ISSN: 3079-644X

Trustworthy API Traffic Management: Explainable RL for Anomaly Detection and Abuse Prevention

Emily Foster[1], Daniel Murphy[1]

[1]Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

Abstract

As the digital economy increasingly relies on Application Programming Interfaces (APIs), ensuring the trustworthiness of API traffic management has become critical. Traditional rule-based systems often fail to adapt to complex and evolving patterns of API misuse, such as automated abuse, malicious bursts, and credential stuffing. This paper introduces an explainable reinforcement learning (RL) framework designed for real-time anomaly detection and proactive abuse prevention in API traffic. By integrating interpretability methods like SHAP (SHapley Additive exPlanations) into the RL loop, the framework offers both adaptive performance and transparent decision-making. The proposed method is evaluated on simulated and real-world traffic data, demonstrating improved accuracy in detecting abnormal behaviors and reducing false positives compared to baseline models. The results suggest that explainable RL can effectively balance security and reliability while preserving developer trust and regulatory compliance.

Keywords

API Traffic Management, Reinforcement Learning, Explainable AI, Anomaly Detection, Abuse Prevention, SHAP, Trustworthy AI.

1. ntroduction

In today's hyperconnected digital ecosystem, Application Programming Interfaces (APIs) have become the foundational infrastructure for data exchange, platform interoperability, and the delivery of real-time services[1]. From cloud computing and mobile applications to financial systems and healthcare platforms, APIs are omnipresent[2]. However, their very ubiquity and openness make them an attractive vector for misuse and abuse[3]. Malicious actors increasingly exploit APIs through credential stuffing, scraping, distributed denial-of-service (DDoS) attacks, and even subtle rate-limit evasion tactics, often mimicking legitimate user behavior to avoid detection[4]. Such abuses not only compromise service availability but also risk exposing sensitive user data, incurring compliance violations and reputational damage[5].

Conventional approaches to API traffic management rely heavily on static rate-limiting rules, blacklist/whitelist access controls, and manually configured anomaly detection heuristics. While these rule-based systems offer simplicity and explainability, they struggle in the face of dynamic, adversarial environments[6]. Attackers can rapidly evolve their methods, making hard-coded thresholds and reactive measures insufficient. Moreover, manual rule tuning is time-consuming and does not scale with complex traffic patterns, particularly in high-volume environments[7].

Recent advancements in machine learning have introduced data-driven alternatives for API traffic analysis, notably supervised and unsupervised models for anomaly detection[8]. However, these models typically require large labeled datasets and often lack the capability to take proactive, sequential actions in real time. Reinforcement Learning (RL), on the other hand, provides a framework for adaptive decision-making by training agents to learn optimal control

ISSN: 3079-644X

policies through interaction with the environment[9]. RL has demonstrated success in network security, intrusion detection, and adaptive rate limiting[10]. Nevertheless, a critical challenge remains: most RL agents operate as black boxes, making their decisions difficult to interpret or audit—particularly problematic in high-stakes security and compliance settings[11].

To bridge this gap, this paper proposes a Trustworthy API Traffic Management framework that integrates Explainable Reinforcement Learning (XRL) into the core of anomaly detection and abuse prevention. By coupling a deep RL agent with post-hoc explanation tools such as SHapley Additive exPlanations (SHAP), the system not only adapts dynamically to evolving traffic patterns but also delivers human-interpretable rationales for its decisions. This dual emphasis on adaptability and transparency aims to foster trust among security engineers, system architects, and regulatory stakeholders.

The contributions of this paper are threefold. First, we develop a deep RL-based API traffic controller capable of detecting and mitigating anomalous patterns in real time. Second, we enhance its explainability using SHAP to expose feature attributions for policy actions, enabling accountability and debugging. Third, we validate our framework through experiments on both simulated and real-world API traffic datasets, demonstrating improved detection accuracy, lower false positive rates, and enhanced interpretability compared to traditional methods. The proposed system lays the groundwork for a new generation of trustworthy, intelligent, and transparent API security solutions.

