intelligence (XAI) techniques in Network Intrusion Detection Systems (NIDS). The research will compare selected machine learning and deep learning models in terms of both intrusion detection performance and the quality of the explanations generated by XAI techniques.

The methodology consists of six main stages: dataset preparation, data preprocessing, development and training of intrusion detection models, application of XAI techniques, performance evaluation, and comparative analysis. This approach is intended to determine whether a highly accurate intrusion detection model can also provide explanations that are stable, faithful, consistent, and computationally practical.

2.3.2 Dataset

The **UNSW-NB15 dataset** will be used as the primary dataset for the experimental study. It is a widely used benchmark dataset for intrusion detection research and contains normal network activities together with several categories of malicious network traffic. The complete dataset contains more than 2.5 million network records and approximately 49 network-related features. It includes attack categories such as Fuzzers, Analysis, Backdoors, Denial of Service, Exploits, Generic attacks, Reconnaissance, Shellcode, and Worms.

The selection of UNSW-NB15 is also supported by recent explainable IDS research. Mohale and Obagbuwa (2025) used the dataset to evaluate several machine learning models with SHAP, LIME, and ELI5, demonstrating its suitability for studying both intrusion detection performance and interpretability.

The main experiment will initially consider **binary classification**, where network traffic is classified as either normal or malicious. Where computational resources permit, a secondary multiclass experiment may be conducted to examine explanations associated with individual attack categories.

2.3.3 Data Preprocessing

Before model training, the dataset will undergo preprocessing to ensure that the data are suitable for machine learning and deep learning analysis. The proposed preprocessing process will include checking for missing or duplicated records, encoding categorical features, scaling numerical features, and examining class distribution.

Categorical attributes such as protocol and service information will be transformed into numerical representations. Numerical features will be normalized or standardized to prevent variables with large numerical ranges from dominating the learning process. Similar preprocessing procedures, including encoding, normalization, feature selection, and treatment of class imbalance, have been used in recent ML-based explainable IDS research.

If significant class imbalance is identified, a balancing technique such as SMOTE will be applied only to the training data to reduce bias toward the majority class while avoiding information leakage into the test set.

2.3.4 Intrusion Detection Models

Three representative detection models will be evaluated:

Random Forest (RF): selected as an ensemble machine learning model that can provide strong classification performance and a useful baseline for feature importance.

XGBoost: selected because gradient-boosting approaches have demonstrated strong performance in intrusion detection and have been used successfully with SHAP and LIME.

Deep Neural Network (DNN/MLP): selected to represent a more complex black-box model in which the need for external explanation techniques becomes particularly important.

The choice of different model types allows the research to investigate whether the effectiveness of an XAI technique changes according to the underlying detection model. Previous research using UNSW-NB15 has evaluated models including XGBoost, Random Forest, and MLP with explainability approaches, while other recent IDS work has demonstrated the use of DNN architectures together with SHAP and LIME.

The models will be trained using the same preprocessing strategy and comparable training and testing partitions to ensure a fair evaluation.

2.3.5 Application of XAI Techniques

After the intrusion detection models have been trained, **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)** will be applied as post-hoc explanation techniques.

SHAP will be used to estimate the contribution of individual network features to model predictions and to generate both global and local information about model behavior. **LIME** will be used primarily to produce local explanations for selected network traffic instances by approximating the behavior of the original model around each prediction. Recent XAI-based IDS research has successfully applied both techniques after model training without modifying the predictive process of the original models.

For each detection model, the same selected test instances will be explained using both **SHAP** and **LIME**. This will allow a controlled comparison of the important features identified by the two methods and help determine whether different explanation techniques provide similar or conflicting interpretations.

2.3.6 Evaluation of Intrusion Detection Performance

The predictive performance of the IDS models will be evaluated using standard classification metrics:

Accuracy: will measure the overall proportion of correctly classified network records.

Precision: will measure how many records predicted as malicious are actually malicious.

