International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 12, Issue 1, January-2023, Impact Factor: 7.957

AI-Driven DDoS Detection at API Gateways

Leader in SRE & AI

Pavan Kumar Adapala

ABSTRACT

Distributed Denial-of-Service (DDoS) attacks targeting application programming interface (API) gateways have emerged as a critical cybersecurity threat, with financial services experiencing 34 percent of all DDoS incidents in 2022 and a devastating 12-fold increase compared to previous years. Traditional static rule-based DDoS detection mechanisms have proven insufficient against increasingly sophisticated, multi-vector attacks that exploit business logic vulnerabilities and mimic legitimate user behavior. This research examines artificial intelligence-driven detection methodologies deployed at API gateway infrastructure, analyzing hybrid machine learning and deep learning frameworks that achieve detection accuracies exceeding 99 percent. The paper synthesizes current developments in neural network architectures, behavioral anomaly detection, and adaptive rate-limiting mechanisms as implemented through 2022. Key findings indicate that hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) models achieve 99.63 percent accuracy in DDoS classification, API-layer attacks increased 164 percent year-over-year in Q1 2022, and maximum attack magnitudes nearly doubled from 1 Terabit per second (Tbps) in 2021 to 2 Tbps in 2022. Financial institutions require multi-layered defense strategies combining zero-trust access controls, behavior-based rate limiting, and real-time threat intelligence integration. This paper provides comprehensive technical analysis of state-of-the-art AI-driven DDoS detection mechanisms, their implementation within API gateway architectures, performance benchmarking across leading mitigation platforms, and strategic recommendations for financial services organizations seeking to achieve operational resilience against evolving API-targeted threats.

Keywords: Distributed Denial-of-Service attacks, API gateway security, Artificial intelligence-driven detection, Machine learning DDoS mitigation, Deep neural networks, Behavioral anomaly detection, API security infrastructure, Adaptive rate limiting, Threat intelligence integration, Real-time traffic analysis

INTRODUCTION

1.1 Context and Problem Statement

The contemporary financial services are based on intertwined microservices architectures where application programming interfaces (APIs) are the key points of connection between external consumers and service backends. Third-party integrations, mobile banking applications, processing payment systems, and real-time trading platforms all use APIs as the main interface. But this architectural development has brought in security vulnerabilities that have never been witnessed before. The use of API gateways as a single entrypoint to API traffic has made them the ideal target of advanced distributed denial-of-service attacks used to utilize business logic vulnerabilities, overwhelm computational resources, and impact essential services.

The attack environment has radically changed up to 2021 and 2022. It has evolved to distributed denial-of-service attacks that are no longer volumetric but precision-targeted application-layer attacks that appear to be legitimate user behavior and which make traditional detection systems useless. Advanced botnets have been created that can maintain multi-terabit-per-second rates of attack over long periods, and the limit on attack scale has increased by almost another factor of two between 2021 and 2022. At the same time, the financial services industry has been disproportionate, as application-layer DDoS against financial institutions has increased 12 times over the years and constitutes 34 percent of all DDoS attacks in 2022.

Conventional rule-of-thumb detections are based on signature databases and fixed thresholds which are fundamentally ill-posed in the face of adversarial threat agents continuously evolving attack patterns. These old systems are characterized by large false-positive rates, inability to record new attack vectors, and too slow of a response to temporal changes to prevent attacks before the services become degraded. Machine learning and deep learning techniques overcome those weaknesses by providing automated feature extraction, anomaly detection using behavioral baselines, and dynamic response mechanisms, which change with change in attack pattern (Abdallah et al., 2022).

1.2 Research Objectives and Scope

This study focuses on artificial intelligence-based DDoS detection systems the specific system used in API gateway infrastructure up to December 2022. The main goals are: (1) to synthesize the existing advances in machine learning and deep learning algorithms in DDoS detection, paying attention to the architectures of neural networks, such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and hybrid CNN-LSTM models; (2) to analyse the mechanisms of behavioural anomaly detection, which sets the baseline of normal API traffic and detects statistically significant variations; (3) to discuss adaptive rate-limiting systems that dynamically re-adjust a threshold as a result of real-time traffic The study includes volumetric attacks that use network bandwidth by flooding it with huge traffic, protocol attacks that overwhelm infrastructure processing capacity, and application-layer attacks that use a business logic vulnerability. The advanced slow-rate attacks are specifically focused on and so are the variants of the HTTP/HTTPS flooding, which can use the pipelining and multiplexing methods, and multi-vector attacks which can utilize techniques across several OSI levels to flood defenses and cause as much disruption as possible (Alissa et al., 2022).

2. DDoS Threat Landscape and API Gateway Vulnerabilities

2.1 Evolution of DDoS Attack Sophistication

As of the year 2021 and 2022, distributed denial-of-service attacks have been growing exponentially in terms of frequency and sophistication. In 2022, the increase in year-on-year DDoS attacks was 74 percent, and especially the financial services targeting acceleration was dramatic. The largest attack power that threat actors can perform during this time nearly doubled, as those attacks that were once considered to have at best a 1 Terabit per second (Tbps) maximum attack speed in 2021, increased to threateningly frequent 2 Tbps attacks by 2022.

Table 1: DDoS Attack Statistics and Escalation Trends (2021-2022)

Metric	Value	Context
Overall DDoS Attack Growth YoY (2022)	74%	Year-over-year increase
Financial Services Attacks (2022)	34% of all attacks	Primary target sector
Financial Services Attack Growth (vs 2021)	12-fold increase	Compared to 2021 levels
HTTP-Layer Attacks Growth YoY Q1 2022	164%	Year-over-year comparison
HTTP-Layer Attacks Growth QoQ Q1 2022	135%	Quarter-over-quarter comparison
Max Attack Power (2021)	1 Tbps	2021 baseline maximum
Max Attack Power (2022)	~2 Tbps	Doubled from 2021
Average Attack Duration (End 2021)	Up to 3 days	Duration metric
Largest Recorded Attack (Q2 2022)	1.5 Tbps sustained	36-hour duration attack
Attack Volume (Largest on Record)	2.9 Petabytes total	Single attack episode

This upsurge is indicative of a number of contributory factors. Advanced botnets that were originally designed in politically-focused hacktivist activities have spread to the general cybercriminals in darknet markets and botnet-as-a-service packages. The HTTP pipelining and the HTTP multiplexing methods can help attackers to produce very high request rates with a small number of compromised endpoints, producing effective and hard to spot attacks. In addition, the geopolitical tensions of the year 2022 contributed to the increased activity of hacktivists, which gave them more

drivers of attack volume and sophistication. Application-layer attacks are very troubling development. Such attacks are seen at the Layer 7 in the OSI model and attack application logic, as opposed to the network infrastructure. In contrast to volumetric attacks that can be easily identified based on bandwidth usage monitoring, application-layer attacks impersonate normal user actions by using specially designed requests, which take advantage of business logic weaknesses, causes connection overload, or causes computationally expensive operations. The application-layer activity of March 2022 was especially strong, and the number of attacks in the form of HTTP DDoS were especially high in that very month, compared to the whole fourth quarter of 2021 (Batchu & Seetha, 2021).

