ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2026.0578

SECURITY THREAT ANALYSIS AND DEFENSE ARCHITECTURE FOR AUTONOMOUS AGENTS VIA ADAPTIVE ADVERSARIAL DETECTION FRAMEWORK SYSTEM

J. Jasmine[1] and Thinanath Ravichandran[2]
[1]*Department of Computer Science and Engineering, Karpagam College of Engineering, India*
[2]*Department of Data Science, Monash University, Australia*

Abstract

Autonomous agents operated across distributed and dynamic environments faced increasing exposure to security threats that disrupted decision integrity and system reliability. Prior studies indicated that adversarial interference, policy manipulation, and input perturbation affected agent behaviour, although systematic protection remained limited. The study addressed this gap by examining structured defence mechanisms for secure autonomous decision processing. The problem focused on inconsistent threat recognition and weak resilience against adversarial manipulation within multi-agent environments. Existing approaches lacked unified modelling for threat isolation and response stability under uncertain conditions. To address this issue, the study proposed an Adaptive Adversarial Isolation Framework (AAIF), which integrated multi-layer threat observation, anomaly scoring, and isolation-based response logic. The framework did not rely on static rule definitions and instead maintained adaptive separation between normal and malicious behavioural patterns. The method followed a structured analytical design where simulated autonomous agent environments underwent controlled adversarial injection scenarios. The AAIF model did analyse behavioural deviations through probabilistic mapping and comparative state evaluation across operational cycles. Performance evaluation was conducted using stability metrics, detection consistency measures, and response latency indicators. Results indicated that AAIF achieved improved threat separation consistency and maintained stable operational outputs under adversarial pressure conditions. The system reduced incorrect behavioural propagation across connected agents and strengthened decision robustness under uncertain inputs. Comparative analysis showed that AAIF sustained higher detection reliability across varied attack scenarios when compared with baseline defensive configurations.

Keywords:

Autonomous Agents, Adversarial Security, Threat Detection, Adaptive Framework, Multi-Agent Systems

1. INTRODUCTION

Autonomous agents have become essential components in modern computational ecosystems where distributed decision-making and self-regulated operations are required. These systems operate across dynamic environments and interact continuously with external inputs and peer agents [1]. The growing reliance on autonomous decision structures has increased the importance of securing operational integrity against external manipulation [2]. In recent developments, multi-agent coordination systems have expanded into domains such as intelligent transportation, industrial automation, and adaptive control systems [3]. These deployments demand consistent reliability even under uncertain or adversarial environmental conditions. Autonomous systems often function through layered perception and decision pipelines that interpret input signals and generate contextual responses. The stability of these pipelines directly influences overall system trustworthiness. As agent networks scale, inter-agent communication also introduces additional exposure surfaces that can be exploited. Therefore, security considerations must extend beyond isolated components and address system-wide behavioural consistency. Several challenges affect secure deployment of autonomous agents. First, adversarial interference often alters input signals in subtle ways that remain difficult to distinguish from legitimate variations [4]. Second, distributed coordination introduces dependency chains where a single compromised agent can propagate incorrect states across the network [5]. Third, variability in operational environments reduces the effectiveness of static detection rules and limits long-term resilience. These challenges collectively reduce the reliability of autonomous decision-making systems. Additional complexity arises from the absence of unified detection strategies capable of handling both known and unknown threat patterns. Existing frameworks often focus on specific attack types and fail to generalize across heterogeneous scenarios. This limitation restricts adaptability and increases vulnerability under evolving attack conditions.

The primary problem addressed in this study involved the lack of a robust and adaptive mechanism for identifying and isolating adversarial behaviour within autonomous agent networks [6]. Current systems did not consistently maintain operational stability under dynamic threat conditions. This gap highlighted the need for a structured approach that could continuously evaluate behavioural deviations and respond without disrupting overall system performance. The study aimed to design a structured defence mechanism that could identify adversarial behaviour in autonomous agents with higher consistency. It also aimed to maintain system stability during threat exposure and ensure reliable decision propagation across multi-agent networks. Another objective focused on developing a scalable framework that could operate across varied environments without requiring frequent structural modification.

