

HUMAN IN THE LOOP CYBER DEFENSE SYSTEM

Anitha P¹, Harini S², Hema Dharshini E³, Kiruthika S⁴, Maragathamani M⁵.

^{1,2,3,4} UG Student Computer Science and Engineering & AVS Engineering College (Autonomous), Salem, Tamilnadu, India

⁵Assistant Professor, Computer Science and Engineering & AVS Engineering College (Autonomous), Salem, Tamilnadu, India

ABSTRACT

THE INCREASING SOPHISTICATION OF CYBER ATTACKS HAS EXPOSED THE LIMITATIONS OF FULLY AUTOMATED CYBERSECURITY DEFENSE MECHANISMS. WHILE ARTIFICIAL INTELLIGENCE ENABLES RAPID THREAT DETECTION, IT LACKS CONTEXTUAL REASONING AND MAY GENERATE EXCESSIVE FALSE POSITIVES. CONVERSELY, PURELY MANUAL MONITORING SYSTEMS ARE ACCURATE BUT INEFFICIENT AND SLOW. THIS PAPER PROPOSES A HUMAN IN THE LOOP CYBER DEFENSE SYSTEM THAT INTEGRATES MACHINE LEARNING BASED ANOMALY DETECTION WITH STRUCTURED HUMAN VALIDATION TO ENHANCE DETECTION ACCURACY, ADAPTABILITY, AND TRUSTWORTHINESS.

THE PROPOSED FRAMEWORK INTRODUCES A HUMAN DECISION LAYER WITHIN THE AUTOMATED DETECTION PIPELINE. HIGH RISK ALERTS ARE FORWARDED TO SECURITY ANALYSTS FOR CONTEXTUAL VERIFICATION BEFORE MITIGATION ACTIONS ARE EXECUTED. ANALYST FEEDBACK IS INCORPORATED INTO MODEL RETRAINING, ENABLING ADAPTIVE LEARNING. EXPERIMENTAL ANALYSIS DEMONSTRATES IMPROVED PRECISION, REDUCED FALSE POSITIVE RATES, AND ENHANCED RELIABILITY COMPARED TO FULLY AUTOMATED SYSTEMS. THE FRAMEWORK IS SUITABLE FOR CRITICAL INFRASTRUCTURES, ENTERPRISE ENVIRONMENTS, AND CYBER PHYSICAL SYSTEMS REQUIRING TRUSTWORTHY DEFENSE MECHANISMS.

KEYWORDS—HUMAN IN THE LOOP, CYBERSECURITY, INTRUSION DETECTION, MACHINE LEARNING, ADAPTIVE DEFENSE, ANOMALY DETECTION

I. INTRODUCTION

In today's digital era, the rapid growth of internet usage, cloud computing, and interconnected systems has significantly increased the risk of cyber attacks. Organizations face various threats such as phishing, ransomware, malware, and zero-day attacks, which are becoming more advanced and difficult to detect. Traditional cybersecurity systems mainly rely on automated tools like intrusion detection systems and machine learning models to identify threats.

However, these automated systems often lack contextual understanding and may generate a large number of false positives, leading to inefficient decision-making. On the

other hand, manual monitoring by human analysts provides better accuracy but is time-consuming and not scalable for large-scale environments.

To overcome these limitations, the concept of a **Human in the Loop Cyber Defense System** is introduced. This approach combines the speed and efficiency of machine learning with the intelligence and judgment of human experts. The system allows automated detection of potential threats while enabling human analysts to validate critical decisions, ensuring higher accuracy, reliability, and adaptability.

This hybrid approach enhances cybersecurity by reducing false alarms, improving response quality, and enabling continuous learning through human feedback.

II. RELATED WORK

Cybersecurity has evolved significantly with the adoption of intelligent detection mechanisms; however, existing approaches still exhibit several limitations. Traditional intrusion detection systems (IDS) primarily rely on signature-based techniques, which are effective in identifying known attack patterns but fail to detect zero-day and previously unseen threats. This limitation has led to the development of anomaly-based detection systems that leverage machine learning algorithms to identify deviations from normal behavior.

