IRSTI 20.23.17 DOI: https://doi.org/10.62687/STJ.7.1.2025.14

INTERPRETABLE AI FOR CYBER THREAT DETECTION IN SMART SYSTEMS

[1]A.A. Abdukarimova, **[2]R.A. Ismailova**, **[3]A.M. Jumagaliyeva***, **[4]V.B. Rystygulova**, **[5]A.E. Koxegen**

[1,3,4]K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan
[2]Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan
[5]S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan
*e-mail: jumagalievaainur.m@gmail.com

A.A. Abdukarimova – PhD, Associate Professor, Department of Information Technology, K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan, e-mail: a.abdukarimova777@gmail.com, https://orcid.org/0000-0002-6932-6282

R.A. Ismailova – PhD, Associate Professor, Department of Computer Engineering, Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan, e-mail: rita.ismailova@manas.edu.kg, https://orcid.org/0000-0003-0308-2315

A.M. Jumagaliyeva – Senior lecturer, Department of Information Technology, K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan, e-mail: jumagalievaainur.m@gmail.com, https://orcid.org/0000-0001-8632-5209

V.B. Rystygulova – Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Information Technology, K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan, e-mail: rystygulovaV@mail.ru, https://orcid.org/0000-0003-3883-5612

A.E. Koxegen – Senior lecturer, Department of Computer Science, S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan, e-mail: a.koksegen@kazatu.kz, https://orcid.org/0000-0002-8994-4096

Abstract. The rapid development of intelligent technologies, IoT infrastructures, and cloud services has led to an increasingly complex cybersecurity landscape. Because intelligent systems generate massive volumes of heterogeneous streaming data in real time, the challenges of accurately detecting threats in the cyber environment have increased significantly. Traditional intrusion detection systems used to identify cyber threats typically rely on static attack signatures and, therefore, have limited ability to detect new forms of cyber threats. This study presents an interpretable artificial intelligence framework for real-time cyber threat detection in streaming intelligent systems using machine learning models, combined with explainable AI methods such as SHAP and LIME to enhance detection capabilities. Security-related data streams and other sources were analyzed to identify potential anomalies and cyberattacks. Based on experimental evaluation, the proposed model, based on combined methodologies, such as hybrid explainable AI model, demonstrated superior performance compared to each of the individual machine learning models across several evaluation metrics, including accuracy, precision, recall, and F1 score. Overall, this work demonstrates how to leverage machine learning and explainable AI methods to improve trust, transparency, and the practical applicability of cybersecurity monitoring solutions in dynamic, intelligent environments.

Keywords: cybersecurity, machine learning, explainable AI, intrusion detection, streaming data, smart systems, anomalies.

СМАРТ ЖҮЙЕЛЕРДЕ КИБЕРҚАУІПТЕРДІ АНЫҚТАУҒА АРНАЛҒАН ИНТЕРПРЕТАЦИЯЛАНАТЫН ЖАСАНДЫ ИНТЕЛЛЕКТ

[1]А.А. Абдукаримова, [2]Р.А. Исмаилова, [3]А.М. Джумагалиева*, [4]В.Б. Рыстыгулова, [5]Ә.Е. Көксеген

[1,3,4]Қ.Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан
[2]«Манас» Қырғыз-Түрік университеті, Бішкек, Қырғызстан
[5]С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан
*e-mail: jumagalievaainur.m@gmail.com

А.А. Абдукаримова – PhD, қауымдастырылған прфессор, Ақпараттық технологиялар кафедрасы, Қ.Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан, e-mail: a.abdukarimova777@gmail.com, https://orcid.org/0000-0002-6932-6282

Р.А. Исмаилова – PhD, қауымдастырылған прфессор, Ақпараттық технологиялар кафедрасы, «Манас» Қырғыз-Түрік университеті, Бішкек, Қырғызстан, e-mail: rita.ismailova@manas.edu.kg, https://orcid.org/0000-0003-0308-2315

А.М. Джумагалиева – сеньор-лектор, Ақпараттық технологиялар кафедрасы, Қ.Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан, e-mail: jumagalievaainur.m@gmail.com, https://orcid.org/0000-0001-8632-5209

В.Б. Рыстыгулова – қауымдастырылған прфессор, Ақпараттық технологиялар кафедрасы, Қ.Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан, e-mail: rystygulovaV@mail.ru, https://orcid.org/0000-0003-3883-5612

Ә.Е. Көксеген – аға оқытушы, Компьютерлік ғылымдар кафедрасы, С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, e-mail: a.koksegen@kazatu.kz, https://orcid.org/0000-0002-8994-4096

Аңдатпа. Ақылды технологиялардың, IoT инфрақұрылымдарының және бұлтты қызметтердің қарқынды дамуы киберқауіпсіздік ландшафтының күрделене түсуіне әкелді. Ақылды жүйелер нақты уақыт режимінде үлкен көлемдегі гетерогенді ағынды деректерді жасайтындықтан, киберортада қауіптерді дәл анықтау қиындықтары айтарлықтай артты. Киберқауіптерді анықтау үшін қолданылатын дәстүрлі басып кіруді анықтау жүйелері әдетте статикалық шабуыл қолтаңбаларына сүйенеді және сондықтан киберқауіптердің жаңа түрлерін анықтау мүмкіндігі шектеулі. Бұл зерттеуде машиналық оқыту модельдерін пайдалана отырып, ағынды интеллектуалды жүйелерде нақты уақыт режимінде киберқауіптерді анықтауға арналған түсіндірілетін жасанды интеллект құрылымы, анықтау мүмкіндіктерін жақсарту үшін SHAP және LIME сияқты түсіндірілетін жасанды интеллект әдістерімен біріктірілген. Қауіпсіздікке қатысты деректер ағындары және басқа да ықтимал ауытқулар мен кибершабуылдарды анықтау үшін талданды. Эксперименттік бағалау негізінде ұсынылған модель біріктірілген әдіснамаларға (гибридті түсіндірілетін жасанды интеллект моделі) негізделген, дәлдік, дәлдік, еске түсіру және F1 ұпайын қоса алғанда, бірнеше бағалау көрсеткіштері бойынша әрбір жеке машиналық оқыту модельдерімен салыстырғанда жоғары өнімділікті көрсетті. Жалпы алғанда, бұл жұмыс динамикалық, интеллектуалды ортада киберқауіпсіздікті бақылау шешімдерінің сенімділігін, ашықтығын және практикалық қолданылуын жақсарту үшін машиналық оқытуды және түсіндіруге болатын жасанды интеллект әдістерін қалай пайдалану керектігін көрсетеді.

Кілт сөздер: киберқауіпсіздік, машиналық оқыту, түсіндірілетін жасанды интеллект, бұзушылықтарды анықтау, деректер ағыны, интеллектуалды жүйелер, ауытқулар.

ИНТЕРПРЕТИРУЕМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ДЛЯ ОБНАРУЖЕНИЯ КИБЕРУГРОЗ В УМНЫХ СИСТЕМАХ

[1]А.А. Абдукаримова, [2]Р.А. Исмаилова, [3]А.М. Джумагалиева*, [4]В.Б. Рыстыгулова, [5]Ә.Е. Көксеген

[1,3,4]Казахский университет технологии и бизнеса им.К.Кулажанова, Астана, Казахстан
[2]Кыргызско-Турецкий университет «Манас», Бишкек, Киргизстан
[5]Казахский агротехнический исследовательский университет им.С.Сейфуллина, Астана, Казахстан
*e-mail: jumagalievaainur.m@gmail.com

А.А. Абдукаримова - PhD, асоциированный профессор кафедры Информационных технологий, Казахский университет технологии и бизнеса им.К.Кулажанова, Астана, Казахстан, e-mail: a.abdukarimova777@gmail.com, https://orcid.org/0000-0002-6932-6282

Р.А. Исмаилова - PhD, асоциированный профессор кафедры Компьютерных наук, Кыргызско-Турецкий университет «Манас», Бишкек, Киргизстан, e-mail: rita.ismailova@manas.edu.kg, https://orcid.org/0000-0003-0308-2315

А.М. Джумагалиева- сеньор-лектор кафедры Информационных технологий, Казахский университет технологии и бизнеса им.К.Кулажанова, Астана, Казахстан, e-mail: jumagalievaainur.m@gmail.com, https://orcid.org/0000-0001-8632-5209

В.Б. Рыстыгулова - к.ф.-м.н. асоциированный профессор кафедры Информационных технологий, Казахский университет технологии и бизнеса им.К.Кулажанова, Астана, Казахстан, e-mail: rystygulovaV@mail.ru, https://orcid.org/0000-0003-3883-5612

Ә.Е. Көксеген – старший преподаватель кафедры Компьютерных наук, Казахский агротехнический исследовательский университет им.С.Сейфуллина, Астана, Казахстан, e-mail: a.koksegen@kazatu.kz, https://orcid.org/0000-0002-8994-4096

Аннотация. Быстрое развитие интеллектуальных технологий, инфраструктур Интернета вещей и облачных сервисов привело к усложнению среды кибербезопасности. Поскольку интеллектуальные системы генерируют огромные объемы разнородных потоковых данных в режиме реального времени, задачи точного обнаружения угроз в киберсреде значительно возросли. Традиционные системы обнаружения вторжений, используемые для выявления киберугроз, обычно используют статические сигнатуры атак и, следовательно, имеют ограниченные возможности для выявления новых форм киберугроз. В этом исследовании будет представлена интерпретируемая структура искусственного интеллекта для обнаружения киберугроз в режиме реального времени в потоковых интеллектуальных системах с использованием моделей машинного обучения, в сочетании с объяснимыми методами искусственного интеллекта, такими как SHAP и LIME, для повышения возможностей обнаружения. Было анализировано потоки данных, связанные с безопасностью, и другие источники для выявления потенциальных аномалий и кибератак. На основе экспериментальной оценки предложенная модель, основанная на комбинированных методологиях (гибридная модель объяснимого искусственного интеллекта), показала лучшие результаты, чем каждая из моделей машинного обучения по отдельности, по нескольким метрикам оценки, включая точность, прецизионность, полноту и F1-меру. В целом, эта работа демонстрирует, как использовать методы машинного обучения и объяснимого искусственного интеллекта для повышения уровня доверия, прозрачности и практического применения решений по мониторингу кибербезопасности в динамичных, интеллектуальных средах.

Ключевые слова: кибербезопасность, машинное обучение, объяснимый искусственный интеллект, обнаружение вторжений, потоковые данные, интеллектуальные системы, аномалии.

Introduction. The speed at which digital infrastructure, smart technologies, and interconnected cyber-physical systems are developing has dramatically increased the number of different kinds of information that exist in today’s modern information environments. Smart

systems, such as IoT platforms, cloud computing infrastructures, intelligent transportation systems, and smart city networks produce large amounts of disparate data streams all day, every day, with no end in sight. While these technologies have great value from a business perspective in automating operations, increasing productivity, and allowing for instantaneous decision-making; they are also opening up new opportunities for exploitation and are increasing the threat exposure of businesses to those who are exploiting new and advanced types of cyber attacks.

Smart environments are now highly susceptible to a multitude of different types and sources of cyber attacks including: distributed denial-of-service (DDoS) attacks, data injection attacks, botnet activity, scanning for reconnaissance purposes, and man-in-the-middle intrusions. These attacks can cause catastrophic service disruptions, threaten sensitive information, and create large-scale system failures. Traditional cybersecurity protection measures (i.e., signature-based intrusion detection systems IDS) rely upon detecting attacks that have already been identified, or have already been documented. While this approach is sufficient for detecting attacks that already exist; it is frequently not capable of detecting unknown or evolving attack methods that are introduced because of the dynamic nature of the networks in which they occur.

To mitigate the inability to detect “new” attacks using traditional cybersecurity protection mechanisms, organizations are beginning to use machine learning ML-based intrusion detection systems (IDS). These ML-based IDS are able to learn complex behaviors of users and/or machines through their interactions with the network, and therefore, can identify anomalous behaviors that may signal a malicious activity. Algorithms for example such as the Random Forest, gradient boosting models (GBM), and deep neural networks (DNN) have all shown to produce high accuracy levels at detecting cyber threats within multi-dimensional data sets. Additionally, deep learning models such as Long Short-Term Memory (LSTM) networks are particularly proficient at analyzing sequential and temporal patterns within the streams of network traffic data they process. However, despite the high predictive accuracy of many ML-based detection systems, a critical challenge remains: the lack of transparency and interpretability of model decisions. Many advanced machine learning models operate as “black-box” systems, providing predictions without clear explanations of the factors influencing these decisions. In cybersecurity contexts, this lack of interpretability can limit the trust of security analysts and complicate incident response processes. As a result, explainable artificial intelligence (XAI) has emerged as an important research direction aimed at improving transparency and interpretability of machine learning models.

Using Explainable AI methods (SHAP and LIME) provide analysts with the ability to identify how much each characteristic impacts a given model's output. Explainable AI allows users to gain insights into the global behavior of a model, as well as the local behavior of an individual detection decision. The use of explainable AI in Cybersecurity Monitoring Systems provides users with an understanding of the sources of the detected anomalies, thereby, increasing the usefulness of AI-based security solutions.

The second area impacting modern Cybersecurity is the requirement to process massive streams of data that are generated on an ongoing basis. Real-time Smart Systems generate Security Events that need to be analysed and reported in as little time as possible. As a result, Cybersecurity Detection Frameworks must be capable of processing vast amounts of Data Streams at High Prediction Accuracy and Computational Efficiency.

However, despite significant progress in machine learning-based intrusion detection and explainable AI techniques, existing studies do not sufficiently address the integration of real-time streaming data processing with interpretable detection models in smart system environments. Most approaches either focus on detection accuracy without interpretability or apply explainable methods without considering dynamic data streams. This gap motivates the development of a unified framework that ensures both high detection performance and model transparency in real-time conditions.

