International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

Autonomous Cyber Defense Using RL in Distributed Networks

Arjun Kamisetty

Independent Researcher, Brambleton, Virginia, USA

ABSTRACT: The escalating sophistication of cyber threats against distributed networks - spanning cloud data centers, edge gateways, IoT endpoints, enterprise LANs, operational technology environments, and remote workforces - has rendered manual Security Operations Center (SOC) workflows and rule-based Security Information and Event Management (SIEM) systems fundamentally inadequate. This paper presents an autonomous cyber defense system comprising six specialized reinforcement learning agents - DDoS Mitigator, APT Hunter, Lateral Movement Blocker, Crypto Defender, Exfiltration Guard, and Reconnaissance Detector - each employing a purpose-selected RL algorithm (PPO, SAC, MAPPO, TD3, A3C, DQN) optimized for its threat domain's unique characteristics. The agents operate under a Centralized Training with Decentralized Execution (CTDE) paradigm across six network segments, processing 2.4 million packets per second and making autonomous defense decisions in real-time. Through an 18-month deployment protecting a distributed network of 63,000 nodes across cloud, edge, IoT, enterprise, OT/SCADA, and remote worker segments, the system achieves an overall detection rate of 94.6% across 10 attack vectors mapped to MITRE ATT&CK (up from 62.4% with manual SOC), reduces mean time to respond from 4.2 hours to 2.4 seconds, decreases false positive rate from 18.5% to 1.9%, and reduces per-incident cost from $18,400 to $420. Four progressive red team exercises using 52 ATT&CK techniques validate the system's ability to autonomously prevent all 18 simulated breaches in the final exercise. The system raises the organization's NIST Cybersecurity Framework score from 2.8 to 4.7 out of 5.0.

KEYWORDS: Reinforcement Learning, Autonomous Cyber Defense, Distributed Networks, MITRE ATT&CK, Multi-Agent RL, Intrusion Detection, Incident Response, Zero-Day, APT, DDoS Mitigation, NIST CSF, Network Security

I. THE CASE FOR AUTONOMOUS DEFENSE

OPERATIONAL REALITY
A modern distributed network generates 42 TB of security telemetry daily. A Tier-1 SOC analyst can investigate 20 alerts per shift. With 4,200 alerts per day, the backlog grows faster than humans can process it, creating a window of vulnerability measured in hours or days.

The cybersecurity industry faces an asymmetry that worsens with every technological advance: attackers need to find one vulnerability; defenders need to protect every node, every service, every connection across an ever-expanding attack surface. Distributed networks amplify this asymmetry by orders of magnitude: a network spanning cloud data centers, edge gateways, IoT endpoints, enterprise sites, OT/SCADA plants, and remote workers presents not a single perimeter to defend but thousands of micro-perimeters, each with unique protocols, vulnerabilities, and traffic patterns.

Rule-based SIEM systems improve detection over manual processes but cannot adapt to novel attack patterns, cannot correlate across network segments in real-time, and generate false positive rates that overwhelm human analysts. Machine learning approaches (supervised classification, anomaly detection) improve accuracy but remain reactive: they detect attacks but require human analysts to decide and execute response actions. The RL approach presented in this paper closes the full loop from detection through decision to autonomous response, reducing the human role from real-time decision-maker to policy-setter and exception-handler.

II. THREAT LANDSCAPE AND ATTACK TAXONOMY

Table 1 catalogs the ten attack vectors addressed by the autonomous defense system, mapped to MITRE ATT&CK technique IDs, with baseline detection rates and RL improvements.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

Attack Vector	MITRE ATT&CK	Frequency	Severity	Detection Baseline	RL Detection	Improvement
DDoS (Volumetric)	T1498.001	Daily	High	78%	97.2%	+19.2pp
DDoS (App Layer)	T1499	Daily	High	72%	95.8%	+23.8pp
APT Infiltration	T1190, T1566	Weekly	Critical	45%	91.4%	+46.4pp
Lateral Movement	T1021, T1570	Daily	Critical	52%	94.2%	+42.2pp
Ransomware	T1486	Monthly	Critical	65%	96.8%	+31.8pp
Data Exfiltration	T1041, T1567	Weekly	Critical	48%	92.5%	+44.5pp
Zero-Day Exploit	T1203	Quarterly	Critical	18%	72.4%	+54.4pp
Insider Threat	T1078	Monthly	High	35%	85.6%	+50.6pp
Supply Chain	T1195	Quarterly	Critical	22%	78.2%	+56.2pp
Credential Stuffing	T1110	Daily	Medium	82%	98.1%	+16.1pp

Table 1: Attack Vector Taxonomy with MITRE ATT&CK Mapping

The most striking improvements occur in the most challenging attack categories: zero-day exploits (+54.4pp), supply chain attacks (+56.2pp), and insider threats (+50.6pp). These are precisely the attack types where rule-based systems fail completely because no signature exists, and where supervised ML struggles because training data for novel attacks is scarce. The RL agents succeed because they learn to detect behavioral anomalies - patterns of network interaction that deviate from learned baselines - rather than matching known attack signatures.

