ISSN: 0974-6471, Vol. 17, No. (1) 2025, Pg. 23-25

Oriental Journal of Computer Science and Technology

www.computerscijournal.org

Artificial Intelligence and Security: Enhancing Cyber Defense Mechanisms

H. TOYA

Department of Computer Science, Caleb University, Imota, Lagos, Nigeria.

Abstract

The rapid evolution of Artificial Intelligence (AI) has significantly impacted cybersecurity, offering both opportunities and challenges. While AI enhances threat detection, anomaly identification, and automated response systems, it also introduces new vulnerabilities, such as adversarial attacks and AI-powered cyber threats. This PhD thesis explores the intersection of AI and security, focusing on machine learning (ML) and deep learning (DL) techniques for cyber defense, the risks posed by malicious AI, and strategies to mitigate these threats. The research proposes novel AI-driven security frameworks, evaluates their effectiveness against emerging cyber threats, and discusses ethical considerations in AI-based security solutions.

Article History

Received: 20 May 2025
Accepted: 29 July 2025

Keywords

Artificial Intelligence;Cyber Defense Mechanisms

Introduction

The increasing sophistication of cyber threats necessitates advanced defense mechanisms. Traditional security systems rely on predefined rules and signatures, making them ineffective against zero-day attacks and polymorphic malware. AI, particularly ML and DL, offers dynamic solutions by learning from data patterns and adapting to new threats in real time. However, adversaries are also leveraging AI to develop more sophisticated attacks, creating an ongoing arms race between attackers and defenders.

This thesis aims to

- Investigate AI-based security models for intrusion detection, malware analysis, and fraud prevention.
- Examine adversarial AI techniques and their implications for cybersecurity.
- Develop robust AI security frameworks resilient to evasion attacks.
- Analyze ethical and regulatory challenges in deploying AI for security.

AI in Cybersecurity: Opportunities and Applications

Threat Detection and Anomaly Identification

AI-powered systems can analyze vast datasets to detect anomalies, identify malicious activities, and predict potential breaches. Supervised and unsupervised learning techniques, such as Random Forests, Support Vector Machines (SVMs), and Neural Networks, improve detection accuracy.

CONTACT H. Toya h.toya@calebuniversity.edu.ng Department of Computer Science, Caleb University, Imota, Lagos, Nigeria.

Doi: http://dx.doi.org/10.13005/ojcst17.01.06

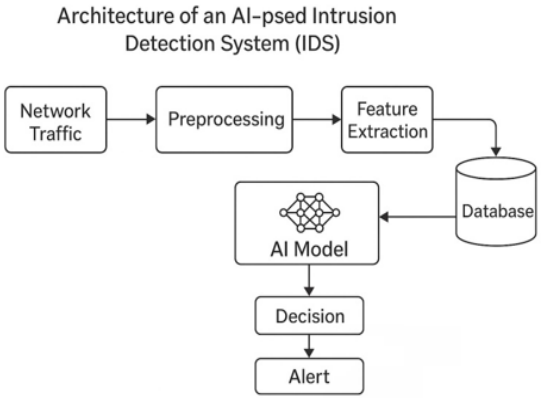

Fig. 1. Architecture of an AI-based intrusion detection system (IDS) (Insert diagram showing data flow from network traffic → feature extraction → ML model → threat classification → alert generation)

Automated Incident Response

AI enables real-time threat mitigation through automated incident response systems. Reinforcement learning (RL) can optimize response strategies, reducing the time between detection and remediation.