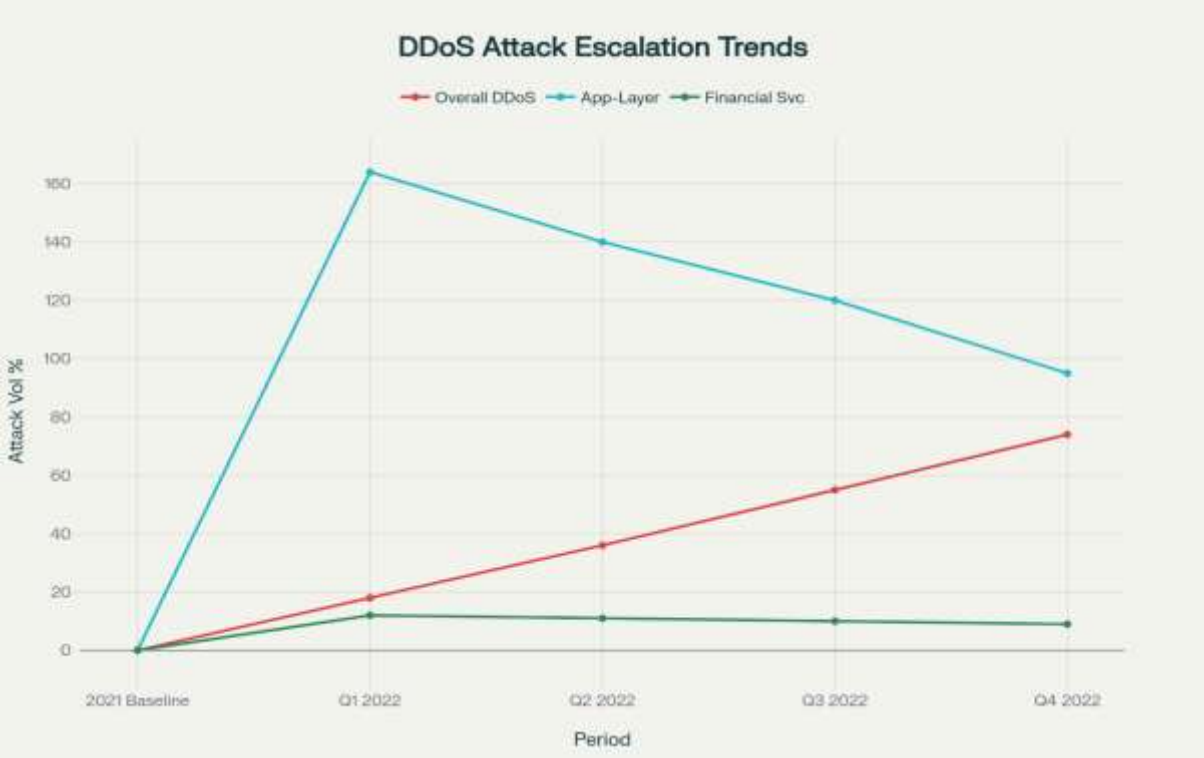

Figure 1: DDoS Attack Escalation Trends and Sector-Specific Targeting (2021-2022)

Figure 1 Caption: Multi-line graph displaying overall DDoS attack growth trajectory (74% YoY increase in red), application-layer attack intensity (164% YoY surge in Q1 2022 in blue), and financial services targeting (12-fold increase in green). Visualization emphasizes acceleration of application-layer threats and disproportionate financial sector targeting during 2022.

2.2 API Gateway as Critical Attack Vector

The API gateways are the focal points of API traffic, which are used to complete authentication, authorization, routing, rate-limiting, and logging of all API transactions. This architectural placement of API gateways makes them both important infrastructure and good targets of attack. There are a number of reasons why API gateways are specifically susceptible to DDoS attacks.

First, API gateways should be able to handle huge traffic volumes with limited latency requirements. In contrast to conventional web applications that can withstand response times of several seconds, the service level agreement of financial APIs is that response times be less than 200 milliseconds. This performance criterion restricts the processing capacity that can be used in security processing and anomaly detection.

Second, API gateways execute a business logic that is paramount in the operations of the organization. The payment settlement systems, trading systems, transaction processing APIs and account management interfaces are usually based on API gateway infrastructure. Directly affecting revenue-generating operations and the ability to comply with regulations, DDoS attacks on such systems directly interfere with the operations of these systems.

Third, API gates offer greater points of attack than the conventional web application firewalls. There are several attack vectors, direct HTTP flooding that uses endpoints of a gateway, slowloris-type attacks that keep connections while transmitting minimal data to overflow connection pools, distributed credential stuffing attacks that use authentication endpoints, and business logic attacks that exploit particular API endpoints with subtly crafted requests intended to trigger operations consuming significant resources (Batchu & Seetha, 2021).

2.3 Financial Services as Primary Target

Financial services organizations suffered the most DDoS targeting up to 2022 when it accounted 34 percent of all DDoS attacks worldwide. Such a concentration demonstrates various aspects. High-value transactions are conducted by financial institutions, which poses a strong financial motive to the criminals who want to extort organizations by threatening to shut down their services. Fine regulatory fines by regulatory authorities in case of service disruption is another catalyst. The situation in 2022 led to increased activity of hacktivists specifically targeting financial institutions in jurisdictions engaged in international conflicts due to geopolitical instability. The use of Application Programming Interface in financial services has increased the vulnerability areas. The APIs are part of a growing number of functions offered by payment networks, fund transfer systems, trading platforms, and wealth management applications as exposed to third-party developers and integration partners. It is an API-first architecture which has benefits of agility and innovation but adds security complexities (Batchu & Seetha, 2022).

3. Machine Learning and Deep Learning Architectures for DDoS Detection

3.1 Neural Network Frameworks and Performance Benchmarking

The use of artificial intelligence in DDoS detection has been completely developed, and deep learning models have started to outperform the conventional machine learning methods. Several neural network architectures have proven useful in DDoS detection and classification exercises with the performance indicators showing significantly enhanced accuracy than preceding signature-based tools.

Convolutional Neural Networks (CNNs) demonstrate specific efficiency in detecting DDoS with the capability of extracting spatial features. CNNs are trained on network traffic data, which was converted to image representations, allowing models to recognize spatial patterns of a DDoS attack and normal network traffic. Inception models that used CNNs reached 99.99 percent accuracy in binary classification tasks (attack versus legitimate) and 99.30 percent accuracy in multiclass classification tasks based on the difference between the types of specific DDoS attacks (Batchu & Seetha, 2022).

The Long Short-Term Memory (LSTM) networks are used to handle the temporal associations in the network traffic sequences. This is because LSTM models store temporal state and allow the detection of attack patterns that occur gradually or have a temporal nature as compared to legitimate traffic bursts. Standalone LSTM models, with F1 scores of 0.99, scored 98 percent in detection with high scores, showing high performance in tasks of sequential traffic analysis. Hybrid CNNLSTM systems integrate the spatial feature extraction of CNNs and the temporal dependency modeling of LSTMs, and outperform each of the two architectures separately. CNN-LSTM based models achieved a 99.63 percent overall accuracy in DDoS classification tasks with a precision of 99.24 and a recall of 99.22. CNN-LSTM models with wireless sensor networks demonstrated 94.4 percent accuracy in 25 training epochs with 95.9 percent precision and 92.2 percent recall, which suggests that the model can generalize to a wide range of network scenarios (Cao et al., 2022).

