DEEP REINFORCEMENT LEARNING BASED HOLISTIC CYBERSECURITY SYSTEM

Project Reference No.: *49S_BE_6793*

College : R.V. Institute of Technology and Management, Bengaluru
Branch : Department of Information Science and Engineering
Guide : Dr. Niharika P Kumar
Student(s) : Ms. Aiman Sabah
Mr. Chandrakant S Devanagaon
Mr. Chethan B L
Ms. Honnu Shree B L

Keywords:

Cybersecurity, Federated Learning, Deep Reinforcement Learning, Intrusion Detection Systems, and Explainable AI

Introduction:

Modern digital infrastructures face increasing cybersecurity risks due to the rapid growth of internet-based systems, cloud computing, and interconnected devices. Traditional security tools such as NIDPS, firewalls, and antivirus systems operate independently, leading to fragmented detection, high false positives, and delayed responses. Advanced cyber threats are adaptive and multi-stage, making rule-based approaches ineffective. To address this, the proposed system integrates AI techniques like Federated Learning, Deep Reinforcement Learning, and sandboxing into a unified framework for intelligent, real-time threat detection and automated response.

Objectives:

1. To design and develop a unified AI-driven cybersecurity architecture integrating multiple security components
2. To implement decentralized detection mechanisms using machine learning and deep learning models
3. To develop a Deep Reinforcement Learning (DRL) model for adaptive threat detection and automated response
4. To reduce false positives and false negatives using sandbox-based validation techniques
5. To establish a centralized intelligence layer for correlating threats across multiple systems
6. To incorporate Explainable AI (XAI) for transparent and interpretable decision-making

Methodology:

The project follows a multi-layered intelligent architecture where decentralized security components such as NIDPS, antivirus, firewall, and WAF independently analyze system behavior using ML/DL models, while suspicious activities are validated in a sandbox to reduce false positives. Federated Learning enables collaborative model training without sharing raw data, and a Deep Reinforcement Learning agent continuously improves defense strategies through reward-based learning. The system also re-evaluates safe data to detect hidden threats, maintains a central logging system for performance analysis, and uses Explainable AI to provide clear, human-readable insights, ensuring adaptive learning, high accuracy, and autonomous cybersecurity decision-making.

Result and Conclusion:

The experimental results, shown in the DIREWOLF XAI dashboard and logs, demonstrate real-time threat detection with low latency (11 ms) and efficient resource usage. Multiple threats like code injection and ransomware are successfully identified and blocked through coordinated modules such as the NIDPS kernel, sandbox, DRL agent, and memory virtualization. Security logs highlight accurate threat classification and real-time mitigation of attacks like SQL injection and memory access, while the AI assistant confirms system integrity, continuous monitoring, and adaptive learning. Overall, the system achieves high accuracy, low false positives, and effectively handles advanced and zero-day threats.

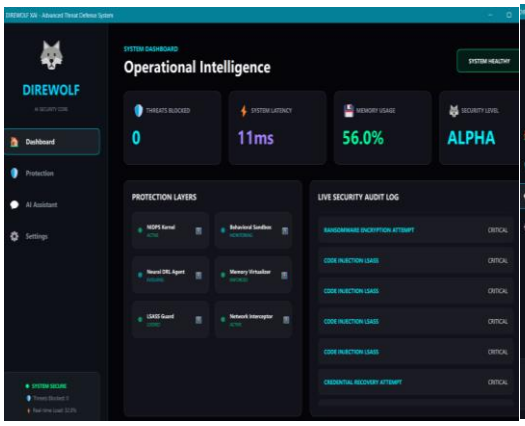

Fig 1 : Dashboard

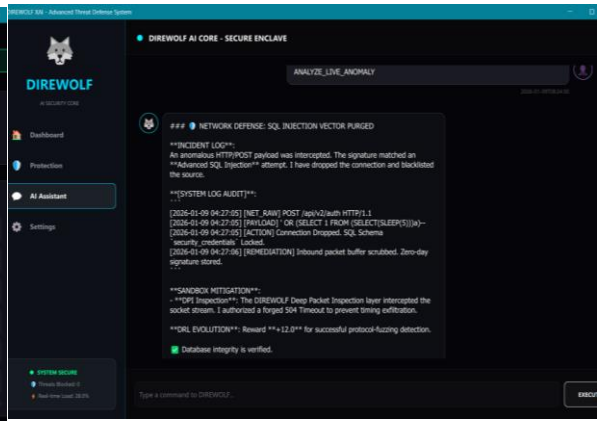

Fig. 2: XAI Chatbot

In conclusion, The project successfully demonstrates that an AI-driven integrated cybersecurity system can overcome the limitations of traditional isolated tools. It provides a scalable, intelligent, and adaptive solution capable of real-time threat detection and automated response, making it suitable for modern cybersecurity challenges.