The novelty of the proposed work lies in the Adaptive Adversarial Isolation Framework, which integrates behavioural observation with dynamic isolation logic. Unlike conventional systems, the framework does not depend on predefined attack signatures. Instead, it evaluates behavioural deviation patterns through adaptive comparison cycles and isolates anomalous agents based on computed risk divergence. This approach strengthens resilience against previously unseen attack strategies. The study contributed two primary advancements. First, it introduced a structured adaptive isolation mechanism that improved threat separation accuracy within autonomous systems. Second, it demonstrated a scalable evaluation model that maintained performance stability under varied adversarial

conditions. These contributions provide a foundation for future research in secure autonomous system design and multi-agent resilience engineering.

2. RELATED WORKS

Previous research on autonomous agent security has explored multiple detection and defence strategies across distributed computational environments. Early studies focused on rule-based monitoring systems that identified predefined anomalies within agent behaviour [7]. These approaches provided limited flexibility and struggled when exposed to unknown or evolving adversarial patterns. Subsequent research introduced statistical modelling techniques that evaluated behavioural deviation using probabilistic thresholds [8]. These models improved detection accuracy but still relied heavily on historical data distributions, which reduced adaptability in dynamic environments. Some studies extended this work by incorporating multi-agent consensus mechanisms to validate behavioural integrity [9]. However, consensus-based systems often suffered from latency issues and vulnerability propagation when multiple agents were compromised. Machine learning approaches later gained attention in autonomous security research. Classification-based models attempted to distinguish between normal and malicious behaviour using training datasets [10]. While these systems improved detection rates, they required extensive labelled data and struggled with generalization across unseen environments. Hybrid models combining statistical analysis and learning-based methods were also introduced [11], yet they still faced limitations in real-time adaptability. Recent research shifted toward adversarial learning frameworks that simulate attack conditions during training phases [12]. These approaches improved robustness but increased computational complexity and were not always suitable for resource-constrained agent systems. Some works explored reinforcement learning strategies for adaptive defence responses [13], although instability during training cycles remained a concern. Graph-based security models have also been applied to multi-agent systems, where relationships between agents are analysed to detect structural anomalies [14]. These methods effectively captured interaction-level threats but required complex graph construction and maintenance. Additionally, graph-based systems often struggled with rapid state changes in highly dynamic environments. More recent studies introduced isolation-based detection techniques that separate suspicious agents from operational networks based on behavioural divergence scores [15]-[18]. These approaches showed promise in reducing threat propagation; however, they often lacked adaptive recalibration mechanisms, which limited long-term effectiveness.

2.1 PROPOSED ADAPTIVE ADVERSARIAL ISOLATION FRAMEWORK

The proposed method defines an Adaptive Adversarial Isolation Framework that models autonomous agents as interacting stochastic decision entities operating under uncertain and adversarial conditions. The framework continuously observes agent behaviour, extracts structural and temporal representations of state transitions, and evaluates deviation from expected operational norms. A probabilistic scoring layer quantifies threat likelihood, while an adaptive threshold module dynamically adjusts sensitivity based on environmental volatility. The final stage isolates compromised agents and propagates corrective constraints across the network. The framework operates in a closed feedback loop, ensuring continuous refinement of detection boundaries and system stability across evolving attack patterns.

3. SYSTEM STATE REPRESENTATION AND AGENT MODELLING

The system models an autonomous multi-agent environment as a structured graph where nodes represent agents and edges represent communication or influence pathways. Each agent maintains a state vector that captures internal decision parameters, sensory inputs, and contextual memory. The global system state evolves over discrete time intervals and reflects both local agent actions and inter-agent dependencies. Let the multi-agent system be defined as a directed graph $G=(V,E)$, where $V=\{a_1,a_2,...,a_n\}$ represents autonomous agents and $E \subseteq V \times V$ represents communication links. Each agent a_i holds a state vector $s_i(t) \in \mathbb{R}^d$, which evolves over time step t. The global system state is represented as: $S(t) = \{s_1(t), s_2(t), \ldots, s_n(t)\}$.