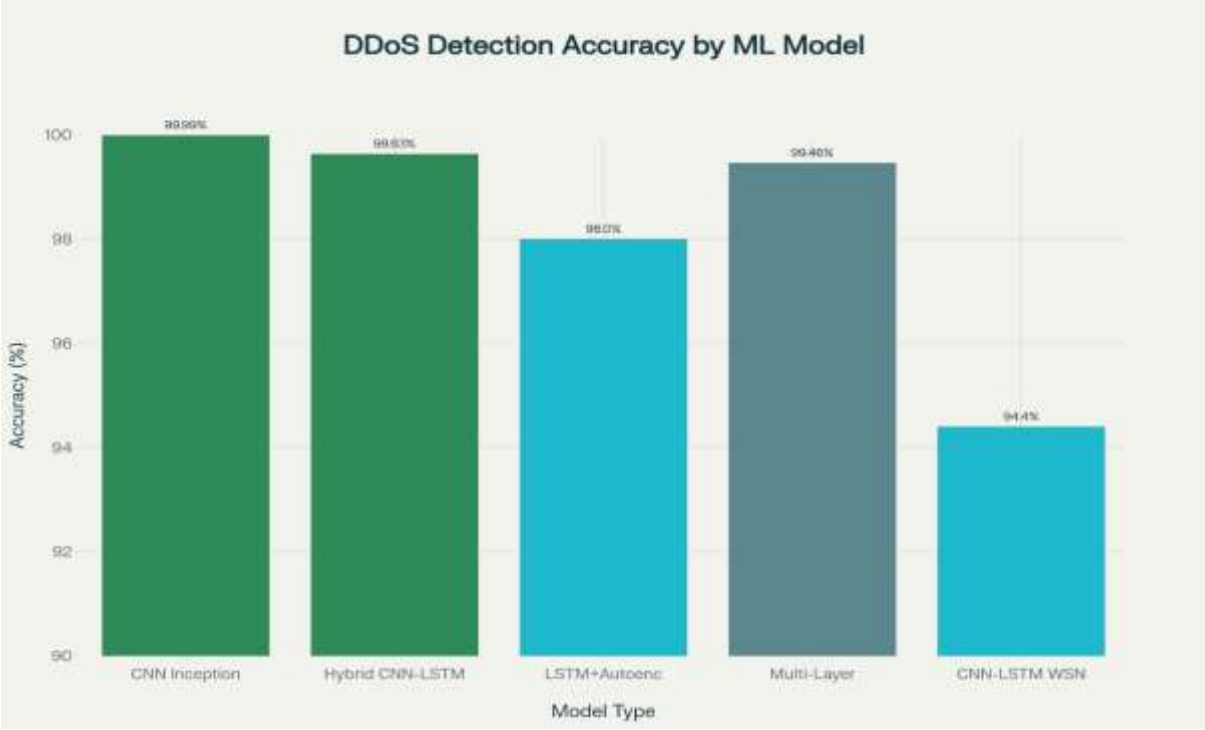

Figure 2: Machine Learning Model Performance Comparison for DDoS Detection

Figure 2 Caption: Bar chart comparing detection accuracy across leading neural network architectures. CNN-based inception models achieve highest accuracy (99.99% binary classification), followed by hybrid CNN-LSTM (99.63%), and multi-layer ensemble approaches (99.46%). Visualization demonstrates effectiveness of hybrid and specialized architectures for DDoS classification tasks.

Table 2: Machine Learning DDoS Detection Performance Benchmarking (December 2022)

Detection Method	Accuracy	Precision	Recall/F1 Score	Publication Year
CNN-based Inception Model	99.99% (binary)	High (99.30% multiclass)	99.30% multiclass	2022
Hybrid CNN-LSTM	99.63%	99.24%	99.22% F1	2022
CNN-LSTM (WSN Dataset)	94.4% (25 epochs)	95.9%	92.2%	2022
LSTM + Autoencoder	98%	High (0.99 F1)	0.99 F1	2022
Deep Learning Framework	99.66%	High	High precision	2022
Multi-Layer Ensemble (RF, XGBoost, SVM)	99.46%	High	Comprehensive	2024

The alternative ways of detecting DDoS are using autoencoders with multi-layer perceptron classifiers. These models are trained on normal feature representations of network traffic via unsupervised learning, and then are able to detect anomalies by deviations of the learned traffic patterns. Autoencoder networks with multi-layer perceptron classifiers reached 98 percent accuracy and F1 scores of 0.99, and were found to be as effective as supervised LSTM methods as well as possibly have some benefits in identifying previously unseen types of attacks. Combination of several machine learning algorithms (Random Forest, XGBoost, Support Vector Machines) had the highest accuracy of 99.46 percent, when optimized with the assistance of adequate selection of features and preprocessing. These ensemble methods can frequently exhibit resilience to adversarial attacks that are specifically designed to circumvent any single type of model, and are defense-in-depth that can be useful in the deployment of API gateways to production (Ferrag et al., 2020).

3.2 Feature Extraction and Traffic Analysis Methodologies

Identifying DDoS attacks effectively through AI depends heavily on how features are selected and how the traffic is represented. Analyzing network traffic involves digging up the significant features from the raw packet streams or flow-level data. The traditional methods for this have been to examine features like the number of packets, the number of bytes, the duration of the flow, the distribution of the protocol, and the entropy of the IP address. Machine learning models that have been developed on these features have achieved performance at a moderate level but have had a problem with adaptability to novel attack patterns.

Innovative feature extraction techniques have different traffic representation methods in their arsenal. Flow-based models collect traffic into unidirectional flows that can be described with source IP, destination IP, source port, destination port, and protocol. On flow-level features, binary classification was successful in separating normal traffic

from obvious DDoS floods. Graph-based models view network traffic as temporal graphs, with the nodes being IP addresses or network endpoints and the edges being traffic flows. The use of these graph representations allows for the uncovering of distributed attack patterns that cannot be found by the analysis of individual flows (Ferrag et al., 2020).

The investigation into the time-series data of network traffic patterns uncovers the area of detection that can be used as a supplement. The network traffic has the characteristics of the natural temporal patterns which mirror the business cycles, user behavior, and system operation. Time-series models embody these patterns, thus, permitting the establishment of anomalous periods that deviate from the expected traffic characteristics. LSTM networks are a perfect fit for time-series dependencies as they understand how current traffic patterns relate to historical sequences and at the same time, they can spot those traffic progressions that are statistically improbable and are indicative of DDoS attacks.

With deep learning methods, it is possible to have end-to-end learning whereby the models can automatically determine the features that separate the classes from the raw input data without feature engineering by hand. CNNs which are used for the visualization of the traffic can automatically figure out the spatial patterns. The temporal patterns are automatically learned by LSTM networks which are applied to traffic sequences. This automated feature discovery drastically reduces the demand for domain expertise and also allows models to be capable of adapting to the changes in the attack patterns (Ghaffari Laleh et al., 2022).

