



Original Article

Explainable Artificial Intelligence for Cybersecurity Risk Assessment and Interpretable Detection of Emerging Network Attacks in Complex Information Systems

Raja Ganesan
Independent Researcher, USA.

Received On: 27/05/2026

Revised On: 24/06/2026

Accepted On: 04/07/2026

Published On: 15/07/2026

Abstract - The growing complexity of complex interconnected information systems, cloud infrastructures and heterogeneous network environments has led to the need for new challenges in cyber security risk assessment and the detection of new network attacks. Traditional rule-based and machine learning security mechanisms are effective in detecting threats but can have trouble identifying new attack patterns, false-positive rates, and lack of transparency in decision making. This study tackles this limitedness by proposing an Explainable Artificial Intelligence (XAI) cybersecurity framework that combines cybersecurity risk assessment and interpretable network attack detection. The framework consists of machine learning based threat identification, risk scoring, anomaly detection, and explainability features to detect suspicious network traffic and give human-readable evidence on why a security decision has been made. The proposed framework is assessed with the representative cybersecurity network datasets and comparative experiments between traditional ML and XAI. Accuracy, precision, recall, F1-score, false-positive rate, latency of detection and measures related to explainability are used to evaluate predictive effectiveness and explainability. The experimental results show that the proposed approach can enhance the attack detection reliability and meaningful explanations for model predictions and cyber security risk factors. The study proposes an integrated and interpretable framework for cybersecurity risk assessment and emerging attack detection, which could assist security analysts with understanding the model decisions, prioritizing threats, and making more transparent decisions in complex information systems from the security perspective.

Keywords - Explainable Artificial Intelligence, Cybersecurity Risk Assessment, Network Attack Detection, Cyber Threats, Detection Systems, Information Systems Security, XAI.

1. Introduction

1.1. Background, Motivation, and Cybersecurity Challenges

The growing adoption of Cloud Computing, Internet of Things (IoT), Edge Computing, Distributed Applications, and Interconnected Enterprise Platforms has resulted in today's Information Systems being highly Heterogeneous,

rapidly changing, and dynamic environments. [1] While this connectivity can help to achieve scalability and operational efficiency, it also increases the attack surface and creates intricate dependencies between applications, networks, devices, users and data. Cybersecurity threats have then changed to be more sophisticated and adaptive, such as distributed denial-of-service attacks, credential-based attacks, ransomware, malware, botnet activities, insider threats and previously unseen attack patterns. Conventional security solutions that rely mainly on pre-defined signatures and static rules might not be adequate to detect behavioral anomalies and new threats. Furthermore, network data is high dimensional and noisy, and constantly generated, making risk assessment a difficult process in cybersecurity, and false positives can overload security analysts and false negatives can allow attacks to proceed without detection. These challenges result in a demand for intelligent cybersecurity mechanisms, which can assess abnormal behavior, measure related risk, and deliver a timely and reliable security decision.

1.2. Limitations of Conventional Detection and Need for Explainable AI

Machine learning (ML) and deep learning (DL) have contributed to the advancement of AI in cybersecurity, allowing models to identify complex patterns within network and security data. Very complex models, however, are often treated as "black-boxes" and offer little insight into the reasoning behind the classification of a given event as malicious or the reason for a given level of risk. [2] The opacity can hinder the security analyst's ability to validate alerts, investigate incidents, prioritize threats and believe in automated recommendations. Explainable Artificial Intelligence (XAI) overcomes this by providing human interpretable explanations using methods like SHAP, LIME, Feature-Importance Analysis, Counterfactual Explanations, etc. However, many of the current methods treat attack detection and cybersecurity risk assessment as independent tasks, while neglecting to explain the models. Thus, there is a need for an integrated solution: to detect emerging attacks on the network accurately, as well as to assess the risk accurately and provide meaningful explanations of the individual decisions made for security.

1.3. Research Problem, Objectives, and Research Questions

The main research problem is that of simultaneously achieving effective detection of emerging network attacks and transparent cybersecurity risk assessment in complex information systems. While the existing AI-based methods can yield high predictive accuracy, they might lack interpretability, and simpler interpretable models may fail to capture the complexity and evolving nature of attacks. The purpose of this study is to create a cybersecurity framework based on XAI that will incorporate network attack detection, cybersecurity risk assessment, and interpretable decision making. In particular, the study is set to create an AI-driven attack detection system, integrate risk assessment as a component of the detection process, determine the key features of the network and behaviour that most affect the predictions of security, assess the performance of the predictions and interpretability of the system, and compare with traditional machine learning and non-explainable models. What is the model's accuracy in detecting new network attacks using XAI? What are the most important characteristics that are most relevant for cybersecurity risk predictions? How much do explanations help in making AI-based security decisions more transparent? and What is the degree of successful compromise between detection performance and interpretation?

2. Background and Related Work

2.1. Cybersecurity Risk Assessment, Emerging Threats, and AI-Based Attack Detection

Cybersecurity risk assessment involves a systematic process of determining threats, vulnerabilities, potential impact, and likelihood of attack to information assets. [3] Complex information systems pose particular challenges for risk assessment, as they have diverse infrastructures, dynamic networks, interdependent services, and continually evolving attack surfaces. New Malware threats like zero-day attack, polymorphic malware, botnets, credential-based attacks, ransomware, and advanced persistent threat can manifest in dramatically different ways from what has been previously seen, and signature-based detection methods are not enough. Therefore, machine learning (ML) and deep learning (DL) have been increasingly used for network intrusion detection, anomaly detection, and attack classification. Random Forest, Support Vector Machine and XGBoost are examples of supervised learning algorithms that can detect discriminative patterns in labeled traffic, while DL architectures like CNN, LSTM, autoencoders, and transformer-based models can identify complex spatial, temporal and behavioral patterns. Yet these methods are still data quality and class-balance dependent, as well as concept drifting and under-representing new attacks.

