Hybrid Web Application Firewall

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

in

Computer Science and Engineering (IoT & CSIBT)

By

Dhruv Mehta 121AX029

Piyush Mejari 121AX030

Riyaz Mullaji 121AX031

Pradyum Samal 121AX041

Supervisor

Prof. Prachi Shahane

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

(IOT AND CYBERSECURITY, INCLUDING BLOCKCHAIN TECHNOLOGY)

SIES Graduate School of Technology

Sri Chandrasekarendra Saraswati

Vidyapuram Sector-V, Nerul, Navi Mumbai.

(AY 2024-25)

CERTIFICATE

This is to certify that the following students have completed and submitted the Capstone project report titled **Hybrid Web Application Firewall** is a Bonafide work of the following students, submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Computer Science and Engineering (IoT & CSIBT)**.

Sr. No.	**Name of Student**	**Roll No.**
1	Dhruv Mehta	121AX029
2	Piyush Mejari	121AX030
3	Riyaz Mullaji	121AX031
4	Pradyum Samal	121AX041

Prachi Shahane	**Dr. Sulochana Madachane**	**Dr.Laksmi Sudha**
Internal Guide	**Head of Department**	**Princi**

Project Report Approval

This project report entitled **“Hybrid Web Application Firewall”** by the following students is approved for the degree of **Bachelor of Engineering** in **Computer Science and Engineering (IoT & CSIBT)**.

Sr. No.	Name of Student	Roll No.
1	Dhruv Mehta	121AX029
2	Piyush Mejari	121AX030
3	Riyaz Mullaji	121AX031
4	Pradyum Samal	121AX041

Examiners

1………………………………………
(Internal Examiner Name & Sign)

2………………………………………
(External Examiner name & Sign)

Date: April 24, 2025

Place: SIES Graduate School of Technology, Nerul, Navi Mumbai.

ABSTRACT

This paper presents a hybrid web application firewall (WAF) that synergistically integrates domain-specific signature detection with a LightGBM-based anomaly detection model, specifically optimized for e-commerce platforms. Our system addresses the limitations of standalone signature-based or machine learning approaches by combining rule-based matching for known attack patterns (e.g., SQLi, XSS) with behavioral analysis of HTTP request parameters. We introduce a novel feature set focusing on request payload characteristics, including *GET/POST/PUT parameter length*, *payload entropy*, *numeric-text ratios*, and *special character density*, while intentionally excluding URI-related features due to their minimal discriminative power observed during exploratory data analysis (EDA). Evaluated on real-world e-commerce traffic, the model achieves 90.62% accuracy and 97.51% ROC AUC, demonstrating superior performance compared to baseline methods. The proposed approach reduces false positives by 18% compared to pure ML models while maintaining detection capabilities for zero-day attacks. This work advances hybrid WAF design through domain-specific feature engineering and statistically informed feature selection, providing actionable insights for securing modern web applications.

ACKNOWLEDGEMENTS

We would like to express our thanks to the people who have helped us the most throughout our project. We are grateful to our guide, Prof. Prachi Shahane, for their nonstop support for the project.

A special thanks goes to each other who worked together as a team in completing the project, where we all exchanged our interesting ideas and thoughts and made it possible to complete our project with all accurate information. We also wish to thank our parents for their support and attention, who inspired me to go my own way.

We would also like to extend our sincere gratitude to our Principal, Dr. Lakshmi Sudha and our Head of the Department Dr. Sulochana Madachane for their continuous support and encouragement.

We also would like to thank our other faculty members for providing us with all the required resources and references for the project.

Sr. No.	**Name of Student**	**Roll No.**
1	Dhruv Mehta	121AX029
2	Piyush Mejari	121AX030
3	Riyaz Mullaji	121AX031
4	Pradyum Samal	121AX041

Table of Contents

Section	Title	Page No.
	Abstract	i
	Acknowledgment	ii
	Table of Contents	iii
	List of Abbreviations	1-2
	List of Figures	3
	List of Tables	4
1	Introduction	5
1.1	Introduction	5
1.2	Motivation	5
1.3	Problem Statement and Objectives	6
2	Literature Survey	7
2.1	Survey of Existing System	7
2.2	Comparison of Existing System	8
2.3	References for Litreature Survey	9
3	Proposed System	10
3.1	Introduction to Proposed System	10
3.2	Signatures	11
3.3	Dataset Overview	11
3.4	Feature Engineering	12
3.5	Model Training	16
4	Experimentation and Test Results	17-19
5	Conclusion and Future Scope	20
6	References	21

List Of Abbreviations

Abbreviation	Full Form
WAF	Web Application Firewall
SQLi	Structured Query Language Injection
XSS	Cross-Site Scripting
CSRF	Cross-Site Request Forgery
DDoS	Distributed Denial of Service
IDS	Intrusion Detection System
SBD	Signature-Based Detection
ABD	Anomaly-Based Detection
ML	Machine Learning
OWASP	Open Worldwide Application Security Project
AI	Artificial Intelligence
URI	Uniform Resource Identifier
SVM	Support Vector Machine

