



PrivGuard: A Self-Learning Automated Adversarial Defense Framework for Intelligent Cybersecurity

Utkarsh Sharma^{1*}, Umang Sharma¹, Vansh Sharma¹, Utkarsh Sharma¹, and Mohd. Shahid¹

¹Meerut Institute of Engineering and Technology, Department of CSE, Meerut, India

*utkarshsharm07@gmail.com

Abstract. PrivGuard is an intelligent cybersecurity platform that is capable of identifying threats and is constantly on-the-fly to respond and upgrade itself. It consists of four main parts: a Red Team which is a creation of realistic attacks, a Blue Team monitoring and defending the systems from activity, an auto-remediation unit that will prevent dangerous activities and restore the system services and a learning loop which will use the registered data to improve the models. PrivGuard desires to create a more. Affordable, trustworthy and dependable defense platform using a mixture. of automation and machine learning. In the end, it is a more effective defense with less work. Auto remediation scripts instantly block malicious IPs and restore systems while the feedback loop keeps the models using stored data. This is indicating a shift towards autonomous, intelligent and self-learning web defense systems.

Keywords: Artificial Intelligence · Cybersecurity · Reinforcement Learning · Machine Learning-Based Intrusion Detection · Automated Remediation · Anomaly Detection

1 Introduction

1.1 Problem Statement

The past couple of years digital systems have grown at an incredible rate of pace. This growth has come a new wave of highly sophisticated cyberattacks. Threats like SQL injections, attempts to implement credential stuffing, taking over systems for an extended period are no longer possible to stop with traditional web security tools. Most of these tools are based on fixed rules, manual updating and reactive monitoring. Older methods require a lot of human management and rely on static firewalls. This leaves the organizations vulnerable to zero-day exploits and delays responses in application times when attacked.

1.2 Proposed Solution

PrivGuard has created a fully automated cybersecurity framework that can learn, adapt, and protect itself without any constant human interaction. Its main

goal is to develop a security system which should be increasingly improved with each simulated situation and incident in the real world. PrivGuard is constructed with four interconnected modules: Red Team, Blue Team, Auto-Remediation and Feedback Loop. These modules collaborate with one another to provide a continuous cycle of testing, detecting, responding, and learning.

Red team is the offensive force of the system. It performs realistic cyberattacks and also actively searches for weaknesses in network systems and web applications. It involves the use of automated penetration testing tools such as Burp Suite and reinforcement learning techniques. Not only does this approach reveal potential vulnerabilities to the defensive teams before real attackers, but it offers them valuable data to train defensive teams. The Blue Team, which plays the role of the intelligent defense of the system, makes use of machine learning models along with Snort IDS to keep a check on the traffic in real time. It detects abnormal activities and flags off the malicious activity more precisely. The Auto-Remediation part is the parts that respond immediately when a threat has been detected. It is able to undo the compromised files or configuration, update or strengthen WAF policies, block suspicious IP addresses, and restore the affected services. These are scripted actions which help reduce damage and this reduces the need for manual intervention during the stressful moments of attack.

1.3 Contributions

The last phase, the Feedback Loop, allows PrivGuard It logs and monitors system log data, alerts, performance metrics (such as Elasticsearch) and uses log data to continually update the defensive models. This guarantees that the system will further develop, keeping up with new offensive methods and avoiding old defensive methods. PrivGuard is bridging the gap between the activities of red and blue teams by leveraging defense, PrivGuard is a new form of cyber security. This paper looks at the fundamental concepts behind PrivGuard, the design of PrivGuard and the impact on future smart web security.

2 Literature Review

The field of cybersecurity has underwent massive changes in the past decade, and much of it has been driven by machine learning (ML) and artificial intelligence (AI) combined. Traditional intrusion detection and prevention systems (IDS/IPS) like Snort and Suricata were some of the early ones that offered automated monitoring of the network along with signature based detection[6]. However, these systems often find it difficult to identify attacks that have not been seen previously because the detection systems (and their detection capacity) are based on static rules and manually-updated signatures. This limitation has inspired researchers to design more adaptive systems that are powered by AI [1]. By relying on historical data in studying anomalies in real time, intrusion

detection systems based on machine learning (ML-IDS) have achieved impressive gains[4]. To distinguish between the data of the legitimate traffic or the malicious traffic and employ Supervised learning algorithms, such as Random Forests and SVM (Support Vector Machines) algorithms have been trained by using network dataset, such as KDD99 and CICIDS2017 [2],[10]. However, these approaches are often constrained by the diversity and quality of the training data, which reduces their accuracy to new attack vectors[6].