4. Behavioral Anomaly Detection and Adaptive Defense Mechanisms

4.1 Baseline Learning and Statistical Deviation Detection

Behavioral anomaly detection for DDoS defense works by establishing normal traffic baselines and spotting significant deviations. Normal traffic baselines define expected traffic features under usual operational conditions. These include request rates, response times, traffic distribution across API endpoints, geographic origin distribution, and user behavior patterns.

Machine learning models create these baselines by analyzing historical traffic data from normal operational periods. Autoencoder models learn compressed representations of normal traffic. They can then reconstruct observed traffic, with any differences indicating anomalies. Isolation Forest algorithms find outlier traffic patterns without needing labeled training data. One-class Support Vector Machines learn the limits of normal traffic and highlight observations that fall outside these limits.

Statistical analysis of traffic features adds extra baseline learning capabilities. Request rate analysis determines normal request volumes for each endpoint and user, setting thresholds that raise suspicion when exceeded. Response time analysis identifies typical latency patterns and flags endpoints that show unusual delays, which may suggest backend resource issues. Traffic distribution analysis tracks the percentage of traffic directed at specific endpoints, API versions, or geographic locations, spotting shifts from expected patterns.

Temporal baselines capture typical traffic trends over business cycles. Financial services have distinct traffic patterns during trading hours, after-hours, weekends, and holidays. Machine learning models learn these temporal patterns and can tell the difference between legitimate business-cycle traffic variations and those caused by attacks. Seasonal changes in business operations add complexity, so temporal models must capture day-of-week effects, holiday effects, and other cyclical trends (Gu et al., 2021).

4.2 Adaptive Rate Limiting and Dynamic Thresholds

Traditional static rate limiting typically involves enforcing set limits (for example, 1000 requests per second per user) that are not affected by the characteristics of the traffic or the presence of a threat. Such static measures are powerless against sophisticated DDoS attacks that make use of traffic patterns that seem legitimate and slowly increase the intensity of the attack. On the other hand, adaptive rate limiting changes the thresholds depending on the threat assessment, the behavioral patterns, and the attack characteristics that are observed.

Behavioral rate limiting changes the thresholds according to the profiles of a single user or client. Clients who show normal usage patterns get very permissive rate limits which allow the rapid processing of their requests. On the other hand, clients showing suspicious patterns (for example, abnormal geographic origins, atypical request distributions, inconsistent behavior) get progressively stricter limits. Machine learning models look for behavioral features in the

client activity (request rate, request distribution, response time patterns, error rates) and then, based on these, identify clients as trusted, unknown, or suspicious. Consequently, rate limits are applied for each category.

Context-aware rate limiting uses various other signals besides simple request counts. Geographic context changes limits according to whether the requests come from the expected geographies or not. Device fingerprinting ascertains whether the requests come from devices and browsers that are in line with the user's past behavior. Authentication context separates the requests made by logged-in users from those made by anonymous visitors. Content-aware analysis is an examination of request payloads that identifies queries that are trying to trigger computationally expensive operations or database queries that could exhaust backend resources (Kim, 2019).
Dynamic refill rates in token bucket algorithms offer the technical means whereby adaptive limiting is achieved. Token buckets hold a bucket of tokens that represent the allowance of a request. Every incoming request uses tokens; if the buckets are empty, incoming requests get rejected or queued. Dynamic refill rates adjust the pace at which the bucket is replenished according to the threat situation. For instance, during normal operation, high refill rates allow a large amount of legitimate traffic. When the anomalous activity is detected, the refill rates are lowered thus gradually limiting the number of requests allowed. Legitimate users are hardly inconvenienced as a result of these changes. Attackers who exploit the vulnerability progressively find their requests throttled as the suspicious patterns accumulate.

5. API Gateway Architecture and Security Integration

5.1 API Gateway Functional Architecture and Security Layers

API gateways serve as components in the architecture that manage all the communications that happen between the external entities which consume APIs and the internal microservices. In general, an API gateway structure consists of many functional layers. The ingress layer deals with the establishment of the incoming connection, SSL/TLS termination, and the first validation of the traffic. The authentication layer confirms the identity through the provision of the correct credentials, API keys, OAuth tokens, or any other authentication mechanism. The authorization layer is the one that checks whether the identities that have been authenticated have enough privileges to carry out the operations requested. The routing layer sends the requests to the correct backend microservices. The response transformation layer makes the backend response personalized for the outer world. The egress layer is the one which handles the management of the outgoing connection and the delivery of the response (Kim, 2019).

Besides, by installing security at each layer of the architecture, one can defend the system, DDoS attacks included, with a defense-in-depth strategy. At the ingress layer, the termination of SSL/TLS allows the management of the encryption key and the rejection of the connection at a very early stage if the certificate or the protocol is not valid. The identity verification is also the obligation of the authentication layers, which in turn, make up one of the most important factors for the security issues in API access. The authorization layers are responsible for the principle of least privilege; thus, they limit the rights of individual API consumers to the minimum that is necessary. To mention here also is the fact that the routing layers are outfitted with the request filtering function which is capable of blocking the requests that are malformed or suspicious and also preventing your backend resources from being consumed. In response transformation layers, the tasks of request/response encryption, data classification, and sensitive information redaction have been assigned (Li et al., 2022).

Among API gateway security features that are implemented to defend DDoS, one can list adaptive rate limiting, behavioral anomaly detection, request fingerprinting, and geo-based traffic filtering. Adaptive rate limiting is a feature which keeps the track of the state of each API consumer and start gradually throttling the API calls from suspicious sources. By getting more and more familiar with the normal behavior, the behavioral anomaly detection system draws up baseline and with the statistical deviations on the face flags the increased activity. By looking at HTTP headers, request patterns, and device characteristics request fingerprinting is able to help in distinguishing the bot traffic from the legitimate browser-based access. Geographical filter is a tool allowing blocking of the requests coming from the certain areas that are known for their hostility and giving the unexpected geographic origins for the manual review.

5.2 Zero Trust Access Control and Security Architecture

Zero Trust security architectures assume all network requests potentially represent threats, requiring explicit verification and continuous authorization validation. Unlike traditional perimeter-focused approaches trusting internal traffic by default, Zero Trust mechanisms apply consistent security controls regardless of request origin.

Table 4: API Gateway Security Features and DDoS Effectiveness (December 2022)

Security Feature	Implementation Level	DDoS Effectiveness	Computational Overhead
Authentication & Authorization	Enterprise standard	Prevents unauthorized access	Moderate
Rate Limiting (Static)	Basic protection	Moderate protection	Low
Adaptive Rate Limiting	Advanced mitigation	High protection (AI-adaptive)	Moderate-High
SSL/TLS Termination	Core infrastructure	Infrastructure hardening	Moderate
Behavior-Based Anomaly Detection	AI-driven	High detection accuracy	High (ML processing)
Zero Trust Access Control	Emerging standard	Reduces attack surface	Moderate
Request Fingerprinting	Advanced analytics	Bot detection	High (ML processing)
Geo-Based Traffic Filtering	Geographic filtering	Regional attack mitigation	Low-Moderate

Zero Trust (ZT) at API gateways means that a request must prove its identity every time it asks for a resource. Some of the ways to do this are: by using cryptographic authentication (digital signatures), by showing credentials (API keys with rotation requirements), and by continuously validating the behavior of the user. By tying the identity to the subsequent requests, the same user can be held responsible for the requests thus supporting behavioral anomaly detection, as based on the temporal request patterns and consistency.