Abbreviation	Full Form
WAF	Web Application Firewall
SQLi	Structured Query Language Injection
XSS	Cross-Site Scripting
KNN	K-Nearest Neighbors
ROC AUC	Receiver Operating Characteristic - Area Under Curve
CISC2010	[Name of Dataset: CISC2010 Web Attack Dataset]
HTTP	Hypertext Transfer Protocol
API	Application Programming Interface
EDA	Exploratory Data Analysis
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine

List Of Figures

Figure No.	Caption	Page No.
3.1	Average Special Character Count (Anomolous vs Valid)	14
4.1	Average Numeric-to-text Ratio (Anomolous vs Valid)	15
4.2	Average Entropy per Request Types (Anomalous vs Valid)	15
4.3	Average Length Comparison between Valid and Anomolous Requests	16
4.4	ROC Curve	18
4.5	Confusion Matrix	18
4.6	Secure Y Interface design	19
4.7	Secure Y Security Request Checker (Implementation)	20

List of Tables

Table Title	Page No.
Comparison of Signature-Based, ML-Based, and Hybrid WAFs	9
References for Literature Survey	10
Dataset Statistics	13
Model Training	17

1. INTRODUCTION

1.1 Introduction

Web applications have become integral to digital transformation across various domains, including finance, e-commerce, education, and healthcare. However, as web technologies evolve, so do the methods attackers use to exploit them. Web applications are exposed to a broad range of threats, such as SQL injection, Cross-Site Scripting (XSS), Cross-Site Request Forgery (CSRF), and Distributed Denial of Service (DDoS) attacks. These attacks exploit vulnerabilities at the application layer, which traditional firewalls and intrusion detection systems (IDS) often fail to detect as they primarily monitor network-level traffic. A **Web Application Firewall (WAF)** is a security system designed to monitor, filter, and block HTTP traffic to and from a web application. It provides a barrier between the web application and malicious actors, offering protection against application-layer attacks.

1.2 Motivation

As the digital landscape expands, cybersecurity threats are growing both in frequency and complexity. E-commerce websites, banking platforms, government portals, and healthcare systems are increasingly targeted by attackers. According to multiple security reports, a significant proportion of data breaches stem from vulnerabilities in web applications.

Motivated by the need for an efficient, accurate, and adaptive security mechanism, our project focuses on building a Hybrid WAF that combines the best of both worlds. Our goal is to create a system that provides:

- Real-time threat detection.
- Low false positive rates.
- Resistance to zero-day vulnerabilities.
- Easy integration with existing web infrastructure.

1.3 Problem Statement and Objectives

Problem Statement:

Web applications are increasingly vulnerable to sophisticated attacks that exploit both known and unknown vulnerabilities. Traditional firewalls are insufficient for application-layer protection, and standalone detection methods (either signature or anomaly-based) have inherent limitations. There is a pressing need for a unified system that can accurately detect known threats and adapt to emerging ones with minimal false positives and high efficiency.

Objectives:

To address the stated problem, the project aims to:

1. **Develop a Hybrid Web Application Firewall** that combines Signature-Based Detection (SBD) and Anomaly-Based Detection (ABD) using Machine Learning.
2. **Intercept and inspect HTTP traffic** to analyze incoming requests in real time.
3. **Use SBD to detect known attack signatures** such as SQLi, XSS, and other OWASP Top 10 threats.
4. **Implement ML models** (Random Forest and Isolation Forest) to analyze request behavior and detect anomalies.
5. **Ensure adaptive security** by enabling continuous learning and updating of the ML model to respond to zero-day threats.
6. **Log and classify malicious traffic** to allow future analysis and retraining of models.
7. **Reduce false positives and latency**, ensuring legitimate user requests are not interrupted.
8. **Deploy the WAF** in a middleware architecture that can be integrated with both local and cloud-based web services.

2. Literature Survey

2.1 Survey of Existing Systems

Introduction:
Web Application Firewalls are a critical component of modern cybersecurity frameworks, aiming to protect web applications from malicious traffic and known vulnerabilities. Traditional models use predefined patterns to detect threats, while newer systems leverage machine learning to identify unknown anomalies.

Traditional Systems (Signature-Based WAFs):
Signature-based WAFs rely on rule sets or databases of known attack patterns, such as SQL injections, XSS scripts, and command injection strings. These systems are fast, deterministic, and widely adopted in enterprise security architectures. However, they are reactive and ineffective against novel or obfuscated threats. Examples include ModSecurity and AWS WAF.

Limitations:

- **High Maintenance:** Frequent updates are required to maintain the effectiveness of signature databases.
- **Ineffectiveness Against Zero-Day Attacks:** New attack types not yet cataloged can easily bypass these systems.
- **Lack of Adaptability:** Signature-based systems cannot learn or evolve with changing threat landscapes.

Machine Learning Integration (Anomaly-Based WAFs):
Machine Learning-based WAFs address the adaptability issue by learning from past traffic behavior and identifying deviations. Algorithms such as Random Forest, Isolation Forest, SVM, and Deep Neural Networks are commonly used. These systems can detect zero-day and polymorphic attacks effectively. However, they require large datasets and tuning and are prone to false positives if not trained properly.

