IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141020

Cyber I: A self-evolving, self-learning, self-protecting AI agent for Autonomous Cyber Threat Detection and Response

Ms. Sneha Bankar[1], Om Kalyankar[2], Rajesh Shinare[3], Nikita Shinde[4], Sakshi Landge[5]

Assistant Professor, Department of AI & DS, Dr. D Y Patil College of Engineering & Innovation, Pune, India[1]

Student, Department of AI & DS, Dr. D Y Patil College of Engineering & Innovation, Pune, India[2]

Student, Department of AI & DS, Dr. D Y Patil College of Engineering & Innovation, Pune, India[3]

Student, Department of AI & DS, Dr. D Y Patil College of Engineering & Innovation, Pune, India[4]

Student, Department of AI & DS, Dr. D Y Patil College of Engineering & Innovation, Pune, India[5]

Abstract: As devices, cloud services, and critical systems connect, cyber threats are increasing. Cyber I am an intelligent security agent that learns, adapts, and defends itself in real time. Using machine learning and reinforcement learning, it detects suspicious activity and reacts instantly to limit damage. Unlike static systems, Cyber I continuously update from new attack patterns and real-time threat data, providing organisations with a dynamic and reliable defence against both known and emerging cyber risks.

Keywords: Artificial Intelligence (AI), Cybersecurity, Autonomous Cyber Defence, Real-Time Threat Detection, Self-Learning Systems, Intrusion Prevention.

I. INTRODUCTION

With the increasing interconnection of devices, cloud services, and mission-critical systems, the number of cyber threats increases. Cyber I is a dynamic security agent that learns, adapts, and protects itself in real time. With the assistance of reinforcement learning and machine learning, it senses anomalies and acts in a matter of a split second to limit damage. Unlike static systems, it continuously renews itself from new patterns of attacks and real-time threat feeds, providing organisations with a dynamic and reliable defence against known threats and new threats. In the current day and digital age, cyberspace threats are quickly increasing as a result of the fast migration of systems, devices, and mission-critical services from offline to online. With the usual security tools, namely firewalls and intrusion detection systems, typically, they cannot keep up with these day-of-attacks of every description, including zero-day attacks, obfuscate sophisticated malware, distributed denial-of-service (DDoS) attacks, daily advanced persistent threats (APTs), and the litany of others. This gives rise to a new entity - a digital guardian born from cyberspace, known as Cyber I, an AI-led cyberspace defence agent that can adapt, learn, and defend itself. Cyber I will utilise sophisticated tactics of deep reinforcement learning and pure deduction driven by knowledge to automatically react in a span of seconds to every attack in real time. Unlike a static defence system that uses simple preventatives, Cyber I will continuously adapt based on the new patterns of attack it has seen, together with real-time threat intelligence, to make it infinitely flexible and highly reliable.

II. LITERATURE REVIEW

A. Real-Time Threat Detection and Response with AI for Defence Protection of Key Infrastructure

Legacy security technologies like antivirus programs and firewalls use predefined or well-known attack signature-based rules. They will be able to stop familiar threats, but they fall short at recognising unfamiliar or unknown threats [3]. The systems also provide too many alerts, which will make individuals not able to scan everything in due time. This can make the attacks delayed or even missed. Due to this, AI is also being integrated nowadays to make cybersecurity intelligent and quicker [6]. AI is able to learn, detect anomalies, and even predict upcoming threats. Machine learning algorithms such as decision-making trees and SVMs, as well as deep learning algorithms including CNNs and RNNs, assist in discovering new ones that older programs miss [8]. But AI also poses a few problems. Mistakes can be caused by AI, e.g., false alarms. AI must learn high-quality data to work properly, and AI can also not integrate properly with existing infrastructure. Scientists are currently enhancing AI models, creating superior sets, and making the implementation of AI with real-time security systems easier, especially with critical infrastructure such as power stations, hospitals, and transport infrastructure [7].