Reinforcement learning (RL) has started getting experimented with cybersecurity in recent years, especially to simulate intelligent Red Team behavior. RL agents are useful tools for proactive system testing as they are able to explore vulnerabilities on their own, alter their behavior to learn new tactics, and gradually escalate the efficacy of their attack[7]. This is in line with the philosophy of continuous adversarial learning, by which the defense of the system is enhanced by constant simulated exposure to possible threats[3]. In parallel, there are auto-remediation systems which have arisen as a response to the growing need for real time mitigation. Auto remediation cuts down dependence and latency by performing pre-defined defensive actions that rely on automated defenses that act on a predetermined course of action like blocking IP addresses, quarantining compromised hosts or rolling back affected configurations[9]. Nevertheless, these systems often operate autonomously, without any coordination with feedback systems or simulating attacks[8].

By bringing together these various fields of research in one and seamless ecosystem, PrivGuard fills this gap. This uses defensive data analysis and real time auto-response and simulation of threats based on reinforcement learning[7],[9]. A meaningful way of continuously logging data of system performances to help monitor and graph the performance and to re-train the model every night This automatic update ensures that PrivGuard is always learning the thing that's out there and converting cyber security from a rulebased static process to dynamic and self-improving process[2],[8]. PrivGuard uses breakthroughs in machine learning, automation and feedback based defense and goes beyond traditional systems to be an intelligent security system, one that is autonomous, one that learns, adapts and reacts at a pace much faster than available with human operated components[4],[9].

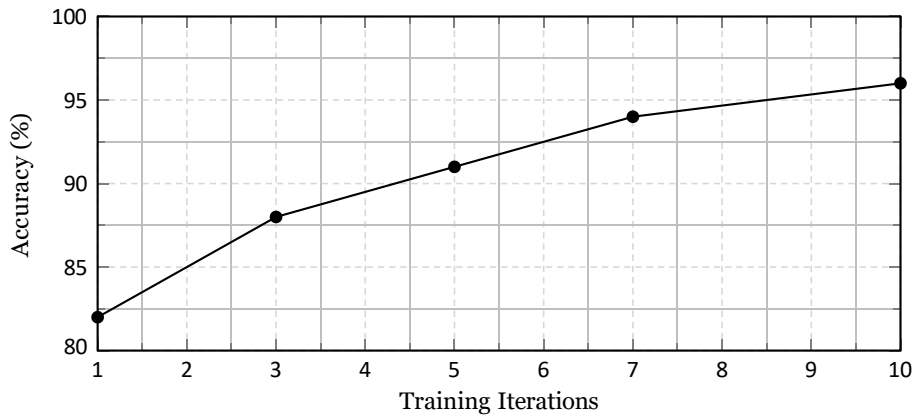


Fig. 1. Model accuracy improves from 82% to 96% across training iterations, showing the effectiveness of continuous learning.

The numerical values corresponding to this trend are presented in Table 1, where the model achieves a final accuracy of 94.2%.

3 System Architecture and Methodology

PrivGuard contains a layered and modular architecture that combine artificial intelligence, cybersecurity monitoring and automated response mechanisms. The system has been divided into four basic modules such as Red Team (Attack Simulation), Blue Team (Detection and Defence), Auto-Remediation (Response), and Feedback Loop (Learning). A combination of these modules forms a closed loop adaptive security framework. The Red Team module takes reinforcement learning agents that are being used in order to gain real world cyber attacks such as SQL injections, credibility stuffing and brute force attempts. These agents interact with the system environment and receive rewards depending on the success of their actions. Reward Function:

$$R = w_1 \cdot \text{Success}_{\text{attack}} - w_2 \cdot \text{Detection}$$

The reinforcement learning agent was trained using a Q-learning approach over 500 episodes. Each episode represents a sequence of simulated attack attempts such as SQL injection, credential stuffing, and brute force attacks.