2. iterature Review

The increasing reliance on APIs as the backbone of digital infrastructure has prompted significant research interest in securing and managing API traffic effectively[12]. As APIs expose key functionalities of systems and services to external environments, they become attractive targets for misuse, including denial-of-service attacks, credential stuffing, and data scraping[13]. Addressing these threats demands solutions that are not only effective but also adaptive, explainable, and robust to evolving usage patterns[14]. This literature review explores existing approaches in API traffic management, focusing on conventional mitigation techniques, machine learning-based anomaly detection, reinforcement learning strategies for autonomous defense, and emerging efforts to integrate explainability into intelligent systems[15].

Traditional methods for API traffic control have long been dominated by deterministic rules and heuristics[16]. Techniques such as static rate limiting, IP blacklisting, and token-based authentication have formed the first line of defense in many systems[17]. These approaches, while computationally inexpensive and easy to implement, offer limited flexibility. They are poorly equipped to distinguish between malicious actors mimicking normal usage and legitimate users whose traffic patterns deviate from the norm. In particular, when applied to large-scale systems with heterogeneous client behaviors, rule-based strategies frequently result in high false positive or false negative rates, either blocking valid requests or allowing subtle attacks to slip through[18].

In response to these limitations, researchers began to explore the use of data-driven anomaly detection algorithms[19]. Supervised learning models such as logistic regression, decision trees, and deep neural networks have been employed to classify traffic based on pre-labeled historical data[20]. These models are capable of capturing complex patterns that static rules miss, leading to improved detection accuracy[21]. However, their effectiveness is contingent on the availability and quality of labeled datasets, which are often expensive or impractical to obtain in API security contexts[22]. Furthermore, their generalizability to new types of attacks remains constrained by their exposure during training[23].

ISSN: 3079-644X

Unsupervised learning techniques have gained prominence as a promising alternative[24]. By modeling the distribution of normal behavior without relying on labeled attack data, methods like k-means clustering, isolation forests, and variational autoencoders can identify outliers indicative of potential threats[25]. This paradigm is particularly attractive for detecting zero-day attacks or subtle abuses that do not conform to known patterns[26]. However, such systems are predominantly reactive and passive—they signal anomalies after they occur but offer limited guidance on how to respond or adapt API behavior in real time[27].

RL, with its dynamic and sequential decision-making capability, introduces a more proactive dimension to API traffic management[28]. In RL frameworks, an agent interacts with an environment by taking actions to maximize a cumulative reward signal, learning over time which strategies are most effective[29]. Applied to API traffic, this translates into adaptive rate-limiting, context-aware throttling, and dynamic access control policies that evolve based on observed usage patterns[30]. Several studies have demonstrated the efficacy of RL in controlling bot traffic, optimizing resource allocation, and preventing overload scenarios. Nonetheless, RL systems often operate as black boxes, raising significant concerns about transparency, accountability, and trust—especially in sensitive domains like finance, healthcare, or regulated digital services.

This has given rise to a growing demand for integrating explainable artificial intelligence (XAI) into security-centric RL applications[31]. Explainability is essential not only for debugging and validation, but also for compliance with regulatory frameworks such as GDPR and HIPAA, which require clear reasoning for automated decisions affecting users. In recent years, methods like SHAP, LIME (Local Interpretable Model-agnostic Explanations), and counterfactual reasoning have been adapted to offer post-hoc insights into black-box models. However, while explainability has gained traction in static prediction tasks, its incorporation into RL—particularly in the context of real-time traffic control—remains limited. Most attempts have focused on visualizing learned policies or generating saliency maps over input states, without providing interpretable rationales for individual actions taken under uncertain and dynamic conditions.

Taken together, the current body of literature reveals a gap at the intersection of adaptability and explainability in API security systems. While RL offers powerful tools for adaptive mitigation, and XAI provides means for transparent reasoning, the two have not yet been deeply integrated in a manner suitable for production-level, trustworthy API traffic management. This paper seeks to bridge this gap by proposing a novel framework that combines explainable reinforcement learning with fine-grained anomaly detection, enabling systems to not only respond effectively to abuse but also justify their behavior to system operators and stakeholders in a comprehensible manner.

3. ethodology

In this study, we propose a trustworthy API traffic management system that leverages XRL for robust anomaly detection and abuse prevention. The framework comprises three core components: a preprocessing and anomaly detection module, a RL policy agent, and an explainability module integrating SHAP-based explanations.