References:

[1] Toyin et al., “Machine Learning-Based Intrusion Detection Using Random Forest and Decision Trees,” *International Journal of Cybersecurity*, 2020.

[2] Shanthi et al., “Hybrid Deep Learning Model (ConvLSTM) for Real-Time Intrusion Detection,” *Journal of Network Security*, 2021.

[3] Dong et al., “Android Malware Detection Using CNN–DNN Hybrid Model,” *IEEE Access*, vol. 8, pp. 123456–123467, 2020.

[4] Jeon et al., “Multi-Feature Malware Detection for IoT Security,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4567–4578, 2020.

[5] Ohtani et al., “Federated Learning for Zero-Day Attack Detection in IoT,” *IEEE Transactions on Information Forensics and Security*, 2023.

[6] Yang et al., “AutoML-Based Optimization for Federated Cybersecurity Systems,” *IEEE Communications Magazine*, vol. 60, no. 3, pp. 90–96, 2022.

[7] Wu et al., “WAFBooster: Automatic Boosting of Web Application Firewall Security,” *IEEE Transactions on Dependable and Secure Computing*, 2021.

[8] Dawadi et al., “LSTM-Based Web Application Firewall for DDoS and Injection Attack Detection,” *IEEE Access*, vol. 9, pp. 112233–112245, 2021.

[9] M. T. Islam, M. K. Syfullah, M. G. Rashed, and D. Das, “Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI,” *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 11, pp. 5337–5360, 2024.

[10] V. Z. Mohale and I. C. Obagbuwa, “A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity,” *Frontiers in Artificial Intelligence*, vol. 8, p. 1526221, 2025.

[11] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, “Transparency and explainability of AI systems: From ethical guidelines to requirements,” *Information and Software Technology*, vol. 159, p. 107197, 2023.

[12] S. Tiwari, V. Sresth, and A. Srivastava, “The role of explainable AI in cybersecurity: Addressing transparency challenges in autonomous defense systems,” *International Journal of Innovative Research in Science Engineering and Technology*, vol. 9, pp. 718–733, 2020.

[13] V. Franzoni, “From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy AI,” in *Proc. International*

Conference on Computational Science and Its Applications, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 118–130.

[14] G. Chaudhary, “Unveiling the black box: Bringing algorithmic transparency to AI,” *Masaryk University Journal of Law and Technology*, vol. 18, no. 1, pp. 93–122, 2024.

[15] W. J. Von Eschenbach, “Transparency and the black box problem: Why we do not trust AI,” *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[16] E. Puiutta and E. M. Veith, “Explainable reinforcement learning: A survey,” in *Proc. International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Cham, Switzerland: Springer International Publishing, Aug. 2020, pp. 77–95.

[17] S. Milani, N. Topin, M. Veloso, and F. Fang, “Explainable reinforcement learning: A survey and comparative review,” *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–36, 2024.

[18] L. Viganò and D. Magazzeni, “Explainable security,” in *Proc. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Sep. 2020, pp. 293–300.

[19] E. Holder and N. Wang, “Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst,” *Human-Intelligent Systems Integration*, vol. 3, no. 2, pp. 139–153, 2021.

[20] V. Z. Mohale and I. C. Obagbuwa, “Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability,” *Frontiers in Computer Science*, vol. 7, p. 1520741, 2025.

Future Scope:

The future scope of this project includes:

1. Connect the system to real-world and cloud environments for better usage.
2. Add support for IoT devices and improve security using blockchain.
3. Protect the system from advanced attacks and improve AI explanations.
4. Integrate with monitoring systems to build a fully automatic cybersecurity solution.