The evolution of each agent state follows a stochastic transition model influenced by local input and neighbour interaction. This transition is expressed as:

$$s_i(t+1) = f_i\left(s_i(t), \sum_{j \in \mathrm{N}(i)} w_{ij} s_j(t), x_i(t) \right) \quad (1)$$

Where $\mathrm{N}(i)$ represents the neighbourhood of agent a_i, w_{ij} represents influence weight, and $x_i(t)$ represents external input.

A second formulation defines the joint system probability distribution over states as:

$$P(S(t)) = \prod_{i=1}^{n} P\left(s_i(t) \mid s_{\mathrm{N}(i)}(t), x_i(t) \right) \quad (2)$$

This representation allows structured dependency modelling across agents while preserving local autonomy. It also enables detection of abnormal divergence patterns when observed transitions deviate from learned distributions. The modelling layer provides a foundational representation for all subsequent anomaly detection and isolation operations.

3.1 FEATURE EXTRACTION AND BEHAVIOURAL ENCODING

The framework extracts behavioural signatures from each agent by transforming raw state transitions into structured feature embeddings. These features capture temporal variation, decision consistency, and interaction stability across communication links. The encoding process transforms high-dimensional state dynamics into compact representations suitable for probabilistic evaluation.

Each agent state sequence is mapped into a feature vector $\phi_i(t) \in \mathbb{R}^k$ using a nonlinear transformation function:

$$\phi_i(t) = \Psi\left(s_i(t), s_i(t-1), \Delta s_i(t) \right) \quad (3)$$

where $\Delta s_i(t) = s_i(t) - s_i(t-1)$ represents state deviation across time intervals.

The aggregated behavioural embedding across the system is defined as:

$$\Phi(t) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(t) \tag{4}$$

A second transformation defines interaction-aware encoding that incorporates neighbour influence:

$$\phi_i^*(t) = \phi_i(t) + \sum_{j \in \mathrm{N}(i)} \gamma_{ij}\phi_j(t) \tag{5}$$

where γ_{ij} represents interaction correlation strength.

The probability density of behavioural conformity is modelled as:

$$P(\phi_i(t)) = \mathrm{N}\left(\mu_\phi, \Sigma_\phi\right) \tag{6}$$

A second equation defines divergence magnitude from expected behaviour:

$$D_i(t) = \left(\phi_i(t) - \mu_\phi\right)^T \Sigma_\phi^{-1}\left(\phi_i(t) - \mu_\phi\right) \tag{7}$$

This encoding process ensures that subtle deviations in behaviour become measurable in a structured feature space, enabling early detection of adversarial influence. The representation also supports robustness by incorporating both temporal and relational dependencies.

3.2 PROBABILISTIC THREAT SCORING MECHANISM

The framework evaluates each agent using a probabilistic threat scoring function that quantifies the likelihood of adversarial behaviour. The scoring mechanism integrates behavioural deviation, interaction inconsistency, and temporal instability into a unified metric.

The threat score for agent a_i at time t is defined as:

$$T_i(t) = \sigma\left(\alpha D_i(t) + \beta R_i(t) + \gamma H_i(t)\right) \tag{8}$$

where $\sigma(\cdot)$ represents a sigmoid normalization function, $D_i(t)$ denotes deviation score, $R_i(t)$ represents relational inconsistency, and $H_i(t)$ represents historical instability.

Relational inconsistency is defined as:

$$R_i(t) = \sum_{j \in \mathrm{N}(i)} \left|\phi_i(t) - \phi_j(t)\right| \tag{9}$$

A second formulation defines temporal instability as variance across time window τ:

$$H_i(t) = \frac{1}{\tau}\sum_{k=0}^{\tau}\left(\phi_i(t-k) - \bar{\phi}_i\right)^2 \tag{10}$$

The normalized threat probability is expressed as:

$$P(T_i(t)) = \frac{1}{1 + e^{-T_i(t)}} \tag{11}$$

A second equation defines comparative risk ratio across agents:

$$\mathrm{R}_i(t) = \frac{T_i(t)}{\sum_{j=1}^{n} T_j(t)} \tag{12}$$

This scoring mechanism enables continuous evaluation of each agent's behavioural integrity. The integration of spatial and temporal components ensures that detection is not limited to isolated anomalies but reflects system-wide contextual deviations.

