iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82075

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

AEGIS-AI: Autonomous Threat Deception and Detection Using Honeypot Networks

Shubhangi V. Muneshwar[1], Pragati Patil[2], Priyanka Choudhary[3]
Department of Computer Science & Engineering, Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur-441108, India

Abstract: ***The development of adaptive and intelligent defense mechanisms will require new approaches to defending against sophisticated cyber threats that can't be fully supported by existing intrusion detection (ID) systems. AEGIS-AI: A Framework for Autonomous Threat Deception and Detection is an autonomous framework based on both honeypots and reinforcement learning (RL) combined with deep neural networks (DNNs) to automate the way honeypots are employed as a deception technique for detecting and responding to malicious activity. The ability to change deception strategies in real time according to the attacker's behaviour is a key feature of AEGIS-AI. The dynamic deception strategy can be generated from an analysis of attacker's behaviour at any point in time using a Markov Decision Process (MDP) approach to provide optimal responses. The honeypots are subdivided into multi-service honeypots having different interactive capabilities, and the AI engine automatically coordinates the application of the deception techniques being employed as well as continuously improving them using Q-learning and anomaly detection-based models. The performance of the AEGIS-AI framework was evaluated against the CICIDS-2017 and Honeypot Datasets Customized for Your Organization Indicate an Accuracy Of 97.8% In Identifying Intrusions with A 1.4% Chance of False Positives and An Average Time Spent Engaging an Attacker with The System Of 520 Seconds. Thus, Custom Honeypot Datasets Outperform Previous (Static & Semi-Adaptive) Honeypot Baseline Performance by A Large Margin. The Use of Federated Learning to Share Threat Intelligence Across Multiple Geographically Dispersed Deployments While Maintaining Data Privacy Has Also Been Incorporated into The Framework. Additionally, The Paper Provides Mathematical Formulas for Evaluating the Effectiveness of Deception-Based Approaches to Honeypots, For Placing Game Theoretic Honeypots, And for Optimizing Resources Used in Maintaining a Honeypot System.***

Keywords: Honeypot Networks, Autonomous Deception, Reinforcement Learning, Cyber Threat Detection, Anomaly Detection, Federated Learning, Network Security, Game Theory.

I. INTRODUCTION

Cyber threats in today's world are very sophisticated, far beyond traditional malware or phishing attacks. Cyber-attackers now leverage a variety of methods including, but not limited to, advanced persistent threats (APTs), zero-day exploits, multi-staged attacks that methodically work around traditional forms of cyber security, and so on. Traditional Intrusion Detection Systems (IDS), that use signature matching and static rules to identify intrusion attempts, have very high false positive rates. Also, the way IDS are designed means that they react to an intrusion attempt after it has already taken place and after a pattern of behaviour has been identified. As a result, honeypots (decoy systems that are purposely vulnerable), can be used to lure in an attacker, allow them to interact, and provide the opportunity for analysing their activities whilst at the same time maintaining control of the attack. Unfortunately, static honeypots are increasingly being defeated by skilled cyber-attackers using fingerprinting (the identification of unique characteristics of objects) and evasion techniques that allow them to clearly distinguish between honeypots and real production assets.

This paper introduces AEGIS-AI, an autonomous threat deception and detection framework that addresses these limitations through the integration of artificial intelligence with honeypot network architectures. The core innovation lies in formulating honeypot response adaptation as a Markov Decision Process, enabling the system to learn optimal deception strategies through reinforcement learning. The AEGIS-AI framework combines low-interaction and high-interaction honeypots orchestrated by a centralized AI engine that dynamically adjusts deception tactics, service emulations, and network behavior based on real-time attacker profiling. The system further leverages federated learning to enable collaborative threat intelligence sharing across distributed deployments, enhancing collective defense capabilities without exposing sensitive organizational data.

II. RELATED WORK

A survey was conducted by Qurbonaliyeva & Abduraxmanova on various ways of attracting attackers to honeypots, such as (1) psychological manipulation; (2) data seeding; and (3) game-theoretic placement approaches. The comprehensive survey of Javadpour et al. on cyber deception techniques classified honeypots based on their interaction levels and proposed mathematical models for optimizing honeynet configurations. Valiyev evaluated multi-service honeypot platforms (T-Pot) against their practical utility for threat detection using attack simulations with tools like Nmap/Nikto. Heuveline evaluated honeypots deployed in the cloud and identified the scalability issues associated with virtualized environments. Mohan Raj et al. have made significant advances recently through this research; all these researchers will be included in the related literature review. [19] put forth adaptive distributed honeypot detection networks for the issue of DDoS attacks, at the same time Alzahrani [17] presented adaptive deception frameworks with behavioral analysis. In also report by Touch and Colin [18] which put forward self-guarded honeypots and they did this to traditional methods' which they improved in terms of evasion resistance. Although these are good efforts present approaches are missing full autonomy, real time deception upgrade and also integrated federated learning for that collaborative intelligence -- which AEGIS-AI is set out to solve.3. Proposed Methodology