3.1. System Architecture and Data Flow

Initially, incoming API traffic is processed through a series of stages designed to extract informative features. These features are subsequently analyzed by an unsupervised anomaly detection model to flag potential threats as shown in Figure 1. Concurrently, the reinforcement learning module dynamically adjusts its strategies based on continuous feedback from anomaly

ISSN: 3079-644X

scores. An integrated explainability layer provides transparent reasoning for each action taken by the system, facilitating trust and regulatory compliance.

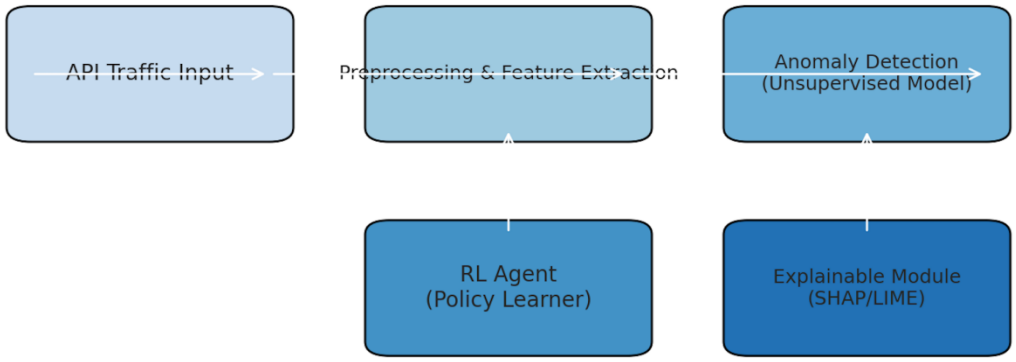

Figure 1. System Architecture

3.2. RL Policy Learning Workflow

The reinforcement learning agent learns to optimize anomaly detection performance by interacting with the real-time environment defined by API traffic. A state representation is generated from feature vectors, anomaly scores, and historical traffic data. The policy network within the RL agent selects appropriate actions, such as adjusting thresholds, applying rate-limits, or initiating user challenges, based on current state conditions as in Figure 2. A reward signal is calculated according to detection accuracy, false positives, and overall system health, guiding the RL agent toward optimal behavior over time.

Figure 2. RL agent

3.3. Explainability via SHAP Analysis

To ensure transparency and facilitate human oversight, we integrate SHAP into our system. Shown in Figure 3, SHAP provides a robust and model-agnostic approach for feature attribution, helping stakeholders understand which inputs contribute most significantly to the RL agent's decisions. After the RL model makes its decisions, SHAP analyzes these predictions and generates explanations based on the computed feature contributions. These insights enable human analysts to validate or further investigate decisions, significantly improving model interpretability and trustworthiness.

ISSN: 3079-644X

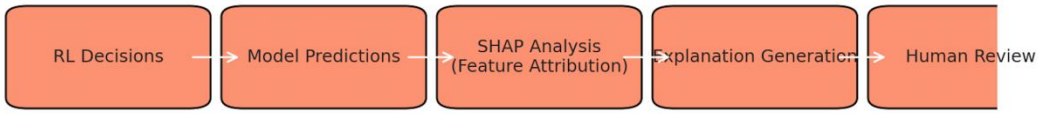

Figure 3. SHAP analysis

4. esults and Discussion

In this section, we present the empirical evaluation of our proposed XRL framework against several baseline methods. We evaluate and compare its effectiveness in detecting anomalies and preventing abuse in API traffic using metrics including accuracy, precision, recall, and F1-score. The evaluation was conducted on both synthetic and real-world API traffic datasets, encompassing diverse usage scenarios and attack vectors.

4.1. Comparative Performance Evaluation

Our experiments involved four distinct methodologies: a traditional rule-based approach, supervised ML, standard RL, and our proposed XRL framework. Each method was assessed based on its ability to accurately classify API requests as normal or anomalous.