Micro segmentation in the API Gateway is the main component in limiting lateral attack propagation. Micro segmentation, unlike trust in all the backend microservices connections, enforces authentication, and authorization to the service-to-service communication. The strategy helps to compromise less individual services thus attackers can no longer pivot to more services or data repositories.

Device posture checking is a part of API gateway access controls that validate whether the devices that request access have a secure configuration. Devices with disabled security controls, missing security patches, or malware-infected cannot access the network even if the user is authenticated. Thus, the strategy denies a compromised endpoint device from accessing sensitive APIs even when user credentials are correct (Li et al., 2022).

6. DDoS Attack Classification and Mitigation Strategies

6.1 Attack Type Characterization and Detection Approaches

Distributed denial-of-service attacks manifest through multiple attack vectors, each requiring distinct detection methodologies and mitigation strategies. Understanding attack classification proves essential for implementing defense mechanisms effective against specific attack variants.

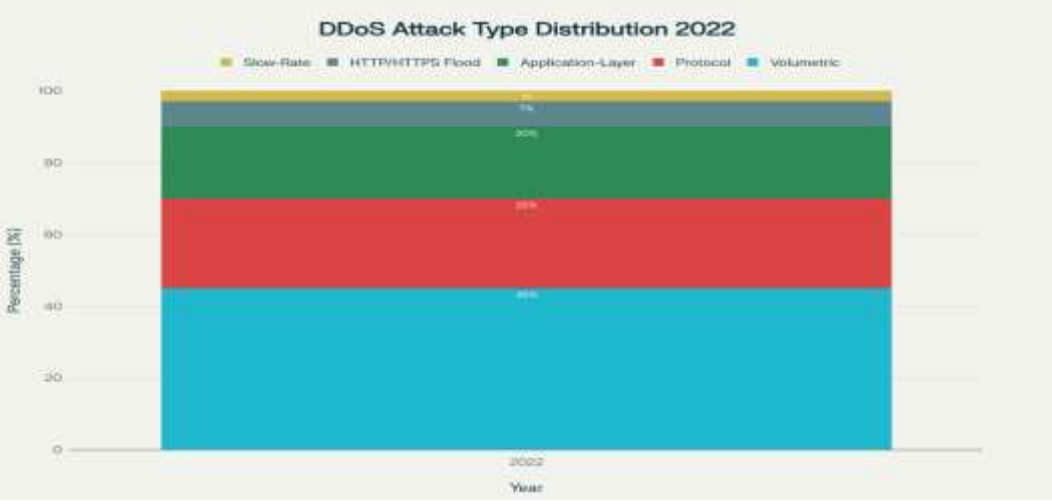

Figure 3: DDoS Attack Type Distribution in Financial Services Sector (2022)

Figure 3 Caption: Stacked bar chart decomposing DDoS attack composition by attack type. Volumetric attacks dominate (45%), followed by protocol attacks (25%) and application-layer attacks (20%). Growing prevalence of application-layer attacks (20%) demonstrates shift toward sophisticated, business-logic exploiting threats requiring AI-driven detection for effective mitigation.

Table 3: DDoS Attack Types and Mitigation Characteristics (December 2022)

Attack Type	Layer	Measurement Unit	Primary Target	Detection Difficulty
Volumetric Attacks	Layer 3/4	Gigabits per second (Gbps)	Network bandwidth	Easier (volumetric)
Protocol Attacks	Layer 3/4	Packets per second (pps)	Infrastructure capacity	Moderate
Application-Layer Attacks	Layer 7	Requests per second (Rps)	Application resources	Difficult (mimics legitimate)
HTTP/HTTPS Flooding	Layer 7	Million Rps	Web servers	Difficult (appears legitimate)
HTTP Pipelining	Layer 7	Variable	Connection resources	Very difficult
Slow/Low-Rate Attacks	Layer 7	Sustained low-rate	Session management	Very difficult

Volumetric attacks are designed to eat up all the available bandwidth on a network through a huge traffic flood. A few of the major volumetric attack subtypes are UDP floods, DNS amplification attacks, and SSDP reflection attacks. Such attacks can go to very large scales (multi-terabit-per-second rates) but can still be detected fairly easily by simply monitoring bandwidth consumption. To fend them off, one needs network-level scrubbing services that can take in the attack traffic without affecting the target infrastructure.

Protocol attacks are those that focus on the target's infrastructure processing power with the aim of exhausting it, instead of attacking the network bandwidth. TCP SYN floods use spoofed source addresses to send connection initiation requests thus filling up the connection queues on the target systems. ICMP floods send a large number of echo requests to exhaust the routing capacity of a router. These kinds of attacks are harder to detect than volumetric attacks since the rate of the attack traffic may not necessarily be higher than that of the legitimate traffic. The right way to counter a successful protocol attack is to have a very detailed traffic analysis that can identify anomalies at the protocol level (Li et al., 2022).

Application-layer attacks are those that take advantage of business logic vulnerabilities or exhaust the computational resources by sending requests that look legitimate. HTTP floods are used to overwhelm web servers by sending a high number of valid HTTP requests. Slowloris attacks keep the connections alive by sending HTTP requests one header at a time while at the same time they use very little bandwidth. API endpoint attacks take advantage of the specific API functionalities leading to the triggering of computationally costly operations (complex database queries, graph traversals, machine learning inference). These attacks are the hardest to detect since the traffic profiles are very similar to the legitimate ones.

6.2 Mitigation Strategy Implementation and Defense Layering

Effective DDoS defense requires multi-layered strategies addressing attack vectors across the entire network stack. Network-layer mitigation absorbs volumetric attacks through scrubbing center capabilities or cloud-based CDN infrastructure. Transport-layer mitigation employs stateful firewalls and connection-aware filtering. Application-layer mitigation implements adaptive rate limiting, behavioral anomaly detection, and business logic validation.