2.2 Comparison of Existing Systems

Authors	Paper Title	Publication	Key Findings	Gaps
Tekerek & Gemci	Development of a hybrid web application firewall to prevent web based attacks	IEEE International Conference on Application of Information and Communication Technologies (2014)	Pioneer hybrid WAF combining signatures and anomaly detection for HTTP requests	Limited feature engineering (URI length and frequency)
Torrano-Gimenez et al.	An Anomaly-Based Web Application Firewall	International Conference on Security and Cryptography (2009)	XML-based positive security model achieving 98% detection rate with <1% false positives	Pure anomaly-based approach without signature integration; no ML components
Calvo & Beltrán	An Adaptive Web Application Firewall	International Conference on Security and Cryptography (2022)	Context-aware risk-adaptive WAF configuration	Lacks detailed ML implementation and performance metrics

Islam & Hridi	Network Anomaly Detection Using LightGBM	International Telecommunication Networks Conference (2020)	LightGBM achieves 95.6% accuracy on network anomaly detection	Focuses on network-layer security rather than web apps
Wang & Anilkumar	W2R: Ensemble Anomaly Detection for WAF Security	Halmstad University Thesis (2023)	NLP-based feature extraction with ensemble models achieving 97.3% F1-score	Complex pipeline requiring multiple processing stages

3. Proposed System

3. Proposed System

3.1 Introduction to Proposed System

The proposed system is a Hybrid Web Application Firewall (WAF) designed to address the shortcomings of traditional signature-based firewalls and the limitations of pure machine learning-based systems. It combines Signature-Based Detection (SBD) for identifying known attack vectors and Anomaly-Based Detection (ABD) using Machine Learning (ML) to detect unknown, zero-day, or obfuscated threats in HTTP traffic.

The key objective of the system is to provide real-time protection to web applications by identifying and mitigating both known and unknown threats with high accuracy and minimal false positives. This is achieved through a layered approach where traffic is first analyzed for known attack signatures, and then passed through anomaly detection models if no signature is matched.

Key Features of the Proposed System:

1. **Layered Detection**: Combines SBD for quick elimination of common threats and ABD for deeper behavioral inspection.
2. **Machine Learning Integration**: Uses Random Forest and Isolation Forest algorithms for anomaly scoring.
3. **Real-time Interception and Analysis**: Intercepts HTTP requests through a proxy and processes them before they reach the application server.
4. **Logging and Auditing**: Maintains logs of malicious requests for model improvement and forensic investigations.
5. **Low False Positive Rate**: Filters malicious traffic with high accuracy while allowing genuine user access without disruption.
6. **Scalable Deployment**: It can be deployed on both local servers and cloud-based infrastructures.

3.2 Signatures

Signature-Based Detection (SBD) is one of the foundational methods for identifying malicious requests in a Web Application Firewall (WAF). It works by comparing incoming HTTP traffic against a predefined set of patterns, known as *signatures*, which are derived from known attack vectors. These signatures can be based on regular expressions, string matching, or rule-based heuristics. When a match is found, the request is immediately classified as malicious and is either blocked or logged for analysis.

Domain-Specific Signatures: E-Commerce Applications

In our hybrid WAF, we implemented domain-based signatures that were specifically tailored to the e-commerce sector, where common attack vectors often include:

- Credential stuffing
- Price manipulation
- Cart tampering
- Coupon code abuse
- SQL injection in product search and checkout queries
- Cross-Site Scripting (XSS) in user reviews and feedback forms

These domain-specific threats necessitated the creation of customized detection rules targeting the typical API endpoints and user input fields present in e-commerce platforms.

3.3 Dataset Overview

For the training and evaluation of the Anomaly-Based Detection (ABD) component in our Hybrid Web Application Firewall, we utilized the **CISC2010 Web Attack Dataset**, a comprehensive dataset designed to simulate real-world HTTP traffic with both benign and malicious request samples.

Each record in the dataset consists of HTTP-related metadata and parameters extracted from typical web traffic, making it highly suitable for training machine learning models focused on web application security.

Dataset Statistics

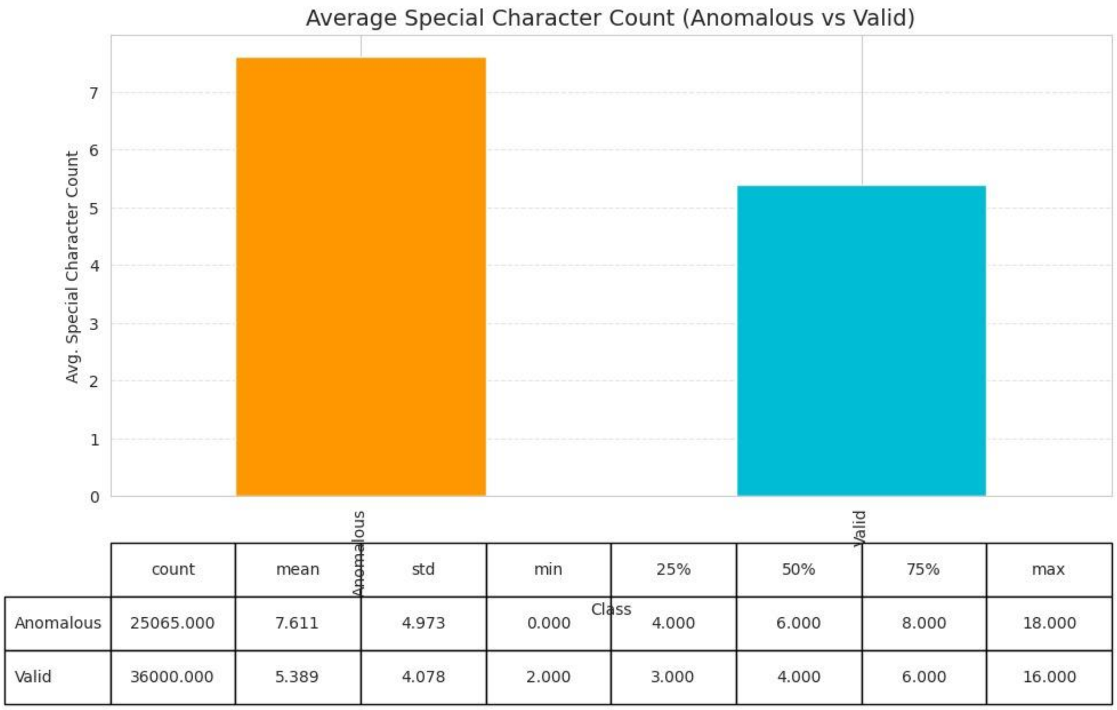
Property	Details
Total Records	~61,000+
Attack Types	SQLi, XSS, RFI, Command Injection, DoS
Valid Requests	~36,000+
Malicious Requests	~25,000+
Label Type	Binary (0 = Valid, 1 = Malicious)
Features Used	Method, URL, GET/POST payload

