

Course Title: Research Methodology in Computer Science

Course Code: AI800

Weighting: 20%

Lecturer: Dr. Mohammed Anbar

Hand-out Date: 20 Jun 2026

Hand-in Date: 04 July 2026

Hand-in Method: Online Submission via E-Learning.

Assignment 2 : Problem Statement for Interpretable Artificial Intelligence for Intrusion Detection Systems in Cybersecurity

Assignment Report: Interpretable Artificial Intelligence for Intrusion Detection Systems in Cybersecurity

Contents

INTRODUCTION.....	2
1. LIST OF CHAPTERS	2
1.1) CHAPTER 1 – PROBLEM STATEMENT	2
1.1.1 PROBLEM STATEMENT	2
2.1) CHAPTER 2 – RESEARCH OBJECTIVES.....	4
2.1.1 General Objective.....	4
2.2.1 Specific Objectives	4
2. REFERENCES	5

Introduction

1. List of Chapters

1.1) Chapter 1 – Problem Statement

1.1.1 Problem Statement

Artificial intelligence (AI) has become a core component of modern intrusion detection systems (IDS), enabling cybersecurity solutions to identify complex attacks with greater accuracy than traditional rule-based methods.

However, despite their impressive detection capabilities, many AI based IDS models, particularly those based on deep learning, operate as black boxes, offering little explanation for their predictions.

Consequently, cybersecurity analysts often struggle to understand, validate, and trust the decisions generated by these systems (Capuano et al., 2022).

This lack of explainability poses a significant challenge in operational cybersecurity environments, where analysts are required to justify alerts, investigate incidents, and make timely security decisions.

The importance of this problem continues to grow as organizations increasingly deploy AI-driven IDS to protect enterprise networks, cloud infrastructures, Internet of Things (IoT) environments, and other critical information systems. In these environments, incorrect or unexplained predictions may increase false positives, reduce analyst confidence, delay incident response, and complicate compliance with security regulations and auditing requirements (Rjoub et al., 2023). Consequently, improving the transparency and interpretability of AI-based IDS has become a critical research priority.

Several Explainable Artificial Intelligence (XAI) techniques have been introduced to address this challenge. Existing approaches such as LIME, SHAP, Grad-CAM, and rule-based explanation methods aim to provide human-understandable explanations while preserving acceptable intrusion detection performance. More recent studies have also proposed integrating Large Language Models (LLMs) with XAI to improve threat analysis, reduce false positives, and assist cybersecurity analysts through natural language explanations (Ghazal et al., 2024). Nevertheless, current studies indicate that these solutions still suffer from several limitations, including the trade-off between explainability and detection accuracy, computational complexity, scalability issues, privacy concerns, vulnerability of explanation methods to adversarial attacks, and the absence of standardized evaluation frameworks for measuring explanation quality (Zhang et al., 2022; Calzarossa et al., 2025).

Therefore, there is a clear need for further research to develop an interpretable AI framework for Intrusion Detection Systems that maintains high detection accuracy while producing reliable, transparent, and computationally efficient explanations. Addressing this research problem would enhance analyst trust, improve incident response, reduce false positives, and facilitate the practical deployment of Explainable Artificial Intelligence in real-world cybersecurity operations.

2.1) Chapter 2 – Research Objectives

2.1.1 General Objective

To propose an interpretable Artificial Intelligence framework for Intrusion Detection Systems that improves transparency while maintaining high cybersecurity detection performance.

2.2.1 Specific Objectives

- To identify the major limitations of existing AI-based Intrusion Detection Systems regarding interpretability and transparency.
- To analyze current Explainable Artificial Intelligence techniques, including SHAP, LIME, and Grad-CAM, for cybersecurity applications.
- To design an interpretable Intrusion Detection System framework that balances explanation quality with intrusion detection accuracy.
- To evaluate the proposed framework using cybersecurity performance metrics, including Accuracy, Precision, Recall, F1-score, False Positive Rate, and explanation quality.
- To compare the proposed framework with existing XAI-based Intrusion Detection approaches in terms of transparency, robustness, computational efficiency, and detection performance.
- To recommend practical strategies that support the adoption of Explainable Artificial Intelligence in operational cybersecurity environments.

2. References

Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). *An assessment framework for explainable AI with applications to cybersecurity*. *Artificial Intelligence Review*, 58, 150.

<https://doi.org/10.1007/s10462-025-11141-w>

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). *Explainable artificial intelligence for cybersecurity: A literature survey*. *Annals of Telecommunications*, 77(11–12), 789–812.

<https://doi.org/10.1007/s12243-022-00926-7>

Ghazal, T. M., Janjua, J. I., Abushiba, W., Ahmad, M., Ihsan, A., & Al-Dmour, N. A. (2024). *Cybersecurity revolution via large language models and explainable AI*. In *2024 IEEE Conference Proceedings*. IEEE

Sindiramutty, S. R. (2023). *Autonomous threat hunting: A future paradigm for AI-driven threat intelligence*. arXiv. <https://arxiv.org/abs/2401.00286>

Ellesel, M., Hennawi, S., Alzubi, S., Kanan, T., & Mughaid, A. (2023). *Unleashing the full potential of artificial intelligence and machine learning in cybersecurity vulnerability management*. *Proceedings of the 2023 International Conference on Information Technology (ICIT)*, 276–283.

<https://doi.org/10.1109/ICIT58056.2023.10225910>

Farag, M., Ndhlovu, M., Tilanyi, N., Cordeiro, L., Debbah, M., & Lestable, T. (2023). *Revolutionizing cyber threat detection with large language models*. *arXiv*.

<https://arxiv.org/abs/2306.14263>

Bujro, G., Bentahar, J., Walab, O., Mizouni, R., Song, A., Cohen, P., Otrók, H., & Mourad, A. (2023). *A survey on explainable artificial intelligence for cybersecurity*. *IEEE Transactions on Network and Service Management*, 20, 5115–5140.

<https://doi.org/10.1109/TNSM.2023.3282740>

Ben Rabah, N., Grand, B., & Pinheiro, M. (2021). IoT botnet detection using black-box machine learning models: The trade-off between performance and interpretability. *2021 IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 101–106.

<https://doi.org/10.1109/WETICE53228.2021.00030>

Alodibat, S., Ahmad, A., & Azzeh, M. (2023). Explainable machine learning-based cybersecurity detection using LIME and SecML. *2023 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 235–242.

<https://doi.org/10.1109/JEEIT58638.2023.10185893>