Recall: will measure the ability of the IDS to identify actual malicious traffic.

F1-score: will provide a balanced measure between precision and recall.

False Positive Rate (FPR): will measure how frequently legitimate network activity is incorrectly classified as malicious.

These metrics are commonly used in intrusion detection research and allow the models to be compared independently of their explanation mechanisms. Recent XAI-IDS studies also evaluate models using accuracy, precision, recall, F1-score, false-positive rate, and related measures.

2.3.7 Evaluation of Explanation Quality

Unlike studies that evaluate only classification accuracy, this research will also assess the quality of the explanations produced by **SHAP** and **LIME**. Four main dimensions will be considered.

Fidelity/Faithfulness: will examine whether the features identified as important by the explanation method genuinely influence the original model's prediction. This can be tested by modifying or removing highly ranked features and observing the resulting change in model output.

Stability: will examine whether an XAI technique produces similar explanations when the same or very similar instances are evaluated repeatedly. The overlap between the most important features across repeated explanation runs can be calculated. Stability has previously been used as an important security-focused criterion for evaluating explanations in NIDS.

Consistency or Agreement: will compare the important features identified by SHAP and LIME for the same detection model and network instances. Feature ranking overlap or rank similarity can be used to determine the extent to which the two techniques agree. The importance of examining disagreement is supported by recent findings showing that different XAI techniques can generate conflicting explanations for the same NIDS model.

Computational Efficiency: will be evaluated by recording the time required to generate explanations. This measurement is important because computationally expensive XAI techniques may not be suitable for intrusion detection environments where alerts need to be processed quickly. Reviews of XAI-IDS research identify computational overhead and real-time applicability as continuing challenges.

2.3.8 Comparative Analysis

After completing the experiments, the results will be organized into two main comparison levels. The first will compare the intrusion detection models using predictive performance metrics. The second will compare **SHAP** and **LIME** using explanation-quality measures.

The final analysis will examine the relationship between these two dimensions rather than selecting the best model based only on accuracy. For example, a model with slightly lower detection performance may be preferable if it provides substantially more stable and faithful explanations, while a highly accurate model may be less useful if its explanations are inconsistent or computationally expensive.

The objective is therefore to identify the **model-XAI** combination that provides the most appropriate balance between intrusion detection effectiveness and explanation quality.

2.3.9 Implementation Tools

The experiments will be implemented primarily using Python. Libraries such as Pandas and NumPy will be used for data preparation and analysis, while Scikit-learn and XGBoost will be used for machine learning models. TensorFlow/Keras may be used for the neural-network model. The **SHAP** and **LIME** Python libraries will be used to generate model explanations.

Experimental results will be presented through comparative tables, confusion matrices, feature-importance plots, and explanation visualizations. The same computing environment and test data will be maintained wherever possible to ensure that differences between models and XAI techniques result from the evaluated approaches rather than changes in experimental conditions.

2.3.10 Proposed XAI-IDS Evaluation Framework

This research proposes an XAI-IDS Evaluation Framework designed to evaluate Intrusion Detection Systems from two complementary perspectives: intrusion detection performance and explanation quality. The framework does not aim to introduce a new machine learning or XAI algorithm. Instead, it provides a structured experimental process for determining how different intrusion detection models interact with selected explainability techniques.

The proposed framework consists of five main stages:

Stage1 Data Preparation and Preprocessing

The UNSW-NB15 dataset will first be prepared for experimentation. This stage will include data cleaning, categorical feature encoding, feature scaling, class distribution analysis, and the separation of training and testing data. If class imbalance is significant, an appropriate balancing technique will be applied only to the training data.

Stage 2 Intrusion Detection Model Development

Selected intrusion detection models, including Random Forest, XGBoost, and a Deep Neural Network (DNN/MLP), will be trained using the processed dataset. The models were selected because they represent different levels of complexity and interpretability, allowing the study to examine whether model architecture affects both detection performance and explanation quality.

