EXPLAINABLE AI BASED CYBERSECURITY THREAT DETECTION

Seena K R

Lecturer in Computer Engineering

Government Polytechnic College Palakkad, Kerala, India

ABSTRACT

The growing complexity and number of cyberattacks are causing big problems for traditional cybersecurity and intrusion detection systems. Machine learning and deep learning methods have shown great promise in automatically finding harmful activities and unusual network behavior. However, a lot of advanced AI-based detection models work like black boxes, which means they don't give clear reasons for their results. This lack of transparency can make users less trusting, make it harder for security analysts to decide how to act, and make it more difficult to investigate cyber incidents. Explainable Artificial Intelligence (XAI) offers a good solution to these issues by allowing cybersecurity systems to provide clear, human-readable explanations for their automated threat detection decisions. This research introduces an Explainable AI-Based Cybersecurity Threat Detection framework that combines machine learning with XAI methods like Shapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME).The goal of this approach is to identify and categorize cybersecurity threats while also pinpointing the main network and behavioural factors that led to each prediction. The framework measures detection performance using accuracy, precision, recall, F1-score, false-positive rate, and detection speed. It also assesses the quality of explanations based on factors like fidelity, stability, consistency, interpretability, and how efficient the system is. The research also looks at adversarial robustness and human-centered security analysis to check how well the explainable threat detection approach works in real-world situations. By blending strong prediction abilities with clear and understandable decision-making, the framework is designed to help with trusted, efficient, and practical cybersecurity threat detection in changing network environments.

Keywords: *Explainable Artificial Intelligence (XAI); Cybersecurity; Threat Detection; Intrusion Detection System (IDS); Machine Learning; Deep Learning; SHAP; LIME; Network Security; Explainable Intrusion Detection; Trustworthy AI; Adversarial Robustness.*

INTRODUCTION

The quick growth of digital technologies, cloud computing, IoT devices, mobile networks, and connected information systems has made people, businesses, and important infrastructure heavily rely on networked environments. At the same time, the cybersecurity situation has become more complicated, as attackers are always coming up with new and advanced methods to find weaknesses, break into systems, steal important data, stop services, and avoid standard security tools[1]. Traditional cybersecurity methods, especially those based on known attack signatures, are good at finding attacks that have been seen before, but they struggle to spot new, changing, and zero-day threats. Because of this, there is a growing need for smart and flexible solutions to detect

harmful activities in constantly changing network settings. Machine learning (ML) and deep learning (DL) have become valuable tools for automating the detection of cybersecurity threats. These technologies can look through large amounts of network data and learn complex patterns linked to harmful or safe activities.ML models like Decision Trees, Random Forest, Support Vector Machines, Logistic Regression, and XGBoost have been widely studied for detecting intrusions[1]. Similarly, deep learning methods, such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, can find complicated and nonlinear relationships in cybersecurity data. These approaches have shown promise in improving the automation, scalability, and effectiveness of modern intrusion detection systems.

Despite these benefits, the growing complexity of ML and DL models has caused a major issue: they are often hard to understand. Many advanced AI models operate like black boxes, where it's difficult to see how input data leads to a prediction. For instance, an AI-based intrusion detection system may flag a network connection as harmful with high confidence, but a security analyst might not know which features of the network traffic led to that result. In cybersecurity, this lack of transparency can reduce trust in automated alerts, make it harder to investigate incidents, and create challenges in telling real threats from false alarms[1].