The purpose of this paper is to develop an Interpretable Artificial Intelligence Framework for Real-Time Cyber Threat Detection in Streaming Smart Systems. The proposed Framework deploys Machine Learning Models and Explainable AI Methods to enable Accurate Cyber Threat Detection, along with providing Transparency and Interpretability of the Model Outputs. The object of this

study is cybersecurity monitoring in streaming smart systems operating in dynamic and data-intensive environments.

The subject of the research is the development and application of interpretable machine learning models for real-time cyber threat detection using streaming security data.

The proposed Framework will integrate the use of Ensemble Learning Algorithms, Deep Learning Models, and Explainable AI Analysis Methods to detect Cyber Threats based on Streaming Security Data collected from three sources; Network Traffic, IoT Devices, and System Logs. The main contributions of this study are:

1. Develop a Conceptual Framework for Interpretable Cyber Threat Detection in Streaming Smart Systems;
2. Design a System Architecture that Integrates ML Detection Models with Explainable AI Methodologies;
3. Conduct experiments using several Classifiers, including Random Forest, XGBoost, LSTM, and a Hybrid Explainable AI Detection Model;
4. Evaluate and Compare each of the Detection Models, with an understanding of the Inputs and Outputs of each Model, by conducting Feature Importance Analyses and using Explainable AI Methodology to Show Model Interpretability.

The novelty of this study lies in the development of a unified framework that integrates machine learning and explainable artificial intelligence for real-time cyber threat detection in streaming smart systems.

In contrast to other techniques, the proposed model balances prediction accuracy with interpretability by providing both global and local feature importance metrics through SHAP and LIME analysis of model predictions.

Authors also use hybrid modeling techniques to help improve the effectiveness of the detection process across a variety of cyber risk factors.

The results will demonstrate that using both Machine Learning Models and Explainable AI Mechanisms produces Improved Cyber Security Monitoring System Detection Performance and Transparency. The Proposed Approach provides the basis for creating Intelligent Interpretable and Scalable Cyber Threat Detection Systems for next-generation Smart Infrastructure.

Literature review. In today's society, we are living through a technological revolution. Smart technology is being developed at lightning speed and the cloud is increasing in importance every day as well. The Internet of Things (IoT) is rampant, and has led to a more complex security landscape than ever before. Traditional intrusion detection systems (IDS), or systems that detect intrusion attempts on a network, largely rely on signature or rule-based methods of detecting known attacks (Al Rawajbeh et al., 2025). While traditional IDS can effectively identify previously known threat types, they are often unable to effectively identify unknown or new threat types in ever-changing environments. In response, a growing number of researchers have begun using machine learning techniques to automatically detect cyber threats according to (Prasad et al., 2025).

In the past decade, supervised learning algorithms have been used most widely for IDS systems because these algorithms are capable of classifying network traffic and detecting anomalous activity related to that traffic as well. SVMs, decision trees, and random forest classifiers are some of the most common machine learning algorithms used in IDS research (Rahmati, 2025). Random Forest classifiers have received much attention because of their ability to accurately classify data, even in the presence of noise, and their capability to process high-dimensional network traffic features (Almheiri et al., 2025). Additionally, gradient boosting algorithms such as XGBoost and LightGBM have received considerable interest in the cyber attack detection area due to their ability to model complex, non-linear relationships that exist within large sets of data according to (Paul, 2025).

In the last few years, researchers have also been focusing on deep learning techniques. A number of different neural network architectures including DNNs, CNNs, and RNNs have been applied to detecting complex cyber threats in large networks (Khalaf et al., 2025). LSTM networks, in particular, have been shown to be effective for temporal analysis and capturing time-based relationships in the behavior of cyber threats (Thiruvenkatasamy et al., 2025). These models can also

be used to detect complex behaviors associated with a multi-stage attack, which may be difficult for traditional machine learning techniques to accomplish (Alshudukhi et al., 2025).

Even though these models perform extremely well when it comes to detecting intrusions, there is a significant lack of transparency in many of the machine-learning based intrusion detection systems that are currently available (Jumagaliyeva et al., 2025). Complex systems such as deep neural networks, ensemble methods and gradient boosting methods work as "black-boxes"; therefore it is extremely difficult for anyone to understand how the machine-learning based intrusion detection system made its final decision (Kalutharage et al., 2025). When conducting investigations on incidents in cybersecurity, a lack of interpretability can reduce the trust of security analysts and make it much more difficult for them to carry out their duties of investigating, responding to and remediating incidents (Mohale et al., 2025).

Recent research has increasingly focused on integrating explainable artificial intelligence (XAI) into the suite of techniques utilized in cybersecurity analytics as a way to address this problem. The goal of explainable AI methods is to enhance transparency by providing interpretable explanations of how machine-learning predictions are generated from the input features (Alabdulatif, 2025). SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are two of the most widely used techniques to explain the decision-making process of machine-learning models. These two techniques are helpful for researchers because they help them understand what features played important roles in the final decision made by the model, while also allowing for the generation of global model interpretation as well as local model interpretations for individual predictions. Thus, they are extremely helpful to researchers involved with the development of cybersecurity monitoring systems (Lee et al., 2024).

An important issue that has been actively explored by researchers is how to analyze the streaming security data that is produced from smart systems. For instance, many smart infrastructure systems continuously produce large amounts of real-time data from various sources such as network traffic flows, Internet of Things (IoT) sensors, system logs, and cloud telemetry (Moustafa et al., 2023). As such, there is a need for real-time analytics on these smart infrastructures with high predictive accuracy and computational efficiency in order to effectively process these streams of security data for intrusion detection purposes. Researchers have proposed stream-based intrusion detection frameworks that utilize online learning algorithms to facilitate real-time threat detection in dynamically changing environments.

Recent studies have shown that using machine learning based detection techniques in conjunction with explainable artificial intelligence (XAI) technologies can significantly improve the ability of the cybersecurity systems to detect threats both accurately and explainably (Patel et al., 2025). By utilizing various combinations of machine learning models such as Random Forest, XGBoost, LSTM Networks and XAI methodologies like SHAP and LIME to explain the predictions produced by these models, cybersecurity professionals can not only identify potential threats but also understand the reasons for the prediction produced by the machine learning models (Akshya et al., 2025). There is enhanced need for integrated cybersecurity frameworks that contain high-performance machine learning models, real-time streaming processing systems and XAI techniques to produce a transparent and trustworthy approach to detecting cyber threats in smart systems 16. (Rystygulova et al., 2025).

Therefore, in order to address this need, this paper proposes an interpretable AI framework for detecting cyber threats using real-time data obtained from smart systems. This framework incorporates the use of Random Forest, XGBoost and LSTM models along with SHAP and LIME XAI methodologies in order to maintain an accurate method of detecting threats in real-time while providing a clear explanation of the predictions made by these models.