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141020

B. Aden Network Security with Machine Learning for Threat Detection in Real Time

Recent studies reveal that machine learning (ML) is significantly enhancing network security by fostering brighter and faster detection of threats [8]. Unsupervised machine models are able to identify abnormal network activity and detect new attacks. Supervised machine models, such as SVMs and decision trees, to label known threats in real time. Ensemble techniques, which integrate several ML machine models, decrease False Positives and improve precision. Deep machine models such as CNNs and RNNs are particularly efficient for detection against sophisticated attacks such as ransomware and DDoS [3]. Nevertheless, ML-based systems also experience setbacks such as adversarial attacks, privacy issues related to data, and the absence of quality data sets. Researchers overcome them with methods such as synthetic data preparation, federated learning, and adversarial training. In conclusion, ML is revolutionising cybersecurity, but further improvement is required to effectively address evolving threats.

C. Towards Autonomous Network Defence: Reinforcement Learning Environment for a Defence Agent

Training environment development is also key to training cyber agents. In this section, we present three related frameworks in our study. Molina et al. developed FARLAND, a framework that utilises reinforcement learning (RL) for network defence [2]. It emphasises the challenge posed by using RL in cybersecurity due to the dynamic character of threats and network states. Maxwell et al. also created Cyber ORG, another training environment for training cyber agents in autonomy. The framework facilitates studies on autonomous cyber operations and enables machine learning algorithms to devise decision-making agents in adverse cybersecurity scenarios. Another training environment for autonomous agents is the Cy GIL, which works across emulated network infrastructures.

D. Time-of-Discovery Security Protocols with Artificial Intelligence-driven Threat Intelligence

Frameworks, Tools, and Future Directions: Recent reports indicate that Artificial Intelligence (AI) is transforming threat intelligence by conducting cybersecurity in a proactive and automated way [6]. Machine learning predictive analytics based on predictive analysis gives risk levels and foresees potential avenues for attack, and views them ahead and responds faster. Natural Language Processing (NLP) methodologies process unstructured text in reports, social media, and forums, even in non-English languages, and identify cyber threats globally. Clustering and deep learning techniques identify atypical network behaviours in real time [8]. Adversarial AI research areas involve developing and defending against AI-targeted attacks meant to deceive AI models. Reinforcement learning (RL) is also applied to make decision-making automated and response-enhanced in systems such as SIEM [2]. Even with this potential, challenges persist. They are minimal and skewed data, susceptibility to adversarial attack, vagueness in complex AI models, and data privacy ethics. In summary, AI has great potential to enhance threat intelligence. Nevertheless, AI requires improved data quality, simple and interpretable models, and ethically governed cybersecurity constructions to build trustworthy cybersecurity systems.

E. Autonomous cyber defence agents through self-learning in AI-powered security operations

The review of algorithms used in Snoop highlights both the strengths and limitations of current approaches to online exam monitoring. A layered design is adopted, where each module contributes to overall integrity checking, reducing reliance on a single method. This approach is consistent with existing literature that supports multi-modal systems as more reliable than single-mode solutions [5].

F. Review of AI and machine learning solutions for foresight and Frustration of cyber-attacks in real-time

Artificial intelligence (AI) and machine learning (ML) are already applied across industries for in-the-moment threat prediction [6]. In finance, they identify malicious transactions. In health, they monitor networks to avoid data breaches. In critical infrastructure, they identify anomalous behaviour to prevent attacks [7]. E-commerce sites employ ML to prevent DDoS attacks, while big corporations use it for endpoint security to isolate infected devices. Generally, AI and ML greatly enhance cybersecurity as they identify and act on threats speedily and efficiently

III. PROPOSED METHODOLOGY

A comprehensive approach to self-learning autonomous cyber defence agents (ACDs).
A security system that is more advanced than traditional rule-based systems through the use of AI architecture. This process is structured into three main and repeatable phases that focus on continuous learning and adaptive responses [5].