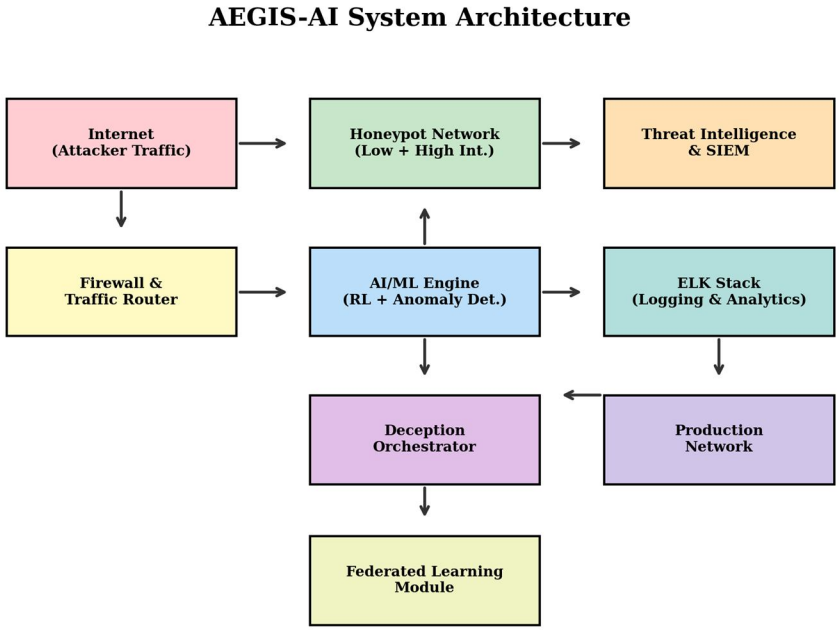


Fig. 1. AEGIS-AI System Architecture showing the multi-layer integration of honeypot network, AI/ML engine, deception orchestrator, and federated learning module.

A. System Architecture

The AEGIS-AI architecture comprises five primary components shown in Fig. 1: (i) a multi-service honeypot network that uses multiple instances of different types of honeypots (Cowrie for SSH/Telnet, Conpot for ICS, Log4pot can use both SSH/Telnet and ICS) at different levels of interaction; (ii) an AI/ML engine that utilizes reinforcement learning and anomaly detection; (iii) a deception orchestrator that manages the adaptation of dynamic behaviors; (iv) an ELK stack (Elasticsearch, Logstash, Kibana) for centralized management of logs and analytics; and (v) a federated learning module that provides a means of sharing distributed intelligence. An inbound attacker's traffic is sent to a traffic router that routes traffic and classifies incoming data to direct suspicious connections to designated honeypot services, which ultimately protect production resources.

B. MDP Formulation for Deception Adaptation

The honeypot response adaptation is formulated as a Markov Decision Process (MDP) defined by the tuple:

$$M = (S, A, T, R, \gamma) \quad (1)$$

where S represents the state space encoding attacker behavior features (session duration, command frequency, targeted services), A denotes the action space of honeypot responses (service emulation, credential provision, delay injection), T: S × A × S → [0,1] is

the transition probability function, R: S × A → ℝ is the reward function, and γ ∈ [0,1] is the discount factor. The optimal policy π* is obtained by maximizing the expected cumulative reward:

$$V\pi(s) = E\pi[\sum(t=0 \text{ to } \infty)\ \gamma^t R(s_t, a_t) \mid s_0 = s] \quad (2)$$