Materials and methods. The amount of heterogeneous, streaming data generated by today's smart systems is extremely high, and includes data that is generated by IoT devices, system logs, network traffic, and data created in the cloud. The environments created by these systems are highly dynamic and susceptible to a wide range of cyber threats, such as DDoS attacks, data injection, botnet activity, and man-in-the-middle attacks. To effectively detect and understand any such threats, there

needs to be an integration of different types of analysis methods or techniques that utilize machine learning algorithms to detect anything from malware to unauthorized access.

The experimental evaluation utilized two benchmark cybersecurity datasets: CICIDS2017 and UNSW-NB15. The CICIDS2017 dataset contains over 2.8 million network flow records. Each record contained 78 features, drawn from real-world network traffic, ranging from benign to malicious activity, such as denial of service (DoS), distributed denial of service (DDoS), botnets, and penetration attacks.

The UNSW-NB15 dataset contains approximately 2.5 million records, representing 49 features reflecting a range of modern attack categories, such as exploits, reconnaissance attacks, backdoor attacks, and shellcode attacks. Using more than one dataset ensures the robustness and generalizability of the proposed framework across various cyberthreat scenarios and network conditions.

Before training the models the datasets were first preprocessed by addressing the missing data points, normalizing the numeric features, and encoding any categorical features prior to feature scaling for consistency between the input variables. The datasets were then split into training (70%) and test (30%) datasets. Cross-validation was also used while training the models for additional robustness and to help avoid overfitting.

Model training was performed using the Random Forest, XGBoost, and LSTM machine learning models with optimum hyperparameters. The final hybrid XAI model combines multiple detection methods in one model to produce a higher overall performance. Model evaluation was performed using standard classification metrics, including accuracy, precision, recall, and the F1-score.

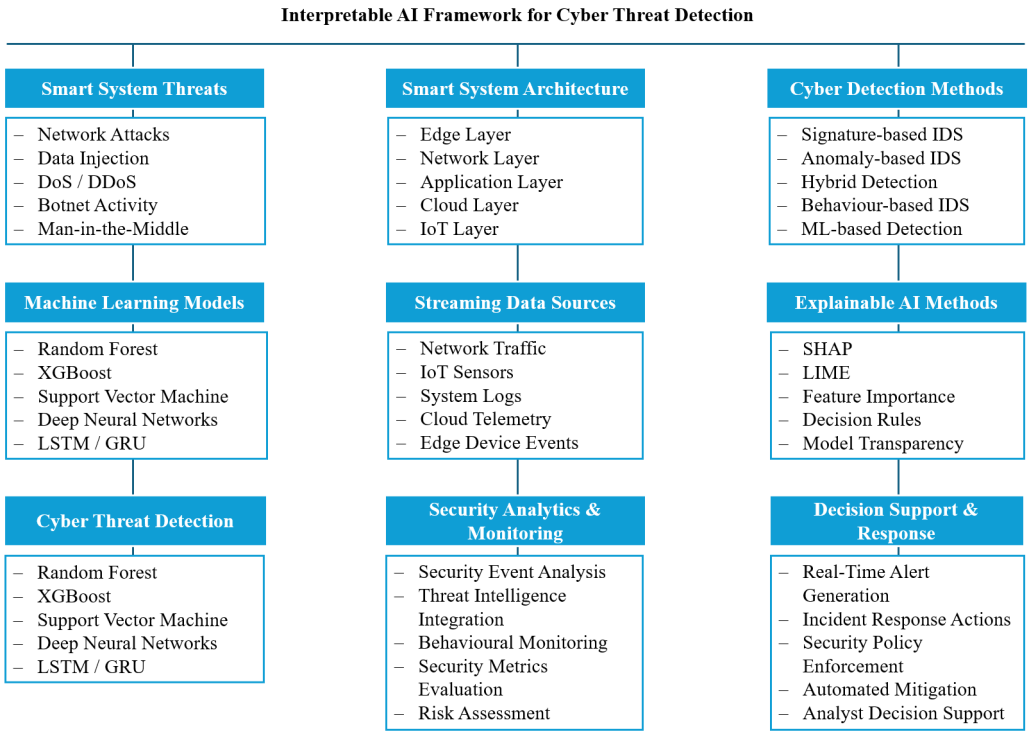

Figure 1. Conceptual framework of AI-based cyber threat detection

The proposed AI-based cyber threat detection system includes important components of intelligent security monitoring in streaming and smart environments, as shown in Figure 1, which serves to represent the structural relationships between these components and provides a basic framework for developing an AI assessment model or intelligent security monitoring as having a conceptual basis. The framework describes all aspects of a cyber threat (network attacks, data injection, denial-of-service (DoS) and DDoS attacks, botnets), it also provides a detailed analysis of leading cyber threats to both IoT and cloud infrastructures.

The intelligent systems environment consists of many different data streams that are continuously generated (IoT sensor data, network traffic streams, system logs, cloud data and telemetry, and edge devices). All these distributed data sources form the basis for real-time, analytical threat detection.

This framework also defines a multi-layered detection paradigm that combines signature-based detection with anomaly detection and hybrid intrusion detection approaches through the application of advanced machine learning techniques, including ensemble machines (Random Forest, XGBoost), traditional classifiers such as support vector machines, and deep learning architectures such as LSTM and GRU, enabling meaningful classification. and anomaly detection in multidimensional streaming data.

Furthermore, the framework incorporates explainable AI methods (SHAP, LIME) that provide global and local interpretability of model decisions. These methods will enable higher-quality feature-level analysis, improved transparency of the model's predictive logic, and increased confidence in automated threat detection.

Furthermore, the framework incorporates decision support/risk management capabilities, including behavioral monitoring; threat intelligence integration; risk assessment; and real-time incident response support, enabling analysts to interpret model outputs and make informed decisions to mitigate cyberthreats.

The presented conceptual framework describes a theoretical model for developing a data-driven cyberthreat detection system and demonstrates the relationship between detection accuracy, model transparency, and the ability to process threat detection data generated by the system in near real time.