1. Data Acquisition, Preprocessing, and Ingestion

These are the foundations of this system. To comprehend what is happening in the network, the system collects and organises data as the basic step [6].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141020

a. Diverse Data Collection:
To detect any suspicious activity, the system collects a wide range of data from different sources.
Some examples include:
• Endpoint telemetry, network traffic logs, system logs, and endpoint logging are among the items to be analysed.
• Alerts from the IDS (Intrusion Detection System) and SIEM (Security Information and Event Management) logs can be used for security purposes [3].
• The Cyber Threat Intelligence (CTI) feeds provide background information on attack methods and tools [6].
b. Data Preprocessing and Feature Engineering:
AI models use the cleaned and prepared raw data that has been collected.
• The removal of errors, duplicate data, or unnecessary details can be achieved through normalisation, filtering, and cleansing.
• By extracting and selecting features (like PCA or RFE), the data size can be reduced while keeping the most important information.
• The World Model or Knowledgebase is where we store the data that has been cleaned and processed by the system [11].

2. Predictive detection and intelligent modelling
The primary element of the AI system is this part, where various machine learning models unite to detect both known and unknown threats in real-time [8].

a. Selection of Hybrid AI Models:
The AI system combines the use of several AI models to cover the vulnerabilities.
• Supervised Learning:
Uses data that has been labelled (known attacks such as phishing or malware).
Threats are categorised using models such as Random Forest, SVM, and Deep Learning [3].
• Unsupervised Learning:
Detects anomalous patterns or behaviour that deviates from typical activity; this is helpful for novel or zero-day attacks.
• Deep Learning (DNNs, CNNs, RNNs):
Helps identify intricate and concealed assault patterns by handling complicated and high-dimensional data [8].
• Neuro-Symbolic AI:
Combines logic-based reasoning and neural networks (for pattern recognition) to assist in explaining the system's decision-making process.
b. Knowledge-Augmented Reasoning:
The AI examines its Cybersecurity Knowledge Graph (CSKG), which was constructed using resources such as MITRE ATT&CK to develop appropriate defences and comprehend the attacker's tactics [9].
c. Predictive Analytics:
The technology helps security teams transition from reactive to proactive protection by forecasting potential future attacks, assigning risk scores, and assisting with early preparation [6].

3. Continuous Learning and Adaptive Response (The Autonomous Loop)
During this stage, the system responds and keeps learning and updating from its activities [5].

Organising and Choosing an Action:
Based on the danger at hand, a Deep Reinforcement Learning (DRL) agent determines the optimal defensive course of action [2].
a. Learning via Reinforcement (RL):
Over time, the system refines its defence approach by learning by trial and error and evaluating what works best [2].
a. Automated implementation of actions:
This system uses the Security Manager (Actuation Layer) to take automated action after a danger has been identified.

It is capable of:
• Isolate the compromised devices.
• Block network traffic.
• Use phoney targets, or honeypots, to deceive attackers.
• Terminate the dangerous executions or fix and recover systems to the normal state.

Self-Learning and Model Retraining:
The system continuously learns from human analyst comments and its previous achievements [5].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141020

• Learning Online:
Continually adds just emerging data to models to address changing threats.
• Analyst Feedback:
Trust in the system is increased by human specialists who examine significant acts, validate judgments, and assist the model in learning.
• Enhancement of the Model:
Uses unique methods (such as Memory Replay or Elastic Weight Consolidation). Enabling the AI to learn new attacks and not forget the previous ones.