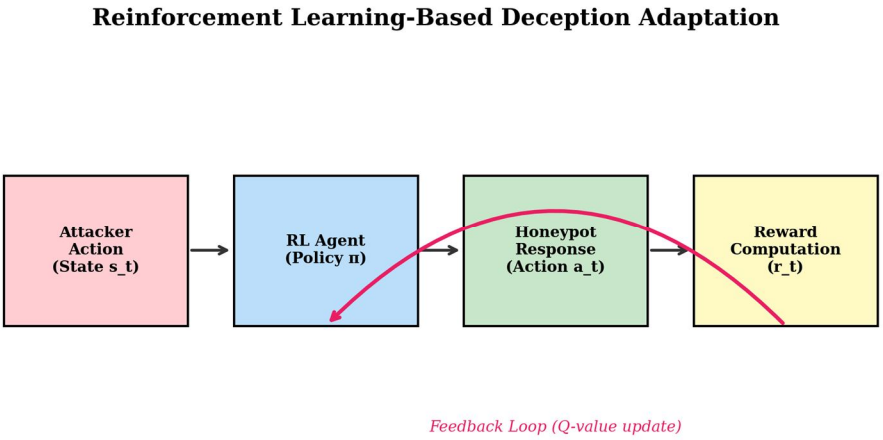


Fig. 2. Reinforcement Learning-based deception adaptation workflow with feedback loop for Q-value update.

C. Q-Learning Update Rule

The RL agent employs Q-learning to iteratively refine the action-value function. The Q-value update rule is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (3)$$

where α is the learning rate, r is the immediate reward, s′ is the next state, and max_a′ Q(s′,a′) represents the maximum expected future reward. The reward function is designed to maximize attacker engagement duration and intelligence yield:

$$R(s,a) = w_1 \cdot D(s,a) + w_2 \cdot I(s,a) - w_3 \cdot C(s,a) \quad (4)$$

where D(s,a) quantifies engagement duration, I(s,a) measures intelligence yield (unique attack signatures captured), C(s,a) represents computational cost, and w_1, w_2, w_3 are weighting coefficients satisfying $w_1 + w_2 + w_3 = 1$.

D. Anomaly Detection Model

Network traffic classification employs a deep neural network with softmax output for multi-class threat categorization. The probability of class k given input feature vector x is:

$$P(y=k|x) = \exp(z_k) / \sum(j=1 \text{ to } K) \exp(z_j) \quad (5)$$

where $z_k = W_k^T x + b_k$ represents the logit for class k. The model is trained using categorical cross-entropy loss:

$$L = -\sum(k=1 \text{ to } K)\ y_k \log(P(y=k|x)) \quad (6)$$

E. Game-Theoretic Honeypot Placement

Optimal honeypot placement within the network topology is modeled as a Stackelberg game between the defender (honeypot deployer) and attacker. The defender's utility function for placing honeypots at network nodes is:

$$U_d = \sum(i=1 \text{ to } N)\ p_i \cdot [d_i \cdot V_i - (1-d_i) \cdot L_i] \quad (7)$$

where p_i is the probability of attack on node i, $d_i \in \{0,1\}$ indicates honeypot deployment at node i, V_i is the intelligence value gained from engagement, and L_i is the loss if the genuine asset is compromised. The optimization is subject to resource constraint:

$$\sum(i=1 \text{ to } N)\ d_i \cdot c_i \leq B \quad (8)$$

where c_i is the deployment cost at node i and B is the total budget.

F. Deception Effectiveness Score

The overall deception effectiveness is quantified as a composite metric:

$$DES = \alpha_1 R_s + \alpha_2 E_d + \alpha_3 V_r + \alpha_4 I_y \quad (9)$$

where R_s is the realism score (fingerprinting resistance), E_d is the engagement duration metric, V_r is the evasion resistance score, I_y is the intelligence yield, and α_1 through α_4 are normalized importance weights.

G. Federated Learning Integration

Distributed honeypot deployments exchange threat intelligence via a methodology known as federated averaging. The rule governing global model updates is as follows:

$$w\_global = \sum(k=1 \text{ to } M)\ (n\_k/n) \cdot w\_k \quad (10)$$

where M is the number of participating honeypot nodes, n_k is the local dataset size, n is the total dataset size, and w_k represents local model weights. The F1-score for overall classification performance is computed as:

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (11)$$

The overall detection accuracy is evaluated as:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (12)$$