2.2. Explainable AI and Interpretable Cybersecurity Decision Making

One such research area of growing significance is Explainable Artificial Intelligence (XAI) which aims to tackle the opacity of intricate ML and DL models in cybersecurity. [4] XAI is not just about prediction results but also understanding what drives the model's decisions, which can be conveyed in a manner comprehensible to security

analysts. Global interpretability is about something that is common to the model, identifying consistently influential features that explain overall model behaviour, while local interpretability is about something that is unique to the network event or security incident, identifying features that are unique that explain individual predictions. Methods like SHAP, LIME, feature importance analysis and counterfactual explanations can be used to uncover network and behavior characteristics that are associated with attack classification or high-risk predictions. Previous studies have shown that interpretable detection can be used to aid in the validation of alerts, investigation of threats, prioritization of risks, and security decision making with human factors. However, the quality of the explanation will be influenced by the fidelity, consistency, stability, and comprehensibility of the explanation, as well as various techniques of XAI may provide different explanations for the same prediction. So, explainability should not be considered a layer of visualization, but rather an element of cybersecurity intelligence.

2.3. Existing Approaches, Research Gaps, and Motivation for the Proposed Framework

The current research indicates significant advancements in AI-supported cybersecurity risk assessment, ML/DL intrusion detection, XAI-driven security analysis, and detection of emerging attacks, but the aforementioned research areas often are studied separately. [5] Many intrusion detection methods focus on classification accuracy but fail to include any explicit cyber-security risk assessment, or risk assessment models can be very rule-dependent or be based on threats from past incidents and may not be able to adapt to new attacks. Likewise, research on XAI for intrusion detection tends to emphasize providing explanations for individual predictions rather than linking these explanations to both the severity of the risk and the probability of an attack, vulnerability, and potential impact. Detection of emerging and zero-day attacks are very difficult since there are not many labeled examples and new attack behaviors can be similar to legitimate anomalies.

The limitations indicate the need for further research to create a unified system capable of attack detection via AI, evolving cybersecurity risks, threat detection of new threats, and explanations in human form. This gap inspires the proposed XAI framework for Cybersecurity that aims to integrate the interpretability and predictive capability to deliver cybersecurity decision support by providing transparent, risk-aware and actionable information in complex information systems.

3. Research Methodology

3.1. Research Design and Overall Workflow

The purpose of this study is to explore the effectiveness of a cybersecurity risk assessment and interpretable detection of emerging network attacks through an Explainable Artificial Intelligence (XAI) framework with a quantitative, experimental research design. [6] The methodology consists of an end-to-end workflow that starts with: selection and collection of datasets, data processing, feature engineering

and feature selection, development of models, classification of attacks, estimation of cybersecurity risk, analysis of explainability, and evaluation of models. First, network traffic and security-event information are structured into a structured analytical representation, followed by training machine learning and deep learning models to differentiate between legitimate and malicious traffic. These predictions

are then subjected to XAI analysis to uncover key factors to the attack classification and risk estimates of individual cases. Last, the proposed approach is compared with selected baseline models to assess the potential for integration of explainability for meaningful improvements in cybersecurity decision support: predictive, operational, robustness, and interpretability metrics are used.

Table 1. Machine Learning and Deep Learning Models

Model	Learning Type	Main Function	Strength	Limitation
Random Forest	Supervised	Classification	Robust	Less effective for temporal patterns
XGBoost	Supervised	Classification	High predictive performance	Parameter sensitivity
SVM	Supervised	Classification	High-dimensional data	Computational cost
LSTM	Deep Learning	Sequential detection	Temporal learning	Training complexity
Autoencoder	Unsupervised	Anomaly detection	Unseen patterns	Reconstruction threshold
Proposed XAI Model	Hybrid	Detection + Risk + Explanation	Integrated analysis	Computational overhead

3.2. Dataset Selection, Data Collection, and Preprocessing

The study employs representative cybersecurity network datasets, including normal traffic and various types of malicious activity, with a special focus on datasets with a wide variety of attack patterns, realistic network characteristics, and enough samples for supervised and anomaly-based learning. [7] Attack diversity, availability of features, distribution of classes, temporal features, data quality, and the suitability of the data for evaluation of emerging attack detection are some of the factors taken into account in the selection of the dataset. Data pre-processing steps performed on the collected data are duplicate elimination, handling of missing values, inconsistent record correction, category encoding, numerical data normalization/standardization and class distribution analysis. The highly correlated or irrelevant attributes are eliminated when suitable, to eliminate redundancy and reduce the computational complexity. The dataset is subsequently split into training, validation and independent testing sets, ensuring the correct distribution of classes and information leakage is prevented. In the evaluation of emerging attacks, the model can be trained with a subset of the attack categories, then tested to see how it performs when evaluating a subset of the remaining categories that the model has not seen before or rarely.

3.3. Feature Engineering, Model Development, and Cybersecurity Risk Assessment

When the raw network observations are transformed to meaningful indicators for detecting attacks and analyzing risks, it is called feature engineering. These features can include packet and flow parameters, source and destination data, protocol behavior, connection length, packet or byte count, errors, authentication-related features, temporal and behavioural patterns. [8] Feature-selection methods are then used to select a subset of the most informative features and to lower the dimensionality of the data and boost the

computational efficiency. Processed features are used to train multiple baseline and proposed models including: Random Forest, XGBoost, Support Vector Machine, neural networks, LSTM, or other appropriate models. The attack-classification component defines if an observation is normal or malicious traffic and if applicable, categorizes the attack type. A cybersecurity risk score is then calculated by aggregating the predicted threat likelihood with the relevant indicators of vulnerability, asset criticality and potential impact. This allows the framework to take a step beyond the detection of binary attacks to a risk-based classification of an attack, such as low-risk, medium-risk, high-risk, or critical-risk attack.