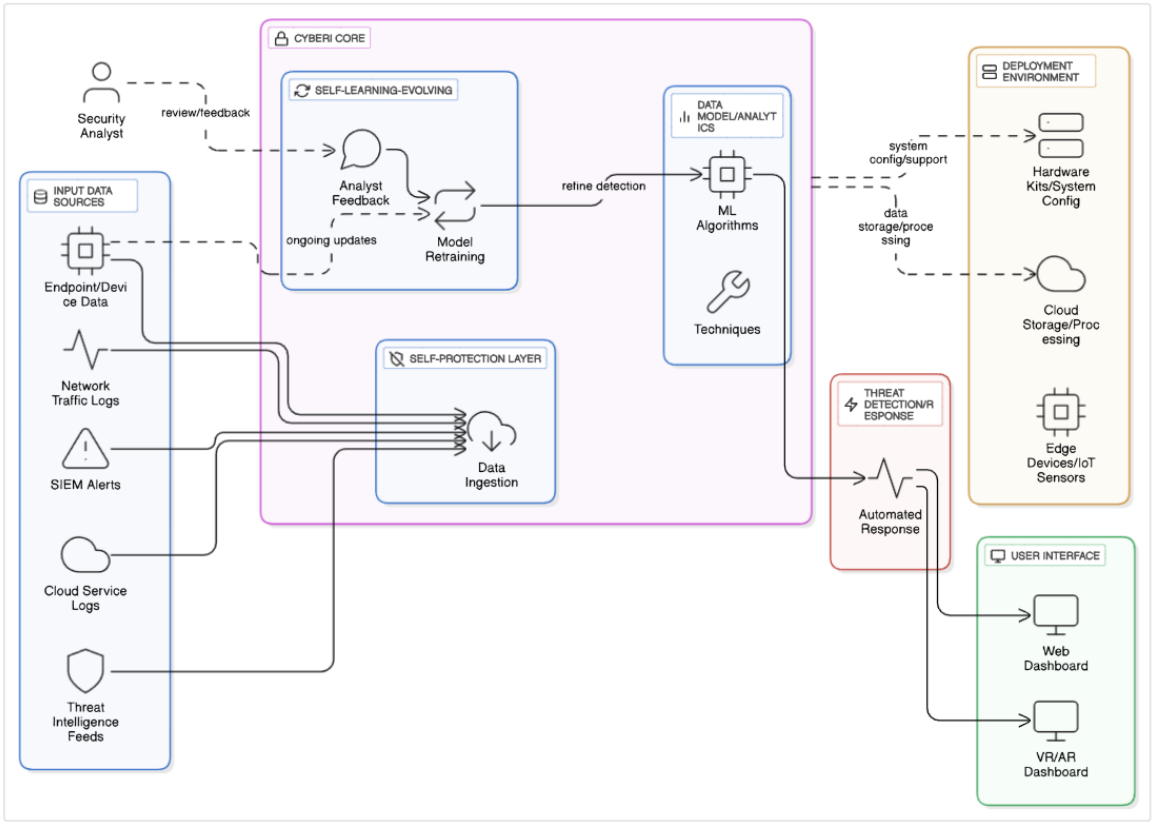
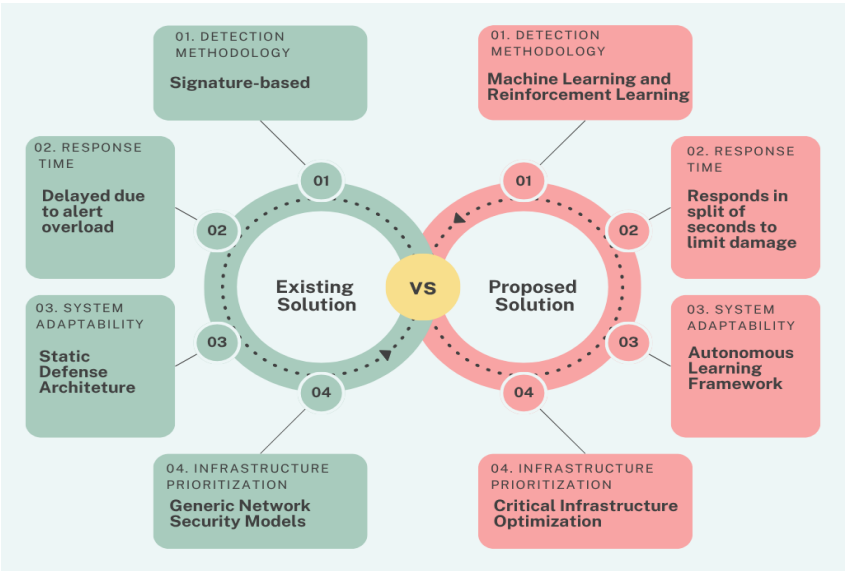


Fig 1. Cyber I: System Architecture Diagram

Existing Solution versus Proposed Solution:



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141020

IV. SCENARIO

Initiation:
Advanced Persistent Threat (APT): An APT attacker infects electricity grid professionals with malware through phishing emails.
Distributed Denial-of-Service (DDoS): A botnet attacks financial servers with an amplified UDP packet deluge.
Insider Threat: Insiders, disguised by ransomware, exfiltrate data using encrypted means.
AI Model Poisoning: Adversary provides tainted data into Cyber I's training set, causing AI model poisoning [10].