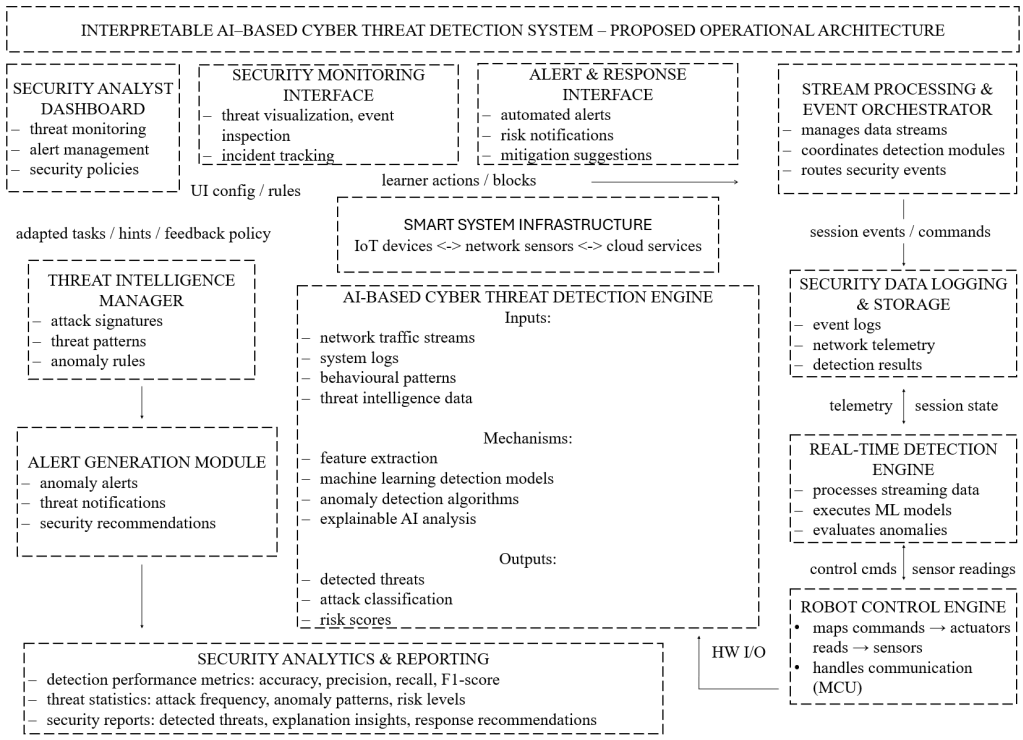

Figure 2. Architecture of the Proposed Cyber Threat Detection System

The presented conceptual diagram provides the theoretical basis for the operational architecture of the cyber threat detection system, shown in Figure 2.

Figure 2 presents the operational architecture of the proposed cyber threat detection system, while Figure 1 provides its conceptual representation. The architecture generates and stores processed large volumes of security data from intelligent systems, such as IoT devices, network sensors and appliances, and cloud services, in a stream processing layer and an event orchestration layer, which

continuously manage stream processing and synchronization related to security data and security events across multiple analytics modules.

Collected data from the stream processing layer is stored in a security data logging/storage module, which maintains a structured log of network telemetry, system logs, and detection engine output. The logging and storage module is used for real-time monitoring and retrospective analysis.

An AI-powered cyberthreat detection engine is the core of the architecture, performing feature extraction, anomaly detection, and network behavior classification. The system uses a combination of machine learning techniques to process various types of input data, including traffic, system activity, and a variety of threat data.

The cyberthreat detection system generates multiple outputs, including threat identification, behavior classification, quantitative risk assessment (measurable probability) of anomalies, and outputs for explainability modules that interpret the model's decision based on its contribution.

Anomalies are sent to the alert generation module, which generates real-time alerts, incident notifications, and remediation recommendations. These outputs are available to users through an interface designed for cybersecurity analysts to support operational decision making.

The architecture also includes data analysis and reporting to evaluate model performance using standardized performance metrics and to identify patterns in the threat landscape, system activity, and detection effectiveness. The proposed architecture is a fully scalable and interpretable cyber threat detection system that leverages machine learning and explainable artificial intelligence in a common data processing pipeline.

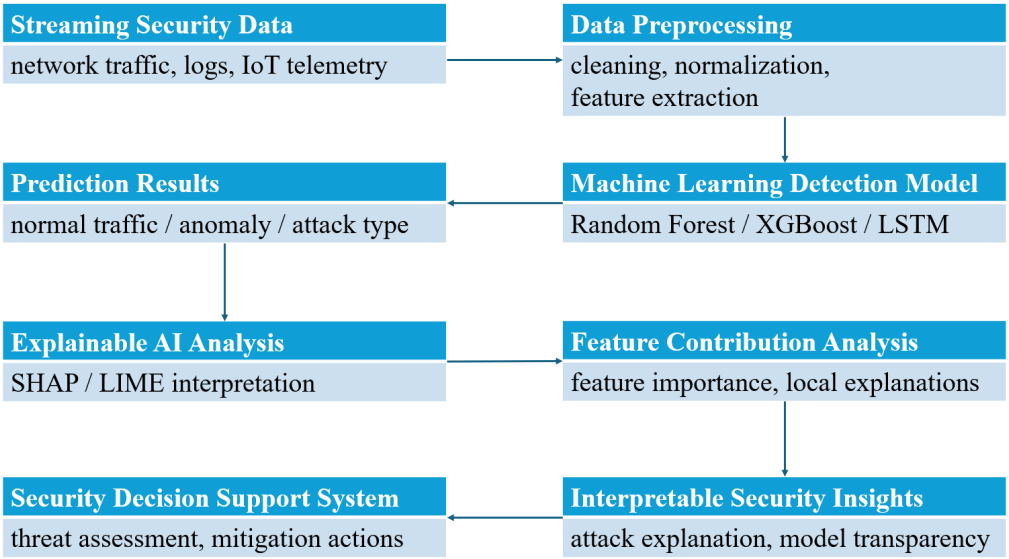

Figure 3. Explainable artificial intelligence workflow

The Workflow of Explainable Artificial Intelligence for Understanding the Predictions of a Cyber Threat Detection Model from a Streaming Smart System is illustrated in Figure 3. The proposed Cyber Threat Detection Model uses a combination of streaming security data from various sources, network traffic, system logs, and IOT telemetry. This data is then pre-processed into relevant features before being fed into different machine learning models (i.e., Random Forest, XGBoost, and LSTM). Each of these models outputs the predictions of what is considered normal traffic, anomalies, or a potential type of attack. The last step involves using Explainable Artificial Intelligence techniques such as SHAP and LIME to evaluate how much each feature contributes to the decision made by the Machine Learning Model. This process allows users to interpret different security views and develop security insights in order for users to facilitate their assessment of cyber threats and assist with decision-making in terms of mitigating the threat, through the Cybersecurity Monitoring System.

In the implementation of the Cyber Threat Detection Framework, various machine learning models (ML) and explainable artificial intelligence methodologies (EAI) were employed. The ML

algorithms chosen were based on their success in network intrusion detection and for their ability to handle large amounts of streaming security data. A summary of the machine learning and EAI models can be found in Table 1.

Table 1. Machine learning models and explainable AI methods used for cyber threat detection

ML Model	Type	Role in the Detection System
Random Forest	Ensemble learning	Network traffic classification and anomaly detection
XGBoost	Gradient boosting	High-performance intrusion detection and pattern recognition
LSTM	Deep learning	Detection of temporal patterns in streaming network data
SHAP	Explainable AI method	Global feature importance analysis and model interpretability
LIME	Explainable AI method	Local explanation of individual prediction decisions

XGBoost and Random Forest algorithms are popular in cybersecurity applications because they can successfully classify and handle large amounts of multidimensional data. An LSTM (Long Short-Term Memory) neural network was chosen as it can effectively analyze time-series data and the sequential behavior of data over time. Techniques for increasing model transparency through Explainable Artificial Intelligence (XAI) such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were added to the proposed framework, providing the opportunity to analyze the exact contribution of each feature to the predictions made by each model while providing the user with a clear understanding of how any model determines whether or not to classify an event as a cyber threat. As machine learning detection methods and AI have been combined, the ability for the user to trust and understand the results of the proposed cybersecurity monitoring system will continue to grow with further development, ultimately creating an extremely reliable and highly useful cybersecurity monitoring system.