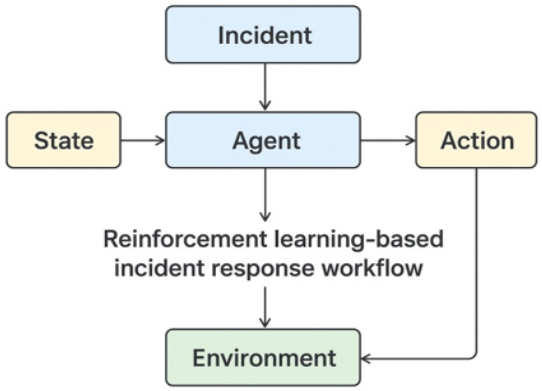

Fig. 2. Reinforcement learning-based incident response workflow (Illustrate states, actions, rewards, and feedback loop in RL for cybersecurity)

Phishing and Fraud Detection

Natural Language Processing (NLP) and deep learning models (e.g., Transformers) enhance phishing detection by analyzing email content, URLs, and user behavior patterns.

Table 1: Comparison of AI models for phishing detection

Model	Accuracy	F1-Score	False Positive Rate
Random Forest	95%	0.94	2.1%
LSTM	97%	0.96	1.5%
BERT	98%	0.97	1.2%

Security Risks in AI Systems

Adversarial Machine Learning

Attackers exploit AI vulnerabilities through adversarial examples—inputs designed to deceive models. Techniques like Fast Gradient Sign Method (FGSM) and Generative Adversarial Networks (GANs) can bypass security systems.

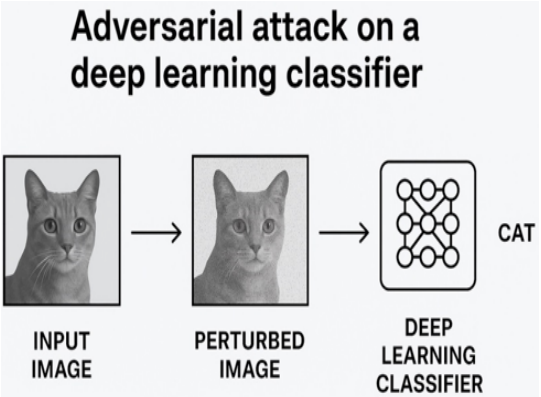

Fig. 3. Adversarial attack on a deep learning classifier (Show original vs. perturbed image and misclassification result)

AI-Powered Cyber Attacks

Malicious actors use AI for automated hacking, deepfake-based social engineering, and AI-driven malware that adapts to evade detection.

Bias and Ethical Concerns

AI models may inherit biases from training data, leading to false positives/negatives in security decisions. Ethical concerns include privacy violations and misuse of AI surveillance.

Proposed AI-Security Framework

This thesis introduces a hybrid AI-security framework combining:

- Deep Learning for Anomaly Detection (e.g., LSTM networks for sequential threat analysis).
- Adversarial Robustness Techniques (e.g., defensive distillation, adversarial training).
- Explainable AI (XAI) for Transparency (ensuring interpretability in security decisions).

Conclusion

AI revolutionizes cybersecurity but introduces new risks. Future research should focus on:

- Developing more resilient AI models against adversarial attacks.
- Establishing regulatory frameworks for ethical AI deployment.
- Enhancing human-AI collaboration in security operations.

This thesis contributes to advancing AI-driven

security solutions while addressing their limitations, paving the way for safer and more intelligent cyber defense systems.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The author declare that we have no conflict of interest.

REFERENCES

1. Goodfellow, I., Explaining and Harnessing Adversarial Examples. arXiv:1412.6572, 2014.
2. LeCun, Y., Bengio, Y., & Hinton, G., *Deep Learning. Nature*., *521*(7553), 436-444, 2015.
3. Papernot, N., The Limitations of Deep Learning in Adversarial Settings. IEEE S&P, 2016.
4. MITRE ATT&CK Framework. Adversarial Tactics, Techniques, and Common Knowledge., 2023.
5. IBM X-Force Threat Intelligence Index. AI in Cyber Attacks: Trends and Countermeasures., 2024.
6. European Union Agency for Cybersecurity (ENISA). Ethical Guidelines for AI in Cybersecurity., 2023.
7. Schneier, B. Click Here to Kill Everybody: Security and Survival in a Hyper-Connected World., 2020.