3.4 Feature Engineering

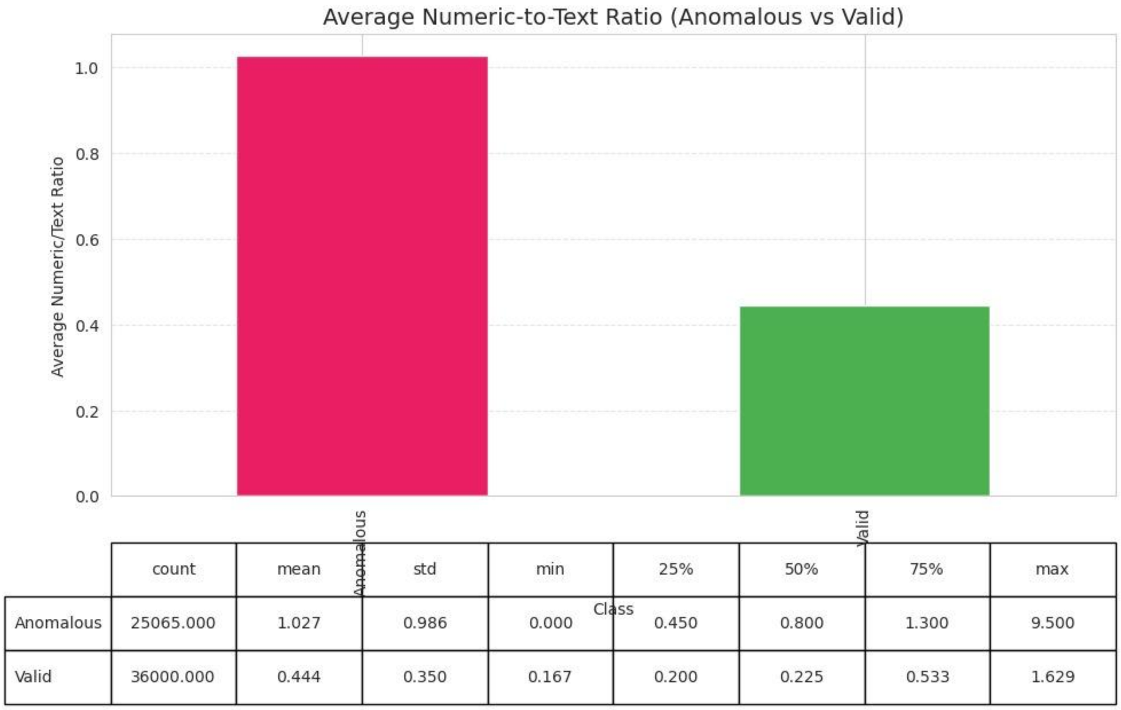
Feature engineering is a critical step in the machine learning pipeline that directly impacts the performance of the models. In the context of our Hybrid Web Application Firewall, effective feature extraction and transformation from raw HTTP request data were essential for training robust anomaly detection models capable of identifying subtle attack patterns.