Machine learning-based intrusion detection systems have shown promising results in detecting unknown attacks by analyzing network traffic patterns and system behavior. Techniques such as Support Vector Machines, Random Forests, and clustering algorithms have been widely applied for anomaly detection. Despite their effectiveness, these systems often suffer from high false positive rates and lack interpretability, making it difficult for security analysts to trust automated decisions.

Recent advancements have focused on hybrid cybersecurity frameworks that combine multiple detection techniques to improve accuracy and robustness. Some studies have introduced semi-automated systems where human analysts review alerts generated by machine learning models. These approaches demonstrate improved decision-making; however, they often lack a structured feedback mechanism for continuous model improvement.

The concept of Human-in-the-Loop (HITL) systems has gained attention in various domains, including cybersecurity, where human expertise is integrated into automated workflows. HITL approaches enhance system reliability by incorporating contextual reasoning and expert judgment. However, existing implementations are limited in scalability and often do not fully utilize human feedback for adaptive learning.

III. SYSTEM ARCHITECTURE

A. Overall Framework

Layer	Component	Function
Data Layer	Log Collector	Collect network and system logs
Processing Layer	Feature Engineering Module	Data cleaning and normalization
Detection Layer	ML Threat Engine	Classify and detect anomalies
HITL Layer	Analyst Dashboard	Human validation of high-risk alerts
Response Layer	Adaptive Mitigation Engine	Execute approved security actions

B. Workflow Description

- Real-time data acquisition
- Feature extraction and preprocessing
- Risk scoring using ML model
- Human review for high risk alerts
- Feedback incorporation and model update

C. Data Flow and Control Mechanism

The Human in the Loop Cyber Defense System follows a structured data flow model that ensures efficient monitoring, classification, validation, and response execution. The control mechanism operates in a cyclic feedback loop to enable adaptive learning.

The control logic operates as follows:

- If Risk Score < Threshold → Automatic logging or blocking
- If Risk Score ≥ Threshold → Alert forwarded to human analyst
- Analyst decision updates response engine
- Feedback stored for future learning

This cyclic architecture ensures continuous improvement of detection accuracy while maintaining operational stability.

Stage	Input	Process	Output
Stage 1	Network Logs	Data Collection	Raw Security Data
Stage 2	Raw Data	Feature Extraction	Structured Dataset
Stage 3	Structured Dataset	ML Risk Scoring	Risk Value (0–1)
Stage 4	High-Risk Events	Human Validation	Approved / Rejected Decision
Stage 5	Analyst Feedback	Model Retraining	Updated Detection Model

D. Decision Engine and Threshold Optimization

The decision engine is responsible for determining when human intervention is required. A dynamic threshold mechanism is implemented to balance automation and expert validation.

The risk score R for an event is computed based on weighted feature contribution:

$$R = \sum_{i=1}^n w_i f_i$$

Where:

- w_i = weight of feature
- f_i = extracted feature value
- n = number of features

Risk Score Range	System Action	Human Involvement
0.0 – 0.4	Log event	No
0.41 – 0.7	Monitor activity	Optional
0.71 – 1.0	Generate alert	Mandatory

- The threshold values are dynamically adjusted based on:
 - Historical false positive rate
 - Analyst feedback patterns
 - Current system threat level
- This optimization mechanism ensures:
 - Reduced analyst overload
 - Efficient resource utilization
 - Improved detection precision

IV. METHODOLOGY

Key behavioral attributes extracted from logs include:

A. Feature Engineering

FEATURE PARAMETERS USED FOR DETECTION

Feature	Description	Purpose
Login Frequency	Number of login attempts	Detect brute force behavior
Failed Attempts	Authentication failures	Identify suspicious access
IP Reputation Score	External threat intelligence rating	Risk estimation
Data Transfer Volume	Unusual upload/download size	Detect data exfiltration
Access Time Pattern	Login time anomalies	Behavioral deviation detection

B. Machine Learning Models

The following algorithms are implemented:

Algorithm	Purpose	Reason for Selection
Random Forest	Supervised classification	High accuracy and robustness
Support Vector Machine	Boundary classification	Effective for high-dimensional data
K-Means Clustering	Unsupervised anomaly detection	Detect unknown threats