3.3 ADAPTIVE THRESHOLD MECHANISM

The framework uses a dynamic thresholding system that adjusts sensitivity based on environmental volatility and system-wide behavioural variance. This prevents excessive false positives during unstable conditions while maintaining detection sensitivity under stable operation.

The adaptive threshold for agent a_i is defined as:

$$\theta_i(t) = \theta_0 + \lambda\,\mathrm{Var}(\Phi(t)) + \mu\Omega(t) \tag{13}$$

where θ_0 represents base threshold, $\mathrm{Var}(\Phi(t))$ represents global behavioural variance, and $\Omega(t)$ represents environmental uncertainty index.

A second formulation defines threshold adaptation over time:

$$\theta_i(t+1) = \theta_i(t) + \eta\left(\bar{T}(t) - \theta_i(t)\right) \tag{14}$$

where $\bar{T}(t)$ represents mean threat score across all agents.

Decision boundary condition is expressed as:

$$\delta_i(t) = \begin{cases} 1, & T_i(t) > \theta_i(t) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

A second equation defines stability constraint of threshold adaptation:

$$\left|\theta_i(t+1) - \theta_i(t)\right| \le \varepsilon \tag{16}$$

This adaptive mechanism ensures that detection sensitivity aligns with real-time system dynamics. It reduces instability caused by abrupt environmental shifts and maintains consistent operational calibration across distributed agents.

3.4 ISOLATION DECISION LOGIC

The isolation module determines whether an agent should be separated from the operational network based on computed threat probability and consistency checks. The decision logic prioritizes containment of high-risk behavioural propagation while preserving system connectivity.

Isolation decision function is defined as:

$$I_i(t) = \mathrm{I}\left(P(T_i(t)) > \theta_i(t) \wedge C_i(t) < \kappa\right) \tag{17}$$

where $\mathbb{I}(\cdot)$ represents indicator function and $C_i(t)$ represents behavioural consistency score.

Consistency score is defined as:

$$C_i(t) = 1 - \frac{1}{|\mathrm{N}(i)|}\sum_{j \in \mathrm{N}(i)}\left|\phi_i(t) - \phi_j(t)\right| \tag{18}$$

A second formulation defines isolation severity level:

$$S_i(t) = \log\left(1 + T_i(t)\left(1 - C_i(t)\right)\right) \tag{19}$$

Isolation action mapping is expressed as:

$$A_i(t) = \begin{cases} \text{Isolate}, & S_i(t) > \delta \\ \text{Monitor}, & \text{otherwise} \end{cases} \tag{20}$$

A second equation defines containment probability:

$$P_c(t) = 1 - \prod_{i \in V} \left(1 - I_i(t)\right) \tag{21}$$

This logic ensures that isolation decisions remain context-aware and avoid unnecessary disruption of stable agents while effectively containing compromised nodes.

3.5 RESPONSE PROPAGATION ACROSS MULTI-AGENT NETWORK

Once isolation decisions are made, the framework propagates corrective updates across the network to maintain global stability. This step ensures that compromised influence does not persist through indirect communication channels.

Propagation influence decay model is defined as:

$$\Delta s_j(t+1) = \rho \Delta s_i(t) e^{-\alpha d(i,j)} \tag{22}$$

where $d(i, j)$ represents graph distance between agents and ρ represents propagation coefficient.

A second formulation defines network stabilization function:

$$\mathrm{S}(t) = \sum_{i=1}^{n} \sum_{j \in \mathrm{N}(i)} \left| s_i(t) - s_j(t) \right| \tag{23}$$

Corrective update rule is expressed as:

$$s_i(t+1) = s_i(t) - \eta I_i(t) \nabla L(s_i(t)) \tag{24}$$

A second equation defines global convergence condition: $\lim_{t \to \infty} \mathrm{S}(t) = 0$. This propagation mechanism ensures that isolation does not create fragmented system states and instead promotes convergence toward stable behavioural equilibrium across all operational nodes.

3.6 CONTINUOUS LEARNING AND MODEL UPDATE

The final stage of the framework involves continuous learning where detection parameters are updated based on feedback from isolation outcomes and system performance metrics. This ensures long-term adaptability against evolving adversarial strategies.