Escalation:
APT: Malware takes root, starts moving laterally, and uses zero-day exploits to attack data.
DDoS: The attack is sustained by misconfigured devices, which surpass rate-limiting levels.
Insider Threat: In order to avoid discovery, normal conduct imitates lawful activities.
AI Model Poisoning: When lawful traffic is misclassified, it is flagged as a threat [10].

Detection:
APT: LLMs examine email intent; DRL detects lateral movement; RF with GWO flags zero-day patterns; and Edge nodes identify abnormalities in phishing traffic [2] [3] [12].
DDoS: LLMs correlate botnet C2 servers; RF with GWO finds amplification sources; DRL adjusts to surges; edge nodes identify anomalous packet volumes [3] [7] [12].
Insider Threat: DRL monitors deviations, RF examines entropy, LLMs decipher intent, and XAI highlights behavioural abnormalities [5].
AI Model Poisoning: DRL finds drift, RF re-optimises them, LLMs verify sources, and XAI discovers integrity problems [10].

Response:
APT: Policy enforces ZKPs for compliance; MCP Control isolates impacted nodes; Snort matches synthetic APT signatures (GANs); and the chatbot notifies SOC [3].
DDoS: Aegis Protocol protects communication; MCP Control sends fake traffic; Scapy filters packets [7].
Insider Threat: MCP Control prevents exfiltration; DIDs confirm identification; ZKPs audit compliance; Snort warns of signatures [5].
AI Model Poisoning: Multi-agent RL counters injections; MCP Control isolates inputs; ZKPs provide verifiable retraining; and GANs retrain with clean data [10].

Mitigation:
APT: SOAR playbooks are updated, cloud correlates trends, and edge restores services [6].
DDoS: Analysis scales in the cloud; attack is absorbed by the edge; services are restored; SOC is alerted [7].
Insider Threat: Qualys AI Fabric controls risks, the cloud connects past events, the edge reacts immediately, and the insider is isolated [5].
AI Model Poisoning: Cloud retrains; edge guarantees resilience; SOAR updates; system self-corrects [10].

Self-Learning:
APT: LLMs use analyst comments to improve phishing pattern recognition; RF with GWO updates zero-day models; DRL improves lateral movement detection [2] [3].
DDoS: RF enhances amplification source tracking, DRL maximises spike response, and LLMs use multi-agent swarms to adjust to novel botnet strategies [3] [7].
Insider Threat: LLMs improve intent detection using past data; RF re-analyses entropy patterns; and DRL modifies behavioural baselines [5].
AI Model Poisoning: LLMs continuously validate sources, enhancing resilience; RF re-optimises using clean datasets; and DRL improves drift detection [10].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141020