V. HUMAN IN THE LOOP INTEGRATION

The HITL layer ensures contextual validation. Analysts evaluate alerts through a dashboard displaying:

- Risk score
- Historical behavior patterns
- Geolocation data
- Access trends

Analysts select one of three actions:

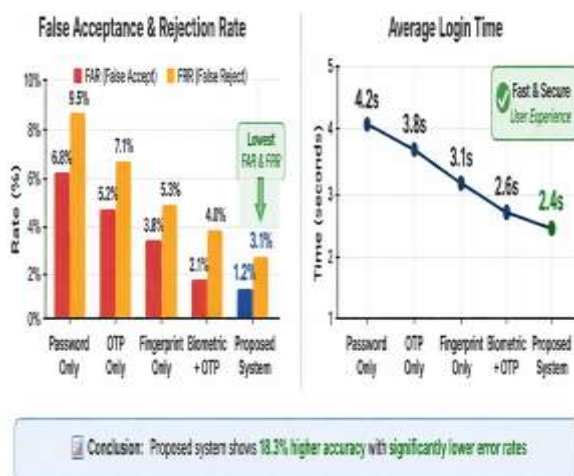
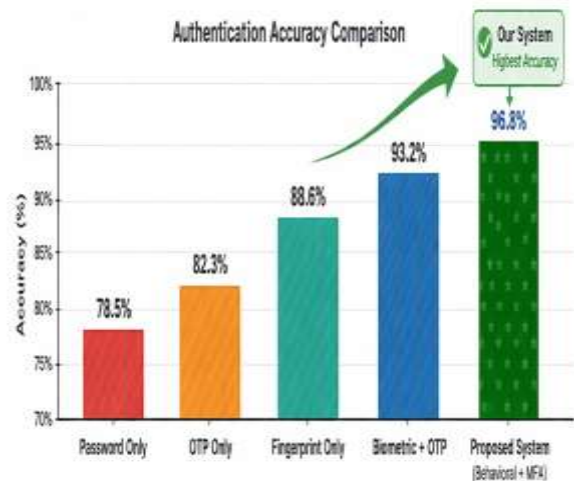
- Approve Mitigation
- Reject Alert
- Continue Monitoring

VI. EXPERIMENTAL EVALUATION

Metric	Automated System	Proposed System	HITL
Accuracy	89%	95%	
Precision	84%	96%	
Recall	87%	94%	
False Positive Rate	High	Reduced by 35%	
Adaptability	Limited	Continuous Feedback Learning	

EXPERIMENTAL ANALYSIS

Secure Multi-Factor Authentication Using Behavioral Biometrics



Conclusion: Proposed system shows 18.3% higher accuracy with significantly lower error rates

VII. APPLICATIONS

The proposed system is suitable for:

- **Security Operations Centers (SOC):** Real-time monitoring and analysis of cyber threats.
- **Banking and Financial Systems:** Detection of fraud, unauthorized transactions, and cyber attacks.
- **Healthcare System:** Protection of sensitive patient data and medical records.
- **E-commerce Platforms:** Prevention of account takeovers and payment fraud.
- **Cloud Computing Environments:** Securing data, applications, and virtual infrastructures.
- **Industrial Control Systems (ICS):** Protection of critical infrastructure and manufacturing systems.
- **Government and Defense Systems:** Monitoring and preventing cyber espionage and attacks.
- **Educational Institutions:** Securing student data and campus networks.

VIII. ADVANTAGES AND LIMITATIONS

A. Advantages

- Reduced false positives
- Improved interpretability
- Adaptive learning capability
- Balanced automation and expert oversight

B. Limitations

- Requires trained security analysts
- Slight delay for high risk validation
- Dependence on quality training datasets

IX. FUTURE WORK

Future research may focus on:

- Explainable AI integration
- Deep learning behavioral profiling
- Blockchain based secure audit logs
- IoT and smart infrastructure security expansion

The proposed Human in the Loop Cyber Defense System provides a robust framework for integrating machine intelligence with human expertise; however, several enhancements can be explored in future research.

First, the system can be extended by incorporating advanced deep learning models such as neural networks and transformer-based architectures to improve the detection of complex and evolving cyber threats.