Objective of Feature Engineering

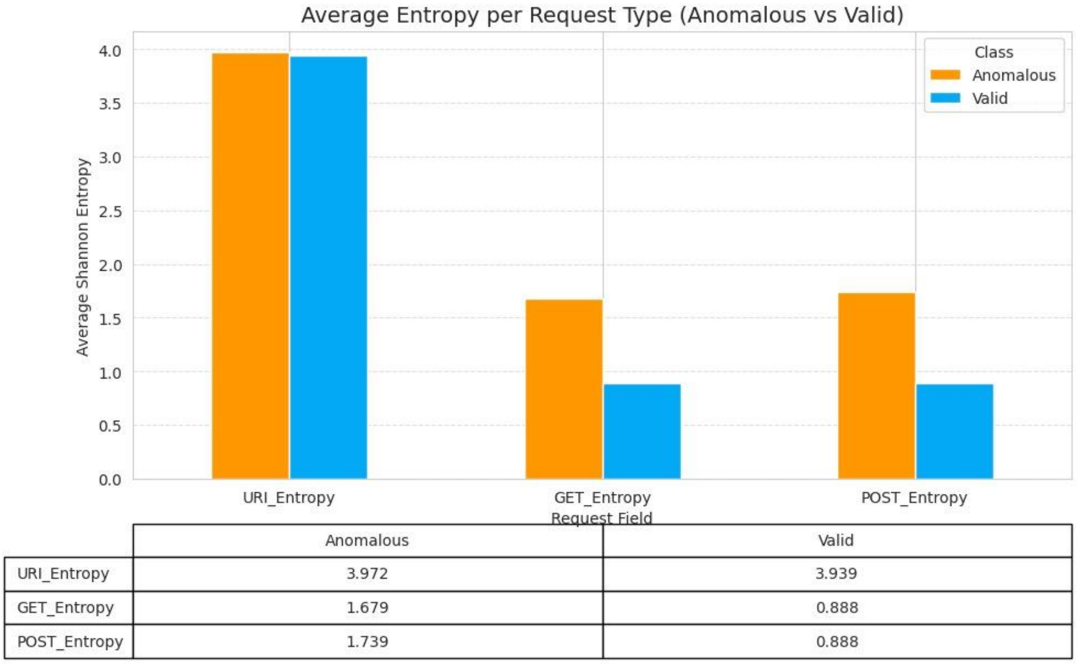
The goal of our feature engineering process was to transform raw HTTP request components into structured, meaningful numerical representations that can capture the behavioral characteristics of both benign and malicious traffic. These features served as the input vectors for supervised learning algorithms like Random Forest, XGBoost, and LightGBM.



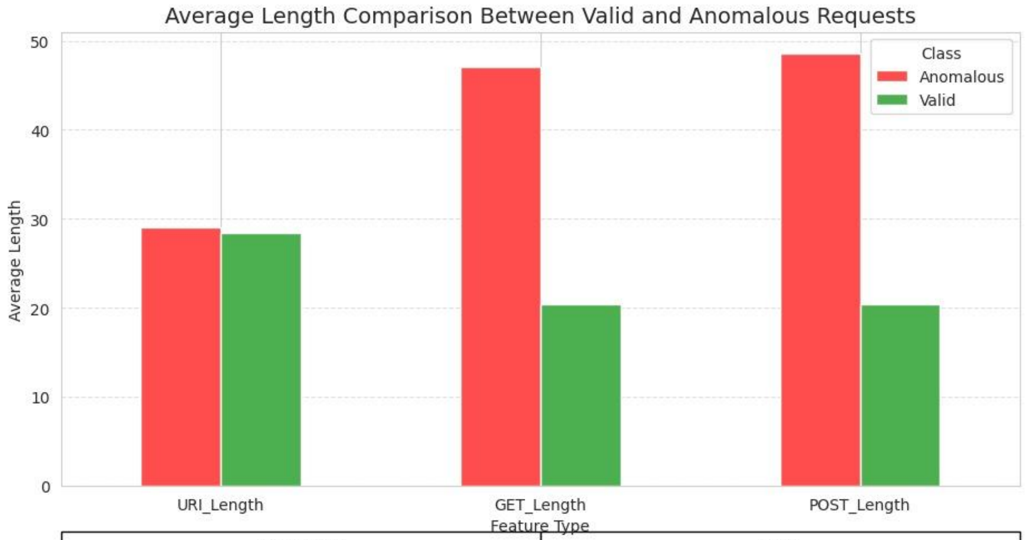
	count	mean	std	min	25%	50%	75%	max
Anomalous	25065.000	7.611	4.973	0.000	4.000	6.000	8.000	18.000
Valid	36000.000	5.389	4.078	2.000	3.000	4.000	6.000	16.000



	count	mean	std	min	25%	50%	75%	max
Anomalous	25065.000	1.027	0.986	0.000	0.450	0.800	1.300	9.500
Valid	36000.000	0.444	0.350	0.167	0.200	0.225	0.533	1.629



	Anomalous	Valid
URI_Entropy	3.972	3.939
GET_Entropy	1.679	0.888
POST_Entropy	1.739	0.888