3.4. Explainability Models and Evaluation Metrics

The interpretability component is used for explaining individual predictions and the behavior of the model in general using XAI methods. Feature-importance and counterfactual techniques are complementary to SHAP and LIME, which can be adapted if relevant for feature-importance and counterfactual explanations for selected network events, respectively. Global explanations identify the network properties that impact model decisions the most overall, while local explanations identify specific properties that lead to particular cases of attack classification or risk scores. Evaluation of the models is based on accuracy, precision, recall, F1-score, ROC-AUC, false-positive rate, false-negative rate, Matthews correlation coefficient, and detection latency, along with other metrics as needed, such as throughput and computational resource utilization. The criteria used to assess explainability include fidelity, stability, consistency, explanation parsimony/comprehensibility. Comparative experiments are carried out using the conventional detection models to see whether the proposed framework can maintain good predictive performance while providing reliable and actionable explanation to cybersecurity analysts.

4. Proposed Explainable AI-Based Cybersecurity Framework

4.1. Framework Overview and System Architecture

The proposed concept is based on the integration of the following concepts within a common architecture of complex information system: Cyber Security Risk Assessment; Cyber Security Risk Detection based on Artificial Intelligence; Emerging Threat Identification; Explainable Artificial Intelligence. [9] The overall structure of the framework is based on five layers: data acquisition, data processing and feature engineering, risk assessment, attack detection, emerging-attack identification,

explainability, threat intelligence and security decision-support. Network traffic, security logs, endpoint events, authentication logs and other relevant security observations are gathered and converted to common representations for analysis. The processed information is concurrently fed into a cybersecurity risk assessment and a malicious activity detection system, and also used in explaining the evidence for each important prediction, appropriately. The architecture is meant to isolate data processing, intelligence, explanation and response functions to allow individual components to be updated without redesigning the overall security system.

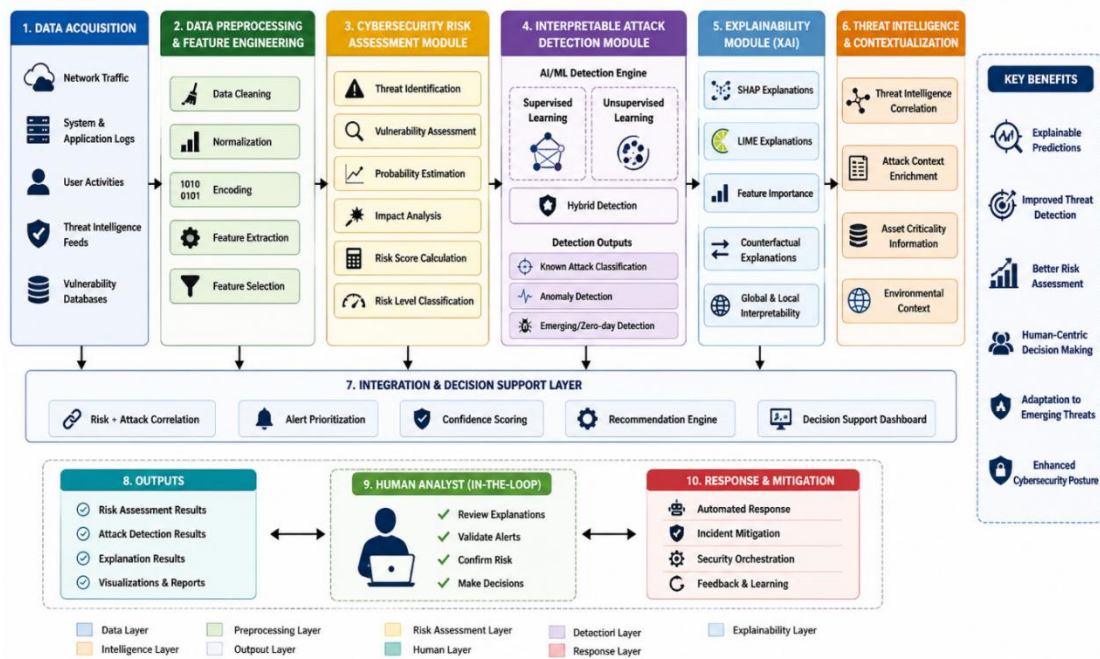


Figure 1. Proposed Explainable AI-Based Cybersecurity Framework

4.2. Data Acquisition, Processing, and Feature Engineering Layer

The data acquisition layer gathers cybersecurity data from various network monitoring systems, intrusion detection sensors, firewalls, endpoint security platforms, authentication services, application logs, and threat intelligence sources. Data processing layer processes data over data integration, normalization, missing value, noise removal, encoding, temporal alignment before analytical processing due to data coming in with different formats, frequencies and levels of granularity. [10] These raw observations are then mapped into security-relevant features such as packet and flow statistics, connection durations, protocol behaviors, source characteristics, destination characteristics, traffic volume, authentication patterns, error rates, and behavioral indicators through the process of feature engineering. Feature-selection mechanisms are used to remove redundant or irrelevant features and keep the attributes that have the highest ability to predict whether an attack has occurred or not and how likely it is to be successful. This layer can then deliver a uniform and high-quality feature representation that can then be analyzed by subsequent layers of AI.

4.3. Cybersecurity Risk Assessment and AI-Based Attack Detection Layer

The core intelligence layer integrates cybersecurity risk assessment with the detection capabilities of AI-based attacks, offering not just threat identification but also the ability to prioritize threats. The attack detection component features trained machine learning or deep learning models to identify observations on the network as normal or malicious and, if training data is adequate, classify specific attack types. [11] Meanwhile, the risk assessment takes place, which assesses threat likelihood, vulnerability indicators, asset criticality, and potential impact, to arrive at a quantitative or categorical risk score. A representative formulation might be $R = P(T) \times I(T)$, where R represents threat, $P(T)$ is the probability of threat, and $I(T)$ is the impact of threat; other contextual factors may be included as needed by the experimental design. This output results in the system being able to differentiate between legitimate and malicious activity and also between the severity of the cybersecurity incidents, giving higher priority to higher risk incidents over lower risk anomalies.