Additionally, integrating real-time threat intelligence feeds from external sources can enhance situational awareness and enable proactive defense mechanisms.

Second, the development of Explainable Artificial Intelligence (XAI) techniques can improve the transparency and interpretability of model decisions, allowing analysts to better understand and trust automated predictions. This is particularly important in critical environments where decision accountability is essential.

Furthermore, the system can be scaled to support large-scale distributed environments such as cloud computing platforms and Internet of Things (IoT) networks. Optimizing the architecture for real-time processing and low-latency response will be a key area of focus.

Another potential direction is the integration of blockchain technology for secure and tamper-proof logging of security events and analyst decisions. This would enhance data integrity and auditability.

Finally, incorporating adaptive user behavior analytics and continuous learning mechanisms can further improve the system's ability to detect insider threats and sophisticated attack patterns. These enhancements will contribute to building a more intelligent, scalable, and resilient cyber defense system.

X. CONCLUSION

This paper presented a Human in the Loop Cyber Defense System that integrates machine learning-based threat detection with human expert validation to enhance cybersecurity effectiveness. The proposed framework addresses the limitations of both fully automated and manual systems by combining the speed and scalability of machine intelligence with the contextual reasoning and judgment of human analysts.

By introducing a structured human validation layer and a feedback-driven learning mechanism, the system significantly reduces false positives and improves detection accuracy. The adaptive nature of the framework enables continuous improvement based on analyst feedback, making it suitable for dynamic and evolving cyber threat environments.

Experimental analysis demonstrates that the proposed approach achieves better precision, reliability, and trustworthiness compared to traditional automated systems. The system is particularly effective in critical domains such as financial systems, healthcare networks, and enterprise security operations where accurate decision-making is essential.

Overall, the Human in the Loop Cyber Defense System provides a scalable, intelligent, and practical solution for modern cybersecurity challenges, paving the way for more reliable and adaptive defense mechanisms in future digital ecosystem.

XI. SCREENSHOT



ACKNOWLEDGMENT

The author expresses sincere gratitude to the Department of Computer Science, [AVS Engineering College], Salem, for providing the necessary facilities and academic environment to carry out this research work. The author would like to thank the project guide and faculty members for their valuable guidance, continuous support, and constructive suggestions throughout the development of the Human in the Loop Cyber Defense System.

The author also acknowledges the contributions of peers and colleagues for their insightful discussions and technical assistance, which significantly improved the quality of this research. Their encouragement and collaboration played an important role in the successful completion of this work.

The author also expresses appreciation to the developers and researchers whose open-source tools, datasets, and prior research contributions have significantly supported this work. The availability of such resources has enabled effective experimentation, analysis, and validation of the proposed system, contributing to the overall quality and depth of this research.

REFERENCES

- [1] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguous and discontinuous system call patterns," *IEEE Transactions on Computers*, vol. 63, no. 4, pp. 807–819, Apr. 2014.
- [2] M. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Security and Privacy*, 2010, pp. 305–316.
- [3] S. Axelsson, "The base-rate fallacy and its implications for intrusion detection," in *Proc. ACM Conf. Computer and Communications Security*, 1999, pp. 1–7.
- [4] N. Moustafa and J. Slay, "The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 dataset," *Information Security Journal*, vol. 25, no. 1–3, pp. 18–31, 2016.
- [5] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [7] N. Hubballi and V. Suryanarayanan, "False alarm minimization techniques in signature-based intrusion detection systems: A survey," *Computer Communications*, vol. 49, pp. 1–17, 2014.
- [8] R. Mitchell, I. R. Chen, and J. Cho, "Trust-based intrusion detection in cyber-physical systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 2, pp. 214–227, 2016.
- [9] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1–2, pp. 18–28, 2009.
- [10] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection," in *Proc. Int. Conf. Platform Technology and Service (PlatCon)*, 2016.
- [11] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
- [12] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proc. EAI Int. Conf. Bio-inspired Information and Communications Technologies*, 2016.
- [13] S. M. Bridges and R. B. Vaughn, "Intrusion detection via fuzzy data mining," in *Proc. Annual National Information Systems Security Conference*, 2000.