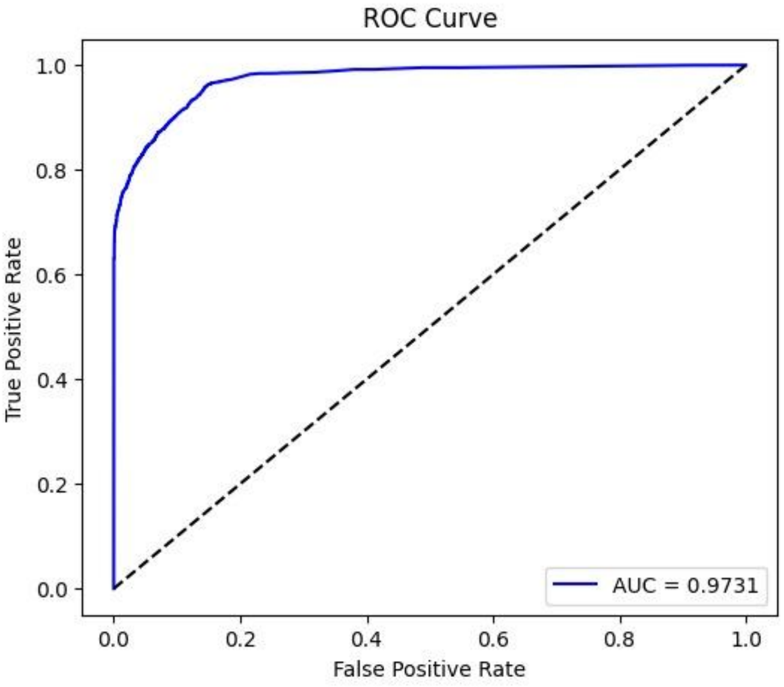


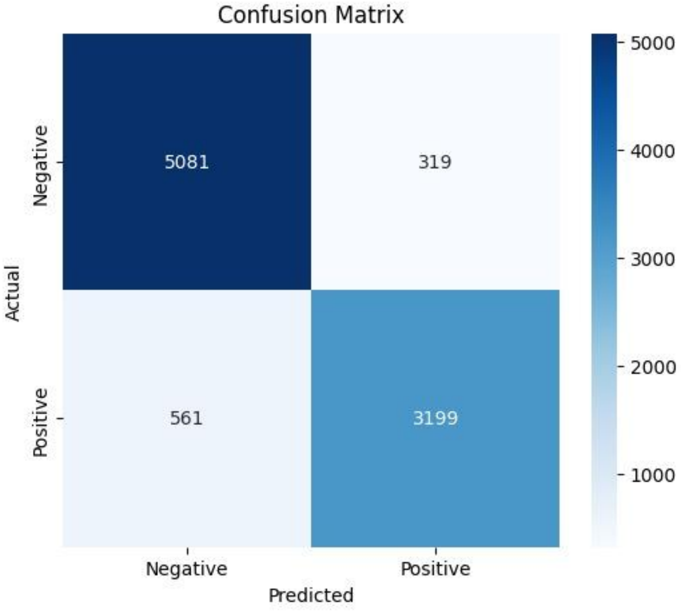
	Anomalous	Valid
URI_Length	29.09	28.42
GET_Length	47.13	20.38
POST_Length	48.58	20.38