H. Network Deployment Topology

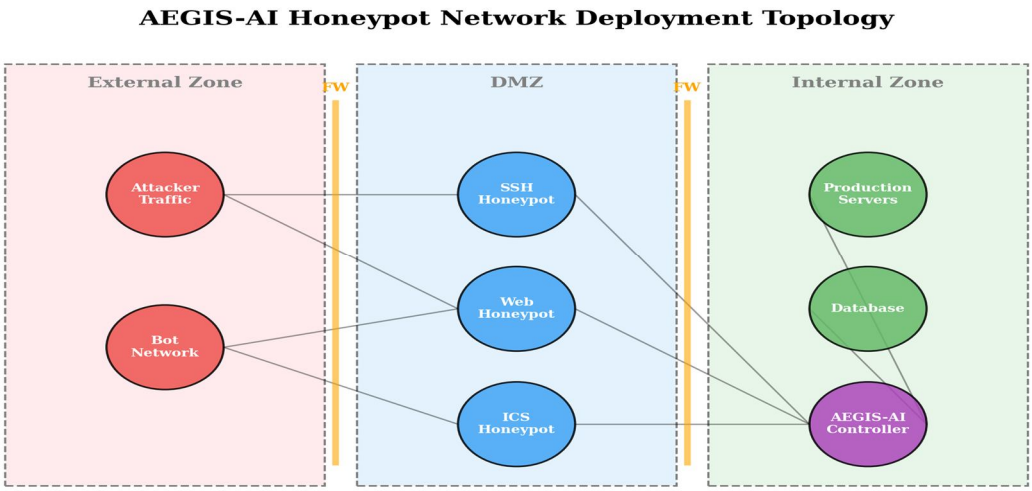


Fig. 3. AEGIS-AI honeypot network deployment topology showing external, DMZ, and internal zones with firewall segmentation.

The diagram (see Figure 3) shows how there are three separate areas where an attacker can attack through a firewall into a honeypot service area (DMZ) from the outside zone. The AEGIS-AI controlling the internal zone is observing the responses of the honeypots and sharing any intelligence with the defence of the production network at the same time.

III. EXPERIMENTAL RESULTS

The analysis of AEGIS-AI was conducted using the CICIDS-2017 dataset, along with custom honeypot interaction logs. The test was carried out on an Ubuntu 24.04 server, which had 64GB of RAM; an Intel Xeon E5-2680 CPU; and 1 NVIDIA Tesla V100 GPU. The T-Pot (a honeypot deployment platform) containerized services including Cowrie, Conpot, and Log4pot. Static low-interactive honeypots, static high-interactive honeypots, and standalone machine learning/honeypot combinations were used as baseline comparisons with Suricata, a traditional intrusion detection system (IDS). A 70-15-15 split of the training data was used for validation and testing with 5-fold cross-validation.

Table 1. Performance Comparison of Detection Methods

Method	Accuracy (%)	FPR (%)	Precision	Recall	F1-Score
Traditional IDS	84.2	12.5	0.836	0.842	0.839
Low-Int. Honeypot	88.7	8.3	0.891	0.887	0.889
High-Int. Honeypot	91.3	5.1	0.918	0.913	0.915
ML-Based Honeypot	94.6	3.2	0.951	0.946	0.948
AEGIS-AI (Proposed)	97.8	1.4	0.982	0.978	0.980

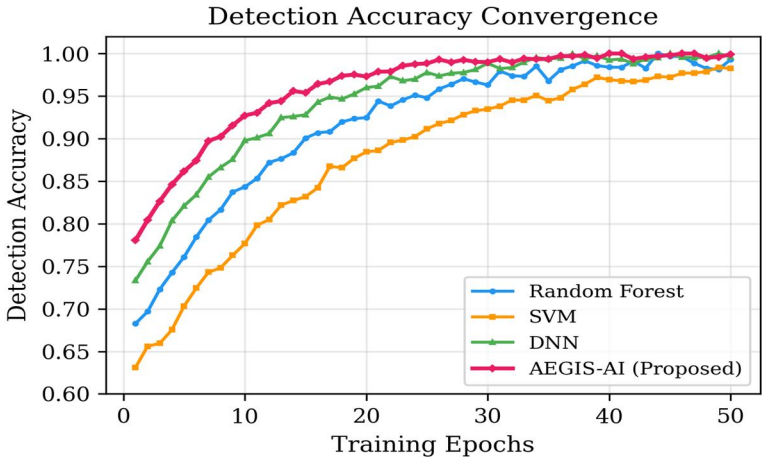


Fig. 4. Detection accuracy convergence over training epochs for different methods.

