

Explainable Artificial Intelligence for Transparent Network Threat Detection Systems

Mr. Aditya V. Jadhav

Adsul's Technical Campus
Chas, Ahilyanagar
Maharashtra, India
adityajadhav7779@gmail.com

Dr. Hemantkumar B. Jadhav

Professor, Adsul's Technical
Campus
Chas, Ahilyanagar
Maharashtra, India
hem3577@gmail.com

Prof. Jagruti R. Mahajan

Assistant Professor, Adsul's
Technical Campus
Chas, Ahilyanagar
Maharashtra, India
mahajan.jagruti@gmail.com

Prof. Pragati B. Chandane

Assistant Professor, Adsul's Technical Campus
Chas, Ahilyanagar
Maharashtra, India
pragatichandane3@gmail.com

Dr. Pradeep M. Patil

Principal, Adsul's Technical Campus
Chas, Ahilyanagar
Maharashtra, India
drpmp66@gmail.com

Abstract—Explainable Artificial Intelligence (XAI) has emerged as a critical solution to improve transparency and trust in AI-based cybersecurity systems [31]. Traditional intrusion detection systems are ineffective against modern cyber threats such as zero-day attacks and advanced persistent threats [23]. Although machine learning models improve detection accuracy, their black-box nature limits interpretability [24]. This paper proposes an XAI-based network threat detection framework that integrates machine learning models with explainability techniques such as SHAP and LIME [28]. The proposed system enhances decision transparency, reduces false positives, and improves trust in automated threat detection. Experimental analysis demonstrates that the approach maintains high detection accuracy while providing meaningful explanations for predictions. Furthermore, this work examines different datasets and machine learning models, including supervised and unsupervised approaches, to assess their effectiveness when combined with explainability techniques. The results demonstrate that incorporating XAI not only improves transparency but also helps in identifying false positives and refining detection strategies. Overall, this research contributes to the development of more trustworthy, efficient, and interpretable network security systems.

Keywords: Explainable AI, Cybersecurity, Intrusion Detection System, Machine Learning, SHAP, LIME

I. INTRODUCTION

In recent years, the rapid growth of digital communication and internet-based services has significantly increased the risk of cyber threats and network attacks. Organizations rely heavily on network security systems to protect sensitive data and maintain operational continuity. Traditional approaches to network threat detection, such as signature-based and rule-based systems, are no longer sufficient to handle the complexity and evolving nature of modern cyberattacks. As a result, machine learning (ML) and deep learning (DL) techniques have been widely adopted to detect anomalies and malicious activities in network traffic [24].

While these intelligent systems provide high detection accuracy, they often operate as black-box models, meaning their internal decision-making processes are not easily interpretable by humans. This lack of transparency creates several challenges, including reduced trust among cybersecurity analysts, difficulty in debugging models, and challenges in regulatory compliance. In critical sectors such as banking, healthcare, and government systems, understanding how and why a system identifies a threat is just as important as the detection itself.

Explainable Artificial Intelligence (XAI) addresses this issue by making the decision-making process of AI models more transparent [31] and understandable. By providing clear explanations for predictions, XAI enables security professionals to validate results, detect errors, and improve overall system reliability. This research focuses on integrating XAI techniques into network threat detection systems to enhance both interpretability and performance. It aims to bridge the gap between high accuracy and human understanding, ensuring that AI-driven cybersecurity solutions are both effective and trustworthy.

Furthermore, the increasing sophistication of cyberattacks, such as zero-day exploits, ransomware, and advanced persistent threats (APTs), demands more intelligent and adaptive detection mechanisms [5]. Traditional systems often fail to detect such attacks due to their reliance on predefined rules and signatures. In contrast, AI-based systems can learn patterns from data and adapt to new types of threats [25], [26]. However, without proper interpretability, these systems may produce decisions that are difficult to justify, especially in high-stakes environments.

The integration of XAI not only enhances transparency but also improves collaboration between automated systems and human experts. By providing feature-level explanations [27], [28] and visual insights, analysts can better understand attack patterns and refine detection strategies. This leads to improved incident response times and more effective threat mitigation.