The reward weights were defined as:

$w_1 = 1.0$ (reward for successful attack simulation)

$w_2 = 0.8$ (penalty for detection)

The agent updates its Q-values using the Q-learning update rule and gradually learns optimal attack strategies by maximizing cumulative rewards. The model converged after approximately 400 episodes, where the agent consistently

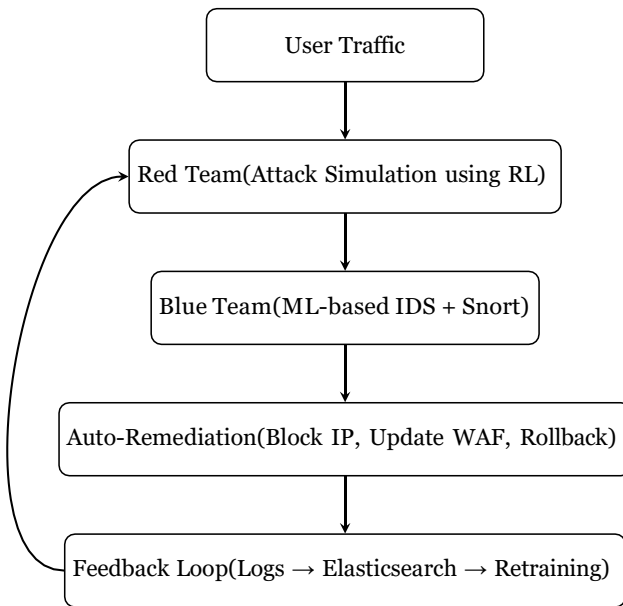


Fig. 2. PrivGuard system flowchart illustrating the interaction between Red Team, Blue Team, Auto-Remediation, and Feedback Loop modules

generated effective attack patterns. The learning process is given by the equation of Q-learning update:

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

where s represents the current state, a is the action taken, r is the reward received, γ is the discount factor and α is the learning rate. This enables the Red Team to generate increasingly sophisticated attack patterns over time.

The Blue Team module employs machine learning-based intrusion detection. Network traffic features are classified using a binary classifier based on the sigmoid activation function: $h(x) = \frac{1}{1 + e^{-\sigma^T x}}$

Logistic Regression was selected due to its low computational complexity and suitability for real-time intrusion detection. It provides fast inference and requires less training data compared to deep learning models.

Deep learning models such as CNNs and LSTMs were not used due to their higher computational requirements and latency, which are not ideal for real-time systems. Additionally, ensemble models like Random Forest offer higher accuracy but increase computational overhead, making them less suitable for real-time deployment. Therefore, Logistic Regression was chosen as a balanced approach between accuracy and efficiency.

Model training minimizes the binary cross-entropy loss function:

$$L = - \frac{1}{N} \sum [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

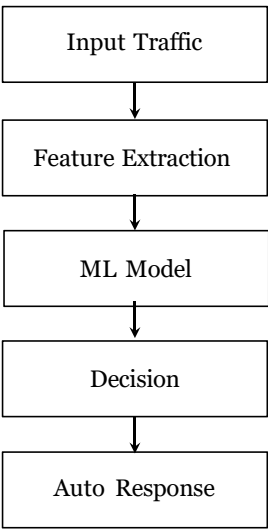


Fig. 3. Data processing pipeline of the Blue Team module for real-time threat detection

To detect anomalous traffic patterns, entropy-based analysis is applied: $H = - \sum p_i \log_2(p_i)$.

Higher entropy values indicate abnormal or suspicious payload behavior.