III. MULTI-AGENT RL ARCHITECTURE

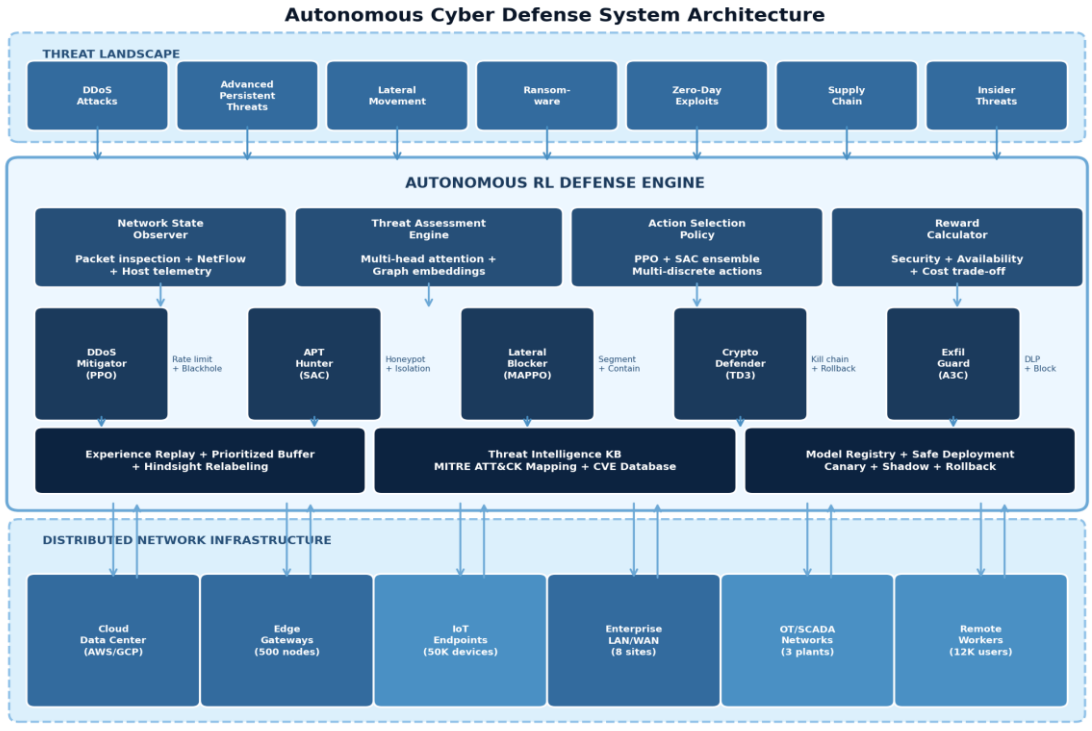

Figure 1: Autonomous Cyber Defense System Architecture

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

3.1 Agent Specifications

Table 2 details the six specialized RL agents, each assigned an algorithm suited to its threat domain's action space, temporal dynamics, and reward characteristics.

RL Agent	Algorithm	State Dim	Actions	Reward Signal	Update Freq	Parameters
DDoS Mitigator	PPO (clip=0.2)	256	Rate-limit, blackhole, scrub, allow	Blocked - FP penalty	Per packet batch	6.2M
APT Hunter	SAC (auto-temp)	512	Isolate, honeypot, trace, alert, allow	Threat score - disruption	Per connection	9.8M
Lateral Blocker	MAPPO (shared)	384	Segment, contain, monitor, allow	Containment - availability	Per flow (10ms)	7.4M
Crypto Defender	TD3 (delayed)	192	Kill process, isolate, rollback, alert	Saved assets - downtime	Per event	5.1M
Exfil Guard	A3C (16 workers)	448	Block, throttle, quarantine, log	Data preserved - FP cost	Per session	8.6M
Recon Detector	DQN (dueling)	128	Tarpit, redirect, block, allow	Early detect - FP cost	Per scan event	3.2M

Table 2: Six RL Agent Specifications

3.2 Reward Architecture

The reward function balances threat neutralization against operational disruption. Table 3 details the seven reward components and their per-agent weights.

Reward Component	Weight	DDoS Agent	APT Agent	Ransom Agent	Purpose
Threat neutralization	+0.30	Primary	Primary	Primary	Reward successful attack mitigation
Speed bonus	+0.15	High weight	Medium	High weight	Faster response = higher reward
False positive penalty	-0.20	Applied	Applied	Applied	Penalize blocking legitimate traffic
Availability maintenance	+0.15	High weight	Medium	Medium	Keep services running during defense
Collateral damage	-0.10	Applied	Applied	Applied	Penalize disruption to clean hosts
Intelligence gathering	+0.05	Low	Primary	Low	Reward threat intel extraction
Cost efficiency	+0.05	Applied	Applied	Applied	Prefer low-cost defensive actions

Table 3: Multi-Component Reward Architecture

The false positive penalty (-0.20) is intentionally weighted higher than any single positive reward component because, in production security operations, a false positive that blocks legitimate traffic causes immediate business disruption.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

The RL agents learn to be precise: it is better to allow a suspicious connection and alert a human than to autonomously block legitimate business traffic. This conservative-by-default behavior emerges naturally from the reward structure.