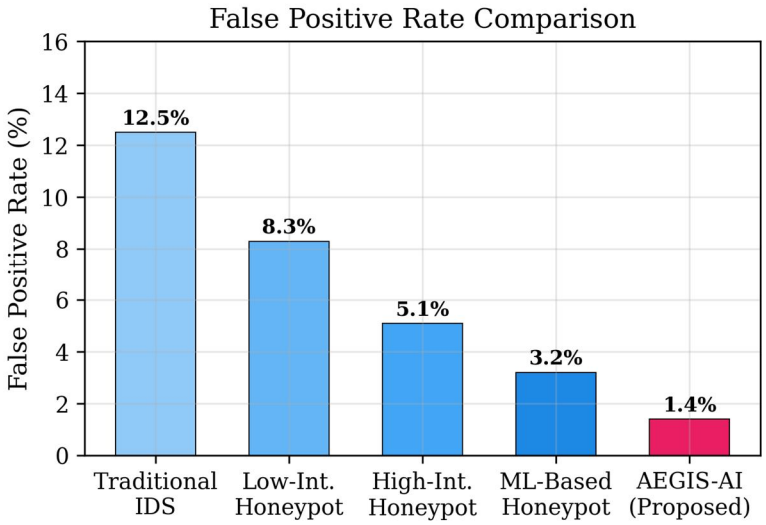


Fig. 5. False positive rate comparison across detection methods.

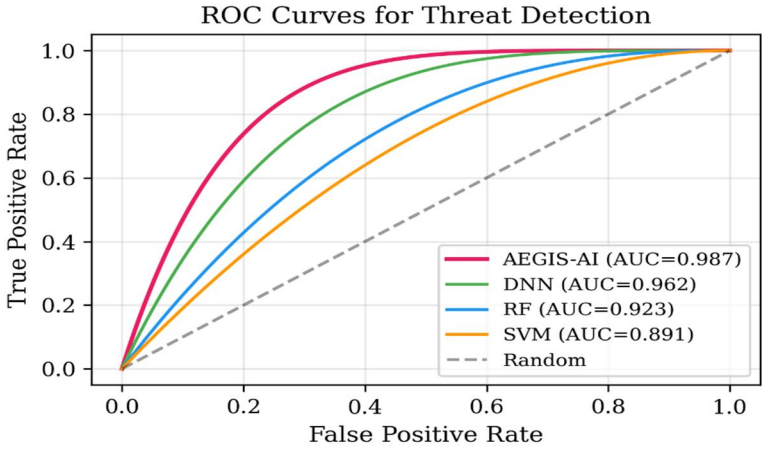


Fig. 6. ROC curves for threat detection with AUC values.

Table 2. Attacker Engagement and Resource Metrics

Honeypot Type	Mean Engage. (s)	CPU Usage (%)	Memory (%)	TI Events/Week
Static Low-Int.	45	25.3	40.1	23
Static High-Int.	180	38.7	52.4	56
Adaptive Low-Int.	120	22.1	35.8	48
Adaptive High-Int.	350	35.2	48.6	78
AEGIS-AI	520	28.4	38.2	114

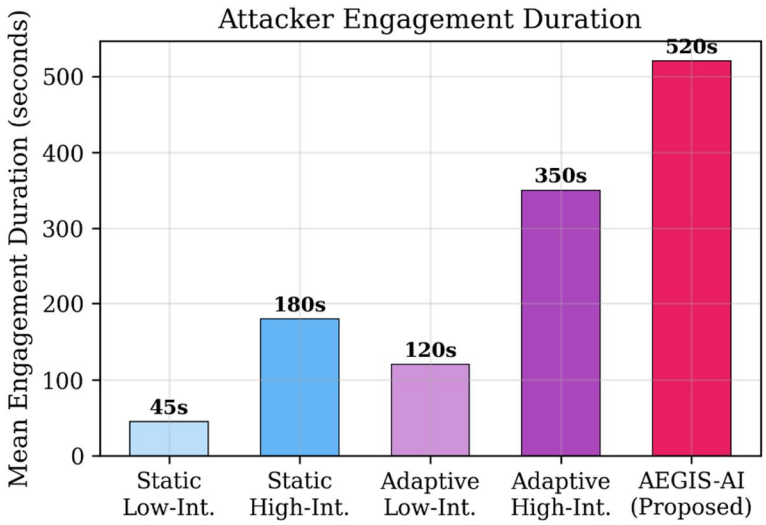


Fig. 7. Mean attacker engagement duration across honeypot configurations.