Stage 3 XAI Integration

After the IDS models have been trained, SHAP and LIME will be applied as post-hoc explainability techniques. Both methods will be used on the same selected test instances wherever possible to ensure a consistent basis for comparison. SHAP will provide feature-attribution information at global and local levels, while LIME will generate local explanations for individual intrusion predictions.

Stage 4 Dual Evaluation

The framework will evaluate the models in two dimensions. The first dimension will measure intrusion detection performance using accuracy, precision, recall, F1-score, and false positive rate. The second dimension will evaluate explanation quality using fidelity, stability, consistency, and computational efficiency.

Stage 5 Comparative Analysis and Selection

The final stage will compare the different IDS–XAI combinations. Rather than selecting the best approach only according to classification accuracy, the framework will consider both

detection effectiveness and the reliability of the explanations. The purpose is to identify the combination that achieves the most appropriate balance between accurate intrusion detection and trustworthy explainability.

The overall framework can therefore be represented as:

UNSW-NB15 Dataset → Data Preprocessing → IDS Models → SHAP/LIME → Detection Performance Evaluation → Explanation Quality Evaluation → Comparative Analysis

This framework is motivated by recent research showing that XAI methods can improve transparency in IDS while also presenting challenges related to explanation stability, computational cost, consistency, and disagreement between different explanation techniques (Mohale & Obagbuwa, 2025; Kumar & Thing, 2025).

2.3.11 Framework Validation

The proposed framework will be validated experimentally using the UNSW-NB15 benchmark dataset. Validation will focus on determining whether the framework can consistently evaluate both the predictive capability of intrusion detection models and the quality of the explanations generated by XAI techniques.

The first level of validation will examine IDS predictive performance. Each selected model will be trained and tested using the same dataset preparation strategy and comparable data partitions. Performance will be evaluated using accuracy, precision, recall, F1-score, and false positive rate. These measurements will determine whether the selected IDS models can reliably distinguish between normal and malicious network traffic.

The second level of validation will examine XAI explanation quality. SHAP and LIME explanations will be generated for selected test instances. Fidelity will be evaluated by examining whether highly ranked features genuinely influence model predictions. Stability will be assessed by determining whether repeated explanations identify similar important features. Consistency will compare the degree of agreement between SHAP and LIME

feature rankings, while computational efficiency will be measured using the time required to generate explanations.

The third level will involve comparative validation. The results obtained from different model–XAI combinations will be compared to determine whether high detection performance corresponds to high explanation quality. This is important because a model may achieve strong predictive performance while producing unstable or inconsistent explanations.

Where possible, experiments will be repeated under the same conditions to reduce the effect of random variation. The same preprocessing procedures, test data, evaluation metrics, and computing environment will also be maintained to improve the reliability of the comparison.

The framework will be considered effective if it can clearly distinguish the strengths and limitations of different IDS–XAI combinations and provide measurable evidence regarding the trade-off between intrusion detection performance and explanation quality.

2.3.12 Ethical Considerations

This research will use a publicly available cybersecurity benchmark dataset and will not directly involve human participants or collect personally identifiable information. Therefore, the study presents relatively limited human-subject ethical risks. Nevertheless, several ethical considerations will be addressed throughout the research process.

First, the UNSW-NB15 dataset will be used only for academic and research purposes and in accordance with its permitted usage conditions. The dataset will not be modified in a way that misrepresents the original data or artificially improves experimental results.

Second, the research will maintain academic integrity and research transparency. All datasets, algorithms, XAI techniques, and previous studies used in the research will be properly acknowledged and cited. Experimental procedures and evaluation criteria will be

documented clearly so that the results can be understood and, where possible, reproduced.

Third, the study will report both positive and negative experimental findings. Results will not be selectively removed or altered to favour a particular machine learning model or XAI technique. This is particularly important when comparing explanation techniques because differences in stability, fidelity, or computational efficiency may reveal weaknesses as well as strengths.