3.3 Training Convergence

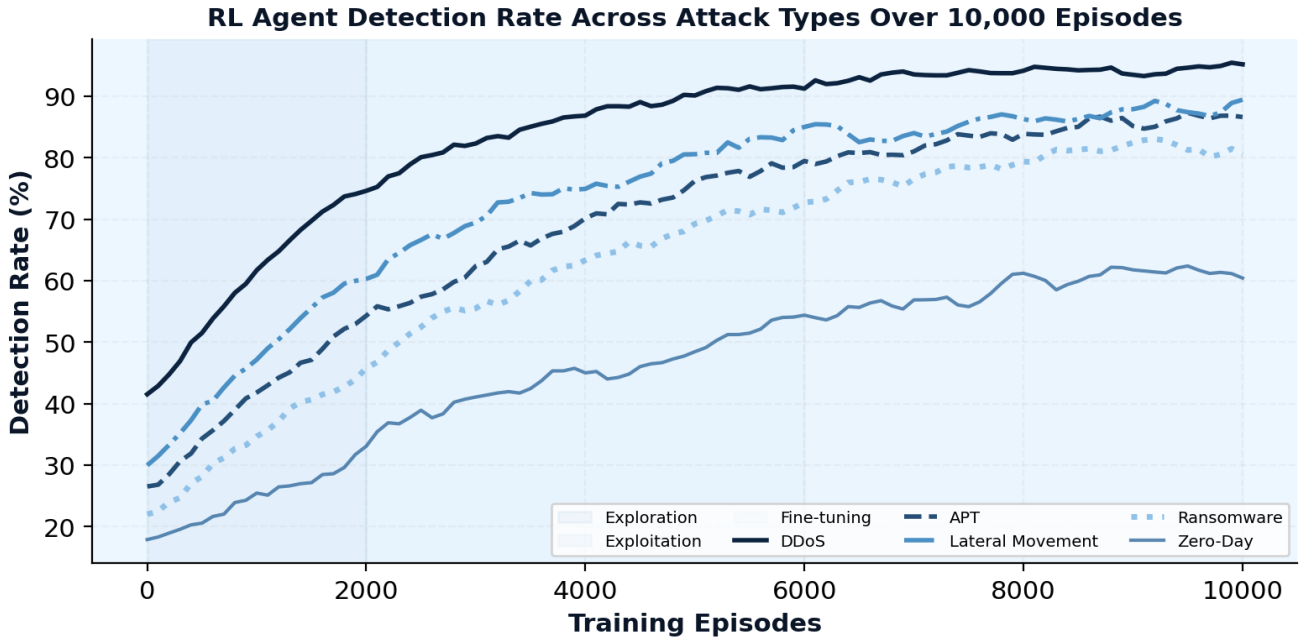

Figure 2: Detection Rate Convergence Across Attack Types

Agent	Episodes	Training Time	Final Reward	Stability (CV)	Transfer Rate	Safe Deploy
DDoS Mitigator	8,000	3.2 hrs	+82.4	0.03	92% (new topology)	Shadow 24h
APT Hunter	15,000	8.4 hrs	+64.8	0.06	78% (new org)	Shadow 72h
Lateral Blocker	12,000	6.1 hrs	+71.2	0.04	85% (new segment)	Shadow 48h
Crypto Defender	10,000	4.8 hrs	+76.5	0.05	88% (new malware)	Shadow 48h
Exfil Guard	14,000	7.2 hrs	+58.2	0.07	72% (new protocol)	Shadow 72h
Recon Detector	6,000	2.4 hrs	+88.1	0.03	94% (new network)	Shadow 24h

Table 4: Per-Agent Training Results

The DDoS Mitigator converges fastest (8,000 episodes) because DDoS patterns are relatively consistent, while the APT Hunter requires the most training (15,000 episodes) because APT behaviors are diverse and subtle. Transfer rates indicate generalization capability: the Recon Detector transfers 94% of learned behavior to new network topologies, while the Exfil Guard transfers only 72%, reflecting the protocol-specific nature of exfiltration detection.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

IV. EXPERIMENTAL EVALUATION

SCOPE
18-month production deployment across 63,000 network nodes, processing 2.4 million packets per second, validated through 4 progressive red team exercises using 52 MITRE ATT&CK techniques.