Parameter update rule is defined as:

$$\Theta(t+1) = \Theta(t) + \eta \nabla_{\Theta} \mathrm{L}(T(t), \delta(t)) \tag{25}$$

where $\Theta(t)$ represents model parameters and $\mathcal{L}$ represents loss between predicted and observed threat states.

A second formulation defines reinforcement-based adjustment:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{26}$$

Adaptive learning stability condition is expressed as:

$$\left\| \Theta(t+1) - \Theta(t) \right\| \le \varepsilon \tag{27}$$

A second equation defines cumulative detection efficiency:

$$E(t) = \frac{1}{t} \sum_{k=1}^{t} \left(\mathrm{TPR}_k - \mathrm{FPR}_k \right) \tag{28}$$

This learning mechanism allows the framework to refine its detection boundaries over time and improve resilience against novel adversarial patterns. The continuous feedback loop ensures that the system evolves in alignment with environmental complexity while maintaining operational reliability across distributed autonomous agents.

4. RESULTS AND DISCUSSION

The experimental evaluation is performed using a Python-based simulation environment implemented in TensorFlow and NetworkX libraries for multi-agent modelling and graph-based interaction analysis. The experiments are executed on a system equipped with Intel Core i7 processor, 16 GB RAM, and NVIDIA GTX 1660 GPU. The simulation environment represents autonomous agents operating in a dynamic adversarial network where attack injection, behavioural drift, and communication interference are systematically introduced. The model is trained and evaluated under controlled stochastic conditions to observe stability and detection consistency.

Table.1. Experimental Configuration Parameters

Parameter	Value
Number of Agents	100
Simulation Iterations	500
Learning Rate	0.01
Threat Injection Ratio	0.25
Communication Density	0.6
Observation Window	20 time steps
Noise Level	0.15
Isolation Threshold Base	0.5

As shown in Table 1, the system configuration defines moderate network density with controlled adversarial injection to evaluate robustness under realistic operational stress.

Performance Metrics

The evaluation uses five metrics. Accuracy measures correct threat classification among all predictions. Precision evaluates proportion of correctly identified threats among predicted threats. Recall measures proportion of detected threats among actual threats. F1-score provides harmonic balance between precision and recall. Detection Latency measures time delay between attack occurrence and identification.

4.1 DATASET DESCRIPTION

Table.2. Dataset Characteristics

Feature	Description
Dataset Type	Synthetic Multi-Agent Simulation Dataset
Data Size	120,000 interaction records
Attributes	State vectors, communication logs, threat labels

Classes	Normal, Adversarial
Generation Model	Stochastic agent interaction simulator
Split Ratio	70% Training, 30% Testing

The dataset represents structured multi-agent behavioural logs generated under controlled adversarial scenarios to simulate realistic attack patterns.

The evaluation considers three baseline methods: Rule-Based Monitoring System, Statistical Threshold Detection Model, and Graph Neural Network Security Model. These methods represent deterministic, probabilistic, and deep learning-based security strategies used for comparison against the proposed adaptive isolation framework.

4.2 RESULTS BASED ON ACCURACY

Table.3. Accuracy Comparison

Iteration	Rule-Based	Statistical Model	GNN Security Model	Proposed AAIF
5	72.1	75.4	78.6	82.3
10	73.5	76.8	80.2	84.7
15	74.2	77.5	81.0	86.1
20	75.0	78.3	82.4	87.6
25	75.8	79.1	83.2	88.9

The results in Table.3 indicate that the proposed AAIF consistently achieves higher accuracy compared to baseline methods across all iterations. The rule-based system maintains lower performance due to its dependency on static decision boundaries. The statistical model shows moderate improvement as it adapts to distributional shifts, yet it fails to capture complex behavioural dependencies. The GNN-based model performs better due to structural learning capability, but it still lacks dynamic isolation logic.

The proposed AAIF demonstrates progressive improvement from 82.3 to 88.9 across iterations. This improvement occurs because the model continuously updates behavioural embeddings and refines adaptive thresholds. The integration of probabilistic scoring with isolation logic improves classification stability under adversarial pressure. The accuracy gap between AAIF and GNN increases from approximately 3.7% at iteration 5 to 5.7% at iteration 25, indicating stronger scalability. The consistency in performance also suggests reduced sensitivity to noise and attack variability. Overall, AAIF maintains superior classification reliability due to continuous feedback-driven adaptation and behavioural divergence modelling.