4.4. Emerging Attack Identification, Explainability, and Threat Intelligence

The emerging-attack identification component enables the detection of behaviour that deviates from previously-defined attack classes. Observations that deviate significantly from a set of normal or malicious observations can be identified by anomaly detection, behavioral profiling, confidence analysis or other appropriate methods. If an attack or uncertainty is detected in a model, then the explainability layer can analyze the model decision, such as using SHAP, LIME, feature importance or counterfactual explanations. The explanation that is generated outlines the network characteristics and behavior that led to the prediction or high probability score. Combined with threat-intelligence information, this analysis can help provide contextual information regarding known indicators, attack techniques, vulnerabilities, and threat actors. By integrating behavioral detection, explainable predictions and contextual threat intelligence, analysts can separate out known threats from the potentially present ones and have a clear understanding of the evidence behind their security decisions.

4.5. Security Decision, Human-in-the-Loop Support, and Framework Workflow

The last layer transforms the detection, risk, explanation and threat intelligence outputs into actionable cybersecurity decisions. [12] Predicted attack type, confidence level, risk score, potential impact and severity can give the system an idea as to the priority of the events, which allows it to recommend an appropriate action, e.g., continued monitoring, escalate alert, investigate further, isolate, or automatically respond. Human In the Loop Decision Support (HWILL) empowers security analysts to both see model predictions and explanations before taking high-stakes actions, providing a hybrid of automated analytical capabilities and human expertise. The overall process starts with the continuous data acquisition, then goes to the data preprocessing and feature extraction, risk estimation and attack classification, emerging threat analysis, explanation generation, threat enrichment and security decision-making. Conversely, feedback from the analysts can be used to update and refine the model and incorporate new knowledge about threats, enabling a continuous cycle of refinement and adaptation to evolving network behaviour and enhancing transparency and reliability of cybersecurity operations.

5. Explainable Cybersecurity Risk Assessment Model

5.1. Risk Assessment Architecture and Threat Identification

The proposed explainable cybersecurity risk assessment model is the development of a structured pipeline that can be used to recognize the threats and to convert the observations of the network into interpretable risk information. The model is fed with network traffic, security logs, endpoint events, authentication data, vulnerability data, and threat-intelligence, and then it processes the information to detect suspicious activities and potential attack conditions. [13]

Threat identification uses AI-driven classification and behavioral anomaly analysis to find out if an observed event matches normal activity, a known category of attack, or if the type of event may be a new attack that hasn't been identified before. Relevant threat characteristics are extracted, such as the nature of the attack, its frequency, the behavior of the source, the assets to be attacked, the communication patterns, and anomalies observed, and are used as evidence in subsequent risk analysis. This architecture allows risk assessment to be continuous and not just periodic manual assessments and forms the basis for being able to explain why a certain event is a potential threat to cyber security.

5.2. Vulnerability, Attack Probability, and Impact Assessment

After identifying threats, it looks at the vulnerability and exposure of affected assets, as well as the likelihood and impacts of an attack. Vulnerability assessment provides a qualitative analysis of the degree of vulnerability, based on asset criticality, known vulnerabilities, configuration conditions, level of exposure, effectiveness of security controls, and the relationship between vulnerable resources, and the security controls that are used to detect attack behaviors. [14] The AI model calculates the probability of an attack based on network characteristics, prior attack history and patterns, deviations from normal behavior and contextual threat data, and outputs the probability for each security event as a normalized value. The following are examples of confidentiality, integrity, availability, operational disruption, data exposure, and service degradation consequences that are taken into account when doing the impact assessment. These dimensions enable the framework to differentiate between events that might seem to be unexpected but don't involve a lot of harm and attacks that pose a more significant security threat to the organization, allowing for more meaningful prioritization of security.

5.3. Explainable Risk Score Calculation and Risk Classification

The overall cybersecurity risk is determined by combining the assessed threat probability, vulnerability/exposure and potential impact. A general statement of risk is $\text{Risk} = \text{Threat Probability} \times \text{Vulnerability} \times \text{Potential Impact}$, where each element is measured on a scale relevant to the event for which it applies so that they can all be compared. [15] This resulting risk score can then be assigned to predefined categories, like low, moderate, high and critical, based on an organizational threshold or an experimentally determined decision boundary. Classification allows security teams to prioritize incidents by their potential impact and make them more serious, rather than treating all detected attacks on equal footing. The framework can also include a breakdown of the individual components of probability, vulnerability, and impact that contribute to the overall risk assessment, to increase transparency so that analysts can determine that a high-risk classification is due to a highly probable attack, highly vulnerable asset, highly impactful target, or a mixture of any of these.