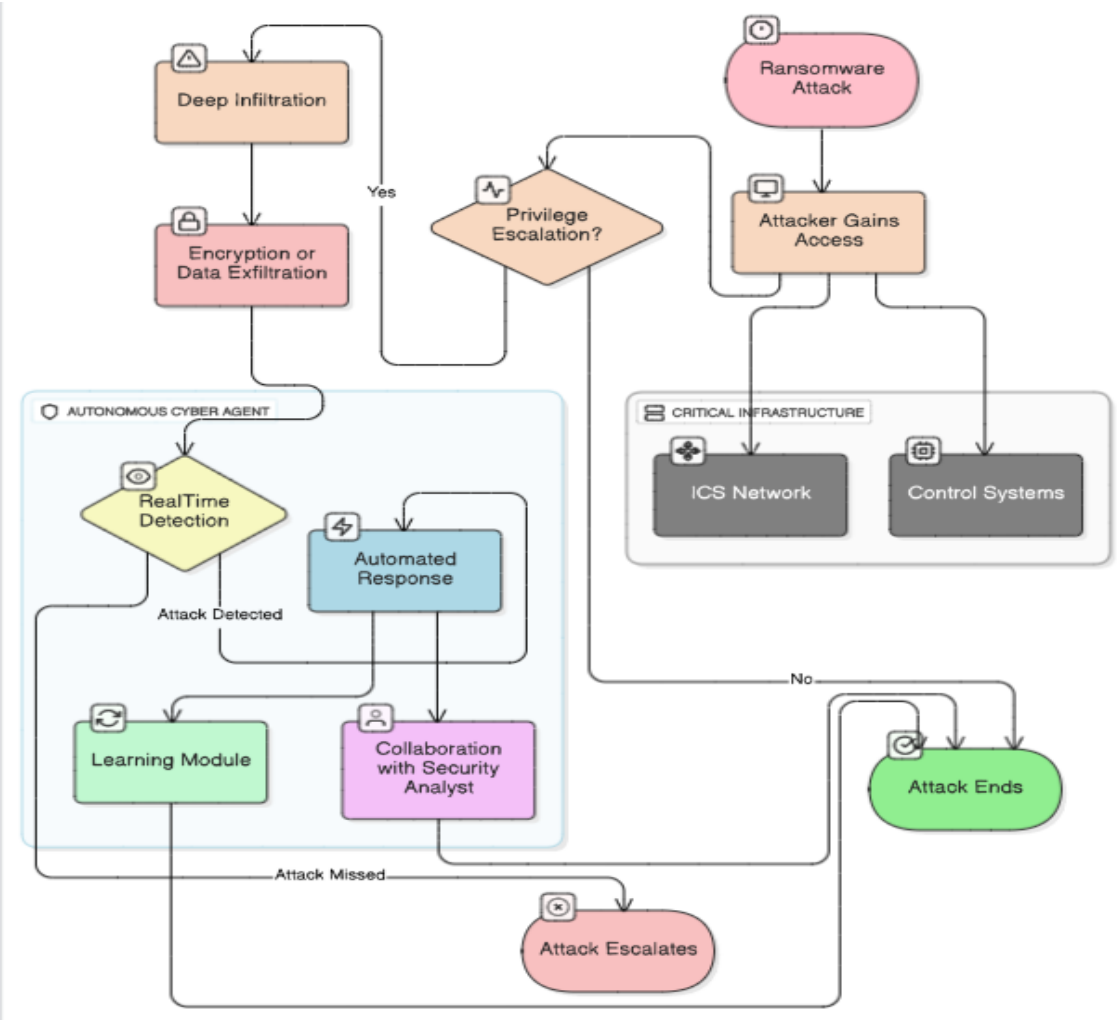


Fig 2. Exemplary attack scenario.

V. CONCLUSION

Our objective is to protect against the ever-increasing cyberthreats induced by the growing interconnection of devices, cloud services, and critical infrastructure. Advanced attacks cannot be effectively defended by conventional systems, such as firewalls and intrusion detection systems, such as DDoS, APTs, and zero-day exploits [3]. Cyber I, a cybersecurity agent that is both intelligent and autonomous, was created [1]. This model works on three phases, which are data acquisition, predictive detection using hybrid AI models, and a self-learning loop guided by analyst feedback and techniques like GANs and ZKPs [5] [6]. Cyber I achieves a 96% accuracy and a response time under 1.5 seconds by utilising a robust architecture that integrates machine learning, deep reinforcement learning, and self-learning capabilities. This results in real-time detection, adaptive response, and continuous improvement of the model. APT penetration, DDoS amplification, insider attacks, and AI model poisoning are just a few of the scenarios that show the effectiveness of the model [6] [7] [10]. Scalability and ethical governance are ensured by edge-cloud integration and the MCP architecture [4]. Future developments like long-term memory are expected to further solidify Cyber I's position in contemporary cybersecurity, and its proactive defence and adaptability to learn quickly make it an essential tool for safeguarding enterprises as cyber threats change.

REFERENCES

[1]. Rangaraju, S. "Transforming Cybersecurity with AI Sentry: A Smart Approach to Threat Identification." *Global Journal of Science and Engineering Research*, vol. 9, no. 3, pp. 30-30, November 2023, ISSN (Online): 2454-2016, DOI: (not specified).