In addition, regulatory frameworks and compliance standards increasingly require transparency in automated decision-making systems. XAI plays a crucial role in meeting these requirements by ensuring that AI models provide explainable and auditable outputs. This makes the proposed approach highly relevant for real-world deployment in secure and regulated environments.

Overall, this research contributes to the development of a next-generation cybersecurity framework that combines the strengths of artificial intelligence with the clarity of explainability. It provides a balanced approach that not only improves detection accuracy but also ensures that the system remains interpretable, reliable, and trustworthy for practical use.

II. PROBLEM STATEMENT

With the increasing use of Artificial Intelligence (AI) in network intrusion detection systems, significant improvements have been achieved in identifying cyber threats. However, despite their high detection accuracy, most AI-based systems operate as black-box models, which limits their practical usability in real-world environments. This lack of transparency creates several critical challenges for cybersecurity professionals and organizations.

- **Lack of Interpretability:** Most machine learning and deep learning models do not provide clear explanations for their predictions, making it difficult for analysts to understand the reasoning behind threat detection.
- **Reduced Trust in Predictions:** Security analysts may hesitate to rely on AI systems when the decision-making process is not transparent, especially in critical applications such as banking, healthcare, and government systems.
- **Difficulty in Debugging Models:** Identifying errors, biases, or misclassifications becomes challenging when the internal workings of the model are not interpretable.
- **High False Positives:** Many existing systems generate false alarms, which increases the workload on security teams and may lead to ignoring genuine threats.

Therefore, there is a need to develop an intelligent network threat detection system that not only achieves high accuracy but also provides clear and understandable explanations for its decisions. The goal of this research is to design and implement an Explainable AI (XAI)-based system that enhances transparency, improves trust, and supports effective decision-making in cybersecurity environments.

III. LITERATURE REVIEW

The application of Artificial Intelligence (AI) and Machine Learning (ML) in network threat detection has been widely studied over the past decade. Researchers have explored various supervised and unsupervised learning techniques, such as Decision Trees, Support Vector Machines (SVM), Random Forests [25], [26], and Deep Neural Networks, to identify anomalies and malicious activities in network traffic. These models have shown high accuracy in detecting known and unknown threats, making them a preferred choice over traditional rule-based systems. However, most of these approaches

focus primarily on performance metrics and often ignore the interpretability [27], [28] of the models.

Recent studies highlight the growing importance of Explainable Artificial Intelligence (XAI) in cybersecurity. Techniques such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and feature importance ranking have been introduced to provide insights into model predictions. These methods help analysts understand which features contribute most to the detection of threats, thereby increasing trust in AI systems. Several researchers have demonstrated that integrating XAI with ML models improves not only transparency but also assists in identifying false positives [31] and refining detection rules.

Despite these advancements, there are still challenges in balancing model accuracy and interpretability. Some highly interpretable models may sacrifice detection performance, while complex models may remain difficult to explain even with XAI tools. Additionally, the lack of standardized evaluation metrics for explainability makes it difficult to compare different approaches. This research builds upon existing work by analyzing and implementing effective XAI techniques specifically tailored for network threat detection systems, aiming to achieve an optimal balance between accuracy, transparency, and usability.

IV. METHODOLOGY

This research adopts a systematic approach to develop and evaluate an Explainable Artificial Intelligence (XAI)-based network threat detection system. The methodology consists of multiple stages, including data collection, preprocessing, model development, integration of explainability techniques, and performance evaluation. Each stage is carefully designed to ensure both accuracy and interpretability of the system.

A. Data Collection and Dataset Selection

The study utilizes publicly available network intrusion datasets such as NSL-KDD [23], CICIDS [29], and UNSW-NB15 [30] which contain labeled instances of normal and malicious network traffic. These datasets include various attack types such as denial-of-service (DoS), probing, and unauthorized access.

B. Data Preprocessing

The collected data is cleaned and prepared for analysis. This includes handling missing values, removing duplicates, normalizing numerical features, and encoding categorical variables. Feature selection techniques are also applied to identify the most relevant attributes affecting threat detection.

C. Model Development

Machine learning models such as Random Forest [25], Decision Tree, and Neural Networks [24] are trained using the preprocessed data. These models are chosen due to their effectiveness in classification tasks and their compatibility with explainability techniques.