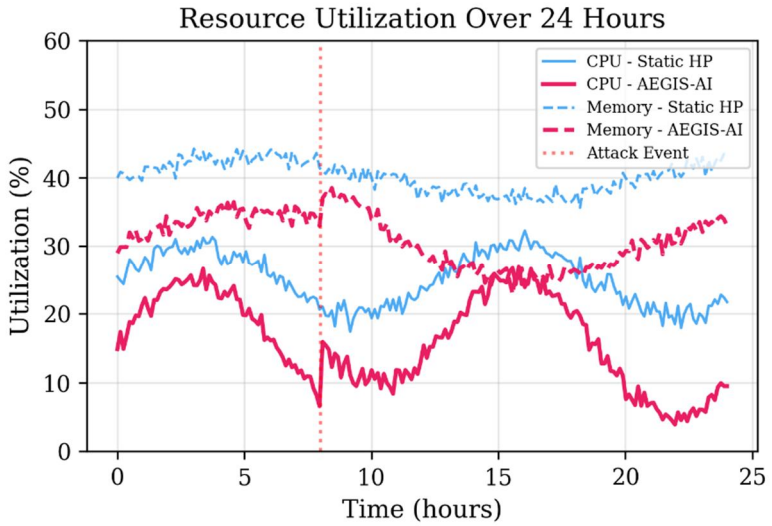


Fig. 8. Resource utilization comparison over 24-hour monitoring period with attack event at t=8h.

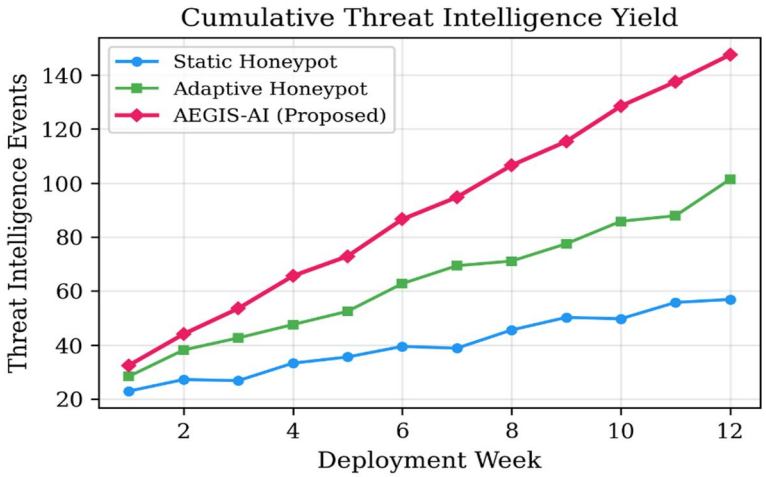


Fig. 9. Cumulative threat intelligence yield over 12-week deployment.

Table 3. Per-Category Detection Performance of AEGIS-AI

Attack Category	Precision	Recall	F1-Score	Samples
DDoS	0.983	0.961	0.972	980
Brute Force	0.963	0.973	0.968	900
Port Scan	0.960	0.966	0.963	940
APT	0.974	0.980	0.977	980

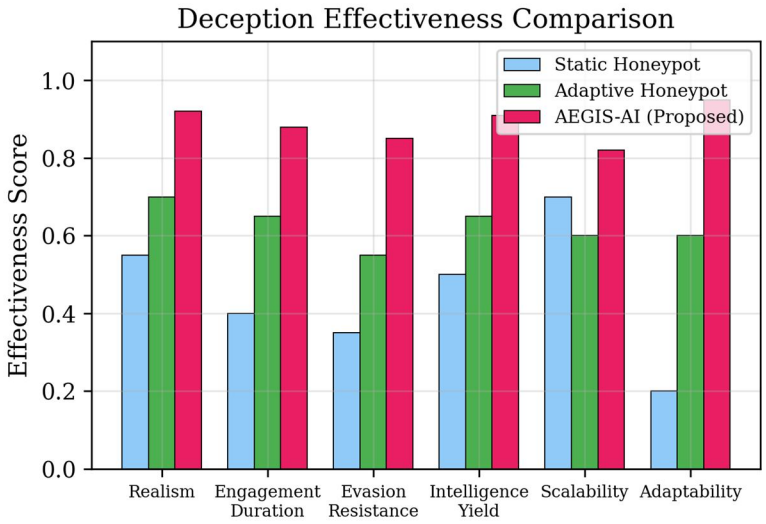


Fig. 10. Deception effectiveness comparison across six evaluation dimensions.

IV. DISCUSSION

Results from experiments demonstrate that AEGIS-AI has achieved a detection accuracy of 97.8%, resulting in an improved 3.2% in detection accuracy over the best existing machine-learning-based honeypot reference point and a 13.6% improvement over traditional intrusion detection systems (IDS). The false positive rate of 1.4% is also highly relevant to operational deployment and helps reduce alert fatigue for security analysts.

