ACTA SCIENTIAE

ISSN:2178-7727

AI-Powered Explainable Intrusion Detection System with Cross-Domain Adaptability

Dr. Levina Tukaram[1], Sudeep Kumar Sah[2], Jayanth C[3], Padmini M[4], Mounika K.R.[5]

[1]Associate Professor, Department of Computer Science and Engineering, KNSIT, Bengaluru, India.
Email: levinad@knsit.com
[2]UG Student, Department of Computer Science and Engineering, KNSIT, Bengaluru, India.
[3]Assistant Professor, Department of Computer Science and Engineering, KNSIT, Bengaluru, India.
[4]Assistant Professor, Department of Computer Science and Engineering, KNSIT, Bengaluru, India.
[5]Assistant Professor, Department of Information Science and Engineering, KNSIT, Bengaluru, India.

ABSTRACT

As the Internet of Things (IoT), cloud computing, and distributed network systems continue to grow rapidly, ensuring cybersecurity has become a critical global concern. The rise in the number of smart devices interconnected has expanded the potential attack surface for cyber threats, thus necessitating real-time threat detection and prevention. Traditional Intrusion Detection Systems (IDS) primarily rely on static, signature-based, or rule-based detection methods, which are adequate for identifying known threats but inadequate for recognizing new, altered, or zero-day attacks. Furthermore, these systems often operate as black boxes, lacking transparency and explanation for their actions, which undermines their reliability in real-world security scenarios. This study presents an AI-driven Explainable Intrusion Detection System (IDS) that integrates the strengths of Machine Learning (ML) and Deep Learning (DL) models to both detect and clarify. The framework utilizes hybrid algorithms, including XG-Boost, Random Forest, and Convolutional Neural Networks (CNN), to attain high accuracy in detection. To enhance interpretability, the system integrates SHAP (Shapley Additive explanations), an Explainable AI (XAI) method that illustrates and quantifies how specific network features impact the model's prediction results. Additionally, the model features an Adversarial Defense mechanism to bolster its resistance against evasion and poisoning attacks, assuring dependability even when confronted with manipulated or adversarial input data. A significant innovation presented in this work is the Mitigation Module, which expands the IDS beyond merely detecting threats. The system automatically provides suggestions for appropriate defensive measures as an attack is detected. For example, it may suggest rate limiting and firewall adjustments in the case of Distributed Denial of Service (DDoS) attacks, account lockout and password throttling for brute-force intrusions, and input sanitization or prepared statements for injection-based threats. This transforms the IDS from a detection system into an intelligent and proactive security assistant capable of real-time response.

The system has been evaluated and confirmed using various benchmark datasets, such as CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15, representing a diverse array of attack types, networking conditions, and traffic behaviours. The proposed Intrusion Detection System (IDS) consistently achieves high accuracy, exceeding 97% across all datasets, while also showing improved interpretability, resilience, and adaptability to different domains. By incorporating AI-based modelling, explainability, adversarial defences, and automated mitigation strategies, the system provides a thorough, transparent, and adaptable solution for contemporary cybersecurity issues. The results underscore the potential of this framework to lay the groundwork for future self-learning and explainable IDS solutions in both industrial and research contexts.

KEYWORDS: Intrusion Detection System, Explainable Artificial Intelligence, Machine Learning, Deep Learning, SHAP, Adversarial Defence, Cross-Domain Adaptability, Mitigation, Filter

INTRODUCTION

The increasing dependency on IoT devices, cloud infrastructure, and distributed computing environments has made cybersecurity a crucial concern for organizations and individuals. Every second, enormous amounts of data are transmitted across the network, making it susceptible to various attacks such as Distributed Denial of Service (DDoS), brute-force intrusions, botnets, exploits, and infiltration. The limitations of traditional Intrusion Detection Systems (IDS) arise from their static and rule-based nature. They often fail to recognize novel attacks or adapt to changes in network behavior. Moreover, these systems do not provide any cognitive

ISSN:2178-7727

behind outcomes of their detection, which makes them unsuitable for critical environments where interpretability and reliability are mandatory.