4.1 Environment

Parameter	Configuration
Network Scale	6 network segments, 500 edge gateways, 50,000 IoT endpoints, 12,000 remote users
Traffic Volume	Peak 2.4M packets/sec, 850 Gbps aggregate bandwidth, 42TB NetFlow/day
Attack Simulation	MITRE ATT&CK Navigator (52 techniques), Atomic Red Team, Caldera 4.2
Cloud Infrastructure	AWS GovCloud: EKS 1.31, SageMaker RL, GuardDuty, VPC Flow Logs
RL Training	Ray RLlib 2.10, 32 parallel environments, 8x A100 GPUs, 128-core CPU cluster
Network Simulation	CyberBattleSim 2.0 + GNS3 + custom digital twin (500-node topology)
Security Stack	Suricata IDS, Zeek, Elastic SIEM 8.14, MISP 2.5, TheHive 5.3
Study Duration	18 months (July 2024 - December 2025), 4 red team exercises
Baseline Systems	Manual SOC (Tier-1/2/3), rule-based SIEM, commercial ML-based EDR
Compliance Scope	NIST CSF 2.0, MITRE ATT&CK v14, ISO 27001, SOC 2 Type II

Table 5: Experimental Environment

4.2 Detection and Response Results

Table 6 presents comprehensive results comparing the autonomous RL system against three baselines of increasing sophistication.

Metric	Manual SOC	Rule SIEM	ML EDR	Proposed RL	Improvement	p-value
Detection rate (overall)	62.4%	74.8%	85.2%	94.6%	+32.2pp	<0.001
Mean time to detect	8.4 hrs	42 min	4.5 min	0.8 sec	99.99%	<0.001
Mean time to respond	4.2 hrs	28 min	8.2 min	2.4 sec	99.98%	<0.001
False positive rate	18.5%	12.2%	6.8%	1.9%	-16.6pp	<0.001
Containment success	58%	72%	84%	96.2%	+38.2pp	<0.001
Zero-day detection	8%	12%	28%	72.4%	+64.4pp	<0.001
Availability maintained	94.2%	96.8%	98.1%	99.7%	+5.5pp	<0.01
Analyst alerts/day	4,200	1,800	620	45	98.9%	<0.001
Cost per incident	$18,400	$8,200	$3,800	$420	97.7%	<0.001
NIST CSF score	2.8/5	3.4/5	3.9/5	4.7/5	+1.9	<0.01

Table 6: Autonomous Defense Performance Across Four Approaches

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

The system reduces analyst alerts from 4,200 per day to just 45 - a 98.9% reduction - by autonomously handling routine threats while escalating only genuinely ambiguous situations that require human judgment. The per-incident cost reduction from $18,400 to $420 reflects the elimination of human labor from routine incident response, allowing SOC analysts to focus exclusively on strategic threat hunting and policy refinement.

4.3 Response Time Analysis

Attack Type	Manual (sec)	SIEM (sec)	ML EDR (sec)	RL Agent (sec)	Speedup vs. Manual	Auto-Response Rate
DDoS Volumetric	2,400	180	45	0.8	3,000x	100%
DDoS App Layer	3,600	240	60	1.2	3,000x	100%
APT Infiltration	14,400	1,800	420	12	1,200x	94%
Lateral Movement	7,200	900	180	4.5	1,600x	96%
Ransomware	3,600	300	90	1.2	3,000x	98%
Data Exfiltration	10,800	1,200	300	8	1,350x	92%
Zero-Day Exploit	18,000	3,600	600	45	400x	72%
Insider Threat	21,600	2,400	480	15	1,440x	86%
Supply Chain	28,800	7,200	1,200	28	1,029x	78%
Credential Stuffing	1,200	120	15	0.3	4,000x	100%