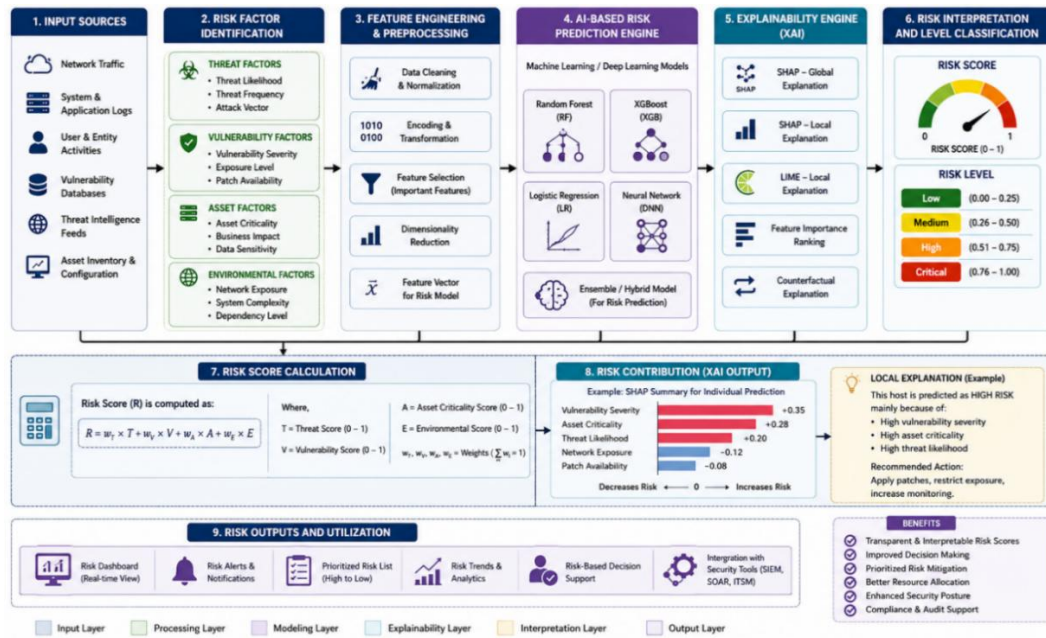


Figure 2. Explainable Cybersecurity Risk Assessment Model

5.4. Explainable Risk Prediction, Factor Attribution, and Security Decision Support

The explainability part gives evidence to the human user that explains what network, behavioral, vulnerability factors and contextual factors most influence the model's decision on the various risk levels. Methods like SHAP and LIME can be used to provide explanations for individual security incidents as well as explanations about the general behaviour of the risk prediction model. For instance, if a certain type of connection was used more often, a high number of packets, repeated authentication failures, high asset exposure and increased attack probability were the primary factors that led to the prediction of high risk. Such explanations allow analysts to confirm the results of their models, explore unusual patterns, differentiate between real threats and false alarms, and prioritize response. This risk score, attack classification, confidence score, and feature attribution are provided as decision support data to aid human analysts to combine AI generated evidence with organizational security policies and expert knowledge to determine if containment or remediation actions should be taken.

6. Interpretable Detection of Emerging Network Attacks

Table 2. Emerging Network Attack Detection Characteristics

Attack Condition	Detection Indicator	Detection Method	Expected Output
Known attack	Recognized pattern	Supervised ML	Attack class
Unknown attack	High anomaly score	Anomaly detection	Suspicious
Zero-day	Novel behavior	Hybrid detection	Potential zero-day
Behavioral deviation	Unusual traffic	Behavioral analysis	Anomaly

High-severity attack	High risk	Risk assessment	Critical
----------------------	-----------	-----------------	----------

6.1. Emerging Attack Detection Architecture and Network Traffic Analysis

The proposed emerging-attack detection architecture continuously analyzes the heterogeneous network and security data, and detects known and unknown malicious behaviors. Relevant monitoring points report the network traffic they see and display it in terms of flow-level and packet-level information including source/destination address, ports, protocols, packet counts, byte volume, connection duration, traffic rates, error conditions, and temporal relationship. [16] Preprocessing operations are used to normalize the numerical variables and to create a uniform representation of the data for AI-based analysis. Preprocessing operations include normalization of numerical variables and the creation of a uniform representation of the data for AI-based analysis. It melds traditional attack classification, behavior based and anomaly detection into a single architecture, with traffic of known attack types being classified immediately, and suspicious observations that haven't been classified according to known types being sent for emerging-threat analysis. This layered architecture enhances the performance of the system to detect new attacks that go beyond just being instances of known attack signatures.

6.2. Behavioral Feature Extraction, Attack Classification, and Anomaly Detection

Behavioral feature extraction helps transform the observable behavior in a network into a description of the deviation from acceptable network communication patterns, and may indicate malicious behavior. Some of these can be abnormalities in connection frequency, unusual distributions of packets or bytes, repeated authentication failures, protocol

usage not expected, accessing unusual destinations, increased or sudden changes in traffic volume, and temporal variations from known users and hosts that access the system. This produces a feature representation that is fed into supervised machine learning or deep learning models which classify known attacks, and anomaly detection mechanisms detect observations that are quite different from learned normal behavior. Classification models can be used to identify types of attack, including denial-of-service, reconnaissance, malware, credential attacks and others; while anomaly detection provides an extra way to detect suspicious activity in the case of network activity lacking representative labeled samples. Classification and anomaly detection can thus be combined to deal with known threats and possibly newer behaviors.

6.3. Unknown, Emerging, and Zero-Day Attack Identification with Severity Estimation

Unknown and Emerging attacks are found by examining observations with important behavioral deviations, which cannot be confidently classified in a known attack category. Anomaly scores, prediction confidence, distance from known feature distributions and temporal behavioural changes are all possible attributes that can be used as part of the framework to determine if an event is a potentially emerging threat. [17] Some observations, especially ones with low confidence or that are very unusual, might be pinpointed for zero day investigation – there's no known signature to classify them. Once a suspicious event was identified, work was done to estimate the severity of the attack based on the following criteria: anomaly magnitude, criticality of attacked asset, probability of attack, persistence of the attack, potential impact, and the number of resources affected. This severity assessment can give events the classification of low, medium, high or critical and enables emerging threats to be prioritised based on their potential cyber security impact, not just on how confident they are that they have found an incident.

6.4. Explainable Attack Classification, Feature Attribution, and Analyst-Oriented Explanation

The explainability part converts predictions made by the model into comprehensible evidence that security analysts can utilize for investigation and response. SHAP, LIME, permutation feature importance or counterfactual analysis can be used with XAI techniques to identify the network and behavioral features that were most important for the classification or anomaly score for each attack detected. For instance, if a high number of connections, unusual packets per connection, repeated failed authentication attempts, or unusual protocol operations were observed, such factors could have a significant impact on the prediction of malicious activity. Global explanations can give you an idea of the features that tend to differentiate between the various classes of attacks, and local explanations can explain why a specific network event was identified as a specific type of attack. The framework shows these explanations together with the attack type, confidence, anomaly score and severity information, allowing analysts to confirm alerts, evaluate new behaviours, minimize false positives, make informed

security decisions and always maintain a line of sight between the automated predictions and the observable network data.