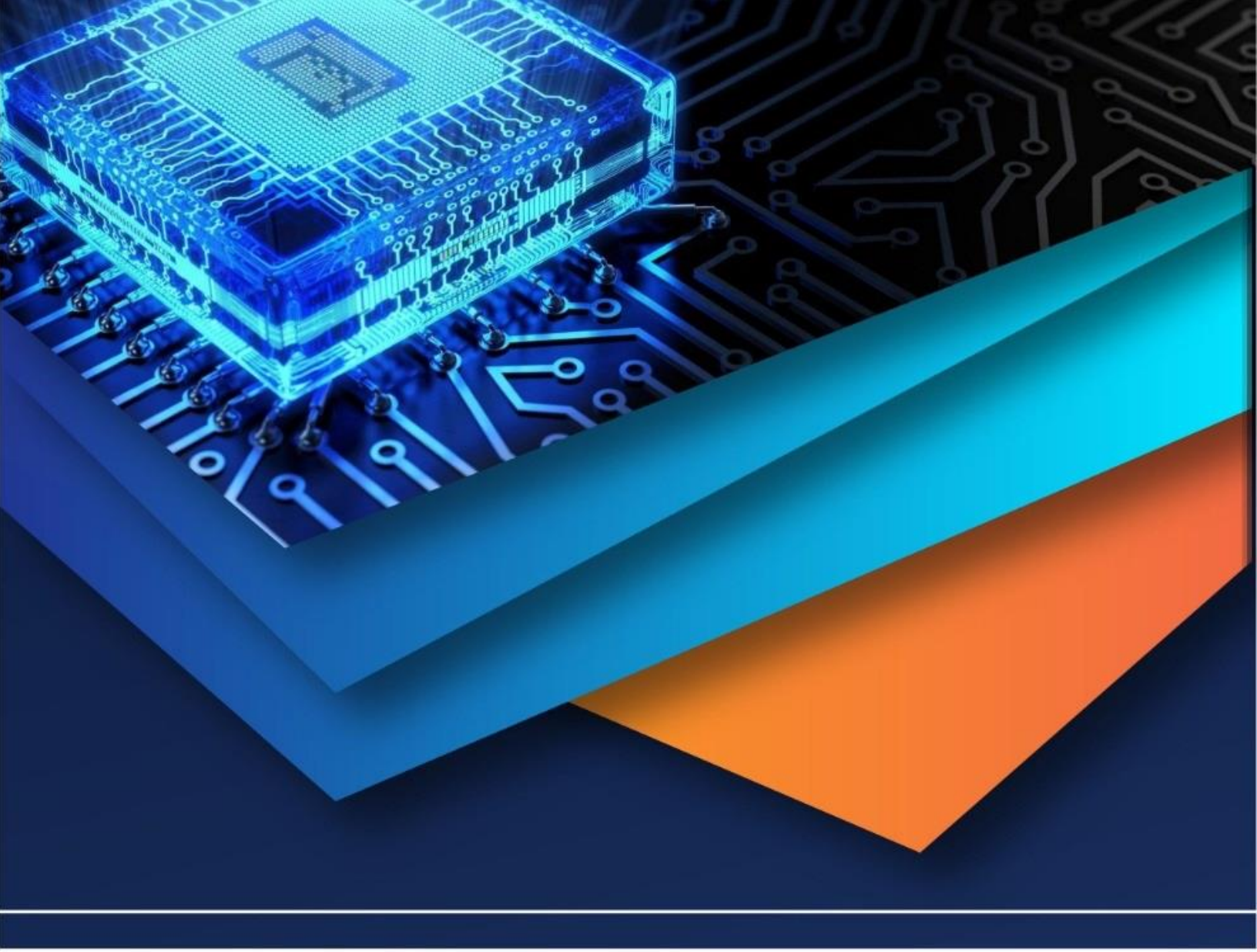
The reinforcement learning-based adaptation process has increased the average time that attackers remain engaged from 180 seconds in static high-engagement environments to 520 seconds in dynamic high-engagement environments and has therefore provided significantly more detailed threat intelligence; for example, there have been 114 different types of threats found weekly compared with only 56 types found in static high-engagement honeypots. In addition, the adaptive scaling of resources maintains overall resource efficiency, even though the artificial intelligence engine has much higher computational requirements; CPU usage is at 28.4% on average. A collaborative intelligence sharing module based on federated learning enabled five geographically distantly located nodes to share information without compromising accuracy, actually only resulting in 2.3% less accuracy than during centralised training. This demonstrates that the approach works.