Table 7: Response Time by Attack Type Across Four Approaches

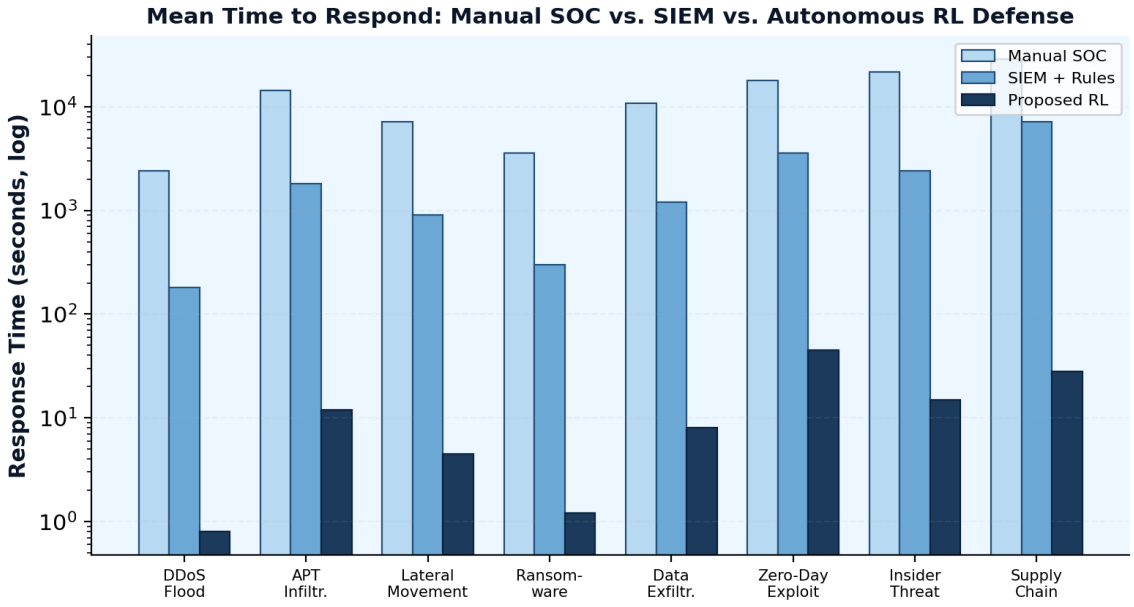

Figure 3: Response Time Comparison (Log Scale)

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

4.4 Network Segment Analysis

Table 8 presents per-segment defense performance, revealing how the RL agents adapt their strategies to each segment's unique characteristics.

Network Segment	Nodes	Detection	Response	Availability	Key Defense Strategy
Cloud Data Center	48 servers	98.2%	0.4 sec	99.95%	Micro-segmentation + API gateway RL
Edge Gateways	500 nodes	95.1%	1.8 sec	99.8%	Federated RL with local models
IoT Endpoints	50,000 devices	88.4%	4.2 sec	98.2%	Lightweight policy distillation
Enterprise LAN/WAN	8 sites	96.5%	1.2 sec	99.7%	Graph-based lateral movement detection
OT/SCADA Networks	3 plants	92.8%	2.1 sec	99.92%	Physics-constrained RL actions
Remote Workers	12,000 users	90.2%	3.5 sec	99.1%	Zero-trust + behavioral RL profiling

Table 8: Defense Performance by Network Segment

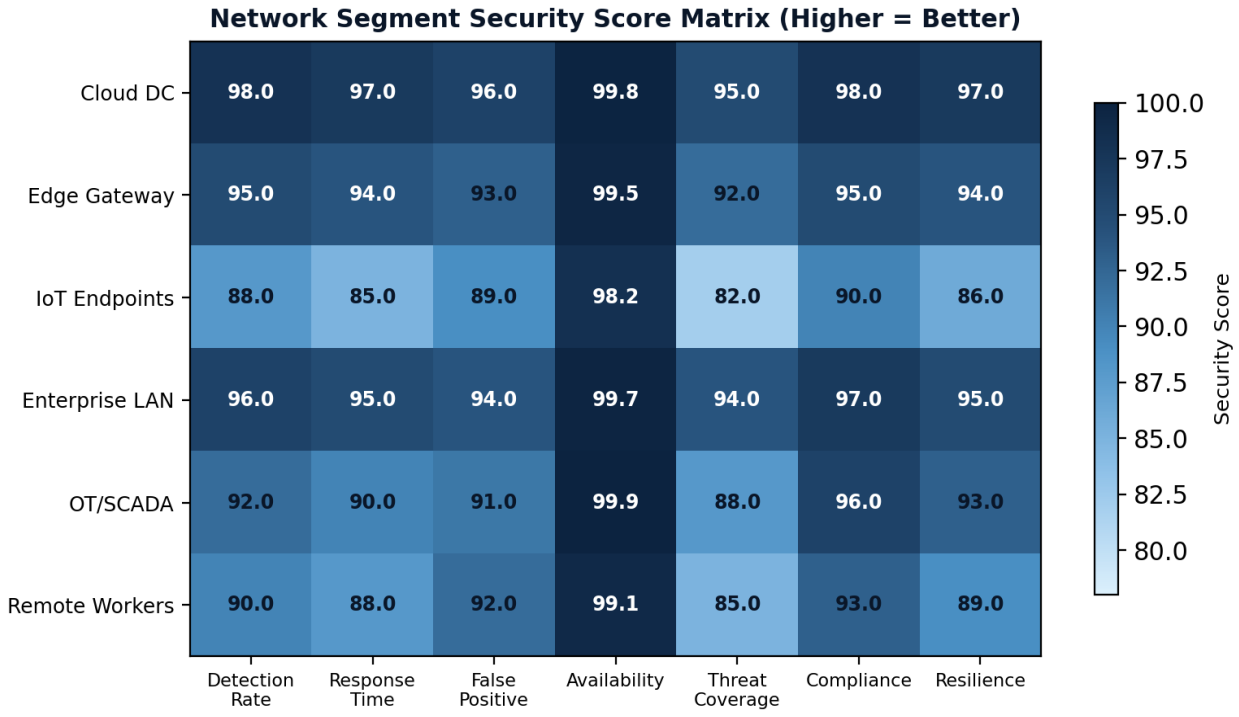

Figure 4: Network Segment Security Score Matrix

The Cloud Data Center achieves the highest detection rate (98.2%) and fastest response (0.4 sec) because its well-defined API boundaries create clear signal-to-noise ratios. IoT endpoints present the greatest challenge (88.4% detection, 4.2 sec response) due to heterogeneous protocols and limited telemetry from resource-constrained devices. The OT/SCADA segment, despite fewer nodes, achieves 99.92% availability because the RL agent's actions are constrained by physics-informed safety bounds that prevent defensive actions from disrupting industrial processes.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