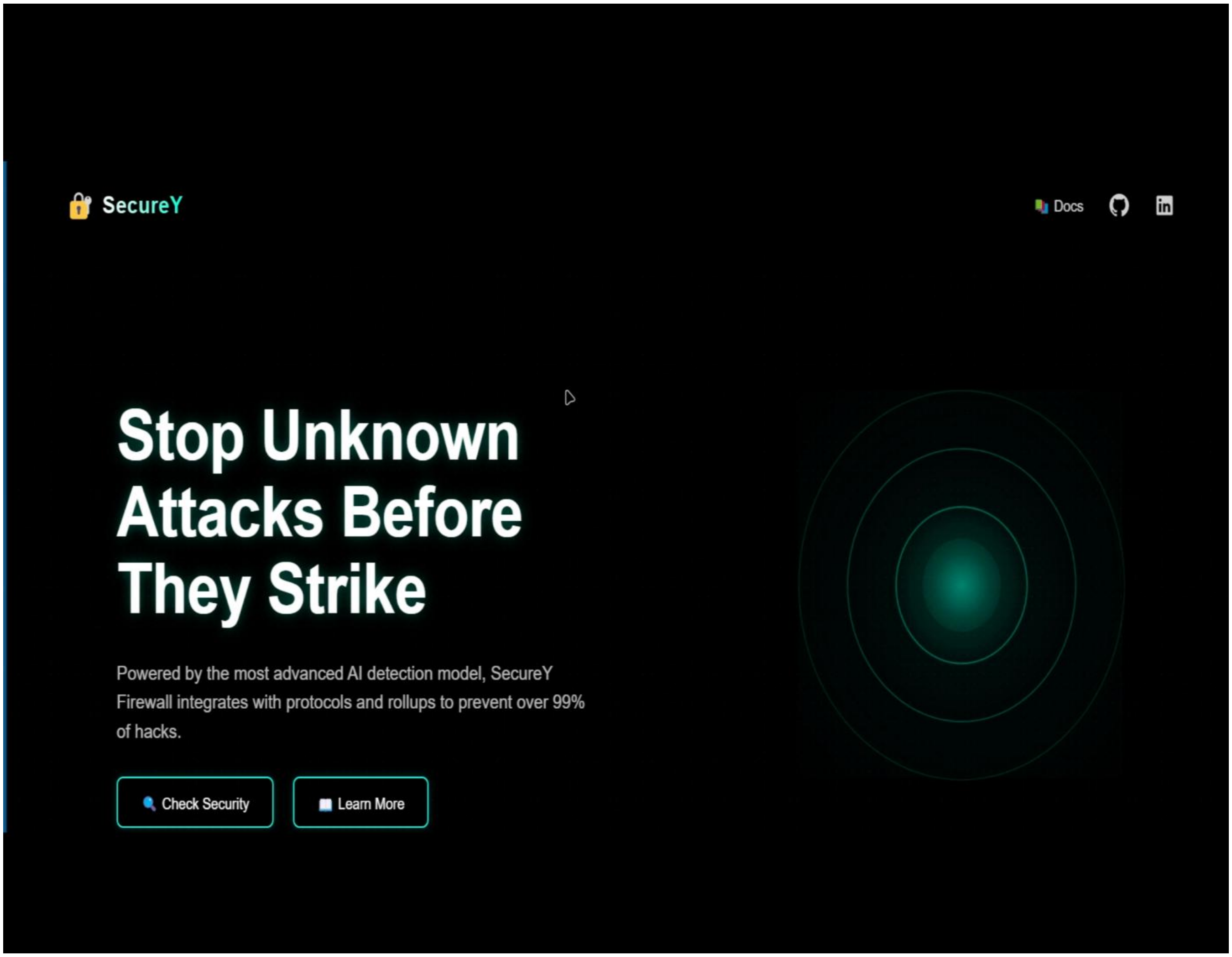
3.5 Model Training

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC
Random Forest	0.903930	0.903670	0.903753	0.903930	0.974469
XGBoost	0.903930	0.903525	0.903896	0.903930	0.974222
LightGBM	0.906223	0.905811	0.906237	0.906223	0.975094
Logistic Regression	0.731878	0.730125	0.729712	0.731878	0.766727
Support Vector Machine (SVM)	0.735808	0.727362	0.736227	0.735808	N/A
K-Nearest Neighbors (KNN)	0.897380	0.897164	0.897160	0.897380	0.959572

4. Experimentation and Results

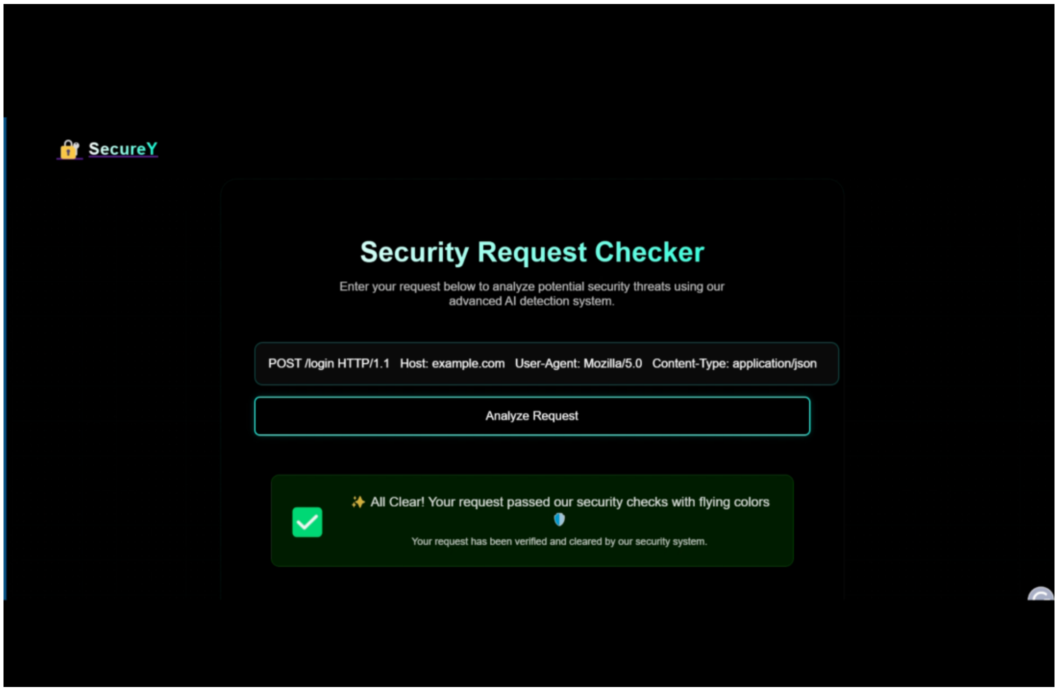






Stop Unknown Attacks Before They Strike

Powered by the most advanced AI detection model, SecureY Firewall integrates with protocols and rollups to prevent over 99% of hacks.