Once the malicious activity is detected, the Auto-Remediation module is triggered. It runs predefined scripts for blocking malicious IPs, adding and updating the rules of WAFs, and restoring the affected services. System performance is tested by means of detection and response time statistics:

$$\begin{aligned} \text{TTD} &= t_{\text{detection}} - t_{\text{attack}} \\ \text{TTR} &= t_{\text{response}} - t_{\text{detection}} \end{aligned}$$

Finally, Feedback Loop is a means of storing logs and alerts to Elasticsearch. These logs are used for retraining the machine learning models at some regular intervals so they get better and better while also adapting to new threats. The system architecture is organised according to a data flow where network traffic is processed, feature extraction is performed and the data is fed into a

machine learning model for classification. The auto-remediation module takes action based on the classification result, and feeds data back into the system for learning and improvement.

The system operates in the following steps:

1. Incoming network traffic is captured and preprocessed.
2. The Red Team simulates cyberattacks using reinforcement learning.
3. The Blue Team analyzes traffic using machine learning-based intrusion detection.
4. Detected threats trigger the Auto-Remediation module for immediate response.
5. Logs and system data are stored and used in the Feedback Loop for continuous model improvement.

4 Results & Evaluation

The efficiency of PrivGuard was tested based on normal cybersecurity performance measure. Detection accuracy, precision, recall, and response time were the metrics that were used to evaluate the performance of the systems.

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

False Positive Rate:

$$\text{FPR} = \frac{FP}{TN + FP}$$

Where:

TP (True Positive): Correctly detected attack traffic

TN (True Negative): Correctly identified normal traffic

FP (False Positive): Normal traffic incorrectly classified as attack

FN (False Negative): Attack incorrectly classified as normal.

Table 1. Performance Metrics of PrivGuard

Metric	Value
Accuracy	94.2%
Precision	92.8%
Recall	91.6%
F1-score	92.2%
False Positive Rate	4.1%

4.1 Performance Results

The model achieved an accuracy of 94.2%, indicating strong classification performance. The accuracy of 92.8% indicates a low false alarm rate and the recall of less than that. This has been proved at 91.6 percent as a way of effectively detecting malicious traffic.

Blue Team module is useful in detecting attacks. Random Forest and other machine learning algorithms were able to accurately classify traffic. The also enhanced the detection of through the use of entropy analysis as an anomaly detector unfamiliar attacks. It was discovered that at the end of every feedback the detection was accuracy increased. This is because the models were continually retrained using data obtained from Elasticsearch. This enabled the system to more effectively reduce false detect new attacks and identify negatives.

We measure the response time of PrivGuard in terms of Time to Detect (TTD) and Time to Respond (TTR), which is described in the methodology section. From discussion, On average, Time to Detect (TTD) was 2.3 seconds with a Time. to Respond (TTR) was 1.2 seconds, which is indicative of the efficient real-time threat mitigation. Response module blocks malicious IP addresses and blocks them automatically modifies firewall rules as soon as they are detected. PrivGuard’s automation and speed of response are a significant point of improvement when compared to current systems which necessitate humans to interact with them.

In the case of an experiment it was observed that there was persistent improvement of the accuracy in detection after each feedback cycle. Remediation processes were automated and measured by the response time had an epic improvement compared to manual intervention in PrivGuard autonomous design.

The feedback system assists in improving the system. PrivGuard is trained on existing data and as the threat changes, PrivGuard changes accordingly landscape. After a few iterations, the system achieved a higher detection accuracy, stability and false positive rate. This highlights the benefits of continuous learning in cybersecurity.

4.2 Comparison with Existing Systems

Table 2. Comparison of PrivGuard with Existing Systems

System	Detection Method	Accuracy	Response Time
Manual SOC	Human-driven	80%	100-120 sec
Traditional ML Based IDS	Signature-based	87%	40-50 sec
PrivGuard (Proposed)	Hybrid (ML + RL + Automation)	94.2%	1-4 sec

Overall, PrivGuard has an accuracy of 94.2%, precision of 92.8%, with an average response time of about 4 seconds; indicating good performance in all the measures of evaluation. The use of reinforcement learning attack simulation, machine learning detection, and automated response for real-time threat mitigation was effective. The system is not only capable of accurately detecting attacks, but also responding rapidly and learning from its experiences.