7. Explainability and Interpretability Analysis

7.1. Global and Local Model Interpretability

The proposed framework measures interpretability in both the global and local context, offering a holistic view of the decisions made in the field of cyber security based on AI. Global interpretability involves understanding how the learned model performs overall, assessing which features in the data have a significant impact on attack classification and cybersecurity risk prediction consistently across the dataset. [18] The relative contribution of network characteristics like packet rate, connection duration, protocol behavior, traffic volume and authentication patterns can be determined using feature-importance analysis and aggregate SHAP values. Local interpretability, by contrast, allows for the explanation of individual predictions by highlighting the specific features that led to a given network event being classified as 'normal', 'malicious' or 'high risk'. This separation matters in cybersecurity in that cyber analysts need to know how the detection model works, as well as specific evidence to analyze each security alert.

7.2. SHAP, LIME, and Feature Attribution Analysis

SHAP and LIME are used as complementary approaches to obtaining interpretations of cybersecurity predictions. SHAP can be computed for an individual feature, depending on its impact on a particular prediction, and can be summed to inform a global picture of the importance of the features. [19] Positive and negative SHAP values can be used to determine if a feature has a positive or negative impact on the attack or risk category prediction. LIME offers local explanations by fitting an interpretable surrogate model around a specific observation from a complex model to gain insights into the factors that impact individual decisions. The feature-importance analysis can also be used to rank features in the model. When applied in combination, the techniques can be used to account for attack-specific and risk-specific predictions and to validate if the model is based on meaningful cybersecurity characteristics or on irrelevant or misinforming ones.

7.3. Attack-Specific, Risk-Specific, and Counterfactual Explanations

The framework yields explanations based on the kind of security decision being studied. Attack-specific explanations describe how the network and behavioral characteristics helped to classify an observation to a specific attack category, while risk-specific explanations describe the factors that led to a specific risk level. [20] Some of those factors could be abnormal traffic volume, repeated failed authentication, high asset exposure and elevated attack probability, to name a few. The counterfactual explanations give another view by pointing to differences in relevant input characteristics that may be responsible for the outcome of model prediction, e.g., decreasing the anomalous connection rate or removing repeated authentication failures. Such explanations can help analysts understand why an event was

labeled malicious or high-risk, and which observable conditions have an impact on a different security conclusion, which will support their investigations, validation, and decision-making.

7.4. Explanation Consistency, Fidelity, and Explainability–Accuracy Trade-Off

Consistency and fidelity measures as well as predictive accuracy are used in addition to the reliability of explanations. Explanation consistency evaluates the consistency of explanations when the network events are similar; and whether the important features identified in the network do not change significantly as the input conditions or models change slightly. Explanation fidelity assesses the fidelity of an explanation to the behavior of the underlying predictive model, especially if an interpretable surrogate is used. The framework also looks at the attack detection performance and interpretability as more complicated models are more likely to detect an attack; however, more complicated models are less likely to generate or understand explanations. Comparative experiments can then compare the accuracy and precision gains with explanation gains in terms of the F1-score. By analyzing this, we ensure that explainability is not a “nice to have” visualisation tool, but a quantifiable element of trustworthy and actionable cybersecurity intelligence.

8. Experimental Setup

8.1. Experimental Environment, Hardware, and Software Configuration

The experimental evaluation is done in a controlled computing environment, designed for reproducible cybersecurity machine learning experiments. The hardware configuration should include the number of cores, CPU type, amount of RAM, storage, and GPU type and memory, if applicable, since these can affect model-training and inference performance. [21] The software environment should identify the operating system and programming language, machine learning frameworks, and data processing and model development libraries and tools that facilitated the model development. The typical implementation might involve Python, libraries like NumPy and Pandas for data processing, conventional machine learning libraries like Scikit-learn, and deep learning tools like TensorFlow or PyTorch, along with libraries such as SHAP and LIME for explainability analysis. The versions of software packages, random seeds and configuration parameters used should be documented to ensure good experimental reproducibility and to facilitate meaningful comparisons with future studies.

8.2. Dataset Description, Distribution, and Experimental Scenarios

The evaluation is based on representative cyber security datasets with a mix of normal network traffic and various types of malicious traffic. The description of the dataset should include the following information: number of observations, feature dimensions, attack categories, class distribution, data source and characteristics of the network environment. The datasets should be analyzed beforehand to look for class imbalance and to make sure that a minority

attack class is well represented when evaluating the dataset. Experiments should contain several scenarios including anomaly detection, multiclass attack type classification, binary attack versus normal classification and emerging attack evaluation (for example, a subset of attack categories should not be included in the training set and be presented at the test time for these experiments). Other scenarios can be used to investigate the effects of noisy data, limited training data, or evolving attack distributions. This experimental design gives the proposed framework a chance to be tested in conventional classification settings as well as challenging settings which simulate previously unknown and changing network attacks.

8.3. Baseline Models, Proposed Model Configuration, and Training Procedure

Representative baseline models are used to compare the improvements that are obtained from the proposed XAI-based framework against the selection of the model. Baselines can consist of traditional classifiers like Decision Tree, Random Forest, Support Vector Machine, XGBoost as well as selected architectures of neural networks or deep learning. [22] The proposed model configuration should include the following: selected learning architecture, input features, output classes, risk-assessment mechanism, explainability method and any components for emerging-attack identification. Hyperparameters (learning rate, tree depth, number of estimators, batch size, number of epochs, regularization parameters, optimization strategy) should be optimized systematically or by using the validation data. The dataset is split into training, validation and independent testing sets, and the preconditioning and feature selection are done without using any information from the test set to guide the process of constructing a model. The training subset is used for training, the validation subset for model selection and hyperparameter tuning, while the test subset, that has not been seen by the model, is used for final performance reporting to get an unbiased evaluation.