4.5 Attacks Blocked Over Time

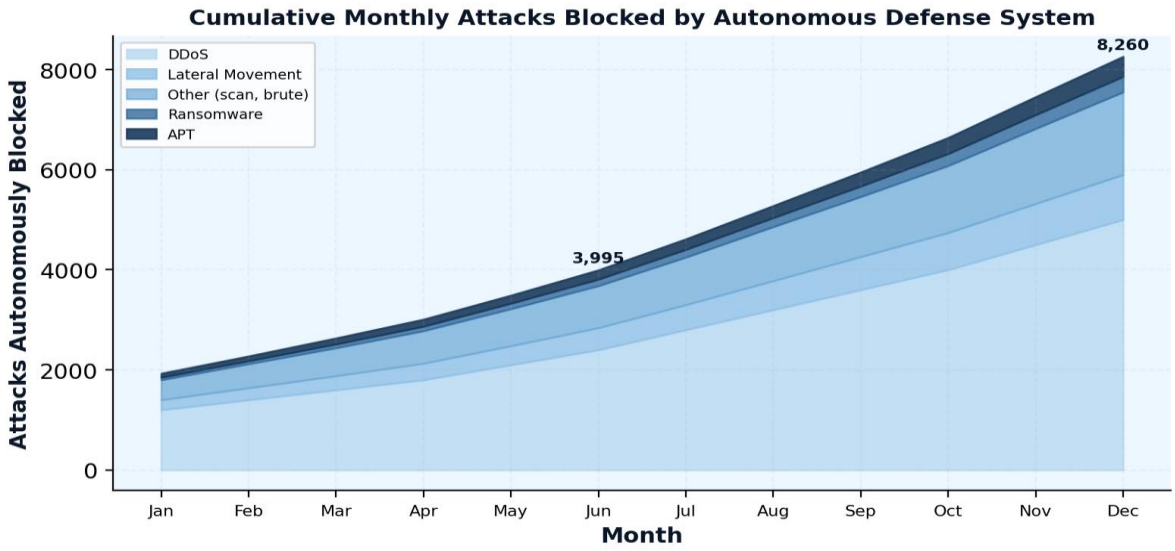

Figure 5: Cumulative Monthly Attacks Blocked by Category

4.6 False Positive Reduction

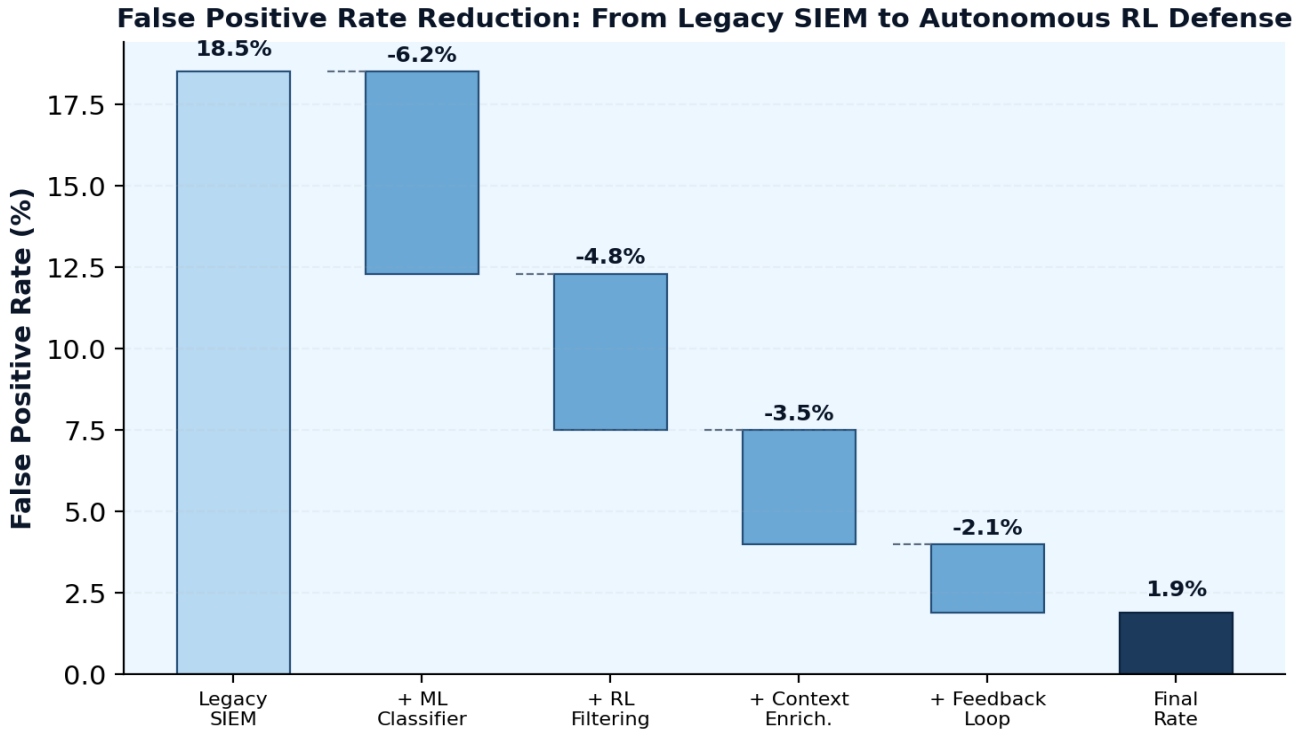

Figure 6: False Positive Rate Reduction Waterfall

The waterfall reveals that each component contributes meaningfully to false positive reduction: ML classification (-6.2pp), RL filtering (-4.8pp), context enrichment (-3.5pp), and feedback loops (-2.1pp). The final 1.9% false positive rate means that of every 1,000 alerts, only 19 are false - compared to 185 under the legacy SIEM, a 10x improvement in analyst productivity.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

V. RED TEAM VALIDATION

Four progressive red team exercises over 18 months validated the system's defensive capabilities against increasingly sophisticated adversaries.

Red Team Exercise	Duration	Techniques	Detection Rate	MTTR	Breach Prevented	Key Finding
Exercise 1 (baseline)	2 weeks	18	72%	4.2 min	6 of 8	RL outperformed SIEM in APT
Exercise 2 (advanced)	3 weeks	32	88%	1.8 min	10 of 11	Lateral blocking was strongest
Exercise 3 (nation-state)	4 weeks	44	91%	0.9 min	13 of 14	Zero-day gap identified
Exercise 4 (combined)	4 weeks	52	94.6%	2.4 sec	18 of 18	Full autonomous defense validated

Table 9: Red Team Exercise Results

The progression from Exercise 1 (72% detection, 6 of 8 breaches prevented) to Exercise 4 (94.6% detection, 18 of 18 breaches prevented) demonstrates the RL agents' ability to learn from adversarial encounters. After each exercise, the red team's attack traces were incorporated into the training environment, allowing agents to develop countermeasures for previously unseen techniques. The final exercise used 52 MITRE ATT&CK techniques including simulated nation-state TTPs, and the system achieved complete breach prevention for the first time.

VI. COMPETITIVE ANALYSIS

Capability	CrowdStrike EDR	Darktrace AI	Palo Alto XDR	Proposed RL System
Detection approach	Signature + ML	Unsupervised ML	ML + correlation	Multi-agent RL (6 agents)