Fourth, cybersecurity-related findings will be presented responsibly. The research focuses on defensive intrusion detection and model explainability rather than the development of offensive techniques. Information that could unnecessarily facilitate malicious exploitation of an IDS will not be emphasized.

Finally, explainability itself creates ethical and security considerations. XAI techniques may expose information about the internal behaviour or important features of an intrusion detection model. While such transparency can improve analyst trust, excessive disclosure could potentially provide useful information to an attacker. Therefore, the study will consider the balance between transparency and security when interpreting the generated explanations.

Overall, the research will aim to maintain responsible data use, accurate reporting, proper citation, reproducibility, and appropriate handling of security-sensitive findings throughout the experimental process.

2.4) Chapter 4 – Research Limitations

Although this research is designed to provide a structured comparison of Explainable Artificial Intelligence techniques in Intrusion Detection Systems, several limitations should be acknowledged.

- First, the experimental evaluation will primarily use the UNSW-NB15 dataset. Although this dataset is widely used in intrusion detection research and contains multiple categories of network attacks, results obtained from a single benchmark dataset may

not fully represent the diversity of modern operational networks. Network characteristics, attack patterns, and traffic distributions may differ across organizations and environments. Previous XAI-based IDS studies have similarly identified reliance on a limited number of benchmark datasets as a constraint on generalizability.

- Second, the study will evaluate a limited number of intrusion detection models, mainly Random Forest, XGBoost, and a Deep Neural Network (DNN/MLP). These models were selected to represent different levels of model complexity, but they do not cover all available machine learning and deep learning approaches. Therefore, the conclusions of the study may not automatically apply to other architectures such as CNNs, RNNs, autoencoders, transformers, or hybrid intrusion detection models.
- Third, the explainability analysis will mainly focus on SHAP and LIME. These techniques are widely used in XAI research, but other explanation approaches, such as TRUSTEE, counterfactual explanations, rule-based methods, and inherently interpretable models, will not be included in the main experimental comparison. Previous studies have demonstrated that different XAI methods may generate substantially different explanations for the same intrusion detection model. Therefore, restricting the experiment to SHAP and LIME means that the study cannot provide conclusions about the entire range of available XAI techniques.
- Fourth, the research will be conducted in a controlled experimental environment using an existing benchmark dataset rather than a live production network. Consequently, the study will not fully reproduce operational conditions such as continuously changing network traffic, concept drift, zero-day attacks, resource limitations, or extremely high-speed data streams. Real-time explainability remains a significant challenge because some XAI techniques can introduce considerable computational overhead.
- Fifth, the evaluation of explanation quality will concentrate on measurable properties such as fidelity, stability, consistency, and computational efficiency. These metrics provide a technical assessment of explanations, but they do not directly determine whether cybersecurity analysts find the explanations understandable, useful, or actionable. A complete human-centered evaluation would require participation from security analysts through interviews, questionnaires, or usability experiments, which is outside the scope of this research.
- Sixth, computational resources may influence the size of the dataset, the complexity of the DNN model, the number of repeated experiments, and the number of instances that can be explained using computationally demanding techniques such as SHAP.

Previous literature has identified computational cost and scalability as continuing limitations of XAI in IDS, particularly when explanations must be generated rapidly.

- Finally, the study will not provide a detailed experimental investigation of adversarial attacks against XAI, privacy-preserving explainability, or information leakage caused by explanations. These issues are important because explanations may reveal characteristics of the underlying detection model, and XAI techniques may themselves be vulnerable to manipulation. They are recognized as important directions for future research but are beyond the present study's objectives.

Despite these limitations, the proposed research can provide a controlled and measurable comparison of selected IDS–XAI combinations and establish a foundation for future studies involving additional datasets, models, explanation techniques, real-time environments, and human cybersecurity analysts.

2.5) Chapter 5 – Expected Outcomes

The proposed research is expected to provide a structured comparison of selected Explainable Artificial Intelligence techniques within Network Intrusion Detection Systems. The outcomes will focus on both the effectiveness of intrusion detection and the quality of the explanations produced by the selected XAI methods.