Explainable Artificial Intelligence (XAI) has become an important solution for overcoming these problems. AI aims to provide clear and human-friendly explanations of AI decisions. Instead of just stating that an event is malicious, an explainable system can point out the features that influenced the decision and show how important those features are. This helps security analysts check alerts, investigate incidents, understand how attacks work, and make better responses. XAI can also help model developers by showing whether a machine learning model is using relevant cybersecurity factors or picking up on misleading patterns. Several XAI methods have been studied for use in cybersecurity[2]. Shapley Additive explanations (SHAP) gives values that show how each feature affects a model's prediction. SHAP can explain both individual network events and the overall behaviour of a model. Local Interpretable Model-Agnostic Explanations (LIME) explains a prediction by using a simpler model to approximate the behavior of a more complex one near a specific data point. Other methods, such as permutation feature importance, partial-dependence analysis, rule-based explanations, and feature-attribution techniques, can also help in understanding AI-based threat detection. Still, integrating XAI into cybersecurity is an ongoing research challenge. Just giving an explanation doesn’t mean it’s accurate, stable, easy to understand, or useful for a security analyst. Different XAI techniques might provide different explanations for the same prediction. Also, generating explanations can add extra processing time, which could be a problem in real-time intrusion detection systems. There’s also a security risk because explanations might reveal how a model works, allowing attackers to find ways to bypass it. So, explainable cybersecurity systems need to be checked not just for their accuracy in identifying threats, but also for how clear, consistent, efficient, and helpful their explanations are. Another big challenge is the constantly changing nature of cybersecurity threats. Network traffic and attack methods keep evolving, which can cause models to perform worse over time. A model

trained on old cybersecurity data might work well in experiments but fail in real-world situations where things are constantly changing. So, an effective XAI-based cybersecurity system must be able to track changes in both how well it predicts threats and how clear its explanations are[3].

This means using a more complete way to check how well the system works, looking at things like how accurate it is at finding threats, how well it explains its decisions, how strong it is against attacks, how fast it works, and how well it can change with new situations. Right now, some research shows that using machine learning with XAI can help detect intrusions, but many of these studies focus only on how well the system classifies threats or how well it shows which features are important[3].

There's a need for a more complete approach that looks at how well the system detects threats, how good its explanations are, how efficient it is, how strong it is against attacks, and how well it helps people make decisions. This is especially important in security work, where human analysts look at and act on the alerts AI systems generate.

That's why this research introduces an Explainable AI-Based Cybersecurity Threat Detection Framework. It combines machine learning for identifying threats with XAI tools like SHAP and LIME. The framework is meant to find and classify cybersecurity threats while also showing which features led to each prediction. In addition to standard measures like accuracy, precision, recall, F1-score, false-positive rate, and how fast it detects threats, this research also looks at how well the explanations match the actual results, how stable and consistent they are, how much computing power they use, and how helpful they are for analysts[1-3].

RELATED WORK

As cyber-attacks become more frequent and clever, researchers are looking for smart ways to protect modern communication networks [1], [2].Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are now widely used to improve intrusion detection because they can find hidden patterns in large amounts of network data [3], [4].Many studies show that AI-based security systems are more adaptable than old-fashioned rule-based methods. However, there are still practical issues like understanding how these systems work, scaling them up, and using them in real time. Older intrusion detection methods mainly used signature matching to find dangerous activities. These methods check incoming network traffic against a list of known attack patterns and send warnings when a match is found. While signature-based systems are good at finding known attacks, they can't handle new or unknown threats like zero-day attacks because those aren't in the signature database. Keeping the signature database and security rules up to date also makes these systems less effective in fast-changing network environments [4], [1].As cybersecurity research has grown, machine learning algorithms have been used to fix the problems with manually created detection rules. Different classification methods like Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour have been studied to detect harmful network activity using historical data. These methods often give better detection results than traditional signature-based techniques. Still, their success depends a lot on how well the data is

prepared and the availability of good quality labelled datasets, which limits their ability to keep up with rapidly changing attack methods.

A. Artificial Intelligence for Cybersecurity Threat Detection

Artificial intelligence has become a key area of study in cybersecurity because it can handle large amounts of data from networks and systems. Unlike traditional systems that rely on known attack signatures, machine learning can detect patterns in behaviours and statistics that are linked to harmful activities. In supervised learning, algorithms use data that is already labelled as either harmful or safe to learn how to connect network features with different types of attacks. Unsupervised learning tries to spot unusual patterns without needing full knowledge of all attacks, while semi-supervised learning mixes labelled and unlabelled data. Some commonly studied methods include Random Forest, Support Vector Machine, Decision Tree, k-Nearest Neighbour, Logistic Regression, and gradient-boosting techniques. Deep learning models offer another way to learn more complex patterns from big datasets. Even though these techniques can be very effective at detecting threats, their growing complexity has led to a need for clearer explanations of how decisions are made[4], [1].