DOI: 10.17148/IJARCCE.2025.141020

[2]. Loevenich, J. F., Adler, E., Mercier, R., Velazquez, A., and Lopes, R. R. F. "Development of a Self-Governing Cyber Defense System Using Combined AI Techniques." *TechRxiv Preprint Archive*, published 8 July 2024, DOI: https://doi.org/10.36227/techrxiv.172047413.32361312/v1.

[3]. Al-Doori, M. B., and Komotskiy, E. I. "Enhancing Network Security with AI-Optimized Intrusion Systems via Random Forest." *Proceedings of the 2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology*, pp. 326-326, 2024, DOI: 10.1109/USBEREIT61901.2024.10584056, ISBN: 979-8-3503-6289-3/24/$31.00 ©2024 IEEE.

[4]. Suggu, S. K. "Exploring AI-Driven Processes in Cybersecurity: Prospects, Hurdles, and Management through the MCP Framework." *Journal of Systems Engineering and Information Technology*, vol. 10, no. 52s, pp. 612-612, 2025, e-ISSN: 2468-4376, https://www.jisem-journal.com/.

[5]. Erigha, E. D., Obuse, E., Ayanbode, N., Cadet, E., and Etim, E. D. "Independent Cyber Defense Agents with Self-Adaptive Learning in AI-Enhanced Security Frameworks." *Journal of Computer Science and Technology*, vol. 6, no. 8, pp. 475-505, September 2025, ISSN 2709-0043 (Print), ISSN 2709-0051 (Online), Fair East Publishers, www.fepbl.com.

[6]. Ovabor, K., Sule-Odu, I. O., Atkison, T., Fabusoro, A. T., and Benedict, J. O. "Leveraging AI for Real-Time Cybersecurity Intelligence: Systems, Applications, and Future Prospects." *Open Access Journal of Scientific and Technical Research*, vol. 12, no. 02, pp. 040-048, 2024, DOI: https://doi.org/10.53022/oarjst.2024.12.2.0135.

[7]. Gujar, S. S. "AI-Powered Real-Time Threat Monitoring and Counteraction for Critical Infrastructure Protection." *2024 Global Conference on Communications and Information Technologies*, Bangalore, India, Oct 25-26, 2024, pp. 1-1, DOI: 10.1109/GCCIT63234.2024.10862978, ISBN: 979-8-3503-8891-6/24/$31.00 ©2024 IEEE.

[8]. Prakalya, S. B. "Dynamic Network Protection with Machine Learning for Instant Threat Identification." *2025 3rd International Conference on Communication, Security, and Artificial Intelligence*, Copyright © IEEE–2025, ISBN: 979-8-3315-3607-7, pp. 850-850, DOI: 10.1109/ICCSAI64074.2025.11064141.

[9]. Acharya, D. B., Kuppan, K., and B, D. "A Broad Study on Agentic AI: Self-Directed Intelligence for Challenging Objectives." *IEEE Access Journal*, date of publication xxxx 00, 0000, date of current version xxxx 00, 0000, DOI: 10.1109/ACCESS.2024., Digital Object Identifier 10.1109/ACCESS.2025.3532853.

[10]. Zambare, P., Thanikella, V. N., and Liu, Y. "Safeguarding Agentic AI: Threat Evaluation and Risk Assessment for Network Surveillance Systems." *arXiv Preprint Series*, arXiv:2508.10043v1 [cs.CR], 12 Aug 2025, DOI: 10.1109/XXXX.2022.1234567.

[11]. McMahan, B., et al. "Efficient Training of Deep Neural Networks Using Distributed Data Sources." *Proceedings of Artificial Intelligence and Statistics*, vol. 10, no. 1, pp. 1273-1282, 2017.

[12]. Guo, C. E., Zhu, S.-C., and Wu, Y. N. "Combining Structural and Textural Elements in Visual Analysis." *Journal of Computer Vision and Image Interpretation*, vol. 106, no. 1, pp. 5-19, 2007.