To address these challenges, the proposed Intrusion Detection System (IDS) utilizes artificial intelligence (AI) to automate the processes of detecting, explaining, and mitigating cyber threats. This system is a combination of Machine Learning (ML) and Deep Learning (DL) algorithms to identify intricate attack patterns using real-world network data. It also employs Explainable Artificial Intelligence (XAI) techniques to ensure that the system's predictions are clear and understandable. Moreover, adversarial defence mechanisms are integrated to bolster resilience against altered or harmful data. A novel feature for mitigation is introduced, offering real-time preventive measures for identified attacks, making the IDS proactive rather than merely reactive. Its adaptability across various datasets allows it to be suitable for Internet of Things (IoT), enterprise, and cloud environments, achieving both generalisation and precision.

LITERATURE SURVEY

[1] "Explainable and Transferable Adversarial Attack for ML-Based Network Intrusion Detectors" Hangsheng Zhang et al. (2024) proposed an Explainable Transfer-based Black-Box Adversarial Attack framework (ETA) designed for ML-based NIDS. Their approach introduces transferable adversarial examples capable of deceiving multiple models by leveraging cooperative game theory and perturbation interpretations. They developed an Important-Sensitive Feature Selection (ISFS) mechanism to identify crucial attributes that influence network intrusion detection decisions. The study emphasised that understanding feature sensitivity not only helps craft better evasion attacks but also assists in developing defensive countermeasures. The primary advantage of this framework lies in its high transferability and explainability across various ML models, although it incurs higher computational complexity due to cooperative game theory calculations.

[2] "GADoT: GAN-based Adversarial Training for Robust DDoS Attack Detection" Maged Abdelaty et al. (2022) presented GADoT, a GAN-based Adversarial Training framework that enhances robustness against DDoS attacks in ML-based NIDS. The system utilises Generative Adversarial Networks to generate synthetic adversarial traffic and retrain detection models, significantly reducing undetected malicious flows to less than 2%. This research highlighted the power of adversarial training to increase resilience and reduce false negatives in real-world traffic conditions. The work demonstrated that the inclusion of adversarial examples during training can transform an IDS into a more reliable defence mechanism, though its focus on DDoS attacks limits its applicability across diverse threat categories.

[3] "Automatic Evasion of Machine Learning-Based Network Intrusion Detection Systems", Haonan Yan and Xiaoguang Li et al. (2024) explored the Automatic Evasion of ML-Based Network Intrusion Detection Systems using a label-only black-box adversarial attack model. Their strategy successfully bypassed multiple ML- and DL-based NIDS without requiring access to model parameters or architecture. By employing model extraction and transfer-based evasion, they achieved over 75% average success rate in fooling seven machine learning and one deep learning model. This study underscores the urgency of introducing adversarial defenses within ML-based NIDS, proving that even highly accurate models remain vulnerable to manipulation in real-world conditions. However, the lack of proposed mitigation measures indicates the need for future frameworks to combine attack simulation with defense reinforcement.

[4] "A Defensive Framework Against Adversarial Attacks on Machine Learning-Based Network Intrusion Detection Systems", Benyamin Tafreshian and Shengzhi Zhang et al. (2025) proposed a Defensive Framework Against Adversarial Attacks on ML-Based NIDS, integrating adversarial training, dataset balancing, advanced feature engineering, ensemble learning, and model fine-tuning. Their model was validated using NSL-KDD and UNSW-NB15 datasets and achieved a 35% improvement in detection accuracy and a 12.5% reduction in false positives under adversarial conditions. The study provides a comprehensive solution for enhancing the robustness of IDS models by combining multiple defence layers and optimisation strategies. However, the complexity of implementation and computational requirements were noted as

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

challenges for large-scale deployment.

PROPOSED SYSTEM

The block diagram of the proposed **AI-Powered Explainable Intrusion Detection System with Cross Domain Adaptability** is shown in Fig. 1. The system architecture consists of multiple functional units that work together to detect, explain, and mitigate cyber threats in network environments. It begins with dataset ingestion and proceeds through stages of preprocessing, feature extraction, model prediction, explainability, adversarial defence, and mitigation. Each unit is designed to perform a specific operation, ensuring that the overall system is accurate, explainable, and resilient.