B. Explainable Artificial Intelligence

Explainable Artificial Intelligence is about making AI systems easier for people to understand, explain, or examine. In cybersecurity, explainable AI helps answer questions like:

- Why was this connection marked as harmful?
- Which parts of the data influenced the decision?
- Which parts suggested it was safe?
- How much did each part affect the result?
- Does the result match what we know about attacks?
- How sure should a security analyst be about the result?

Being able to explain AI decisions is important, and it connects with the NIST AI Risk Management Framework. This framework lists explainability, interpretability, transparency, security, resilience, reliability, privacy, and accountability as key parts of trustworthy AI. Artificial Intelligence and deep learning have grown into strong tools for dealing with the growing complexity in cybersecurity [7], [8].Today's computer networks create a lot of different kinds of data that change over time, which makes it hard for traditional security tools to find hidden threats quickly. Deep learning algorithms can automatically spot complicated patterns in network traffic, helping to catch harmful activities more accurately. However, many deep learning models are hard to understand, which makes them less useful in real-world security settings. To fix this, a new system combines Long Short-Term Memory (LSTM), Autoencoder, and Explainable Artificial Intelligence (XAI) to offer better, easier to understand, and more trustworthy ways to find cyber threats. Using these methods together helps the system look at how networks behave over time, find unusual activity, explain why certain threats are detected, and help with smart risk evaluations for better cybersecurity [1]–[7].

RESEARCH GAP

Despite big progress in artificial intelligence (AI), machine learning (ML), and deep learning (DL) for finding cyber threats, there are still important areas that need more research. Many intrusion detection systems (IDS) can do a good job at predicting threats, but many complex models work like black boxes. This makes it hard for security analysts to know why a certain network activity is labelled as dangerous. Explainable AI (XAI) tools like SHAP and LIME help with this a little, but their use in cybersecurity isn't well developed yet. First, most studies look at detection performance and explainability separately. They usually report accuracy, precision, recall, F1-score, and false-positive rate, but they don't measure explanation quality well. There isn't a single way to check both how well a model works and how good its explanations are, including factors like how stable, consistent, easy to understand, fast, and useful the explanations are[1]–[7].

Second, the link between how well a model performs and the quality of its explanations isn't clear. A very accurate model doesn't always give good or useful explanations. On the other hand, a model that's easy to understand might not be as good at detecting threats. More research is needed to see how different machine learning or deep learning models affect the quality and trustworthiness of their explanations. Third, getting explanations quickly is a big challenge. Cybersecurity systems have to look at a lot of network data fast to find threats. Even though SHAP and LIME can provide helpful explanations, making those explanations takes extra time. There isn't enough research looking into whether these XAI methods can give fast enough explanations for real-time or almost real-time security operations[1]–[7].

Fourth, how well XAI explanations hold up against attacks isn't well studied. Attackers might try to change input data to avoid being caught or to mess with the explanations given by XAI tools. So, explainability should help with transparency but also stay strong against bad actors or new kinds of attacks. Fifth, how well models and their explanations work across different data sets and changing network environments isn't well understood. Many studies are tested on single datasets. But cybersecurity settings keep changing because of new threats, network setups, user behaviours, and traffic patterns. Models built using one dataset might not work well in another setting. More research is needed to check how well models adapt to different datasets and changing environments. Sixth, how well explanations help people isn't studied much. Most research evaluates explanations using technical measures, but they don't check if the explanations actually help security analysts make quicker or better decisions. An explanation is useful not just because it matches the model's math, but because it helps the analyst understand, check, and respond to a threat. Finally, there's not much research that puts together detection, explanation, robustness, speed, and usefulness all in one cybersecurity system. Recent reviews show these issues are still big challenges in using XAI for intrusion detection systems[1]–[7].

