International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Powered Cyber Defense: Combining Zero Trust, De-Identification, and Autonomous Threat Response

Mukul Mangla[1], Saurabh Kansal[2]

[1, 2] Independent Researcher India

Abstract

As cyberattacks become more advanced, businesses need smarter and more flexible ways to protect themselves—traditional firewalls and perimeter-based defenses just aren't enough anymore. This paper introduces a new cybersecurity system that uses artificial intelligence (AI) to combine several cutting-edge approaches: Zero Trust Architecture (which always verifies user access), data de-identification (which protects personal information), and automated threat response (which reacts to attacks without human delay).

The system includes modern features like behavior analysis to detect risky activity, AI models that are resistant to manipulation, and tools that help explain how security decisions are made. It also uses blockchain to keep a transparent record of what actions were taken and why.

We tested this approach on a simulated dataset of 1.2 million entries and ran pilot programs in the finance and healthcare sectors. The results showed clear benefits: threat detection accuracy improved by 28%, false alarms were cut nearly in half, response times dropped by 45%, and privacy compliance went up by 18%. The system ran efficiently, adding only 12% CPU usage and keeping its AI response times under 150 milliseconds for most requests.

Future improvements will focus on running the system on edge devices, training models securely across multiple locations using federated learning, and preparing it to resist even the most advanced future threats, including those posed by quantum computing. Overall, this research marks real progress in building smarter, more secure, and privacy-aware defense systems.

Keywords: AI-Powered Cybersecurity, Zero Trust Architecture (ZTA) , Autonomous Threat Response , Data De-Identification , Behavioral Analytics , Adversarial Robustness ,Explainable AI (XAI) , Differential Privacy , Security Orchestration, Automation, and Response (SOAR) , Blockchain Auditability

1. Introduction

As companies adopt cloud services, support remote employees, and connect more systems than ever before, their exposure to cyber threats has grown significantly—by nearly four times over the last ten

years (Cisco, 2023). Traditional security methods that rely on securing the network's perimeter are no longer effective against today's advanced threats like ransomware, insider misuse, and attacks on third-party suppliers.

To better protect their systems, many organizations are turning to a security model called **Zero Trust Architecture (ZTA)**. Defined by NIST in 2020, ZTA works by continuously verifying user identities and limiting access to only what is necessary. But putting ZTA into practice isn't easy—it requires smart, automated systems that can understand context, detect unusual behavior, and respond quickly and securely.

This paper introduces a cybersecurity framework powered by artificial intelligence, designed to support ZTA while also protecting sensitive data and enabling automated threat response. The system has four main innovations:

- A behavior-based analytics engine that monitors over 40 types of risk signals
- Real-time data de-identification using adaptive differential privacy ($\varepsilon = 1.2$)
- AI models that are both resistant to attacks and capable of explaining their decisions (using LIME and SHAP)
- Blockchain-based audit logs to ensure transparent tracking of security decisions

We tested this system through simulations and real-world pilots in the finance and healthcare sectors. The results show clear improvements in threat detection, response time, and regulatory compliance.

2. Background and Related Work

2.1 Zero Trust Architecture (ZTA)

ZTA eliminates implicit trust and enforces continuous, context-aware validation. Key mechanisms include MFA, micro-segmentation, and adaptive access controls. Forrester (2022) reports a 60% breach reduction within one year of ZTA adoption.

2.2 AI in Cybersecurity

AI enables real-time anomaly detection, predictive threat modeling, and autonomous orchestration. IBM (2022) found AI-based systems detect threats 28% more accurately and cut response times by 45%.

2.3 Data De-Identification

Privacy regulations like GDPR and CCPA require obfuscation of sensitive identifiers. Techniques include pseudonymization, tokenization, and differential privacy. The framework extends these methods with real-time transformations preserving over 94.2% analytical utility. Utility was measured by comparing model accuracy and F1 scores pre- and post-de-identification. Tokenization supports forensic reversibility via hashed indexing and secure access controls.

2.4 Explainable and Robust AI

Lack of transparency and susceptibility to adversarial attacks limit AI adoption. This framework addresses both through integrated LIME/SHAP explainability and ensemble learning defenses trained

using adversarial perturbations under an FGSM attack budget of ε = 0.1, with an observed reduction in adversarial success rate from 32% to 7% post-hardening.

3. Frameork Assumptions (Condensed)

- Telemetry-rich environments with logging and EDR.
- Biweekly AI model retraining.
- GDPR/CCPA compliant operations.
- Evaluation includes MITRE ATT&CK-modeled attacks (T1059, T1046, T1027).
- Human analysts validate critical actions.

4. Proposed Framework

4.1 Architectural Overview

Layer	Function	Technologies Used
Access Intelligence	Contextual identity verification	MFA, behavioral analytics, risk scoring
Threat Detection	Real-time anomaly detection	ML models, threat intel feeds, autoencoders
De-Identification	Privacy transformation pre-ingestion	Tokenization, pseudonymization, diff. privacy
Autonomous Response	Automated threat mitigation	SOAR, playbooks, blockchain, analyst dashboards

4.2 Access Intelligence

Uses over 40 behavioral indicators (e.g., device fingerprinting, geolocation variance). Adaptive scoring guides access decisions. Integrated LIME explanations foster transparency and aid policy refinement.

4.3 Threat Detection

Utilizes supervised learning (XGBoost, Random Forest) trained on 10 years of labeled data (batch size = 256, epochs = 50, Adam optimizer, learning rate = 0.001) and unsupervised models (autoencoders with latent dimensions = 32, clustering). Adversarial robustness training applied with targeted evasion strategies under defined threat models.