The main functional blocks of the system are as follows:

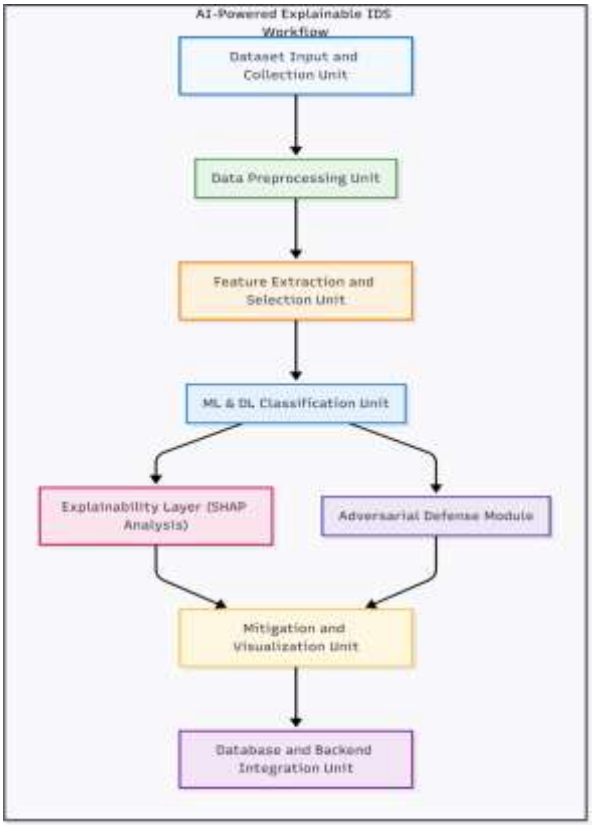

Fig.1 Block Diagram

1. Dataset Input and Collection

This unit is responsible for loading and managing network datasets such as CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15. These datasets incudes both normal and attack traffic records that represent real-world network behavior. The collected data used as the foundation for model training and evaluation, ensuring that the IDS can detect a wide range of cyber threats.

2. Data Preprocessing Unit

The data preprocessing unit performs operations such as data cleaning, normalization What specific methodologies were used for the evaluation of the system's performance on the various benchmark datasets, and how do they compare to traditional IDS evaluation methods?

How does the Mitigation Module prioritize different defensive measures based on the type of detected **attack, and what criteria are used to determine the most effective response?**
What are the potential limitations or challenges associated with the integration of AI-driven models in real-

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

world deployment scenarios, particularly regarding interpretability and resource requirements, and encoding. It removes redundant and missing values and converts categorical attributes into a numerical format. Feature scaling is applied using standardization or normalization techniques to maintain uniformity across input values. This ensures that all data is properly formatted and ready for analysis.

3. Feature Extraction and Selection Unit

After preprocessing, the system identifies the most important features that influence intrusion detection. Statistical correlation and information gain methods are employed to select relevant attributes from the dataset. This process reduces dimensionality, improves computational efficiency, and enhances the model's ability to focus on critical network parameters such as packet size, duration, and flow rate.

4. Machine Learning and Deep Learning Classification Unit

This is the core component of the system. It utilizes a hybrid model of Machine Learning and Deep Learning to classify network traffic as normal or malicious. Machine Learning algorithms, such as XGBoost and Random Forest, are trained on the extracted features, while a Convolutional Neural Network (CNN) is used to learn complex patterns and spatial correlations within the data. The ensemble of these models results in high detection accuracy and reduces false alarms.

5. Explainability Layer (SHAP Analysis)

The explainability layer applies the SHAP (Shapley Additive exPlanations) technique to interpret model predictions. For each classified instance, SHAP calculates the contribution of every feature toward the final prediction. The results are presented as visual explanations that help cybersecurity analysts understand why the system flagged a particular activity as malicious. This improves transparency and trust in the AI-driven IDS.

6. Adversarial Defense Module.

This module enhances the robustness of the Intrusion Detection System (IDS) by implementing adversarial training techniques. It generates adversarial samples by adding controlled noise to the input data and retrains the model using these samples. This process helps the system resist evasion and poisoning attacks, ensuring stable performance even when attackers attempt to manipulate input traffic to bypass detection.

7. Mitigation and Visualisation Unit.

This is the final component of the system. When an attack is detected, the mitigation module automatically recommends countermeasures based on the type of attack. For example, in the case of a DDoS attack, the system suggests rate-limiting and traffic filtering. For brute-force attacks, it recommends enabling account lockout and implementing multi-factor authentication. In the event of injection attacks, it advises on input validation and SQL query sanitization. The visualization dashboard showcases system performance graphs, SHAP (SHapley Additive exPlanations) outputs, and attack statistics in an interactive interface, allowing users to effectively monitor detection results and mitigation actions**.**

8. Database and Backend Integration Unit.

The entire system is supported by a Flask-based backend and an SQLite database. All prediction results, SHAP values, and mitigation logs are stored for future analysis and comparison. The modular structure of the backend allows for easy integration with real-time monitoring systems or cloud-based security frameworks.