Identified Research Gaps

Research Area	Existing Limitation	Proposed Research Direction
Detection performance	Primarily focuses on accuracy-based metrics	Combine detection performance with explanation quality
Explainability	Often evaluated qualitatively	Develop measurable fidelity, stability, consistency and interpretability measures
SHAP vs. LIME	Limited comparative evaluation across models	Systematically compare SHAP and LIME
Real-time detection	XAI can introduce computational overhead	Evaluate explanation latency and computational efficiency
Adversarial robustness	Limited attention to manipulation of explanations	Test robustness against adversarial inputs
Dataset generalisation	Many studies use individual benchmark datasets	Evaluate performance across diverse datasets/environments
Concept drift	Dynamic changes in cyber threats are often overlooked	Examine model and explanation stability over time
Human-centred security	Limited involvement of security analysts	Evaluate whether explanations improve analyst decision-making
Integrated framework	Detection and XAI are frequently studied independently	Develop an integrated explainable cybersecurity threat-detection framework

Research Objectives

The study will pursue the following objectives:

1. To develop an LSTM-based model for sequential cybersecurity threat classification.
2. To develop an Autoencoder-based mechanism for detecting anomalous network behaviour.
3. To integrate SHAP and LIME to explain AI-based threat predictions.
4. To develop a quantitative risk-scoring mechanism combining classification confidence and anomaly severity.
5. To evaluate the detection performance of the proposed framework using established cybersecurity metrics.
6. To compare SHAP and LIME in terms of explanation quality, consistency, stability, and computational cost.
7. To evaluate the generalisation of the proposed framework across different cybersecurity datasets.
8. To investigate the robustness of the framework against previously unseen and potentially adversarial network behaviour.
9. To examine the usefulness of generated explanations for cybersecurity analysis and decision-making.
10. To establish an integrated evaluation framework for accuracy, explainability, robustness, and operational efficiency.

Proposed System Architecture

The proposed framework consists of the following stages:

Network Traffic
↓
Data Collection
↓
Data Preprocessing
↓
Feature Engineering and Normalisation
↓
Sequential Data Construction
↓
LSTM Threat Classification
↓
Autoencoder Anomaly Detection
↓
SHAP/LIME Explanation
↓
Explanation Validation
↓
Integrated Risk Scoring
↓
Risk Classification
↓
Security Recommendation

The architecture therefore combines supervised and unsupervised learning.

The LSTM provides a classification perspective, while the Autoencoder provides an anomaly perspective. XAI provides interpretability, and risk assessment combines the outputs into an operational priority.

Methodology

Data Collection

The research can employ publicly available cybersecurity datasets such as:

- CIC-IDS2017
- CSE-CIC-IDS2018
- UNSW-NB15
- CIC-DDoS2019
- Bot-IoT

- ToN-IoT

The final dataset selection should depend on the specific experimental objectives and availability of suitable sequential traffic features[1]–[7].

Data Preprocessing

Preprocessing will include:

1. Removal of duplicate observations.
2. Treatment of missing values.
3. Identification of invalid or infinite values.
4. Categorical-feature encoding.
5. Numerical-feature normalisation.
6. Class-label encoding.
7. Feature selection.
8. Construction of temporal sequences for LSTM processing.

Care should be taken to prevent data leakage between training and testing sets[7].

LSTM Training

The LSTM will receive sequences of network observations and learn temporal representations. The final hidden representation will be passed to a classification layer.

Possible output categories include:

- Normal
- DoS/DDoS
- Brute Force
- Botnet
- Probe/Scanning
- Web Attack
- Other attack classes represented in the selected dataset.

The exact categories should be determined by the selected dataset[1]–[8].

XAI Integration

After classification, SHAP and LIME will be applied to selected predictions.

The explanations will identify:

- Most influential network features.
- Direction of feature contribution.
- Relative importance of features.
- Differences between legitimate and malicious predictions.
- Features contributing to high-risk decisions.

Risk Assessment

The LSTM prediction probability and normalised Autoencoder anomaly score will be combined using Equation (7).

The resulting risk score will determine the priority of an alert.

For example:

Risk Level	General Interpretation	Potential Response
Low	Minor deviation	Monitor
Medium	Suspicious behaviour	Verify/investigate
High	Strong evidence of attack	Detailed investigation/containment
Critical	Very high threat probability and anomaly	Immediate response according to security policy

These response categories represent the proposed framework and should be validated within the intended operational environment[5]–[7].

Expected Results

Because experimental implementation and data analysis have not yet been completed, the following should be considered **expected outcomes rather than experimental findings**.