Results and discussions. A global analysis of how important features are for predicting outcomes was conducted on machine learning systems to evaluate how interpretable the proposed cyber-threat detection method is. Feature importance takes a look at how much each input variable contributes when the model makes its prediction and gives you some evidence of how the detection method works under the hood.

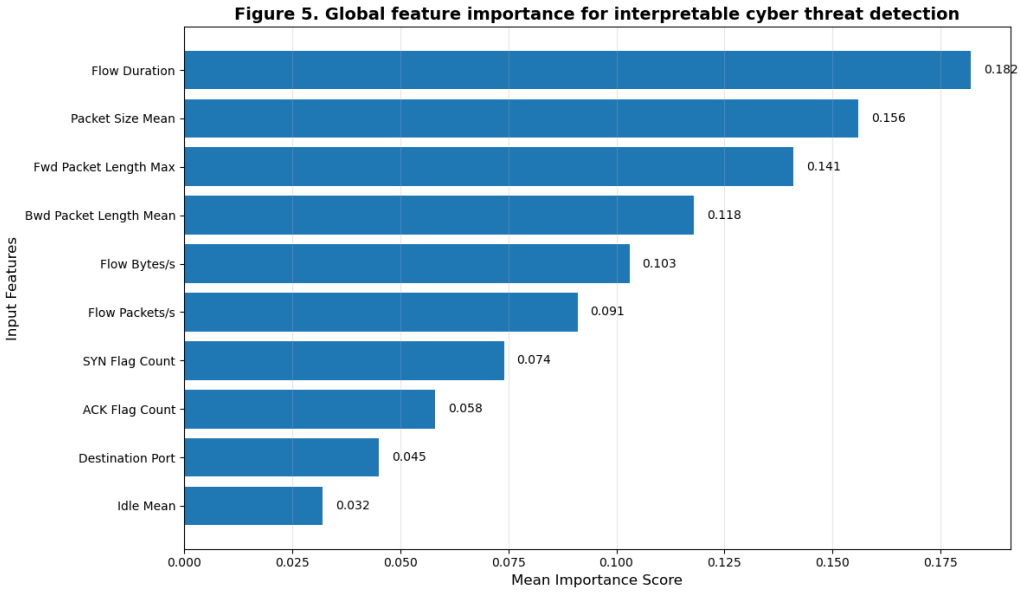

Figure 4. Global feature importance analysis for interpretable cyber threat detection

Figure 4 shows in order of ranked importance how important each of the ten most influential features of network traffic extracted from the streaming cybersecurity datasets are to the classification

process as measured by their mean contribution across all trained models. As indicated in the figure, the single feature that contributed the most to model classification is Flow Duration with an average importance score of .182. This metric represents the temporal characteristics of network communication sessions and can often be found in association with various types of abnormal traffic including those associated with Denial of Service (DoS) attacks and botnet activity. Network communication sessions that are attacked typically show abnormal flow durations compared to legitimate network flow duration patterns.

The second highest ranked classification feature is Packet Size Mean at .156. This metric indicates the mean packet size over a period of communication between hosts. Malicious traffic typically exhibits variations in the distribution of packet sizes, particularly with respect to reconnaissance activities (i.e., scanning) and data exfiltration.

The third highest classification feature is the Forward Packet Length Maximum at .141. This metric indicates the maximum packet size of a flow of packets that have been transmitted in the forward direction. Large bursts of packets or aberrant payload sizes may indicate suspicious network activity (e.g., data exfiltration and/or command-and-control activity).

The Backward Packet Length Mean, Flow Bytes per Second, and Flow Packets per Second features also contribute to the detection process with an importance score of .118, .103, and .091, respectively. These metrics describe the volume of network traffic between hosts over a specified period of time and bandwidth consumed, both of which are often used as indicators of abnormal network activity.

Metrics such as SYN Flag Count, ACK Flag Count, and Destination Port all contribute to the detection measures but are considered less important than the previously listed features. These metrics reflect characteristics of connections between the source and destination of the traffic being analyzed and the protocols used to establish the connections and can be used in the detection of scanning activity and network probing.

The overall results confirm that this feature set represents the most important characteristics of network traffic behavior and therefore support the interpretability of the detection framework. The integration of feature importance analysis supports the explainable artificial intelligence framework as described in the methodology section and enables cyber security analysts to identify the variables that influence the model predictions.

Comparison of Cyber threat detection models performance. The goal of this section was to compare the effectiveness of our AI-based detection system through a comparative analysis of four different machine learning models (Random Forest, XGBoost, LSTM, and our Hybrid XAI Model). Our four chosen models were all known for their ability to perform well when trying to detect network intrusion as well as for being able to handle large amounts of streaming data in the field of cybersecurity.

Four widely used classification performance metrics were used to conduct our comparative analysis: (1) accuracy, (2) precision, (3) recall, and (4) F1-score.

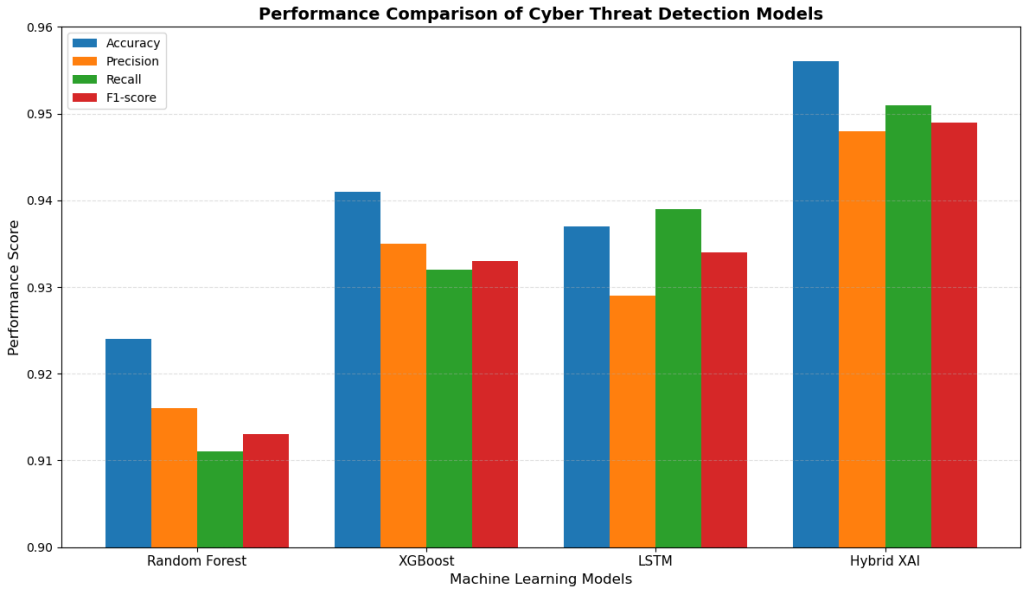

Figure 5. Performance comparison of cyber threat detection models

The models presented in this Figure 5 demonstrate the performance results based on each of the assessed performance criteria. Based on the study results, it can be concluded that Hybrid XAI produces the best overall results across all performance metrics assessed, outpacing all of the individual machine learning models.