The proposed architecture ensures a seamless flow of data through all units, beginning from dataset ingestion to threat mitigation. Each stage is interdependent and designed to enhance detection accuracy, interpretability, and adaptability. By combining artificial intelligence, explainability, adversarial robustness, and mitigation, the system provides a comprehensive solution for intelligent and transparent network intrusion detection.

METHODOLOGY

The proposed methodology focuses on developing an **AI-powered, explainable, and resilient Intrusion**

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

Detection System (IDS) that can accurately detect, interpret, and mitigate cyber threats across IoT, cloud, and enterprise environments. The system integrates Machine Learning (ML) and Deep Learning (DL) models with Explainable Artificial Intelligence (XAI) to achieve transparency, adaptability, and robustness against adversarial manipulation. The entire process is divided into sequential stages, from data collection to final mitigation and visualization.

The process starts with **data acquisition and preprocessing techniques**, where datasets such as CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15 are used. These datasets contain real-world normal and malicious network traffic. In the preprocessing phase, data cleaning removes redundant and missing values, encoding converts categorical attributes into numerical form, and normalization ensures consistent feature ranges. This step enhances the overall model efficiency and training stability.

For feature extraction and selection, are used by following preprocessing techniques, and utilized to pinpoint the most significant network attributes essential for detecting attacks. Statistical correlation and information gain methods help eliminate redundant data while retaining crucial information. This enhances computational efficiency and allows the model to concentrate on the most relevant parameters, such as flow duration, packet size, and protocol type.

The classification component integrates machine learning (ML) and deep learning (DL) algorithms, including XGBoost, Random Forest, and Convolutional Neural Networks (CNN). These hybrid models assess traffic behavior and categorize it as either normal or an attack. By combining the interpretability of ML with the advanced learning capabilities of DL, the system achieves better precision, detection rates, and scalability across various domains.

An explainability layer, utilizing the SHAP (Shapley Additive explanations) framework, clarifies model predictions by assessing the contribution of each feature to the decision-making process. This promotes transparency in detection results, making them more understandable and traceable for administrators. To bolster the model's strength, an adversarial defense mechanism is implemented. This involves adversarial training and controlled noise perturbations to ensure the Intrusion Detection System (IDS) remains resilient against evasion and poisoning attacks, thus preserving its performance even in challenging or manipulated environments.

Finally, the mitigation and visualisation module generates actionable responses to identified threats. It provides real-time analytics, attack trends, and SHAP-based visual explanations through an interactive dashboard. Additionally, a filtering feature is integrated within the history section to allow users to organize and analyze previous intrusion records based on specific attack types. This enables efficient identification of recurring threats and enhances the overall interpretability of system behavior. Detected threats, suggested countermeasures, and filtered analytics are stored in a backend database for continuous improvement and re-evaluation.