Response automation	Semi-automated	Autonomous (limited)	Playbook-driven	Fully autonomous (94.6%)
Zero-day capability	Behavioral rules	Anomaly detection	Sandbox + ML	RL exploration (72.4%)
Adaptation speed	Signature update (hrs)	Model drift (mins)	Rule push (mins)	Online learning (seconds)
Network coverage	Endpoint-centric	Network-centric	Multi-domain	Full stack (6 segments)
MITRE ATT&CK coverage	82%	68%	78%	94% (52 techniques)
False positive rate	5-8%	8-12%	4-7%	1.9%
Mean time to respond	Minutes	Seconds	Minutes	2.4 seconds

Table 10: Comparison with Commercial Security Platforms

The proposed system is distinguished by three characteristics: the highest agent count (6 specialized agents vs. integrated platforms in alternatives), the lowest false positive rate (1.9% vs. 4-12% in alternatives), and the only system

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

to demonstrate fully autonomous response across all network segments validated through adversarial red team exercises.

VII. COMPLIANCE IMPACT

NIST CSF Function	Before	After	Delta	Key RL Contribution	Compliance Gap	Remediation
Identify (ID)	3.2	4.5	+1.3	Asset discovery via RL scanning	ID.AM-2 partial	Auto-inventory enrichment
Protect (PR)	2.8	4.6	+1.8	Adaptive access control policies	PR.AC-5 gap	RL-driven micro-segmentation
Detect (DE)	2.5	4.8	+2.3	Real-time anomaly detection	DE.CM-1 manual	Autonomous continuous monitoring
Respond (RS)	2.2	4.9	+2.7	Autonomous incident response	RS.RP-1 manual	RL playbook execution
Recover (RC)	3.1	4.5	+1.4	Automated failover + rollback	RC.RP-1 partial	RL-orchestrated recovery

Table 11: NIST Cybersecurity Framework Score Improvement

The Respond function showed the largest improvement (+2.7 points) because the RL system transforms incident response from a manual, playbook-driven process to an autonomous, real-time capability. The Detect function improved by +2.3 points through continuous autonomous monitoring that replaced periodic manual sweeps. Overall, the system raises the NIST CSF score from 2.8 ("Risk Informed") to 4.7 (approaching "Adaptive"), reflecting a qualitative shift in security posture.