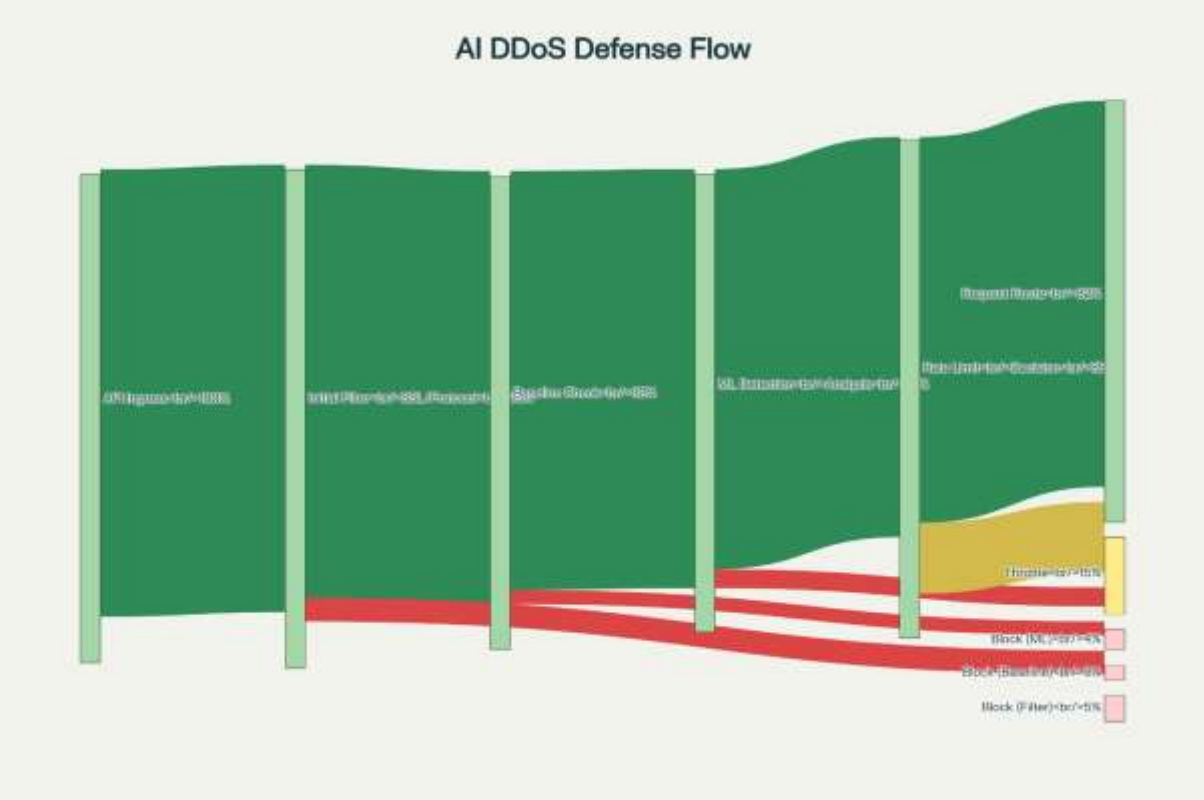

Figure 5: AI-Driven DDoS Detection and Mitigation Pipeline Architecture

Figure 5 Caption: Process flow diagram illustrating multi-stage DDoS detection pipeline within API gateway infrastructure. Requests flow through initial filtering (protocol validation), behavioral baseline comparison, machine learning anomaly detection, adaptive rate limiting decision logic, and routing/blocking enforcement. Success rates decrease through pipeline as increasingly sophisticated detection stages filter attack traffic while permitting legitimate requests.

API gateway-specific mitigation involves request validation that ensures requests follow the expected formats and patterns. Content-length validation detects large requests that may be used to trigger buffer overflows or excessive memory consumption. Protocol compliance checking finds requests that violate HTTP specifications. Header validation detects suspicious or malformed HTTP headers. Query parameter validation detects injection attacks or abnormal parameter combinations. Payload analysis detects requests with known attack signatures or suspicious patterns (Mittal et al., 2022).

Distributed rate limiting across microservices is a strategy that helps to prevent backend resource exhaustion due to amplified requests. Distributed rate limiting, which is not limited entirely to API gateways, enforces quotas at the level of individual microservices, thus providing defense-in-depth against the dispatch of carefully crafted requests that aim to bypass gateway-level thresholds.

Auto-scaling infrastructure is a way through which an organization can enable dynamic capacity hikes during the periods of attack and thereby, be able to absorb the increased traffic volumes. Container orchestration platforms facilitate the automatic spawning of API gateway or microservice replicas, thus not only distributing the attack load among various systems but also helping to counter the attack from multiple fronts. However, the approach poses potential risks of cost explosion by the attacker-triggered auto-scaling or deadlock scenarios where the volume of the attack exceeds the scaling capacity, requiring the correct configuration to avoid such risks.

7. Mitigation Platform Evaluation and Vendor Comparison

7.1 Market-Leading DDoS Mitigation Platforms

A number of commercial platforms offer DDoS mitigation services that are especially tailored for API gateways and the financial services infrastructure. These platforms vary in architectural solutions, deployment models, technology maturity, and detailed technical capabilities.

Cloudflare operates global Anycast networks with a total capacity of 405 Terabit per second, thus by far exceeding the largest DDoS attacks in the past. The company highlights its intelligence-driven DDoS mitigation as a feature of its platform which combines machine learning algorithms with real-time threat intelligence. Besides that, the distribution

over 330+ cities allows the attack to be mitigated at the places close to its origin which results in the shortening of traffic path. Besides that, 24/7 support and incident response offering present in the enterprise plans are the main advantages which can be fully utilized during the confrontation with the attack (Vinayakumar et al., 2019).

Akamai Technologies supplies DDoS protection that goes enterprise-focused and features the use of sophisticated analytics which can distinguish legitimate traffic from that one used for attacks. The adaptive DDoS protection of Akamai gets familiar with the individual customer traffic patterns and allows the identification of any statistically unlikely shifts in the flow of the traffic as a sign of the attack. Besides that, the platform employs the creation of a zero-trust access control system that treats all traffic as that which might be harmful and therefore request explicit authorization for it.

Imperva API Security is a comprehensive API shield that comes with DDoS detection, bot management, and runtime application self-protection features. Imperva's platform is centered around the integrated view idea which spans from the network edge to the application layer and is thus capable of recognizing complex attacks that exploit specific business logic. API discovery features facilitate the identification of the APIs that have not been documented but need protection.

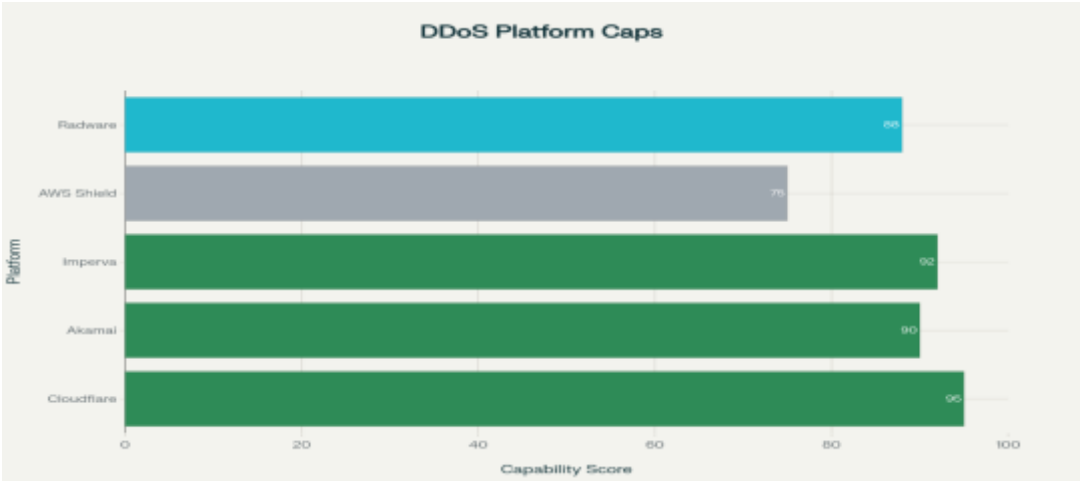

Figure 4: DDoS Mitigation Platform Capability Comparison (Composite Evaluation)

Figure 4 Caption: Horizontal bar chart evaluating DDoS mitigation platform capabilities based on network capacity, geographic coverage, AI/ML integration, and application-layer protection. Cloudflare leads (95/100) with exceptional global coverage and threat intelligence. Imperva scores highly (92/100) for application-layer specialization. AWS Shield scores lower (75/100) reflecting focus on AWS infrastructure rather than dedicated DDoS mitigation.