4.3 RESULTS BASED ON PRECISION

Table.4. Precision Comparison

Iteration	Rule-Based	Statistical Model	GNN Security Model	Proposed AAIF
5	70.4	73.9	77.5	83.1
10	71.6	74.8	78.9	85.0
15	72.3	75.6	79.6	86.4
20	73.0	76.2	80.5	87.2
25	73.8	77.1	81.3	88.5

The Table.4 shows that precision improves steadily for all models, but the proposed AAIF achieves the highest values across all iterations. The rule-based system suffers from false positives due to rigid decision boundaries. The statistical model reduces false positives slightly but remains limited by distributional overlap between normal and adversarial behaviours. The GNN model improves precision by capturing relational structure, yet misclassification persists in dynamic attack scenarios.

AAIF achieves precision growth from 83.1 to 88.5, which reflects stronger reduction in false positive identification. This improvement arises from adaptive threshold tuning that adjusts sensitivity according to environmental variance. The relational inconsistency measure also contributes to improved discrimination between normal and malicious agents. The precision improvement margin over GNN ranges from 5.6% to 7.2%, indicating consistent superiority. The isolation mechanism reduces cascading misclassification by preventing compromised nodes from influencing neighbouring evaluations. As a result, AAIF maintains higher predictive confidence and reduces unnecessary alarm generation across iterations.

4.4 RESULTS BASED ON RECALL

Table.5. Recall Comparison

Iteration	Rule-Based	Statistical Model	GNN Security Model	Proposed AAIF
5	68.9	72.5	76.8	84.0
10	69.7	73.4	78.1	85.6
15	70.5	74.2	79.0	86.9
20	71.2	75.0	80.2	88.1
25	72.0	75.8	81.0	89.3

The Table.5 indicates that recall performance improves across all models, but AAIF achieves significantly higher detection coverage of adversarial instances. The rule-based system fails to detect subtle or evolving attacks due to predefined detection logic. The statistical model improves recall by capturing distributional anomalies, but it still misses complex coordinated attacks. The GNN model enhances recall by leveraging structural dependencies, yet it struggles under high noise and dynamic behaviour shifts.

AAIF achieves recall values from 84.0 to 89.3, showing strong ability to detect adversarial behaviour consistently. The improvement is attributed to continuous behavioural deviation tracking and adaptive scoring mechanisms. The framework reduces missed detections by integrating temporal instability measures with relational divergence scoring. The recall advantage over GNN ranges from 7.2% to 8.3%, which indicates strong detection coverage. The system also benefits from feedback-based threshold adjustment, which reduces under-detection during high volatility phases. This leads to improved sensitivity without significantly increasing false alarms.

4.5 RESULTS BASED ON F1-SCORE

Table.6. F1-Score Comparison

Iteration	Rule-Based	Statistical Model	GNN Security Model	Proposed AAIF
5	70.1	74.1	77.9	83.6
10	71.0	75.0	79.0	85.2
15	71.8	75.8	80.0	86.7
20	72.5	76.6	81.2	87.9
25	73.2	77.4	82.1	88.9

The Table.6 presents F1-score comparison where AAIF consistently outperforms baseline models. The rule-based system shows lowest balance between precision and recall due to rigid classification rules. The statistical model achieves moderate balance but lacks adaptability under dynamic adversarial conditions. The GNN model improves harmonic balance due to structural learning but still exhibits instability under high variability.

AAIF achieves F1-score improvement from 83.6 to 88.9, indicating balanced enhancement in both precision and recall. The improvement reflects effective trade-off management between detection sensitivity and false alarm reduction. The adaptive isolation mechanism contributes to stabilizing classification performance by preventing propagation of corrupted state information. The margin over GNN ranges between 5.7% and 6.8%, demonstrating consistent superiority. The results confirm that adaptive behavioural modelling enhances classification equilibrium under uncertain conditions.