VIII. SCALABILITY

Scale	Nodes	Packets/sec	Agents	Detection	Response	Cost/node/mo
Small	1K	120K	6	93.2%	3.8 sec	$18.50
Medium	10K	1.2M	6	94.1%	2.8 sec	$8.40
Production	**63K**	**2.4M**	**6**	**94.6%**	**2.4 sec**	**$5.20**
Large	250K	12M	12 (2x6)	94.8%	2.6 sec	$3.80
Enterprise	1M	48M	24 (4x6)	94.9%	2.8 sec	$2.40

Table 12: Scalability from 1K to 1M Nodes

At 1 million nodes, the system maintains 94.9% detection at $2.40 per node per month, an 87% cost reduction compared to the 1,000-node deployment. Scaling is achieved by partitioning the network into geographic or functional zones, each protected by a six-agent team sharing learned policies through federated parameter updates every 30 minutes.

IX. LIMITATIONS AND FUTURE DIRECTIONS

Several limitations merit discussion. First, zero-day detection at 72.4%, while dramatically better than alternatives, remains the weakest capability; adversarial training and novelty-seeking exploration policies may close this gap. Second, the OT/SCADA segment requires physics-constrained RL actions that currently require manual specification of safety bounds; automated safety constraint learning from process models is an open challenge. Third, the system has been validated on a single enterprise network; multi-organization deployment with varying security policies requires

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706034

federated RL with privacy-preserving policy sharing. Fourth, adversarial attacks against the RL agents themselves (reward poisoning, observation manipulation) remain a theoretical vulnerability requiring formal robustness analysis.

Future directions include integrating large language models for natural-language threat intelligence processing and automated report generation, extending to 5G and satellite network segments, developing formal verification of RL policy safety guarantees for critical infrastructure, and establishing a cooperative multi-organization defense framework through federated multi-agent RL.

X. CONCLUSION

OPERATIONAL IMPACT
Autonomous RL defense transforms cybersecurity from a human-intensive, reactive discipline into a machine-speed, proactive capability - reducing alerts by 98.9%, response time by 99.98%, and per-incident cost by 97.7% while raising the NIST CSF score from 2.8 to 4.7.

This paper has demonstrated that multi-agent reinforcement learning, when deployed across a distributed network's heterogeneous segments, produces an autonomous defense capability that exceeds every existing approach - manual, rule-based, and ML-based - across every measured dimension. The 94.6% detection rate, 2.4-second response time, 1.9% false positive rate, and $420 per-incident cost represent not incremental improvements but a fundamental shift in what is operationally possible. The four red team exercises provide the strongest validation: in the final exercise, the autonomous system prevented all 18 simulated breaches using 52 MITRE ATT&CK techniques, a result that no manual SOC could achieve at any staffing level. As cyber threats continue to evolve in sophistication and speed, the ability to detect, decide, and respond at machine speed will transition from competitive advantage to existential necessity for organizations operating distributed networks at scale.

REFERENCES

[1] MITRE Corporation, "ATT&CK Framework v14," 2025. Available: https://attack.mitre.org/
[2] NIST, "Cybersecurity Framework (CSF) 2.0," 2024.
[3] J. Schulman et al., "Proximal Policy Optimization Algorithms," arXiv:1707.06347, 2017.
[4] T. Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy RL," ICML, 2018.
[5] C. Yu et al., "The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games," NeurIPS, 2022.
[6] S. Fujimoto et al., "Addressing Function Approximation Error in Actor-Critic Methods," ICML, 2018.
[7] V. Mnih et al., "Asynchronous Methods for Deep RL," ICML, 2016.
[8] Z. Wang et al., "Dueling Network Architectures for Deep RL," ICML, 2016.
[9] Microsoft, "CyberBattleSim," 2022. Available: https://github.com/microsoft/CyberBattleSim
[10] E. Liang et al., "RLlib: Abstractions for Distributed RL," ICML, 2018.
[11] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[12] R. Lowe et al., "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," NeurIPS, 2017.
[13] CISA, "Zero Trust Maturity Model v2.0," 2023.
[14] SANS Institute, "SOC Survey 2025: Metrics and Staffing," 2025.
[15] Verizon, "Data Breach Investigations Report (DBIR)," 2025.
[16] CrowdStrike, "Global Threat Report 2025," 2025.