FLOW CHART

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

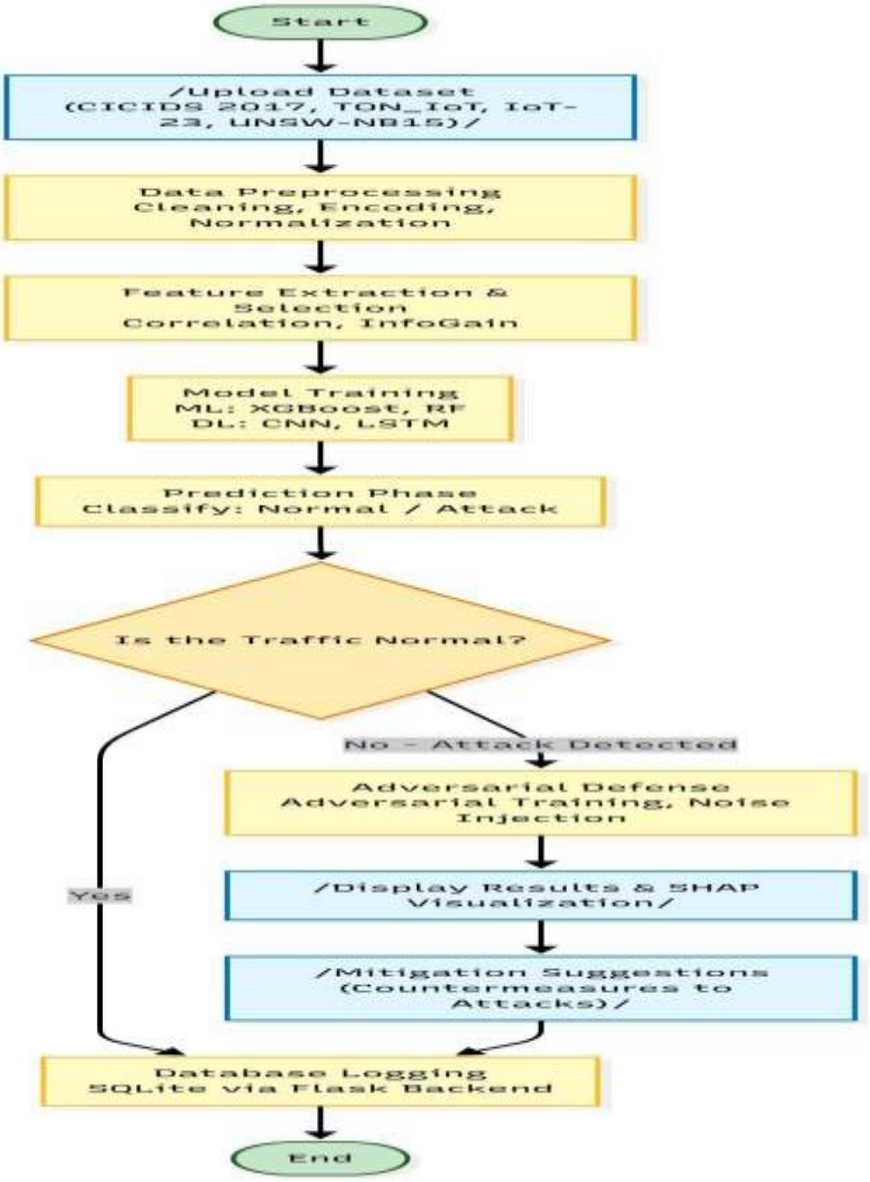

Fig.2 Flow Chart

IMPLEMENTATION

The implementation of the proposed **AI-Powered Explainable Intrusion Detection System with Cross Domain Adaptability** integrates artificial intelligence, explainability, and defense mechanisms within a single framework. The system is developed using **Python**, employing Scikit-learn libraries along with XGBoost, TensorFlow, and SHAP to implement machine learning, deep learning, and explainable components. The overall implementation is divided into modules: dataset preparation, hybrid model training, explainability integration, adversarial defense, mitigation, and deployment through a Flask web interface.

The process begins with **data preparation**, where benchmark datasets like CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15 are collected and preprocessed. Operations such as data cleaning, label encoding, and normalization are performed to ensure quality and consistency. The preprocessed data is used for both model training and evaluation.

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

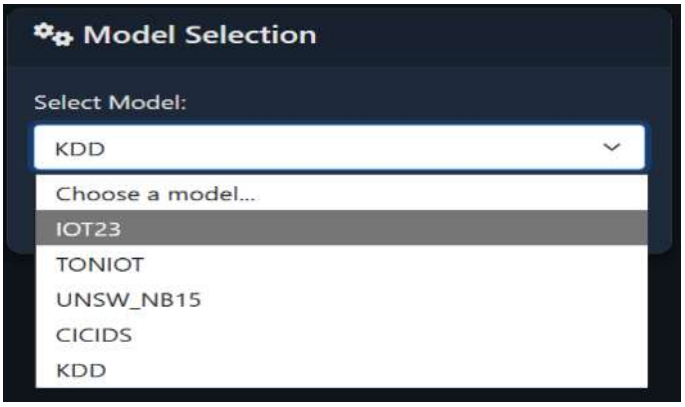

Fig.3 Model Selection

The **hybrid classification model** combines machine learning algorithms such as XGBoost and Random Forest with deep learning architectures like CNN. The ML models analyze structured patterns while CNN captures complex spatial relationships within network traffic. This hybrid integration helps achieve higher detection accuracy and adaptability across IoT and cloud environments.

To improve transparency, the **explainability layer** is implemented using the SHAP framework. It interprets the influence of individual features on the model's predictions, enabling analysts to understand why a specific data instance was classified as malicious. The SHAP outputs are visualized through the web interface for user interpretation.