8.4. Evaluation Metrics and Comparative Evaluation

The experimental assessment assesses the performance of the cybersecurity detection mechanism as well as the explainability mechanism proposed. The performance of classification is evaluated by the accuracy, precision, recall, F1-score, ROC-AUC, and imbalanced datasets based on the balanced accuracy, Matthews correlation coefficient and precision-recall AUC. Security specific parameters are false-positive rate, false-negative rate, detection latency, and throughput and resource utilization can be measured by CPU utilization, memory consumption, and inference time. In the case of numerical risk labels, cybersecurity risk assessment performance can be further evaluated using an appropriate error or agreement measure. Explanation fidelity, the stability of the explanation, whether it is consistent, its sparsity and feature-attribution agreement, if applicable, are used as criteria for assessing explainability. It is then compared with the baseline models using the same set of experimental data and scenarios; statistical significance or confidence intervals should be reported, where appropriate, to help establish whether any performance difference is

statistically significant and therefore might be considered meaningful, or merely random variation.

9. Results and Performance Evaluation

9.1. Cybersecurity Risk Prediction and Network Attack Detection Results

The first part of the assessment focuses on the ability of the proposed framework to detect malicious activities over the network, and to estimate the associated risk of cybersecurity. The results should present the accuracy, precision, recall, F1-score, ROC-AUC, and (when applicable) Matthews correlation coefficient and precision-recall AUC of the binary and multiclass attack detection. [23] The results of risk predictions should also show the capability of the framework to differentiate between risk levels by incorporating threat probability, vulnerability and potential impact. Known and emerging attack classes should be taken into account in analysis and confusion matrices should be reported to highlight classes which are likely to be mis-classed. Special focus needs to be directed towards false positives and false negatives, as too many false alarms may make analysts' job more difficult and not getting alerts on an attack may result in huge security exposure. The results should provide a measure of whether adding to the classification portion with a risk assessment is more informative for security outcomes than doing a classification alone.

9.2 Emerging Attack Detection, Latency, and Resource Utilization

The second stage analyzes the framework when new attacks are introduced, but have not been encountered before, to see how well it adapts to new and different tasks from standard supervised classification. The analysis should include the number of detections of attack classes withheld from the analysis and the number of false positives and false negatives, as well as a measure of the performance of the anomaly detection algorithm, prediction confidence, and the ability to differentiate between emerging malicious activity and legitimate anomalies. Other performance metrics of operations include detection latency, inference time, throughput, CPU usage, memory usage and, if available, GPU usage. The measurements answer the question of whether the proposed framework will be able to detect the events at a timely manner without using too much computational resources. The performance could also be studied with robustness experiments which may look at the performance of the system on noisy data, varying distributions of attacks, smaller training sets, and different intensities of attacks. The goal is to determine if the framework is robust enough to withstand the challenges and uncertainties of complex cybersecurity environments.

9.3. Explainability Performance and Comparative Evaluation

The last step is to determine if the proposed framework offers plausible and consistent explanations, as well as predictive power. Several automated techniques are available to detect the most important factors in attack classifications and cybersecurity risk predictions, both globally and locally:

SHAP, LIME, feature-importance, counterfactual analyses. The degree of explanation should be evaluated based on fidelity, consistency, stability, sparsity and other appropriate measures of interpretability, and illustrative examples of attacks should be shown to illustrate the role that explanations can play in the security analyst's confirmation and prioritization of alerts. The proposed framework is then evaluated in comparison with baseline machine learning and deep learning frameworks with similar setups on the same sets of data, experimental conditions and evaluation metrics. When comparing, differences in detection accuracy, F1-score, false-positive rate, emerging-attack detection capability, latency, resource utilization and explainability should be highlighted. Any statistical analyses, confidence intervals, or significance tests should be included when appropriate to assess if an observed improvement is statistically significant and to determine if the proposed approach is a viable compromise between predictive accuracy, emerging threat detection, computational speed, and interpretability.

10. Discussion

10.1. Interpretation of Experimental Findings and Effectiveness of the Proposed Framework

The results from the experiments should be evaluated from the perspective of the potential of the framework to enhance the cybersecurity detection, risk assessment and transparency of decision making. Greater precision, recall and F1 scores would suggest that the proposed approach can more effectively identify malicious network behavior, and a decrease in false positives and false negatives would suggest an increase in usefulness for security monitoring operations. The emerging-attack results should be analysed to see if supervised classification and behavioural anomaly analysis can increase the number of attack types detected for those not seen in the past. Likewise, the importance of incorporating threat probability, vulnerability, and potential impact to the attack classification alone for a more meaningful prioritization should be discussed. Overall, the discussion should conclude if the experimental evidence suggests that the central premise that the addition of AI-based detection along with cybersecurity risk assessment and XAI offers a more complete security capability than traditional detection methods holds true.

10.2. Explainability, Analyst Trust, and Accuracy–Interpretability Trade-Off

Beyond reporting on SHAP or LIME visualisation of explanations, it is important to explore the role of 'explainability' and consider whether explanations generated are reliable and actionable evidence for cybersecurity analysts. The ability to consistently identify meaningful network and behavioral characteristics enables analysts to validate alerts, gain insight into the characteristics of attacks, investigate suspicious events, and determine which are really threats and which are false positives. Explanation fidelity, stability and comprehensibility should also be discussed since an easy to understand explanation that does not accurately represent the underlying model will have a limited practical use. However, a tradeoff between accuracy and

interpretability should be explored as well. While more complex models could be more effective for detection, they could have higher explanation complexity, and more simpler models could have higher explainability, but lower detection performance. The results should thus evaluate whether the proposed framework achieves an appropriate balance among the accuracy, the interpretability and the operational usefulness of the proposed framework and not an optimum for each of the characteristics alone.