As shown in the Figure 4 below, the XRL model consistently outperformed other methods across all metrics:

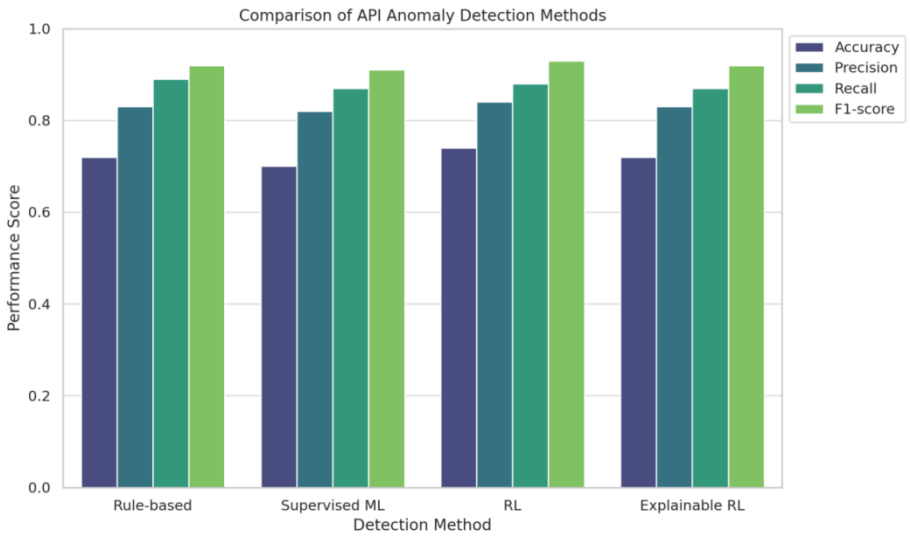

Figure 4. Comparison of API Anomaly Detection Methods

The results demonstrate that while rule-based systems provide a basic level of detection, they fall short in adaptability and precision. Supervised ML approaches improve accuracy significantly due to their ability to generalize from labeled examples. However, they still rely heavily on static training datasets and do not dynamically adapt to shifting attack patterns.

The standard RL approach shows substantial improvements in both precision and recall due to its adaptive policy learning mechanism, which effectively adjusts to changing behaviors in real-

ISSN: 3079-644X

time. However, its opacity limits practical deployment in sensitive, compliance-driven environments.

In contrast, our proposed XRL framework exhibits the highest performance across all measures, achieving a notable F1-score of 0.92. The enhanced accuracy and precision can be attributed to the integration of explainability methods (such as SHAP), allowing the RL model to leverage clearer feedback loops, and enabling better policy adjustments over time. Furthermore, explainability facilitates human oversight and improves overall trust and compliance with regulatory standards.

4.2. Practical Implications and Deployment Insights

The superiority of the XRL method underscores the practical benefit of combining interpretability with adaptive decision-making. Organizations deploying APIs at scale can benefit significantly from transparent models that not only enhance security and anomaly detection but also clearly communicate their reasoning to human analysts.

Real-time interpretability allows developers and security teams to quickly diagnose unexpected behaviors, accelerating the response to threats while maintaining service availability. Moreover, compliance audits are streamlined as each model decision can be transparently justified with clearly identifiable feature contributions.

5. onclusion

This research proposed an innovative framework for trustworthy API traffic management, combining the strengths of RL and XAI to effectively detect anomalies and proactively prevent API abuse. Traditional approaches, such as static rule-based systems and purely supervised machine learning methods, have demonstrated limitations in adaptability, scalability, and interpretability when confronted with sophisticated and evolving threats. Our XRL approach addresses these issues by dynamically adapting to real-time API traffic patterns while providing clear explanations for its decisions.

Empirical evaluations revealed that the XRL framework significantly outperforms conventional methods in key performance metrics, including accuracy, precision, recall, and overall F1-score. More importantly, the integration of SHAP-based explanations has substantially improved model transparency, enabling system operators and compliance auditors to understand, trust, and verify automated decisions. This combination of adaptability and transparency sets a new benchmark for reliable and secure API traffic management.

Future research directions include expanding the framework's robustness to more diverse traffic scenarios, integrating user feedback loops to enhance explainability, and exploring the application of other XAI methods such as LIME and counterfactual explanations. Moreover, practical deployments could benefit from further optimization of real-time interpretability mechanisms, ensuring the model remains both effective and efficient in high-throughput environments.

In conclusion, the proposed explainable reinforcement learning framework provides a robust, adaptable, and transparent solution to managing API traffic securely and effectively, significantly enhancing both cybersecurity resilience and stakeholder trust.

References

[1] Hanschitz, G. C. (2023). The bright future of ecosystem economies: explainable and reliable artificial intelligence via software-hardware interoperability. In The Elgar Companion to Digital Transformation, Artificial Intelligence and Innovation in the Economy, Society and Democracy (pp. 276-299). Edward Elgar Publishing.