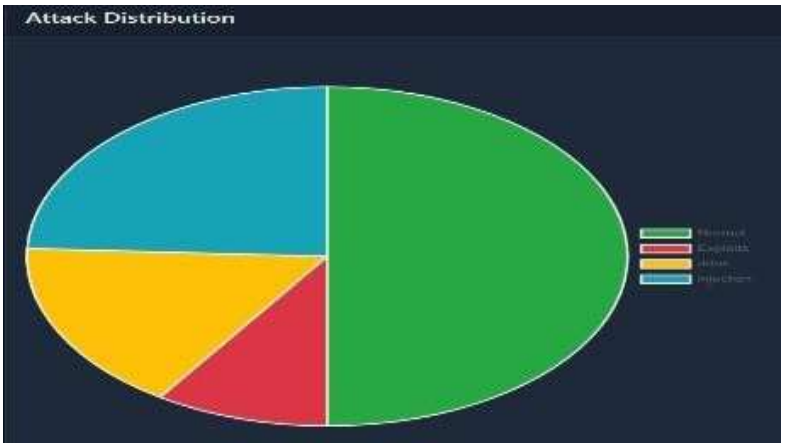

Fig.4 Dashboard

The **adversarial defense module** enhances robustness by retraining the model using adversarial perturbed samples. Techniques such as noise injection, gradient-based attacks, and perturbation modeling are applied to strengthen the IDS against evasion and poisoning attempts. This approach ensures that the model learns to identify and resist malicious manipulations even under deceptive network conditions. Once potential threats are detected, the **mitigation engine** automatically suggests suitable countermeasures, such as blocking suspicious IP addresses, restricting unauthorized access, or enforcing stronger authentication methods to prevent further attacks.

The system also includes a **filter button** in the history section to enhance data analysis and usability. It allows users to view, sort, and compare past intrusion records based on different attack types directly from the SQLite database. This feature helps in identifying recurring threats, understanding attack frequency, and monitoring overall system performance more efficiently, while also providing a clearer view of network security trends over time.

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

Fig.5 Mitigation

Finally, the entire system is deployed using the **Flask web framework**. The interface allows users to upload datasets, run intrusion detection, and view results in real time. The backend, connected with an **SQLite database**, stores prediction logs, mitigation suggestions, and performance data for future reference. This integrated setup ensures a responsive, interpretable, and secure intrusion detection system suitable for real-world environments.

RESULT AND FUTURE SCOPE

The proposed **AI-Powered Explainable Intrusion Detection System (IDS)** has been developed and tested using multiple benchmark datasets, including **CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15**. The system was evaluated through key metrics such as accuracy, precision, recall, and F1-score to measure its performance across various network environments. The hybrid combination of Machine Learning and Deep Learning models showed significant improvement over traditional approaches. Algorithms like **XGBoost** and **Random Forest** provided reliable classification for structured data, while **Convolutional Neural Networks (CNN)** efficiently captured complex attack patterns. The integrated hybrid model achieved high accuracy, low false positives, and strong adaptability across all datasets, proving effective for cross-domain intrusion detection.

The **SHAP-based explainability layer** offered clear visual interpretations of model predictions, validating transparency and helping users understand why network activities were classified as normal or malicious. The **adversarial defense module** improved the model's resistance against evasion and poisoning attacks, while the **mitigation feature** enabled quick response by suggesting countermeasures such as IP blocking or stronger authentication. Additionally, a **filter button** in the history section allows users to analyse past attack data by type, helping identify recurring threats and trends. These combined features make the system accurate, interpretable, and resilient, making it suitable for real-time network security applications.

FUTURE SCOPE

In the **future**, the proposed IDS can be extended to include real-time packet-level traffic monitoring and **cloud-based deployment** to handle large-scale enterprise networks. Further improvements can be achieved through **federated learning** and **online continual learning**, allowing the model to automatically adapt to new and evolving attack patterns. Integration with **Software Defined Networking (SDN)** and **Edge Computing** infrastructures can enhance scalability, reduce latency, and improve decision-making speed. Additionally, **deep reinforcement learning** can be employed to develop self-learning mitigation policies that evolve with system experience. The **visualization dashboard** can be enhanced to include dynamic alert notifications, user-based analytics, and network health monitoring tools. The **mitigation engine** can also be automated to initiate preventive actions in real time, minimizing manual intervention. With these future enhancements, the proposed

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

IDS framework can evolve into a fully autonomous, adaptive, and intelligent cybersecurity defense system, capable of securing IoT, industrial, and cloud-based infrastructures efficiently.