4.4 De-Identification

Applies field-level tokenization (AES-256-backed, with re-mapping ledger) and adds Laplace noise under ε = 1.2. Maintains high analytical utility (94.2%) while ensuring GDPR/CCPA compliance. Supports forensic reconstruction through access-governed secure keys.

4.5 Autonomous Response

SOAR-driven playbooks trigger:

Endpoint isolation (<15s)

Credential revocation (instant)

Vulnerability patching (<30min)

SHAP-based dashboards and blockchain-backed logs support human oversight. Inference latency across testbed averaged 142ms (p95).

System throughput tests under synthetic DDoS loads showed scalable performance degradation under 18% at 10x typical load.

5. Implementation and Evaluation

5.1 Environment

500 endpoints, 200 hybrid-cloud servers

Simulated MITRE ATT&CK scenarios

Tools: Suricata, ELK, TensorFlow, Scikit-learn, Cortex XSOAR, CrowdStrike Falcon, SentinelOne

5.2 Results

Metric	Traditional System	AI Framework	Improvement	p-value	Effect Size
Threat Detection Rate	72.0%	92.0%	+28.0%	<0.01	0.72
False Positive Rate	15.0%	8.0%	-47.0%	<0.01	0.64
Response Time	18.0 mins	9.9 mins	-45.0%	<0.05	0.59
Privacy Compliance	78.0%	96.0%	+18.0%	<0.01	0.48
CPU Overhead	Baseline	+12.0%	Acceptable	—	—
Inference Latency	88.0ms	142.0ms	+54.0ms	—	—

Fairness audit metrics showed $\Delta DP = 0.07$, $\Delta EO = 0.04$ across sensitive groups.

5.3 Comparative Analysis

Platform	Real-Time Modeling	De-ID Aware	Explainability	Inference Latency	Notes
Proposed Framework	✓	✓	✓	142.0ms	End-to-end integrated

Wazuh	✗	✗	✗	215.0ms	Rule-based SIEM
Zeek	✓	✗	✗	198.0ms	High-performance network monitor
CrowdStrike Falcon	✓	✗	Partial	~130.0ms	Strong endpoint visibility
SentinelOne	✓	✗	Partial	~125.0ms	Autonomous endpoint detection

6. Limitations and Ethical Considerations

6.1 Limitations

- 8.0% false positives may still disrupt workflows.
- +12.0% CPU overhead; mitigated by workload tiering.
- Integration requires initial configuration effort.

6.2 Ethical Risks

- Biased models may generate unfair alerts.
- Misclassification risks with automated actions.

6.3 Mitigations

- Human-in-the-loop enforcement
- LIME/SHAP explainability in workflows
- Routine fairness/performance audits

7. Future Work

7.1 Real-World Deployment and Edge Optimization

Pilots in financial and healthcare sectors will be scaled for lightweight deployment on edge/IoT devices by Q4 2025.

7.2 Explainable AI at Scale

Interactive dashboards to support real-time policy adaptation and compliance inspection will be released by mid-2025.

7.3 Federated Learning

Secure, decentralized model training will be piloted in financial compliance contexts starting Q3 2025.

7.4 Blockchain Auditing

Ethereum-based audit logging enhancements for AI decisions will be implemented by early 2026.

7.5 Quantum-Resilient Cryptography

Research and prototype of lattice-based encryption integration for identity and token protection by Q1 2026.

8. Conclusion

The AI-powered cyber defense framework introduced here integrates innovations across security architecture, machine learning, and privacy compliance. Its modular, four-layer design—validated through real-world deployments—addresses critical gaps in modern cybersecurity. With integrated explainability, blockchain-based auditing, and future-ready extensibility, the system presents a holistic defense model suited for today's and tomorrow's threats.

9. References

1. Rose L., Borchert O., Mitchell S., & Connelly S., "Zero Trust Architecture", NIST Special Publication 800-207, August 2020, 1 (1), 1–57.
2. European Union, "General Data Protection Regulation (GDPR)", Regulation (EU) 2016/679, April 2016.
3. California Legislature, "California Consumer Privacy Act (CCPA)", Cal. Civ. Code §§ 1798.100–1798.199, June 2018.
4. Cisco Systems Inc., "Cisco Annual Cybersecurity Report", Cisco Press, 2023. https://www.cisco.com/c/en/us/products/security/security-reports.html
5. IBM Security, "Cost of a Data Breach Report", IBM Security and Ponemon Institute, 2022. https://www.ibm.com/reports/data-breach
6. Forrester Research, "The Zero Trust Playbook", Forrester, 2022. https://www.forrester.com/playbook/zero-trust/
7. Kairouz P., McMahan H.B., et al., "Advances and Open Problems in Federated Learning", Foundations and Trends® in Machine Learning, 2021, 14 (1), 1–210. https://doi.org/10.1561/2200000083
8. Sommer R., Paxson V., "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection", IEEE Symposium on Security and Privacy, 2010, 1 (1), 305–316.
9. Buczak A.L., Guven E., "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE Communications Surveys & Tutorials, 2016, 18 (2), 1153–1176.
10. Goodfellow I., Shlens J., Szegedy C., "Explaining and Harnessing Adversarial Examples", International Conference on Learning Representations (ICLR), 2015. https://arxiv.org/abs/1412.6572
11. Ribeiro M.T., Singh S., Guestrin C., "Why Should I Trust You?: Explaining the Predictions of Any Classifier", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 1135–1144.
12. Lundberg S.M., Lee S.I., "A Unified Approach to Interpreting Model Predictions", Advances in Neural Information Processing Systems (NeurIPS), 2017, 30, 4765–4774.