4.6 RESULTS BASED ON DETECTION LATENCY

Table.7. Detection Latency Comparison (Time Units)

Iteration	Rule-Based	Statistical Model	GNN Security Model	Proposed AAIF
5	12.5	10.8	9.6	7.4
10	12.1	10.5	9.2	7.0
15	11.8	10.2	8.9	6.8
20	11.5	9.9	8.5	6.5
25	11.2	9.6	8.2	6.3

The Table.7 shows that AAIF achieves the lowest detection latency among all compared methods. Rule-based systems exhibit highest latency due to sequential rule evaluation. The statistical model reduces latency slightly by leveraging probabilistic shortcuts, but computation still depends on full distribution analysis. The GNN model introduces moderate delay due to graph propagation and message passing operations.

AAIF reduces latency significantly from 7.4 to 6.3 time units due to early-stage anomaly filtering and isolation-driven pruning of computation paths. The framework avoids unnecessary propagation of compromised state evaluations, which reduces processing overhead. The latency reduction compared to GNN ranges between 22% and 27%, indicating strong efficiency improvement. The adaptive scoring mechanism also prioritizes high-risk nodes, which accelerates decision convergence. This demonstrates that security enhancement does not compromise computational efficiency in AAIF.

4.7 DISCUSSION

The overall experimental evaluation demonstrates that the proposed AAIF consistently outperforms baseline methods across all performance metrics. Accuracy improves by approximately 5% over GNN-based models, while precision shows reduction in false positives by nearly 6% to 7%. Recall improvement reaches up to 8%, indicating stronger detection coverage of adversarial events. F1-score analysis confirms balanced performance gain across both detection sensitivity and reliability. Detection latency analysis shows that AAIF reduces processing delay by approximately 25% compared to graph-based models, which indicates improved computational efficiency. The rule-based and statistical models consistently underperform due to lack of adaptive mechanisms, while GNN models show moderate improvements but remain limited under dynamic adversarial conditions. The consistent performance gain across iterations from 5 to 25 demonstrates scalability of AAIF under increasing system complexity. The adaptive thresholding mechanism plays a critical role in stabilizing detection boundaries, while isolation logic prevents propagation of corrupted states. The integration of behavioural deviation scoring ensures robust detection under uncertain environments. Overall, the system demonstrates both high detection reliability and computational efficiency, which makes it suitable for real-time autonomous multi-agent security applications.

5. CONCLUSION

The study presents an Adaptive Adversarial Isolation Framework designed for secure operation of autonomous multi-agent systems under dynamic adversarial conditions. The framework integrates behavioural modelling, probabilistic threat scoring, adaptive thresholding, and isolation-based response mechanisms to enhance system resilience. Experimental evaluation demonstrates that the proposed method consistently outperforms rule-based, statistical, and graph neural network models across accuracy, precision, recall, F1-score, and latency metrics. The system achieves up to 88.9% accuracy, 88.5% precision, 89.3% recall, and 88.9% F1-score, while maintaining reduced detection latency close to 6.3 time units. These results indicate strong improvement in both detection quality and computational efficiency. The adaptive threshold mechanism ensures stability under fluctuating environmental conditions, while the isolation strategy effectively prevents propagation of adversarial influence across agent networks. The continuous learning component further strengthens long-term adaptability against evolving threats.

REFERENCES

[1] A. Chhabra, S. Datta, S.K. Nahin and P. Mohapatra, “Agentic AI Security: Threats, Defenses, Evaluation and Open Challenges”, *Proceedings of International Conference on Artificial Intelligence*, Vol. 3, pp. 1-29, 2025.

[2] I. Adabara, B.O. Sadiq, A.N. Shuaibu, Y.I. Danjuma and V. Maninti, “Trustworthy Agentic AI Systems: A Cross-Layer

Review of Architectures, Threat Models and Governance Strategies for Real-World Deployment", *F1000Research*, Vol. 14, No. 9, pp. 1-11, 2025.

[3] A. Lopez Pellicer, P. Angelov and N. Suri, "Securing (Vision-based) Autonomous Systems: Taxonomy, Challenges and Defense Mechanisms against Adversarial Threats", *Artificial Intelligence Review*, Vol. 58, No. 12, pp. 1-9, 2025.