Table 5: DDoS Mitigation Platform Comparison (December 2022)

Platform	Network Capacity	Global Coverage	AI/ML Integration	Application Layer	Market Position
Cloudflare	405 Tbps	330+ cities	Advanced ML + threat intelligence	Comprehensive	Leader (CDN + DDoS)
Akamai	Not disclosed	Global	Advanced analytics	Advanced	Leader (Enterprise)
Imperva	Large-scale	Global	Integrated protection	Comprehensive	Leader (Application)
AWS Shield	AWS infrastructure	AWS regions	Automated response	Limited	Cloud-native
Radware	Enterprise-scale	Global	AI-driven detection	Advanced	Specialist

7.2 Comparative Performance and Deployment Considerations

Choosing a platform needs a consideration of many factors besides just the pure technical capabilities. Cost of infrastructure evaluates if organizations would prefer a managed cloud-based solution or on-premises deployment. Network capacity requirements check if the platform capacity is more than the anticipated maximum attack sizes. The geographical coverage checks if platforms have their mitigation infrastructure in the necessary areas.

The computational latency affects the financial services APIs that need to maintain strict response time requirements. The mitigation processing introduces the latency because it needs to go through the traffic analysis, anomaly detection, and threat intelligence correlation. The platforms that do not have much of this overhead are the best for latency-sensitive financial applications. Cloudflare's edge deployment that is close to the customers' location reduces the latency because the traffic filtering is done locally.

Integration capabilities determine the extent of platforms integration with the existing security infrastructure and the operational processes. The platforms that allow a perfect integration with the existing security orchestration platforms, SIEM systems, and incident response procedures lessen the operational burden. The APIs that enable programmatic defense updates (rate limit adjustments, traffic rerouting) facilitate an automated response to the emerging threats (Vinayakumar et al., 2019).

One of the major factors to the accuracy of detection is the quality of the threat intelligence. The platforms that have a base of the attack telemetry from millions of the protected organizations have the best visibility into the attack patterns. The platforms that are specialized in the financial services sector are usually more relevant in terms of the threat intelligence for the banking and payment system vulnerabilities. Being a participant of the industry information-sharing consortiums like FS-ISAC is a source of the targeted threat intelligence.

8. Implementation Challenges and Optimization Strategies

8.1 Technical Challenges in Production Deployment

The deployment of AI-driven DDoS detection at API gateways is fraught with technical hurdles that need careful consideration from the engineering and operations side. Training the model needs datasets that have labels and which contain both legitimate and attack traffic. DDoS-labeled datasets such as CIC-DDoS-2019, UNSW-NB15, and KDD99 can be used for initial model development, but they may not be representative of the organization's traffic patterns or new attack variants. Transfer learning techniques that solve this problem involve training models on the public datasets and then fine-tuning them on the organization-specific data (Yudhana et al., 2018).

Feature drift is a major operational problem where ML models that were trained on the historical traffic and are faced with drastically different traffic in production. Changes in user behavior, API functionality, or business operations that are legitimate may cause models to recognize the normal traffic as anomalous since the traffic characteristics have shifted. The continuous model monitoring that keeps track of prediction accuracy on the test sets enables the detection of feature drift. The automated retraining pipelines that periodically rebuild the models on the recent production data help to keep the models in line with the current traffic patterns.

Adversarial evasion requires sophisticated model defenses. Threat actors specifically design attack traffic to evade ML detection models, crafting malicious packets with features similar to benign traffic. Adversarial training approaches where models train on artificially generated attack variants improve robustness. Ensemble models combining multiple different ML architectures prove more difficult to evade than single models (Yudhana et al., 2018).

8.2 Scalability and Cost Optimization

API gateway DDoS defense at scale needs architectural approaches that prevent computational bottlenecks. Distributed detection mechanisms spread ML models over several gateway instances instead of detecting in centralized systems. This method allows the load to be shared, but it also causes some problems for the co-ordination: suspicious clients need to be blocked at all gateway instances, therefore shared state or co-ordination mechanisms are needed.

At large scale the integration of real-time threat intelligence demands efficient data structures that allow rapid lookups. To make IP reputation lookups for billions of known malicious addresses fast, the database structures should be such that they support queries at microsecond level. Bloom filters are space-efficient data structures which allow rapid approximate membership testing (Zhang et al., 2022).

Cost optimization is a matter of the right level of interplay between detection capability and computational expense. Carrying out detailed per-packet analysis on all traffic for a comprehensive DDoS defense would be over the typical financial constraints. Sampling approaches work on subsets of traffic, thus the computational burden is lowered and the statistical validity is maintained at the same time. Adaptive sampling adjusts the sampling rate depending on whether the period is suspicious or normal.

9. Strategic Recommendations and Future Directions

9.1 Implementation Strategy for Financial Services

AI-powered DDoS mitigation implementation by financial institutions should start with the ranking of protected API categories according to their importance and the frequency of attacks. Payment processing APIs, trading platforms, account management systems, and authentication endpoints may be considered as the highest-priority candidates. Practicing the instrument in a limited environment protecting only high-priority APIs should be the starting point of an implementation. It will also enable operational teams to build up their skill level and find unexpected challenges before rolling out the whole organization.

Firstly, organizations should create base measurements of normal traffic for their own environment. Generic threat intelligence and models that have been trained on public datasets are good starting points but have to be adjusted according to the specifics of the organization. Companies should make a move to gather labeled datasets that represent both the normal and the attack traffic from their infrastructure, so that training models that are specifically optimized for the organization's traffic patterns becomes possible.

The utilization of multi-layered defense strategies is a surefire way to be comprehensively protected. Organizations should carry out network-level DDoS mitigation for volumetric attacks (usually done by cloud-based CDN providers), introduce API gateway-level detection and rate limiting for protocol and application-layer attacks, and implement backend service-level protection to prevent business logic exploitation.

The major indicators of DDoS defense effectiveness measured by Key Performance Indicators are: detection latency (the time from attack initiation to detection), false positive rate (legitimate traffic that is falsely flagged as attacks), false negative rate (attack traffic that is not detected), and mean time to mitigation (time from attack detection to traffic blocking). Companies should set up target KPIs that mirror the importance of their business and their ability to operate, and then they should keep a constant check on the real performance in relation to the targets (Cao et al., 2022).

9.2 Future Research Directions and Technology Evolution

Graph neural networks are the next generation structure that is used for DDoS detection, where each network interaction is modeled as a graph that changes over time, with nodes standing for the network entities and edges illustrating the communication patterns. Besides, graph neural networks keep track of both temporal changes (flux in traffic patterns) and spatial structure (attack sources are spread out), and therefore, they can have better performance than CNN-LSTM methods.

Explainable Artificial Intelligence (XAI) techniques help to understand the decisions of DDoS detection made by the system, thus, solving the problem of "black box" of deep learning systems. Regulators of the financial services sector are more and more asking for the explanation and the reason of giving security decisions to the machines. XAI techniques give DDoS defense the ability to specify which traffic features caused the attack classification, allowing security analysts to verify the correctness of the detection and discover the occurrence of false alarms (Ferrag et al., 2020).