D. Integration of XAI Techniques

Explainability methods such as SHAP [27] and LIME [28] are applied to the trained models to interpret their predictions. These techniques help identify feature contributions and provide local and global explanations for detected threats.

E. Performance Evaluation

The system is evaluated using metrics such as accuracy, precision, recall, and F1-score. Comparative analysis is conducted to assess the trade-off between model performance and explainability.

- Accuracy: Measures the overall correctness of the model.
- Precision: Indicates the proportion of correctly identified positive instances.
- Recall: Measures the ability of the model to detect actual threats.
- F1-score: Provides a balance between precision and recall.

This structured methodology ensures the development of a reliable, transparent, and efficient network threat detection system.



Fig. 1: Proposed Methodology for XAI-based Threat Detection System

V. PROPOSED SYSTEM ARCHITECTURE

The proposed system architecture for Explainable AI-based network threat detection is designed as a multi-layered framework to ensure efficient data processing, accurate threat detection, and transparent decision-making. Each layer performs a specific function, contributing to the overall robustness and interpretability of the system.

A. Data Collection Layer

This layer is responsible for gathering raw network traffic data from various sources, such as routers, firewalls, intrusion detection systems, and servers. It captures both real-time and

historical data, including packet-level information, flow statistics, and protocol details. The quality and diversity of collected data are crucial, as they directly impact the performance of the detection models. This layer ensures continuous monitoring of network activities to identify potential threats at an early stage.

B. Preprocessing Layer

The preprocessing layer prepares the collected data for analysis by removing noise and inconsistencies. It involves data cleaning (handling missing or corrupted values), normalization (scaling numerical features), and encoding categorical variables into machine-readable formats. Additionally, feature extraction and selection techniques are applied to identify the most relevant attributes, reducing dimensionality and improving model efficiency. This step is essential for enhancing both accuracy and computational performance.

C. AI Model Layer

In this layer, machine learning and deep learning models are implemented to classify network traffic as normal or malicious. Algorithms such as Random Forest, Support Vector Machine (SVM), and Neural Networks are trained using labeled datasets. These models learn patterns and behaviors associated with different types of cyberattacks. Ensemble methods and hybrid approaches may also be used to improve detection accuracy and reduce false positives.

D. Explainability Layer

The explainability layer integrates Explainable AI techniques such as SHAP and LIME to interpret the predictions made by the AI models. Instead of providing only a classification result, this layer explains why a particular network activity is flagged as a threat. It identifies key features influencing the decision, enabling cybersecurity analysts to understand model behavior, validate results, and build trust in the system.

E. Decision Layer

The final layer is responsible for generating alerts and initiating response actions based on model predictions. When a potential threat is detected, the system triggers notifications to administrators and may automatically perform mitigation actions such as blocking IP addresses, isolating affected systems, or updating firewall rules. This layer ensures timely response and helps minimize the impact of cyberattacks.

VI. RESULTS AND DISCUSSION

The results of the proposed Explainable AI-based network threat detection system demonstrate significant improvements in both detection accuracy and interpretability compared to traditional methods. Multiple machine learning models, including Random Forest, Support Vector Machine (SVM), and Neural Networks, were evaluated using benchmark datasets such as NSL-KDD [23] and CICIDS [29]. Among these, the Random Forest model achieved the highest overall accuracy, while Neural Networks showed strong performance in detecting complex and previously unseen attack patterns.



Fig. 2: Proposed XAI-based Threat Detection System

Performance evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of the models. The system achieved high recall values, indicating its strong capability to correctly identify malicious activities with minimal false negatives [27], [28]. Precision was also maintained at a satisfactory level, reducing the number of false alarms generated during detection.

TABLE I: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	96.5	0.95	0.97	0.96
SVM	93.2	0.92	0.94	0.93
Neural Network	95.1	0.94	0.96	0.95

The integration of Explainable AI techniques, specifically SHAP and LIME, provided valuable insights into model decisions. These techniques helped identify the most influential features contributing to threat detection, such as packet size, connection duration, and protocol type. By visualizing feature importance, security analysts were able to understand why certain traffic was classified as malicious, thereby increasing trust in the system.