10.3. Comparison, Practical Implications, Scalability, and Generalizability

The proposed framework should be contrasted with current cybersecurity AI and XAI solutions to find any differences in terms of detection efficacy, ability to adapt to emerging attacks, risk assessment, explainability, computational needs and applicability in operations. The improvements should be considered relative to the methodological features of the previous studies and not only on the basis of the performance values taking into account variability of the data sets, type of attacks, evaluation procedure, and experimental environment. For the practical application, the framework can help security operations by prioritizing high risk events, providing meaningful evidence for investigation, and helping analysts to deal with emerging threats. Its scalability needs to be taken into account with regard to the volume of network traffic, the mix of data sources, distributed architectures, and the varying distributions of attacks, and deployment considerations need to be addressed with respect to the latency of inferences, computational resources, model updates, and integration with the existing security infrastructure. Generalizability needs to be evaluated with care, and the extent to which the framework can be applied to other datasets, organizations, network environments and attack types, as well as cross-dataset validation, additional real world traffic, and continuous learning experiments can provide additional evidence that the proposed approach is not overly reliant on a particular dataset or experimental setup.

11. Conclusion

This study proposed an Explainable Artificial Intelligence (XAI) cybersecurity framework, incorporating cybersecurity risk assessment and interpretativeness in the detection of emerging attacks in complex information systems. In the proposed approach, AI-based attack classification, behavioral anomaly detection, risk estimation and XAI techniques will be combined to offer predictive security capabilities and understandable explanations of the model's decisions. The framework assesses the threat probability, vulnerability and potential impact for the purpose of prioritising threats based on risk, whereas explainability mechanisms shed light on network and behavioral characteristics that account for threat classifications and threat predictions. Comparative evaluation, robustness testing, and ablation analysis are adopted as the research methodology to analyze detection effectiveness, emerging-attack identification, the computational performance, and explanation quality. The overall contribution is a cyber-security decision support

approach, which is a holistic solution that overcomes the restrictions of existing detection systems by linking automated threat identification with the human-readable risk assessment and evidence.

The outlined framework has real-world security implications as it gives security analysts the ability to shift from just alerting on attacks to making security decisions based on risk and interpretation. The framework can help validate alerts, prioritize threats, investigate incidents, and select the right response actions, while the new-attack detection functionality adds the ability to detect new malicious behaviors. Incorporating explainability can also enhance the transparency and enable more informed human oversight of automated security systems. However, many factors such as dataset representativeness, computational requirements, model adaptability, explanation reliability, and integration with existing security infrastructure need to be taken into account if a dataset is to be successfully adopted in an operational context. The results of the study show that the fusion of AI-based detection, cybersecurity risk assessment, and explainable decision support presents a promising solution to address the dynamic nature of cyber threats in complex information systems and lays the groundwork for future research in the field of adaptive, robust, and human-centered cybersecurity intelligence.

References

- [1] Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., ... & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management*, 20(4), 5115-5140.
- [2] Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P. F., Han, Y., Jmila, H., ... & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, 77(11), 789-812.
- [3] Guo, Y. (2023). A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer communications*, 198, 175-185.
- [4] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, 12, 30164-30175.
- [5] Hariharan, K. B., Saravanan, S., & Parkunam, N. (2020). Life time improvement of D7 tool steel by cryogenic treatment. *Materials Today: Proceedings*, 21, 619-621.
- [6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [7] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.
- [8] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.

- [9] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- [10] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- [11] Alenezi, R., & Ludwig, S. A. (2021, December). Explainability of cybersecurity threats data using shap. In *2021 IEEE symposium series on computational intelligence (SSCI)* (pp. 01-10). IEEE.
- [12] Islam, M. T., Syfullah, M. K., Rashed, M. G., & Das, D. (2024). Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI. *International Journal of Machine Learning and Cybernetics*, 15(11), 5337-5360.
- [13] Wali, S., Farrukh, Y. A., & Khan, I. (2025). Explainable AI and random forest based reliable intrusion detection system. *Computers & Security*, 157, 104542.
- [14] Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. *Information*, 16(12), 1036.
- [15] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0).
- [16] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1(2018), 108-116.
- [17] Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). Ieee.
- [18] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications* (pp. 1-6). Ieee.
- [19] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157-175). Springer, New York, NY.
- [20] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- [21] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [23] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 55(5), 3503-3568.
- [24] Dhanushkodi, K., & Thejas, S. (2024). Ai enabled threat detection: Leveraging artificial intelligence for advanced security and cyber threat mitigation. *IEEE access*, 12, 173127-173136.
- [25] Trillo, J. R., González-López, F., Morente-Molinera, J. A., Magán-Carrión, R., & García-Sánchez, P. (2025). Evaluation of explainable, interpretable and non-interpretable algorithms for cyber threat detection. *Electronics*, 14(15), 3073.
- [26] Seknametla, P. R., & Sunkara, R. (2025). Applying AIOps for Predictive Incident Management in DevOps-Driven Cloud Infrastructure. *International Journal*, 12(6).
- [27] Gantikota, S. (2026). Securing Microservice Communication across WCF, JAX-RS, and Spring Boot: Authentication, Authorization, and Audit Patterns for Healthcare Interoperability. *American International Journal of Computer Science and Technology*, 8(2), 15-20. <https://doi.org/10.63282/3117-5481/AIJCS> V8I2P102
- [28] Sandra, K. (2024). THE REGULATED BANKING AI LAKEHOUSE. INDO-CONTINENTAL ACADEMIC PUBLISHERS.
- [29] Brahmandam, L. M. K. (2025). Design Patterns and Empirical Evaluation of Reusable Terraform Modules Encoding Audit-Ready Defaults for Multi-Account AWS Deployments: A Cross-Team Study across EC2, S3, RDS, EKS, IAM, and Cloud Watch. *International Journal of Emerging Research in Engineering and Technology*, 6(2), 133-142. <https://doi.org/10.63282/3050-922X.IJERET-V6I2P116>