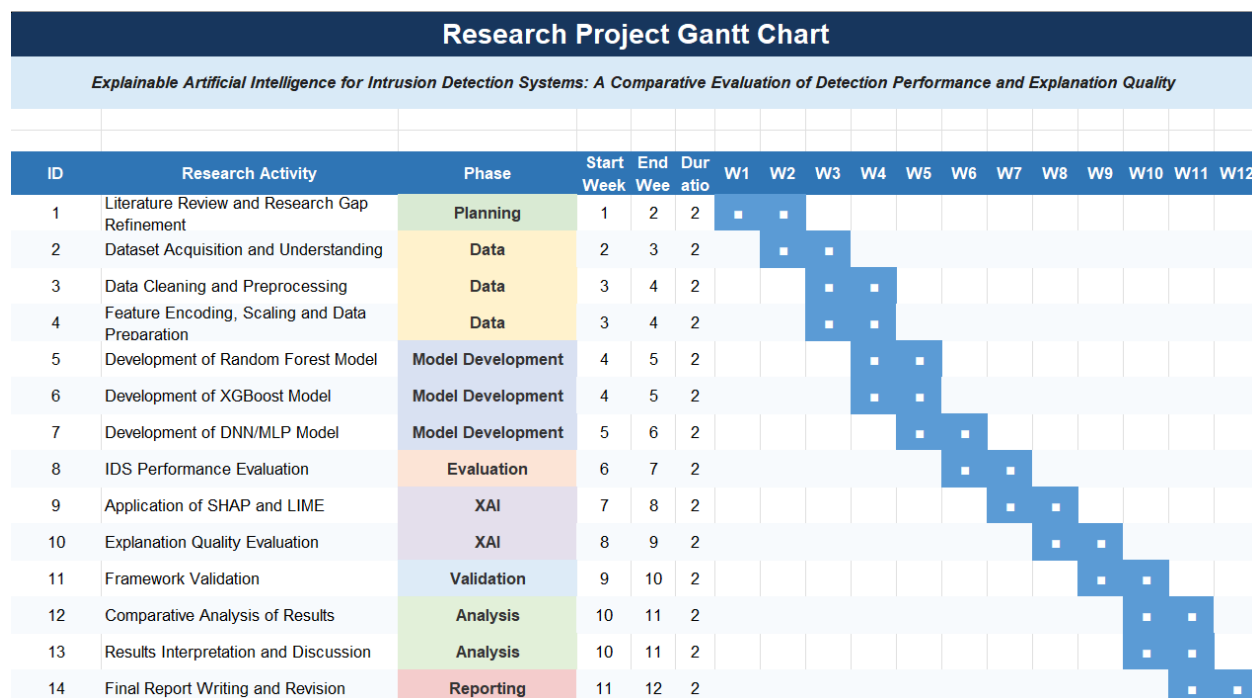
- First, the study is expected to identify the relative performance of the selected intrusion detection models, including Random Forest, XGBoost, and DNN/MLP, using standard evaluation metrics such as accuracy, precision, recall, F1-score, and false positive rate. This will provide a clear understanding of which model offers the most effective detection performance on the UNSW-NB15 dataset. Recent research has shown that different ML models can achieve noticeably different intrusion-detection results, making comparative evaluation necessary.
- Second, the research is expected to demonstrate how SHAP and LIME can improve the transparency of IDS predictions by identifying the network features that contribute

most strongly to the classification of normal and malicious traffic. This should provide a clearer understanding of the decision-making process of complex intrusion detection models and help determine whether the explanations correspond to meaningful network characteristics. Previous studies have demonstrated the usefulness of SHAP and LIME for explaining IDS predictions and identifying influential network features.