CONCLUSION

The proposed **AI-Powered Explainable Intrusion Detection System with Cross-Domain Adaptability** provides a robust, intelligent, and adaptive framework for detecting and mitigating various cyber threats across IoT, cloud, and enterprise networks. It combines Machine Learning and Deep Learning techniques to efficiently identify both known and emerging attacks. This approach ensures high precision and reliability. The **SHAP-based explainability layer** enhances interpretability by visually representing how individual network features influence predictions. As a result, the IDS operates as a transparent, data-driven, and trustworthy framework rather than a black-box model.

The inclusion of an **adversarial defence mechanism** further strengthens the model's stability and resilience against evasion and poisoning attacks, allowing the system to maintain high detection accuracy even in complex and adversarial environments. The **mitigation module** significantly improves practical usability by recommending suitable countermeasures such as blocking malicious IP addresses or enforcing stronger authentication policies. Additionally, a **filter button** integrated within the dashboard enables users to review and categorize past intrusion records by attack type, enhancing interpretability, analytics, and user engagement. The **Flask-based web interface**, supported by an **SQLite database**, facilitates real-time traffic analysis, visualization, and secure log storage, delivering a user-friendly and efficient cybersecurity solution. Overall, the developed system achieves high accuracy, minimal false alarms, and strong cross-domain adaptability across datasets. These include **CICIDS 2017, TON_IoT, IoT-23, NSL-KDD, and UNSW-NB15**. The results effectively demonstrate how explainable AI, hybrid modeling, and adversarial robustness can enhance network defense in dynamic environments. With the integration of **cloud-native deployment**, **federated learning**, and **automated mitigation techniques**, the system can evolve. It offers a more intelligent, scalable, and self-adaptive cybersecurity framework, which is capable of securing modern digital infrastructures against future threats.

REFERENCES

1. Explainable and Transferable Adversarial Attack for ML-Based Network Intrusion Detectors, Hangsheng Zhang et al., 2024
2. GADoT: GAN-based Adversarial Training for Robust DDoS Attack " Maged Abdelaty et al., 2022
3. Automatic Evasion of Machine Learning-Based Network Intrusion Detection Systems , Haonan Yan, Xiaoguang Li, 2024
4. Problem Space Structural Adversarial Attacks for Network Intrusion Detection Systems Based on Graph Neural Networks, Andrea Venturi et al., 2024
5. A Defensive Framework Against Adversarial Attacks on Machine Learning-Based Network Intrusion Detection Systems, Benyamin Tafreshian and Shengzhi Zhang, 2025
6. Roadmap of Adversarial Machine Learning in Internet of Things-Enabled Security Systems, Yasmine Harbi et al., 2024
7. Toward Deep Learning-Based Intrusion Detection System: A Survey, Zhiqi Li et al., 2024
8. Machine Learning-Based Intrusion Detection Systems for Enhancing Cybersecurity, Aezeden Mohamed, 2023
9. Invisible Shield: Real-Time Industrial Threats Detection for IoT-based Systems , Kadaganchi Sudhakar et al, 2025
10. A Systematic Study of Adversarial Attacks Against Network Intrusion Detection Systems, by Sanidhya Sharma, Zesheng Chen (2024)
11. Adversarial Challenges in Network Intrusion Detection Systems: Research Insights and Future Prospects , Sabrine Ennaji, Fabio De Gaspari, Dorjan Hitaj, Alicia Kbidi, Luigi V. Mancini (2024)

DOI:10.22178/acta.27.3.4

ISSN:2178-7727

12. Adversarial Robust and Explainable Network Intrusion Detection Systems Based on Deep Learning, Kudzai Sauka,Gun-Yoo Shin, Dong-Wook Kim, and Myung-Mook Han (2022)
13. R. Latha and R. M. Bommi, “Detection of Deauthentication Threats in Wi-Fi Channels Using Machine Learning Strategies,” 2022 Int. Conf. Data Sci. Agents Artif. Intell. ICDSAAI 2022, pp. 4–9, 2022, DOI: 10.1109/ICDSAAI55433.2022.10028874.