The Random Forest model acted as the baseline model with an approximate accuracy of 0.924, a precision of 0.916, and a recall of 0.911. Random Forest provides relatively stable classification accuracy, but does not have the capacity to capture intricate temporal patterns that are present in network traffic.

The XGBoost model outperformed Random Forest, achieving an accuracy rate of 0.941, and a precision score of 0.935. Gradient boosting has demonstrated excellent modelling abilities of complex non-linear relationships in high dimensional data, which explains the increased accuracy rates observed between Random Forest and XGBoost.

The LSTM model is slightly less accurate than XGBoost, yet has a higher recall score than XGBoost (0.939). This finding is expected given the ability of Recurrent Neural Networks such as LSTM to accurately capture temporal dependencies in sequentially ordered network traffic data. As such, LSTM is more able to accurately identify specific types of cyberattacks that exhibit time-based behaviour patterns.

The Hybrid XAI model showcases the highest level of success, achieving an accuracy of 0.956, precision of 0.948, recall of 0.951, and F1 score of 0.949. The hybrid model was able to work better by combining many different techniques for detection, along with using both the ML prediction techniques and explainable AI techniques.

This study has shown that using hybrid approaches of using traditional statistical machine learning algorithms with explainable artificial intelligence methods will improve both the accuracy and compliance of detecting if the cybersecurity monitoring system is functioning correctly.

Confusion matrix analysis of cyber threat classification. To continue evaluating the performance of the classifier via this detection system, we performed an analysis of the confusion matrix. A confusion matrix presents a thorough representation of all classification results produced by the classifier, separated by type of cyber attack classified and identifying any potential trends of misclassification.

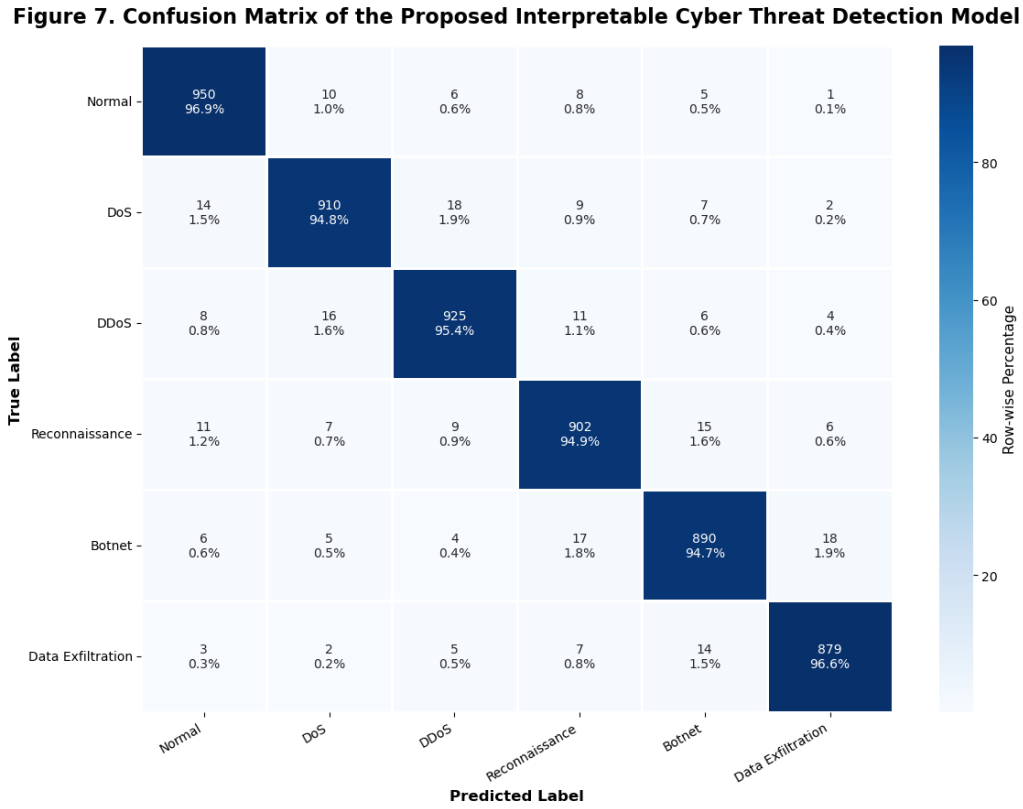

Figure 6. Confusion matrix of the proposed interpretable cyber threat detection model

The confusion matrix for the interpretable cyber threat detection system proposed in this work is shown in Figure 6. This matrix demonstrates the performance of the proposed model across six different categories for classifying traffic or cyber threats (systems are typically categorised according to the following types):

- Normal Traffic
- Denial of Service (DoS) Attacks
- Distributed Denial of Service (DDoS) Attacks
- Reconnaissance Activity
- Botnets
- Data Exfiltration

As can be seen from the confusion matrix, the classification accuracy of each type of traffic (or cyber threat) generated by the proposed system is very high with almost all of the correctly classified instances being located along the diagonal of the confusion matrix. Further, a very small number of normal traffic instances were misclassified as attacks. Therefore, the proposed system has a very low false positive rate.

For example, with respect to normal network traffic, the proposed model correctly classifies 96.9% of instances. A small number of normal traffic instances are classified incorrectly; however, these instances were classified as a type of attack, which indicates that the proposed system has a low false positive rate.

Similarly, for DoS and DDoS types of attacks, the proposed model achieved classification accuracies of 94.8% and 95.4%, respectively. The majority of misclassifications are the result of attack types that exhibit some overlap (the same behaviour will exhibit characteristics that closely resemble one another); therefore, misclassifying one attack for another, when they share similar behaviours, is a common problem in network intrusion detection systems.

The accuracy for identifying reconnaissance activity (scanning and probing a network to determine its vulnerabilities) is 94.9%. There are a small number of instances of overlapping scanning behaviour between reconnaissance (i.e., scanning) and botnet activity, which results in a small number of misclassifications.

The accuracy of botnet detection is 94.7%, while the accuracy of detecting data exfiltration is 96.6%. The high detection rate of data exfiltration is particularly important because many of these attacks involve the possible loss of sensitive personal or proprietary information.