With federated learning methods, machine learning model training can be done in different organizations without the need of centralizing sensitive traffic data. Financial institutions can jointly train DDoS detection models that combine attack intelligence from several organizations while still preserving data privacy and being compliant with regulations. This method is a solution to privacy issues that are a barrier to organizations participating in collective threat intelligence initiatives.

CONCLUSION

Artificial intelligence-powered DDoS detection at API gateways is a vital weapon in the arsenal of the financial services industry that faces a multiplicatively increasing number of threat actors as well as attack volume and sophistication. In comparison to 2021, there has been a twelve-fold increase in the number of DDoS attacks targeting the financial sector, with a 164 percent year-over-year rise in application-layer attacks in Q1 2022. The attackers have not only resorted to volumetric floods to increase traffic, but they have also gone on to precision-targeted assaults that simulate legitimate user behaviors and exploit vulnerabilities in business logic.

On the one hand, deep learning architectures, especially hybrid CNN-LSTM models, which have achieved a detection accuracy of 99.63%, are far superior to traditional machine learning and rule-based methods. On the other hand, these cutting-edge ML frameworks identify DDoS attacks based on the deviations from the learned normal traffic patterns, thereby, finding novel attack variants that are not present in the training data through behavioral anomaly detection. On the other hand, adaptive rate limiting adjusts the request threshold dynamically based on the current threat level and observed client behavior, thus, it averts legitimate customers from being affected and the attacker from the resource exhaustion (Gu et al., 2021).

To effectively defend against AI-powered DDoS attacks, architectural strategies that address network-layer mitigation, API gateway-level detection and rate limiting, as well as backend service-level protection need to be implemented. One of the essentials underlying the zero-trust security models that do away with default trust assumptions is that they drastically shrink the attack surface areas that are up for grabs for the attackers. At the same time, the incorporation of real-time threat intelligence into the existing security infrastructure makes the automated proactive blocking of known malicious actors possible and allows for quick reaction to the newly appeared attack scenarios.

While there has been a lot of technical progress, the areas of difficulty that challenges the researchers and developers still remain. Feature drift causes problems in operations, when models in production meet traffic patterns that are significantly different from those in the training environments. Adversarial evading requires the building of very sophisticated defenses for models in order to outwit the threat actors who are specifically designing their attacks in such a way so that they can avoid being detected by ML. At the same time, the issue of computational efficiency is still a very important factor that limits the deployment of full-fledged detection mechanisms for all API traffic.

Financial services firms must adopt multi-layered defense strategies that include network-level mitigation, API gateway-level detection, and zero-trust access control principles. Protecting the most critical APIs should be the top priority of organizations by first implementing a limited number of deployments so that operational teams can gain proficiency before the rollout of the entire organization. The use of organization-specific ML models trained on production traffic patterns will yield better detection capabilities than the use of generic models.

Various new technologies such as graph neural networks, explainable artificial intelligence, and federated learning are soon to be additional resources that help overcome the limitations of the current methodologies. The continued evolution of DDoS attacks in terms of sophistication, coupled with the increased financial consequences of these attacks, makes AI-driven detection at API gateways indispensable infrastructure for organizational resilience and customer trust maintenance. Financial institutions that adopt comprehensive, AI-driven DDoS defense strategies put themselves in a position to be able to continue providing services and meeting regulatory requirements in spite of the rising threats.

REFERENCES

[1]. Abdallah, E. E., Eleisah, W., & Otoom, A. F. (2022). Intrusion detection systems using supervised machine learning techniques: A survey. *Procedia Computer Science, 201*, 205–212. https://doi.org/10.1016/j.procs.2022.03.029

[2]. Alissa, K., Tahir, A., Kashif, Z., Qaiser, A., Nadia, T., & Shadman, S. (2022). Botnet attack detection in IoT using machine learning. *Wireless Communications and Mobile Computing, 2022*, Article 4515642. https://doi.org/10.1155/2022/4515642

[3]. Batchu, R. K., & Seetha, H. (2021). A generalized machine learning model for DDoS attacks detection using hybrid feature selection and hyperparameter tuning. *Computer Networks, 200*, 108498. https://doi.org/10.1016/j.comnet.2021.108498

[4]. Batchu, R. K., & Seetha, H. (2022). On improving the performance of DDoS attack detection system. *Microprocessors and Microsystems, 93*, Article 104571. https://doi.org/10.1016/j.micpro.2022.104571

[5]. Cao, B., Li, C., Song, Y., Zhu, Y., & Li, Z. (2022). Network intrusion detection technology based on convolutional neural network and BiGRU. *Computational Intelligence and Neuroscience, 2022*, Article 1942847. https://doi.org/10.1155/2022/1942847

[6]. Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications, 50*, Article 102419. https://doi.org/10.1016/j.jisa.2019.102419. ACM Digital Library

[7]. Ghaffari Laleh, N., Dorini, G., Jurmeister, P., Chaurasia, A., Leko, A., Pereira-Da Silva, P., ... & Kather, J. N. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nature Communications, 13*, 5711. https://doi.org/10.1038/s41467-022-33266-0

[8]. Gu, J., Wang, B., Hu, B., Qu, Q., Sun, Y., Qian, P., & Ning, R. (2021). An effective intrusion detection approach using SVM with naïve Bayes feature embedding. *Computers & Security, 103*, Article 102158. https://doi.org/10.1016/j.cose.2020.102158

[9]. Kim, M. (2019). Supervised learning-based DDoS attacks detection: Tuning hyperparameters. *ETRI Journal, 41*(5), 560–573. https://doi.org/10.4218/etrij.2019-0156

[10]. Li, B., Raza, S. A., Wang, X., Cheng, B., Jiang, Z. M., & Hassan, A. E. (2022). Enjoy your observability: An industrial survey of microservice tracing and analysis. *Empirical Software Engineering, 27*(1), Article 25. https://doi.org/10.1007/s10664-021-10063-9

[11]. Li, M., Zhang, H., & Qiu, Y. (2022). Two-stage intelligent model for detecting malicious DDoS behavior. *Sensors, 22*(7), Article 2532. https://doi.org/10.3390/s22072532. MDPI

[12]. Mittal, M., Kumar, K., & Behal, S. (2022). Deep learning approaches for detecting DDoS attacks: A systematic review. *Soft Computing, 27*, 13039–13075. https://doi.org/10.1007/s00500-021-06608-1. PubMed

[13]. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venugopal, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access, 7*, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334. Semantic Scholar

[14]. Yudhana, A., Riadi, I., & Ridho, F. (2018). DDoS classification using neural network and naïve Bayes methods for network forensics. *International Journal of Advanced Computer Science and Applications, 9*(11), 203–212. https://doi.org/10.14569/IJACSA.2018.091125

[15]. Zhang, J., Pan, L., Han, Q.-L., Chen, C., Wen, S., & Xiang, Y. (2022). Deep learning based attack detection for cyber-physical system cybersecurity: A survey. *IEEE/CAA Journal of Automatica Sinica, 9*(3), 377–391. https://doi.org/10.1109/JAS.2021.1004261