Furthermore, the explainability layer assisted in detecting model biases and improving decision-making. It enabled faster incident response by providing clear and interpretable outputs. Overall, the results indicate that combining AI with explainability not only enhances detection performance but also improves system transparency, making it more suitable for real-world cybersecurity applications.

VII. LIMITATIONS

The proposed Explainable Artificial Intelligence (XAI)-based network threat detection system offers improved transparency and accuracy; however, it has certain limitations that need to be addressed for real-world implementation.

1) Computational Complexity

The integration of machine learning models along with explainability techniques such as SHAP and LIME increases computational overhead, which may affect system performance.

2) Scalability Issues

Scaling the system to handle large enterprise networks with high traffic volumes can be challenging and requires optimized infrastructure.

3) Dependency on Data Quality

The system heavily relies on high-quality training data. Poor or biased datasets can lead to inaccurate predictions and unreliable results.

4) Trade-off Between Accuracy and Interpretability

Improving model interpretability may reduce accuracy, while highly accurate models may be harder to explain.

VIII. FUTURE SCOPE

The proposed Explainable Artificial Intelligence (XAI)-based network threat detection system provides a strong foundation for building transparent and efficient cybersecurity solutions. However, there are several areas where further improvements and advancements can be explored to enhance its capabilities and real-world applicability.

1) Real-Time Implementation

Future work can focus on developing a fully real-time threat detection system that processes live network traffic with minimal latency. This would improve the system's responsiveness and enable faster mitigation of cyberattacks.

2) Integration with Advanced Deep Learning Models

Incorporating advanced deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can improve the detection of complex and evolving attack patterns.

3) Scalability for Large-Scale Networks

The system can be extended to handle large-scale enterprise or cloud-based networks. Optimizing computational efficiency and resource management will be essential for such environments.

4) Enhanced Explainability Techniques

Future research can explore more advanced and user-friendly XAI techniques that provide deeper insights and visual explanations, making it easier for non-experts to understand model decisions.

5) Adaptive Learning Mechanisms

Implementing continuous learning or online learning methods will allow the system to adapt to new and emerging cyber threats without requiring complete re-training.

In summary, the future scope of this research lies in enhancing scalability, adaptability, and usability, making the system more robust and effective for modern cybersecurity challenges.

IX. CONCLUSION

In this research, an Explainable Artificial Intelligence (XAI)-based network threat detection system has been pro-

posed to address the limitations of traditional and black-box AI models in cybersecurity. The study highlights the importance of combining high-performance machine learning [25], [26] techniques with interpretability to build trustworthy and efficient security systems. By integrating models such as Random Forest, Support Vector Machine (SVM), and Neural Networks with explainability tools like SHAP and LIME, the system is able to not only detect threats accurately but also provide clear justifications for its decisions.

The experimental results demonstrate that the proposed approach achieves strong performance in terms of accuracy, precision, recall, and F1-score, while significantly improving transparency. The inclusion of the explainability layer [27], [28] allows cybersecurity analysts to understand the reasoning behind threat classifications, making it easier to validate results, reduce false positives, and improve response strategies. This is especially important in real-world scenarios where decision accountability and trust are critical.

Moreover, the multi-layered architecture ensures efficient data handling, from collection and preprocessing to detection and response. It supports real-time monitoring and automated mitigation, making it suitable for deployment in modern network environments.