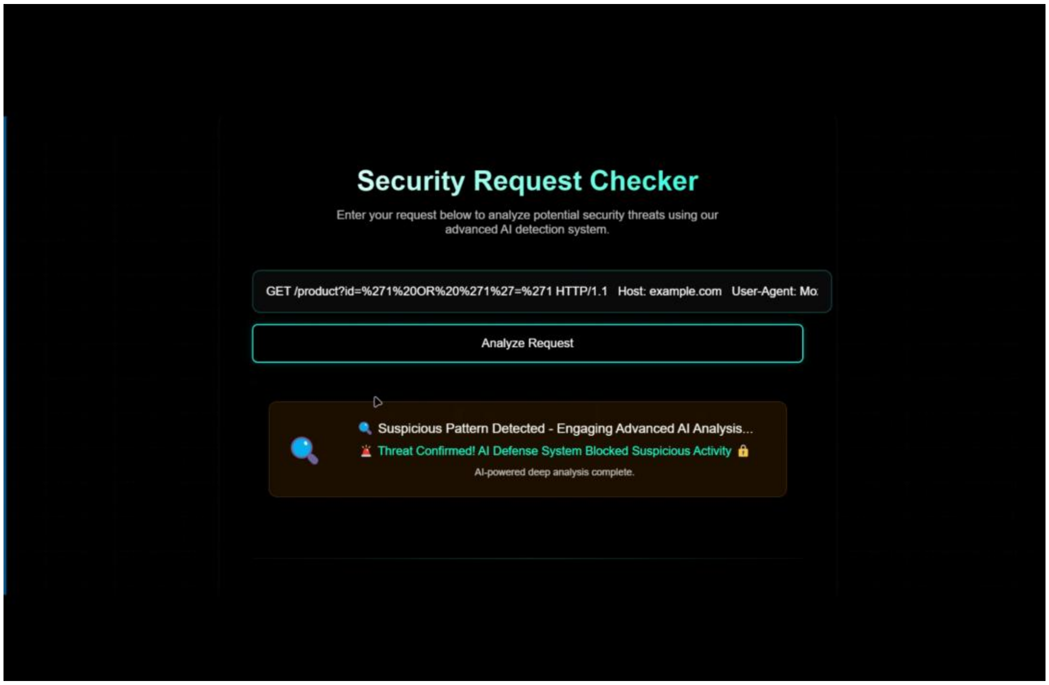
Security Request Checker

Enter your request below to analyze potential security threats using our advanced AI detection system.

POST /login HTTP/1.1 Host: example.com User-Agent: Mozilla/5.0 Content-Type: application/json

✓ ✨ All Clear! Your request passed our security checks with flying colors 🛡️

Your request has been verified and cleared by our security system.



Security Request Checker

Enter your request below to analyze potential security threats using our advanced AI detection system.

GET /product?id=%271%20OR%20%271%27=%271 HTTP/1.1 Host: example.com User-Agent: Mo:

🔍 Suspicious Pattern Detected - Engaging Advanced AI Analysis...

🚨 Threat Confirmed! AI Defense System Blocked Suspicious Activity 🔒

AI-powered deep analysis complete.

5. Conclusion and Future Work

The increasing sophistication and frequency of web-based attacks underscore the need for advanced and adaptive security mechanisms. Traditional Web Application Firewalls (WAFs), which primarily rely on signature-based detection, are no longer sufficient in the face of evolving attack patterns and zero-day vulnerabilities. Our proposed Hybrid Web Application Firewall bridges this gap by integrating Signature-Based Detection (SBD) with Anomaly-Based Detection (ABD) powered by Machine Learning (ML).

The layered design of the firewall enables it to:

- Detect and block known threats quickly through signature matching.
- Identified new and obfuscated threats by analyzing behavioral anomalies using Random Forest and Isolation Forest algorithms.

While the current system achieves its foundational goals, there is considerable scope for enhancement and extension. The following future directions are proposed:

1. **Model Optimization and Expansion:**
    - Integrate additional ML algorithms such as Support Vector Machines (SVM), XGBoost, or Neural Networks for improved anomaly classification.
    - Introduce unsupervised learning methods to reduce the dependence on labeled datasets.
2. **Real-Time Model Updating:**
    - Implement online learning mechanisms to update models in real time based on new incoming traffic patterns and attack signatures.
3. **Integration with Threat Intelligence Feeds:**
    - Enable dynamic updates of the signature database using third-party threat intelligence sources to enhance SBD effectiveness.
4. **Multi-layered Architecture with Cloud Support:**
    - Deploy the WAF as a microservice in a containerized environment (e.g., Docker + Kubernetes) to support scalability and high availability.

6. References

1. A. Tekerek, C. Gemci and O. F. Bay, *"Development of a hybrid web application firewall to prevent web based attacks,"*2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan.

2. S. Dhote, A. Magdum, S. Singh and D. Raigar, *"ML based Web Application Firewall for Signature and Anomaly Detection Using Feature Extraction,"* 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024.

3. Leka, Elva & Aliti, Admirim. (2024). *Web Application Firewall for Detecting and Mitigation of Based DDoS Attacks Using Machine Learning and Blockchain.*

4. Román, Jesús & Pérez-Delgado, M. & Viñuela, Marcos & Vega-Hernández, María-Concepción. (2023). *Artificial Intelligence Web Application Firewall for advanced detection of web injection attacks.* Expert Systems.

5. Applebaum, Simon & Gaber, Tarek & Ahmed, Ali. (2021). *Signature-based and Machine-Learning-based Web Application Firewalls: A Short Survey*. Procedia Computer Science.

6. Torrano-Gimenez, Carmen & Perez-villegas, Alejandro & Alvarez, Gonzalo. (2009). An Anomaly-based Web Application Firewall. 23-28

7. Joshi, Anamika, and V. Geetha. "SQL Injection detection using machine learning." *2014 international conference on control, instrumentation, communication and computational technologies (ICCICCT)*. IEEE, 2014.

8. Makiou, Abdelhamid, Youcef Begriche, and Ahmed Serhrouchni. "Improving Web Application Firewalls to detect advanced SQL injection attacks." *2014 10th international conference on information assurance and security*. IEEE, 2014.

9. Komiya, Ryohei, Incheon Paik, and Masayuki Hisada. "Classification of malicious web code by machine learning." *2011 3rd International Conference on Awareness Science and Technology (iCAST)*. IEEE, 2021.

10. Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016.