[4] S. Narayanan, "Autonomous Cyber Sovereignty: A Dual-Control Architecture for Agentic Artificial Intelligence in Offensive Defensive Security Ecosystems", *World Journal of Advanced Research and Reviews*, Vol. 25, No. 3, pp. 2538-2546, 2025.

[5] P.K. Chakrabarty, "Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts and Defense Strategies", *International Journal of Science and Research*, Vol. 14, No. 4, pp. 1367-1369, 2025.

[6] A. Chhabra, S. Datta, S.K. Nahin and P. Mohapatra, "Agentic AI Security: Threats, Defenses, Evaluation and Open Challenges", *IEEE Access*, Vol. 3, pp. 1-10, 2026.

[7] A.S. Rangappa, "Agentic AI and Cyber Security: Autonomous Threat Hunting, Intrusion Detection and Adaptive Defense Mechanisms in a World of Increasingly Sophisticated Cyber Attacks", *Journal of Digital Security and Forensics*, Vol. 2, No. 1, pp. 128-138, 2025.

[8] B. Vijetha, "Agentic Intelligence for Unified Cyber Defense: A Self-Adaptive Framework for Threat Detection across Cloud, Edge and IoT Systems", *IEEE Access*, Vol. 14, pp. 5104-5118, 2026.

[9] N. Addla, "Autonomous AI Agents for Cybersecurity Threat Detection and Response: A Multi-Agent Architecture Framework using AWS Frontier Agents", *American International Journal of Computer Science and Technology*, Vol. 11, pp. 76-89, 2026.

[10] B. Tekeste, K. Al-Hussaeni, B.C. Fung, I. Alawadhi and C. Fachkha, "Adversarial Machine Learning: A 20-Year Survey of Attacks, Defenses and Standards", *IEEE Access*, Vol. 7, pp. 1-11, 2026.

[11] N. Sharma, A. Jaggi, P. Takkalapally and K. Hudani, "Analyzing Adaptive Intrusion Detection Systems for Improved Network Security", *Proceedings of International Conference on Advances in Computation, Communication and Information Technology*, Vol. 1, pp. 425-430, 2024.

[12] N.K. Chakravarthy, K.N.G. Veerappan and J.A.I. Syed Masood, "Privacy-Preserving Framework using Automated Security Orchestration and Response (Asor) in E-Health Systems", *Congress on Intelligent Systems*, Vol. 87, pp. 57-71, 2021.

[13] B. Gobinathan, M.A. Mukunthan, S. Surendran, K. Somasundaram, S.A. Moeed, P. Niranjan and V.P. Sundramurthy, "A Novel Method to Solve Real Time Security Issues in Software Industry using Advanced Cryptographic Techniques", *Scientific Programming*, Vol. 2021, No. 1, pp. 1-9, 2021.

[14] A. Sheth, A. Patel, C. Upadhyay, H. Ragothaman, B. Patil and S.K. Udayakumar, "Agentic AI for Autonomous Cyber Threat Hunting and Adaptive Defense in Dynamic Security Environments", *Proceedings of International Conference on Electro Information Technology*, Vol. 21, pp. 316-321, 2021.

[15] S. Konyeha, C.C. Konyeha, E. Mintah, O. Ukpebor, O. Sokoya and T. Jessa, "AI-Driven Threat Detection and Response: Toward Autonomous Cyber Defense Systems", *Scientific Reports*, Vol. 32, pp. 1-27, 2025.

[16] T.L. Yasarathna and N.A. Le-Khac, "SoK: Systematic Analysis of Adversarial Threats against Deep Learning Approaches for Autonomous Anomaly Detection Systems in SDN-IoT Networks", *Journal of Information Security and Applications*, Vol. 94, pp. 1-11, 2025.

[17] A.D.M. Ibrahum, M. Hussain and J.E. Hong, "Deep Learning Adversarial Attacks and Defenses in Autonomous Vehicles: A Systematic Literature Review from a Safety Perspective", *Artificial Intelligence Review*, Vol. 58, No. 1, pp. 1-11, 2025.

[18] R.R. Kethireddy, "AI-Augmented Threat Response Systems with Real-Time Adaptive Defense", *International Journal of Artificial Intelligence Research and Development*, Vol. 1, No. 1, pp. 62-71, 2021.