REFERENCES

- [1] H. K. B. Jadhav, D. Goyal, T. Singh, Y. Davda, S. K. Swarnkar, and D. Dhabliya, "Infinite-dimensional Hamiltonian systems in quantum computing simulations," *Journal of Interdisciplinary Mathematics*, vol. 28, no. 6, pp. 2359–2366, 2025.
- [2] R. Tambe, V. Rathod, N. Vikhe, H. Jadhav, K. Sharma, and D. Deotale, "A Multiple Mechanism for Identifying Green Land Cover Using Auto Encoder in Satellite Image Processing," *Panamerican Mathematical Journal*, vol. 35, no. 2S, pp. 428–439, 2025.
- [3] S. Sonavane, G. Prashantha, J. Deshmukh, M. Salunke, H. Jadhav, and P. Nikam, "Design of a Blockchain-Based Access Control Model with QoS-Awareness Via Bioinspired Computing Techniques," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 631–639, 2023.
- [4] N. Sable, V. Rathod, M. Salunke, H. Jadhav, R. Tambe, and S. Kothavle, "Enhancing Routing Performance in Software-Defined Wireless Sensor Networks through Reinforcement Learning," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 10s, pp. 73–83, 2023.
- [5] M. Salunke, S. Rathod, H. Jadhav, M. Yashwante, V. Rewaskar, and P. Deshmukh, "Implementation of Decoy Deception based Detection System for Ransomware Attack," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 714–719, 2023.
- [6] S. Rathod, M. Salunke, H. Jadhav, D. Ajalkar, D. Satre, and D. Bonde, "Role Based Secure Data Access Control for Cost Optimized Cloud Storage Using Data Fragmentation While Maintaining Data Confidentiality," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7s, pp. 316–323, 2023.
- [7] H. Jadhav and S. Pawar, "A Novel Approach for Preservation of Privacy by using Location," in *Proc. IEEE 10th International Conference on Communication Systems and Network Technologies (CSNT)*, 2021, pp. 700–703.
- [8] N. Bhokan, M. Murde, D. Thakur, and H. Jadhav, "Disaster survival training using virtual reality," in *Proc. International Conference on Computing Communication and Security (ICCCS)*, 2020.
- [9] H. Jadhav and A. Jadhav, "Systematic Approach Towards Sentiment Analysis in Online Reviews," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 31, pp. 358–369, 2020.
- [10] J. R. Mahajan and N. N. Patil, "Alpha Channel for Integrity Verification using Digital Signature on Reversible Watermarking QR," in *Proc. ICCUBE4*, IEEE Pune Section, 2015.
- [11] J. R. Mahajan and N. N. Patil, "QR Code Application Based Reversible Watermarking in Image Processing," in *Proc. ICAST*, IEEE Bombay Section, 2014.
- [12] J. R. Mahajan and N. N. Patil, "Permeating Information Transfer across Different Platforms Using QR Tradition Text: A Review," in *Proc. ICRTET*, Elsevier, 2014.
- [13] J. R. Mahajan and N. N. Patil, "A Review on Data Hiding Using Mobile Computing Application - QR with Digital Information," *International Journal of Computer Applications*, 2014.
- [14] D. Rath, S. Shinde, S. Dhoka, and J. R. Mahajan, "A Survey on Conceptual Advancements in Sentiment Analysis," *International Journal of Research in Engineering and Technology*, vol. 5, no. 1, pp. 234–236, 2016.
- [15] D. P. Nimase, S. S. Londhe, S. U. Phatangare, and J. R. Mahajan, "A CS based Secure Watermark Detection and Privacy Preserving Storage Framework," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 2, pp. 1482–1486, 2016.
- [16] A. Kolhe, V. Vadlik, B. Kedar, P. Shinde, and J. R. Mahajan, "Reversible Data Hiding Using Image-Based Authentication," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 5, no. 1, 2025.
- [17] V. Kadam and J. R. Mahajan, "Implementation of DarkReaper Automated OSINT Tool," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, 2025.
- [18] P. B. Chandane and S. G. Joshi, "Review of Statistical Robust Glaucoma Detection Using Fundus Images," *International Interdisciplinary Multilingual Research Journal*, vol. 10, no. 1, 2024.
- [19] P. B. Chandane and S. G. Joshi, "A Statistical Robust Glaucoma Detection Using Fundus Images," *International Journal of Research Publication and Reviews*, vol. 6, no. 4, 2025.
- [20] P. B. Chandane et al., "Enhancing Security of Image Steganography using Visual Cryptography," *International Research Journal of Innovations in Engineering and Technology*, vol. 8, no. 11, 2024.
- [21] P. Chandane et al., "Virtual Treatment and Consultation System," *International Journal of Advanced Research in Computer and Communication Engineering*, 2023.
- [22] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017.
- [23] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.
- [24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [26] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [27] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
- [28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [29] CICIDS2017 Dataset, Canadian Institute for Cybersecurity, University of New Brunswick, 2017.
- [30] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in *Proc. Military Communications and Information Systems Conference*, 2015, pp. 1–6.
- [31] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges," *Information Fusion*, vol. 58, pp. 82–115, 2020.