- Third, the study is expected to produce a comparative assessment of explanation quality using criteria such as fidelity, stability, consistency, and computational efficiency. This will allow the research to determine whether an XAI method produces explanations that are not only understandable but also sufficiently reliable and repeatable. This outcome is particularly important because recent research has shown that different explanation techniques may generate conflicting interpretations for the same intrusion detection model.
- Fourth, the proposed XAI-IDS Evaluation Framework is expected to reveal the relationship between detection performance and explainability. The research may show that the model with the highest classification accuracy is not necessarily the model with the most stable, faithful, or computationally efficient explanations. The framework is therefore expected to support a more balanced evaluation in which predictive effectiveness and explanation quality are considered together rather than independently.
- Fifth, the study is expected to identify the most suitable IDS–XAI combination among the evaluated alternatives. The final comparison should provide evidence regarding which model and explanation technique offer the most appropriate balance between intrusion detection effectiveness, explanation reliability, and computational practicality.
- Finally, the research is expected to provide recommendations for the future development of explainable intrusion detection systems. These recommendations may address the need for standardized XAI evaluation metrics, more efficient explanation techniques, testing across multiple datasets, real-time validation, and human-centered evaluation involving cybersecurity analysts. These areas have also been identified as important research directions in recent XAI-IDS literature.

Overall, the expected outcome of this research is not only to determine which intrusion detection model performs best, but also to establish whether the explanations supporting its predictions are sufficiently accurate, stable, consistent, and practical for use in cybersecurity environments.

2.5) Chapter 6 – Gantt chart



Research Milestones		
Milestone	Description	Target Week
M1	finalized	2
M2	UNSW-NB15 dataset prepared	4
M3	IDS models developed and trained	6
M4	completed	7
M5	SHAP and LIME explanations generated	8
M6	Explanation quality evaluation completed	9
M7	XAI-IDS framework validated	10
M8	Comparative analysis completed	11
M9	Final research report completed	12

3. Appendix

3.1.1 Table 1 Synthesis Matrix for Literature Review.

3.1.2 Tables Critical Analysis.

Noted : all appendixes (will be sent via email upon request)

3.2) Attachments.

3.2.1 Cybersecurity Revolution via Large Language Models and Explainable AI.

3.2.2 A Survey on Explainable Artificial Intelligence for Cybersecurity.

3.2.3 Explainable Artificial Intelligence in Cybersecurity: A Survey.

3.2.4 Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research.

3.2.5 An assessment framework for explainable AI with applications to cybersecurity.

3.2.6 Explainable artificial intelligence for cybersecurity: a literature survey.

3.2.7 An Excel file containing a Gantt chart

Noted : all attachments (will be sent via email upon request)

3.2) References.

Alabbadi, A., & Bajaber, F. (2025). An intrusion detection system over the IoT data streams using eXplainable Artificial Intelligence (XAI). *Sensors*, 25(3), 847.

<https://doi.org/10.3390/s25030847>

Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review*, 58, 150.

<https://doi.org/10.1007/s10462-025-11141-w>

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10, 93575–93600.

<https://doi.org/10.1109/ACCESS.2022.3204171>

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: A literature survey. *Annals of Telecommunications*, 77(11–12), 789–812.

<https://doi.org/10.1007/s12243-022-00926-7>

Ghazal, T. M., Janjua, J. I., Abushiba, W., Ahmad, M., Ihsan, A., & Al-Dmour, N. A. (2024). Cybersecurity revolution via large language models and explainable AI. In *2024 IEEE Conference Proceedings*. IEEE.

Kumar, A., & Thing, V. L. L. (2025). *Evaluating the explainability of deep learning-based network intrusion detection systems*. arXiv preprint arXiv:2408.14040.

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, 8, 1526221.

<https://doi.org/10.3389/frai.2025.1526221>

Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. *Frontiers in Computer Science*, 7, 1520741. <https://doi.org/10.3389/fcomp.2025.1520741>

Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrouk, H., & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management*, 20(4), 5115–5140.

<https://doi.org/10.1109/TNSM.2023.3282740>

Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10, 93104–93139. <https://doi.org/10.1109/ACCESS.2022.3204051>