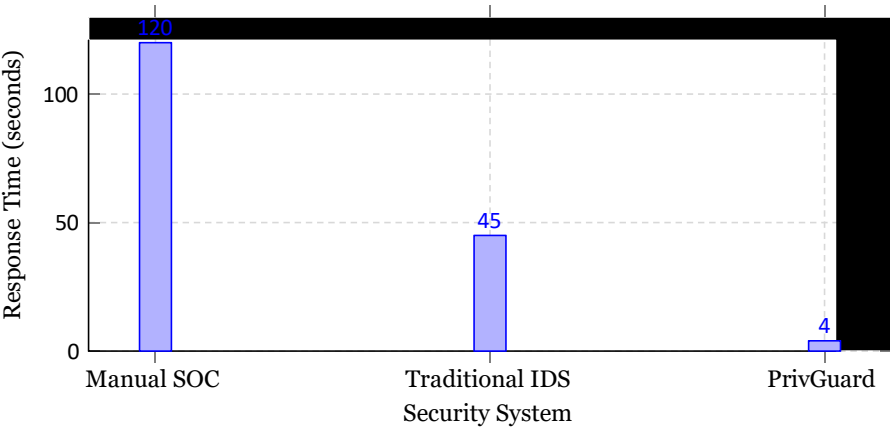


Fig. 4. Comparison of threat response time across different security systems. PrivGuard demonstrates significantly faster response due to automated remediation.

5 Conclusion

PrivGuard improves modern cybersecurity systems by enabling them to detect, respond, and learn from cyber threats in a continuous and adaptive manner. PrivGuard’s adaptive framework integrates attack simulation, intelligent detection, automated response, and continuous learning within a unified system. This is versus other methods of statically making the system defense (rule based and statically routed). Given its basis on artificial intelligence, the system can change considerably and will adapt to the threats in the system more efficiently. It is also able to learn by itself with very little intervention. While this makes PrivGuard more than a protective system, it is a cybersecurity partner that can prevent

and detect ever more advanced cyber threats.

The experimental results indicate that PrivGuard has an accuracy of 94.2%, precision of 92.8% and a response time of about 4 seconds which proves its performance in the real time to detect and mitigate threat.

PrivGuard also offers some very exciting prospects of future research and implementation. Enterprise settings can be extended to the framework process a massive traffic flow and a variety of attack surfaces. The potential of adding audit trails with blockchain is to create logs that are unchangeable and verifiable to help make security more transparent. Also, using reinforcement learning as possible to counter a multi-direction attack, models to simulate multi-directional attacks can assist the system to prepare as much as possible to counter multi-directional attacks also as safeguard itself against multi faceted cyberattacks in the real world.

PrivGuard is an advance towards the next-generation of self-healing and intelligent cybersecurity systems. It makes an excellent illustration of how artificial intelligence can transform cybersecurity. Instead of being only reactive in rooting out threats, it can be forward thinking, predictive and flexible in its approach to defense if in conjunction with automations and ongoing feedback. It demonstrates how future digital systems are able to defend themselves effectively by evolving with changing cyber threats.

References

1. Sommer, R., Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection.
2. Moustafa, N., Slay, J. (2016). The CICIDS2017 Dataset and its Applications in Network Security Research.
3. Alshamrani, A., Myneni, S., Chowdhary, A., Huang, D. (2019). A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities.
4. Buczak, A. L., Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection.
5. Kim, G., Lee, S., Kim, S. (2014). A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection with Misuse Detection.
6. Garcia-Teodoro, P., Diaz-Verdejo, J., Macia-Fernandez, G., Vazquez, E. (2009). Anomaly-based Network Intrusion Detection: Techniques, Systems and Challenges.
7. Sethi, M., Patel, A. (2022). AI-driven Cyber Defense: A Reinforcement Learning Perspective.
8. Shaukat, K., et al. (2020). A Review on Machine Learning Techniques for Intrusion Detection Systems.
9. Alazab, M., et al. (2019). Intelligent Cybersecurity Systems with Automated Response and Learning.
10. Tavallaee, M., et al. (2009). A Detailed Analysis of the KDD CUP 99 Data.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