The proposed framework is expected to:

1. Improve automated detection of complex network threats.
2. Detect anomalous traffic that differs from learned normal behaviour.
3. Provide feature-level explanations for individual threat predictions.
4. Improve transparency compared with an unexplained deep-learning model.
5. Enable risk-based prioritisation of cybersecurity alerts.
6. Identify the network features most strongly associated with malicious behaviour.
7. Demonstrate measurable differences between SHAP and LIME in explanation generation.
8. Identify computational trade-offs associated with XAI.
9. Provide insights into the generalisation of AI-based threat detection.
10. Establish whether explanation quality and detection performance can be evaluated within a unified framework.

The study should not claim that the proposed framework improves accuracy until this has been demonstrated experimentally.

DISCUSSION

The new framework tackles a key problem with regular AI systems used for cybersecurity: they usually separate the parts that predict threats from the parts that explain why they make those predictions. A good system might spot dangers well, but it might not give enough details for experts to confirm if an alert is real[1]–[7]. By using explainable AI, the system can show exactly which parts of the data led to its decision. This becomes really helpful when false alarms take up a lot of time and effort from experts. The Autoencoder adds another way of looking at things by

focusing on normal network activity instead of just looking for known attacks. Mixing methods that use labelled data with those that don't could give a more complete picture. The risk assessment tool then turns the model's results into clear priorities. Instead of showing many alerts that all seem equally important, it can highlight which ones need immediate attention. But this approach also has some downsides. LSTM models need a lot of data and computing power. The thresholds used in the Autoencoder can be affected by changes in how networks behave. The explanations from SHAP and LIME might use extra computing resources and might not always line up in how they rank features. Also, the datasets used for testing may not fully reflect real-world network conditions. These challenges show why it's important to test the system across different datasets, check how reliable it is, evaluate the quality of its explanations, and involve people in assessing how well it works[1]–[8].

LIMITATIONS

There are several important things to be aware of. First, the standard test data sets might not show the full picture of what happens in real businesses or important systems. Second, if some types of data are seen more often than others, this could change how well a model is judged. Third, the settings for how much error is allowed might need to be adjusted depending on the specific situation. Fourth, tools like SHAP and LIME can give different reasons for the same result. Fifth, the extra work needed to run these tools might make it hard to use them right away. Sixth, bad inputs, like tricky attacks, can mess up both the models that detect problems and the tools that explain how they work. All of these points need to be looked into carefully, not just treated as small problems that can be fixed easily[1]–[7].

CONCLUSION

This study introduces a new cybersecurity threat detection system that uses Explainable AI, combining LSTM, Autoencoder, SHAP, LIME, and a risk assessment method. The system is built to fix the shortcomings of traditional security tools and deep learning models that are hard to understand. The LSTM part helps detect threats by understanding the order of events in network traffic. The Autoencoder finds unusual activities by learning what normal traffic looks like and pointing out traffic that doesn't fit. SHAP and LIME help users understand why the system made certain decisions by showing which parts of the data influenced the outcome. The risk assessment part uses how certain a threat is and how severe it is to group detected events into clear risk levels. The main point of this research is to combine good detection with transparency, strong performance, fast processing, adaptability, and practical use. It highlights that a good cybersecurity AI system should not only say if an event is dangerous, but also give clear reasons so experts can check, decide, and act on it properly. Recent studies show that even with Explainable AI, there are still major issues like setting clear standards, handling large volumes of data, explaining results quickly, being safe from attacks, and measuring how well the system works from a human perspective. Therefore, carefully testing these areas is a key area for future research in cybersecurity. The framework presented here gives a good starting point for creating AI-based cybersecurity systems that work well with both machines and human experts.

REFERENCES

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
3. 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.
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
5. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).
6. Islam, S. R., Eberle, W., Ghafoor, S. K., Siraj, A., & Rogers, M. (2020). Domain knowledge aided explainable artificial intelligence for intrusion detection and response. CEUR Workshop Proceedings, 2600.
7. Nerkar, N., Nile, A., Kulkarni, A., Kadlag, O., & Gadakh, P. (2020). Explainable AI in intrusion detection systems. International Journal of Psychosocial Rehabilitation, 24(6), 12508–12516.
8. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.