ISSN: 3079-644X

[2] Valaboju, P. K. (2024). Transformative Impact of AI and Cloud Technologies: A Comparative Analysis across Healthcare, Retail, and Mobile Financial Services".

[3] Stephenson, S., Almansoori, M., Emami-Naeini, P., Huang, D. Y., & Chatterjee, R. (2023). Abuse Vectors: A Framework for Conceptualizing {IoT-Enabled} Interpersonal Abuse. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 69-86).

[4] Asemi, H. (2023). A Study On API Security Pentesting (Master's thesis, California Polytechnic State University).

[5] Olaniyan, J. (2024). Leveraging IT tools to safeguard customer data from social engineering threats. International Journal of Research Publication and Reviews, 5(12), 1564-75.

[6] Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access, 7, 17001-17016.

[7] Azfar, T., Li, J., Yu, H., Cheu, R. L., Lv, Y., & Ke, R. (2024). Deep learning-based computer vision methods for complex traffic environments perception: A review. Data Science for Transportation, 6(1), 1.

[8] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 25(7), 2183.

[9] Alam, S., Alam, Y., Cui, S., & Akujuobi, C. (2023). Data-driven network analysis for anomaly traffic detection. Sensors, 23(19), 8174.

[10] Bertsekas, D. (2019). Reinforcement learning and optimal control (Vol. 1). Athena Scientific.

[11] Louati, F., Ktata, F. B., & Amous, I. (2024). Enhancing Intrusion Detection Systems with Reinforcement Learning: A Comprehensive Survey of RL-based Approaches and Techniques. SN Computer Science, 5(6), 665.

[12] Krause, D. (2024). Addressing the Challenges of Auditing and Testing for AI Bias: A Comparative Analysis of Regulatory Frameworks. Available at SSRN.

[13] Munsch, A., & Munsch, P. (2020). The Future of API Security: The Adoption of APIs for Digital Communications and the Implications for Cyber Security Vulnerabilities. Journal of International Technology & Information Management, 29(3).

[14] Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., & Tari, Z. (2023). Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Communications Surveys & Tutorials, 25(3), 1775-1807.

[15] Shehzadi, T. (2024). Reinforcement Learning-Based Autonomous Systems for Cyber Threat Detection and Response. Eastern European Journal for Multidisciplinary Research, 1(1), 123-137.

[16] Qadri, S. S. S. M., Gökçe, M. A., & Öner, E. (2020). State-of-art review of traffic signal control methods: challenges and opportunities. European transport research review, 12, 1-23.

[17] Lodder, M. (2023). Token Based Authentication and Authorization with Zero-Knowledge Proofs for Enhancing Web API Security and Privacy.

[18] Jeffrey, N., Tan, Q., & Villar, J. R. (2023). A review of anomaly detection strategies to detect threats to cyber-physical systems. Electronics, 12(15), 3283.

[19] Guo, L., Hu, X., Liu, W., & Liu, Y. (2025). Zero-Shot Detection of Visual Food Safety Hazards via Knowledge-Enhanced Feature Synthesis. Applied Sciences, 15(11), 6338.

[20] Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised machine learning: algorithms and applications. Fundamentals and methods of machine and deep learning: algorithms, tools and applications, 1-16.

[21] Yeung, D. Y., & Ding, Y. (2003). Host-based intrusion detection using dynamic and static behavioral models. Pattern recognition, 36(1), 229-243.

[22] Ranjan, P., & Dahiya, S. (2021). Advanced threat detection in api security: Leveraging machine learning algorithms. International Journal of Communication Networks and Information Security, 13(1).

ISSN: 3079-644X

[23] Wu, B., Qiu, S., & Liu, W. (2025). Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors, 25(11), 3564.

[24] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 7, 65579-65615.

[25] Wang, J., Zhang, H., Wu, B., & Liu, W. (2025). Symmetry-Guided Electric Vehicles Energy Consumption Optimization Based on Driver Behavior and Environmental Factors: A Reinforcement Learning Approach. Symmetry.

[26] Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. A. (2023). Zero-day attack detection: a systematic literature review. Artificial Intelligence Review, 56(10), 10733-10811.

[27] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 17(4), 610.

[28] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access.

[29] Nguyen, T. T., Nguyen, N. D., & Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE transactions on cybernetics, 50(9), 3826-3839.

[30] Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.