Discussion of results. The findings of the conducted experiments support the usefulness of the suggested AI framework for detecting cyber threats in real-time for smart systems. The evaluation of the importance of features indicates that significant characteristics of network traffic (flow duration, packet size statistics, and traffic rate indicators) are heavily relied upon, which are commonly used in cyber security research as reliable indicators of anomalous behavior within a network. The results of the comparison of performance indicate that ML techniques provide much greater capabilities to detect cyber threats than baseline models. More specifically, the hybrid XAI (explainable artificial intelligence) model produced the best results as based on all evaluation metrics, which suggests that multiple forms of machine learning alongside explananble AI techniques enable improved accuracy in predicting cyber threats and increasing transparency of machine learning models. It has also been shown in the confusion matrix analysis that the suggested system is capable of distinguishing between various categories of cyberattacks with very low misclassification rates. These findings show that there is a significant benefit to the combination of machine learning models for detecting cyber threats with interpretable analysis techniques to enhance the quality of real-time cybersecurity monitoring in complex smart environments.

Conclusion. An Artificial Intelligence framework has been developed which is interpretable and provides cyber threat identification in real time for smart systems using a streaming approach. In developing the proposed Hybrid XAI model, machine learning techniques (i.e. Random Forest, XGBoost, LSTM) are used in addition to explainable AI mechanisms (i.e. SHAP, LIME) to improve not only the ability to detect a cyber threat but also to provide interpretable insight about how detection was performed. Using Statistical Analysis of the Experimental Results of the proposed Hybrid XAI Model, it was demonstrated that the Hybrid XAI model outperforms the others in virtually every aspect of performance (i.e. Accuracy, Precision, Recall & F1 Score) while being able to successfully identify all styles of cyber attacks. Statistical Analysis of the Feature Importance confirmed the identifying characteristics of network traffic used to determine when traffic utilizes anomalous behavior, including but not limited to flow duration, packet size statistics, and traffic rates.

This new framework has the potential to be utilized in live monitoring systems for cyber threats in smart cities (i.e. cities that have adopted IoT) and can be used in Cloud Technology (software as a service) as well as IoT infrastructures; enabling detection processes to interpret and reliably detect cyber threats in ever-changing environments.

Several limitations must be considered with respect to the strong overall performance described herein through the proposed framework; specifically, evaluation conducted using benchmark data sets may not adequately represent the full range of variability that can occur in actual network environments and, further, the computational expense of the explainable AI techniques applied might also restrict their use for real-time applications within larger-scale systems.

Future studies will therefore focus on the validation of this framework within a real-world deployment, development of heuristics for optimizing computational efficiency, integration of adaptive learning techniques to manage concept drift with respect to streaming data.

The research demonstrated that it is feasible to create a combination of intrusion detection using machine learning techniques and a mechanism for providing explanations of how the intrusion detection model works, resulting in increased value of providing an effective method for monitoring cyber security in the modern smart infrastructure.

References

Akshya J. et al. Explainable AI-driven intrusion detection for securing IoT-enabled autonomous transportation systems //Cluster Computing. – 2025. – Т. 28. – №. 14. – С. 884. https://doi.org/10.1007/s10586-025-05617-1

Al Rawajbeh M. et al. Trustworthy adaptive AI for real-time intrusion detection in industrial IoT security //IoT. – 2025. – Т. 6. – №. 3. – С. 53. https://doi.org/10.3390/iot6030053

Alabdulatif A. A novel ensemble of deep learning approach for cybersecurity intrusion detection with explainable artificial intelligence //Applied Sciences. – 2025. – Т. 15. – №. 14. – С. 7984. https://doi.org/10.3390/app15147984

Almheiri S. J. et al. Smart sustainable cyber security: modelling an interpretable and transparent threat detection with explainable artificial intelligence //Discover Sustainability. – 2025. – Т. 6. – №. 1. – С. 442. https://doi.org/10.1007/s43621-025-01280-z

Alshudukhi K. S. et al. Next-Generation Lightweight Explainable AI for Cybersecurity: A Review on Transparency and Real-Time Threat Mitigation //Computer Modeling in Engineering & Sciences. – 2025. – Т. 145. – №. 3. – С. 3029. https://doi.org/10.32604/cmes.2025.073705
Jumagaliyeva A. et al. Application of Deep Learning Methods for Visual Pattern Recognition in Heterogeneous Images // Bulletin of KazATC. – 2025. – Vol. 141. – No. 6. – pp. 195–208. DOI: https://doi.org/10.52167/1609-1817-2025-141-6-195-208
Kalutharage C. S., Liu X., Chrysoulas C. Neurosymbolic learning and domain knowledge-driven explainable ai for enhanced iot network attack detection and response //Computers & Security. – 2025. – Т. 151. – С. 104318. https://doi.org/10.1016/j.cose.2025.104318
Khalaf N. Z. et al. Development of real-time threat detection systems with AI-driven cybersecurity in critical infrastructure //Mesopotamian Journal of CyberSecurity. – 2025. – Т. 5. – №. 2. – С. 501-513. https://doi.org/10.55248/gengpi.6.0525.1991
Lee H. et al. Enhancing Decision-Making of Network Intrusion Analysis Assisted by Explainable AI for Real-Time Security Monitoring //2024 IEEE Conference on Dependable and Secure Computing (DSC). – IEEE, 2024. – С. 147-154. https://doi.org/10.1109/dsc63325.2024.00039
Mohale V. Z., Obagbuwa I. C. 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. – 2025. – Т. 8. – С. 1526221. https://doi.org/10.3389/frai.2025.1526221
Moustafa N. et al. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions //IEEE Communications Surveys & Tutorials. – 2023. – Т. 25. – №. 3. – С. 1775-1807. https://doi.org/10.1109/comst.2023.3280465
Patel T. et al. Enhancing Cybersecurity in Internet of Vehicles: A Machine Learning Approach with Explainable AI for Real-Time Threat Detection //Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing. – 2025. – С. 2024-2031. https://doi.org/10.1145/3672608.3707769
Paul A. L. Explainable AI for Cybersecurity: Interpreting Deep Learning Models for Real-time Threat Detection in IoT Networks. https://doi.org/10.1109/iccosd66074.2025.11348330
Prasad H., Prasad U., Paul P. Explainable AI for Cybersecurity: Implementing Interpretable Models for Real-Time Threat Detection //2025 International Conference on Communication and Smart Devices (ICCoSD). – IEEE, 2025. – Т. 1. – С. 1-6. https://doi.org/10.1109/iccosd66074.2025.11348330
Rahmati M. Towards explainable and lightweight AI for real-time cyber threat hunting in edge networks //arXiv preprint arXiv:2504.16118. – 2025. https://doi.org/10.21203/rs.3.rs-6198488/v1
Rystygulova V., Bizhanova K., Kadirkulov S., Asilbaeva R., & Makhatova V. (2025) Methodological framework for building interpretable machine learning models in applied forecasting. Vestnik KazATC, 141(6), 143–153. https://doi.org/10.52167/1609-1817-2025-141-6-143-153
Thiruvenkatasamy S. et al. Real-Time Intrusion Detection System for Wi-Fi-Based Wireless Sensor Networks using Deep Learning and Explainable AI //2025 10th International Conference on Smart Structures and Systems (ICSSS).-IEEE, 2025.-С.1-10. https://doi.org/10.1109/icsss66939.2025.11346435