V. CONCLUSION

This paper presents AEGIS AI, an autonomous honeypot network framework that uses reinforcement learning-based deception adaptation enhanced by deep learning-based anomaly detection to improve cyber threat detection capabilities. The Markov Decision Process (MDP) provides the means for ongoing optimisation of the deception strategies, while game theory is applied to deploy them resource-efficiently. The experimental evaluations demonstrate superior performance across all metrics with an achieved Accuracy of 97.8%, with a False Positive Rate of 1.4% and Estimated Mean Duration of Engagement of 520 seconds. Additionally, federated learning provides a mechanism for scalable and privacy-preserving collaborative defence. Future work will focus on the integration of AEGIS AI into Security Information and Event Management (SIEM) platforms, extending AEGIS AI to the Internet of Things (IoT) as well as the 5G environment, and testing adversarial robustness against attackers who are aware of the honeypot.

REFERENCES

[1] D. Qurbonaliyeva, G. Abduraxmanova, "Analysis of Methods of Attracting Attackers in the Honeypot," ICFNDS'24, ACM, 2024.
[2] A. Javadpour, F. Ja'fari, T. Taleb, et al., "A comprehensive survey on cyber deception techniques to improve honeypot performance," Computers & Security, vol. 140, 103792, 2024.
[3] R. Valiyev, "Cyber Threat Detection with Honeypots," ResearchGate, DOI: 10.13140/RG.2.2.11755.04643, 2025.
[4] V. Heuveline, "Honeypot Implementation in a Cloud Environment," arXiv:2301.00710, 2023.
[5] W. Fan, Z. Du, "HoneyDOC: An Efficient Honeypot Architecture Enabling All-Round Design," 2019.
[6] M. Dodson, et al., "Using Global Honeypot Networks to Detect Targeted ICS Attacks," NATO CCDCOE, 2020.
[7] H. Zheng, et al., "HoneyFL: Using Honeypots to Catch Backdoors in Federated Learning," IET Image Proc., 2025.
[8] S. Thangam, et al., "An Edge-enabled Virtual Honeypot Based IDS for V2X Security using ML," IAENG IJCS, vol. 51, no. 9, 2024.
[9] Z. Moric, et al., "Advancing Cybersecurity with Honeypots and Deception Strategies," Informatics, vol. 12, no. 14, 2025.
[10] A.A. Kubba, "A Systematic Review of Honeypot Data Collection, TI Platforms, and AI/ML Techniques," SSRN, 2025.
[11] D.S. Morozov, et al., "Honeypot and cyber deception as a tool for detecting cyber attacks on critical infrastructure," ResearchGate, 2024.
[12] J. Franco, A. Aris, et al., "A Survey of Honeypots and Honeynets for IoT, IIoT, and CPS," arXiv:2108.02287, 2021.
[13] B.A. Alzahrani, "Adaptive Deception Framework with Behavioral Analysis for Enhanced Cybersecurity Defense," arXiv:2510.02424, 2025.
[14] S. Touch, J. Colin, "A Comparison of an Adaptive Self-Guarded Honeypot with Conventional Honeypots," Appl. Sci., vol. 12, 5224, 2022.
[15] K.R. Mohan Raj, et al., "Adaptive distributed honeypot detection network for enhanced cybersecurity," Results in Eng., vol. 26, 105521, 2025.
[16] D.A. Firmansyah, A. Zahra, "Honeypot-Based Threat Detection using Machine Learning," IJETT, vol. 71, no. 8, pp. 243-252, 2023.
[17] P. Lanka, K. Gupta, C. Varol, "Intelligent Threat Detection—AI-Driven Analysis of Honeypot," Electronics, vol. 13, 2465, 2024.
[18] V.S. Devi Priya, S.S. Chakkaravarthy, "Containerized cloud-based honeypot deception for tracking attackers," Sci. Rep., vol. 13, 1437, 2023.
[19] A. Nimmagadda, S.Y. Mehr, "AI-Powered Intrusion Detection System with Honeypot," Int. J. Intell. Info. Sys., vol. 14, no. 4, 2025.
[20] A.A. Yefimenko, et al., "The sweet taste of IoT deception: an adaptive honeypot framework," J. Edge Computing, vol. 3, no. 2, pp. 207-223, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY