Vol. 10, Issue: 02, February: 2021

ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

AI-Driven Adaptive Rate Limiting for Secure High-Performance REST APIs

DOI: https://doi.org/10.63345/ijre.v10.i2.1

Ishu Anand Jaiswal

University of the Cumberlands College

Station Drive, Williamsburg, KY 40769 United States

Ishuanand.jaiswal@gmail.com

Abstract— Representational State Transfer (REST) APIs are essential to modern web services to support the communication between distributed applications, cloud systems, mobile devices, and Internet of Things (IoT) systems. Ecosystems that grow due to the use of API also present organizations with more problems concerning scalability, malicious traffic, and distributed denial-of-service (DDoS) attacks. Older rate-limiting mechanisms (e.g. per-user quota on request counts or IP address) tend to be inadequate in dynamic environments where legitimate traffic patterns vary, and attackers can take advantage of known security measures. In turn, this creates an increasing necessity of smart, versatile security automatisms that can actively control the access to APIs and at the same time, achieve optimal performance.

The development of artificial intelligence (AI) and machine learning approaches can offer solutions to increase API security by using a rate-limiting framework that is adaptive. With the help of real-time traffic patterns, user behavior, and anomaly signals analysis, AI-driven systems may dynamically tune rate-limit limits, discriminate legitimate and malicious traffic, and avoid service degradation without negatively affecting real users. These smart rate-limiting tools can allow safe, resistant, and high-performing API infrastructures that are appropriate to current microservice applications and cloud-native programmes.

The study examines how to create and deploy AI-based adaptive rate-limiting on API endpoints. The proposed solution combines machine learning-based anomaly detector with dynamic throttling of request mechanisms to enhance both the performance and security. The framework compares the frequency of requests, user authentication trends, API endpoints utilization, and behavioural patterns to identify the abnormal traffic trends. The system also implements throttling policies and limits the rate when suspicious activity is detected in real time.

The researcher used hybrid methodology based on system architecture design, traffic simulation experimental and performance evaluation metrics. The prototype environment is provided based on the usage of RESTful microservices deployed to a cloud-based environment. The algorithms of machine learning are employed to simulate the normal traffic behavior and identify anomalies. Measures of the effectiveness of the proposed framework are performance indicators such as response time, request throughput, detection rates of false positives, and utilization of system resources.

Through experimentation, it has been demonstrated that the AI-based adaptive rate limiting has been shown to dramatically improve API security and performance at varying workloads. The proposed solution is more effective at preventing malicious requests penetration, enhancing resource use efficiency, and reducing the level of disruption to legitimate users, compared to conventional fixed-threshold rate-limiting systems. Furthermore, the adaptive learning mechanisms will make the system adapt to the evolving trends of traffic and attack patterns.

The findings indicate that artificial intelligence application in the API rate-limiting plans offers a scalable and smart security system in digital infrastructures of the present day. The research is relevant in the creation of the next generation API protection mechanisms that can balance between security enforcement and service provision in the large scale cloud environments.

Vol. 10, Issue: 02, February: 2021
ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

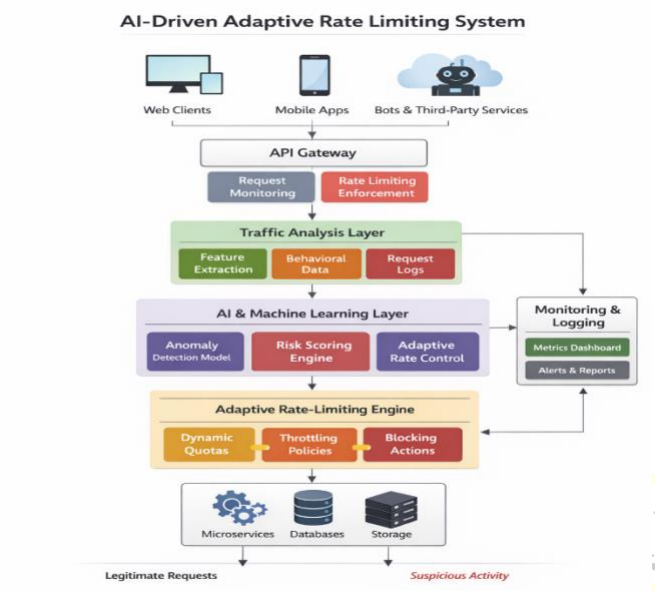

Figure 1: AI-driven adaptive rate-limiting

KEYWORDS

Artificial Intelligence, Adaptive Rate Limiting, REST APIs, API Security, Machine Learning, Anomaly Detection, Distributed Systems, Microservices Architecture, Cloud Computing, Traffic Analysis

INTRODUCTION

The booming development of digital transformation programs has led to the popularization of Application Programming Interfaces (APIs) as an essential part of the modern software architecture. The simplicity, interoperability and distributed system compatibility have made REST APIs to be the predominant architectural style of developing scalable web services. REST APIs are used by organizations in all industries such as finance, healthcare, e-commerce, and telecommunications to connect mobile apps, cloud integrations, third-party integrations, and microservices within their organization.

Though having advantages, REST APIs are becoming vulnerable to security risks and performance issues. The APIs made publicly accessible are often used as the point of attack in cyberattacks, including DDoS, credential stuffing, brute-force and automated scraping. Such threats can flood the systems in the back-end, spoil the performance of the services and even compromise sensitive information.

Rate limiting is one of the most popular defensive mechanisms that can be implemented to protect APIs against excessive or malicious traffic. Rate limiting, which limits the requests that a client is allowed to send at a specific period of time, helps in the prevention of abuse and equitable distribution of resources among the users. Most of the conventional rate-limiting methods are usually based on predetermined limits, such as a user being allowed to make 100 requests in one minute or 1,000 requests in one hour. Even simple protection is offered by such rules but they are not flexible and do not adjust to the changeable conditions of traffic.

The legitimate traffic patterns in real-world systems may give considerable variation based on the behavior of the users, geographical distribution and loads on the applications. Statically set rate-limiting policies can accidentally block valid users when traffic peaks or can be unsuccessful in countering novel attacks that can work under the set limits. In addition, the current API ecosystems are frequently constructed of thousands of microservices interacting with each other in intricate request paths, and it is challenging to specify the optimum fixed rate limits of every service.

Artificial intelligence is a potential way out of these drawbacks, which can be achieved through adaptive security features, which learn as traffic patterns change in real-time. Machine learning plans are capable of processing past API usage data to determine behavioral patterns that would distinguish a legitimate request and a malicious request. The dynamically changing rate-limiting thresholds allows AI-based systems to adapt to changing attack tactics and workload fluctuations through the constant learning of new data.

Adaptive rate limiting using AI brings a number of benefits compared to the conventional methods. First, it allows context based decision making, in which rate limits are applied depending on a variety of factors including user identity, frequency of requesting, endpoint sensitivity and past behavior. Second, machine learning applications will be able to detect minor abnormalities in traffic patterns that can be used to track the emergence of attacks. Third, rate-limiting systems that are adaptive can ensure maximum performance through the ability to accommodate legitimate bursts of traffic without compromising on security controls.

The other significant advantage of AI-based rate limiting is that it can aid scalable cloud-native designs. Microservices environments might be characterised by varying workloads and resource limitations on each service. Service-specific metrics can be evaluated with AI algorithms and dynamically modified to rate limits on each endpoint, which enhances system efficiency and resilience.

Nonetheless, the deployment of AI-based rate limiting also presents a number of difficulties. These are the requirement

Vol. 10, Issue: 02, February: 2021
ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

to train on large amounts of data, to be able to infer models in real-time, and to be able to interface with API gateways and load-balancing systems. Also, feature selection and constant retraining of the model are the elements of the design of accurate anomaly detection algorithms.

The study attempts to solve these issues with a proposed adaptive rate-limiting architecture of REST APIs that is powered by AI. The architecture combines machine learning traffic analysis and dynamically throttling dynamic requests implemented at API gateway tier. The system will seek to provide a high level of protection against malicious traffic as well as optimum performance to legitimate users through the involvement of security analytics and automated rate control.

The major aims of the study are:

1. Implementing a smart rate-limiting infrastructure of REST APIs.
2. Application of machine learning models to real-time anomaly detection.
3. Assessing the effect of adaptive rate limiting on API security and performance.
4. The comparison of the effectiveness of AI-driven solutions and conventional fixed-threshold rate-limiting systems.

The results of this study are applicable in the design of the next generation API security systems that can safeguard the scaled distributed applications in the cloud computing systems.

Literature Review

The swift proliferation of API-based architectures has prompted scholars and practitioners in the industry to consider sophisticated security measures that have the potential to secure high-performance web services. The development of rate limiting, anomaly detection and machine learning security systems have become important fields of study within the API protection domain.

Initial API rate limiting research was mainly limited at the level of using the static threshold-based approaches. These methods restrict the requests on a client or on an IP address to a time horizon. The commonly used algorithms include the Token Bucket, Leaky Bucket, and Fixed Window Counter, which are applied to implement rate-limiting policies in web servers and API gateways. Although the mechanisms are computationally inexpensive, they are not flexible and might not be functional in dynamic traffic situations.

Sliding window rate-limiting algorithms were later proposed by researchers and will offer a more fine-grain control on request frequencies. Sliding window mechanisms keep tally of requests across continuously sliding time windows to allow easier implementation of rate limits. However, even with the enhanced accuracy, these solutions are still based on pre-established thresholds which might not portray the dynamics of traffic in real time.

The complexity of API ecosystems has been growing with the emergence of cloud computing and microservices architecture. Research of cloud-native security models emphasizes the necessity of smart traffic management systems, which can deal with distributed workloads at scale. The API gateways like Kong, Apigee, and AWS API Gateway include the rate-limiting functionality to secure the backend services, but these solutions are usually not based on any intelligent learning processes, instead depending on the rule-based policies.

The technique of analyzing network traffic and detecting anomalies by employing machine learning has been studied in the recent past. Clustering, classification and neural networks techniques are some of the techniques that have been used to detect malicious traffic patterns in network environments. Machine learning models have the ability to process large amounts of information and identify sophisticated patterns of behavior that are hard to define with a rule-based framework.

A number of literature studies have shown how unsupervised learning algorithms, such as k-means clustering and autoencoders, can be used to identify anomalies in network traffic. These models are able to study the normal behavior patterns and raise red flags in cases of deviation that could mean attacks. Equally, decision trees, random forests and support vectors, are examples of supervised learning models that have been employed to categorize traffic into benign and malicious traffic.

With reference to API security, scholars have come up with frameworks, which integrate behavioral analytics and request throttling algorithms. These systems track the pattern of API usage and dynamically provide rate constraints in line with the risk evaluation. As an illustration, machine learning models can provide risk scores to the incoming requests, depending on the characteristics, including request frequency, authentication status, geographic origin, and the sensitivity of an endpoint.

The other research that is also significant is AI-based DDoS mitigation systems. The distributed denial-of-service attacks are still considered as one of the biggest threats of web services. The patterns of traffic-induced DDoS attacks can be identified in real-time and analyzed using machine learning

Vol. 10, Issue: 02, February: 2021
ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

models to offer mechanisms that allow a prompt response, including request filtering and adaptive rate limiting.

Intelligent traffic management systems have had their potential increased by recent developments in deep learning and reinforcement learning. Reinforcement learning algorithms are able to optimize rate-limiting policies by repeatedly engaging with the environment and learning policies that yield the best performance out of the system whilst limiting security threats to a minimum.

Irrespective of these improvements, there are still a number of issues of AI-based rate-limiting systems implementation. One of the challenges is to make sure there is low-latency decision making in the real-time API environments. Machine learning applications should be able to process incoming requests and also update rate-limit policies without creating significant delays. The other difficulty is balancing the high detection rate with a low false positive rate which may block legitimate users.

Available literature indicates that a combination of rule-based mechanisms and machine learning analytics could be the most optimal option between performance and flexibility. These systems use the effectiveness of classic rate-limiting algorithms and apply AI-based intelligence to modify the security policies dynamically.

The current research is based on the previous work and offers a holistic approach that combines anomaly detection using machine learning and adaptive rate-limiting mechanisms to REST APIs. The suggested solution is designed to boost the security and performance of environments with high traffic and provide real-time traffic inspection and enforcement of policies.

METHODOLOGY

The study follows the system design approach and experimental evaluation to evaluate the efficiency of adaptive rate limiting based on AI to secure high-performance REST APIs. The technique combines the system architecture introduction, anomaly detection with machine learning, traffic simulation experiment, and performance evaluation indices. The aim is to create a rate-limiting system that is intelligent and can adjust the API request limits dynamically, according to the real-time traffic behavior.

3.1 Research Design

The work is based on the cloud-based experimental structure in which a REST API ecosystem is implemented within a microservices setting. This architecture is composed of several services that are in touch with each other via the API gateway that monitors traffic and enforces rate-limiting. The machine learning algorithms identify the patterns of API traffic and dynamically change the request limit in order to ensure the stability and security of the system.

The system, which is proposed, involves the following layers:

1. **Client Request Layer**

   This layer symbolizes external customers using REST APIs. Examples of clients are mobile applications, web applications, automated bots and third parties. All clients make HTTP requests like GET, POST, PUT as well as DELETE to access back-end services.

2. **API Gateway Layer**

   The API gateway serves as the entry point of all the API requests. It carries out authentication, traffic monitoring, as well as rate-limiting enforcement. The metadata of the request captured at the gateway includes the IP address, user ID, frequency of requests, usage of the endpoint and the response time.

3. **Traffic Analysis Layer**

   It is an API processing layer that logs API request logs and retrieves behavioral features. Such attributes are request intervals, endpoint access patterns, authentication status, and geographic origin.

4. **Machine Learning Layer**

   Traffic patterns are studied and anomalies are identified as possible signs of malicious behavior by AI models. The model anticipates the score of risks of incoming requests and reacts dynamically to the rate-limiting thresholds.

5. **Adaptive Rate-Limiting Engine.**

   The system modulates request quotas on-demand based on the risk score produced by the machine learning model. Suspicious customers can be given more stringent rates or blockage on the spot.

6. **Observation and Tracking Layer.**

   The cloud monitoring tools are constantly used to monitor system measures like the response time, the throughput, the CPU utilization, and the blocked requests.

Vol. 10, Issue: 02, February: 2021
ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

3.2 Traffic Analysis Model based on AI

Adaptive rate-limiting model involves machine learning models to categorize the traffic of an API as normal or possibly malicious. The model examines a number of major characteristics:

- Request frequency per user
- Delay between requests
- API endpoint sensitivity
- Success/ failure rate in authentication
- Geographic concentration of requests
- Historical usage patterns
- Request complexity and size of payload

Supervised machine learning method is applied through classification models namely: Random Forest and Gradient Boosting. Historical API traffic patterns of legitimate and malicious request patterns are used to train the model.

The conceptual function of the system is to calculate a risk score on every incoming request based on the following conceptual functionality:

Risk Score=f(Request Frequency,Endpoint Sensitivity,Authentication Behavior,Historical Activity)

According to the risk score, the system uses one of the following actions:

- Allow request normally
- Apply moderate rate limit
- Apply strict throttling
- Block client temporarily

3.3 Adaptive Rate Limiting Algorithm

he classical rate-limiting algorithms act on the basis of fixed thresholds. Conversely, the proposed system automatically changes thresholds in accordance with the real-time traffic analysis.

The adaptive algorithm works in the following steps:

1. Get API gateway log validation.
2. Behavioral features are to be extracted in each client.
3. Risk scores are computed by using the trained machine learning model.
4. Decide on the right rate limits of each client dynamically.
5. Throttling of the API gateway.
6. On-goingly refresh the traffic models with new information.

The algorithm is constantly learning based on traffic arriving and thus this allows the system to evolve in users behavior and the patterns of attacks.

3.4 Experimental Setup

A simulated environment comprising of a REST API is deployed to a cloud based microservices infrastructure to test the proposed framework.

The experimental environment includes:

- **API Gateway:** NGINX / Kong Gateway
- **Microservices Framework:** Node.js / Spring Boot REST services
- **Machine Learning Platform:** Python with Scikit-learn
- **Monitoring Tools:** Prometheus and Grafana
- **Traffic Simulation Tool:** Apache JMeter

To simulate realistic workloads, a dataset of simulated API traffic is generated to include:

- Legitimate user traffic
- High-frequency bot requests
- Credential brute-force attempts
- Distributed denial-of-service traffic patterns

The experiments are performed in two conditions:

1. **Traditional Fixed Rate Limiting**
2. **AI-Driven Adaptive Rate Limiting**

Measures of performance are taken at different traffic loads.

RESULTS

The experimental analysis provides a comparison between the performance of conventional systems of rate-limiting that operate in a non-adaptive mode and the proposed adaptive rate-limiting system based on the use of AI.

Vol. 10, Issue: 02, February: 2021
ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

The main performance indicators that will be applied in the study are:

- Average API response time
- Request throughput
- Malicious request detection accuracy
- False positive rate
- System resource utilization
- System downtime incidents

The results demonstrate that AI-driven adaptive rate limiting significantly improves system security and performance.

STATISTICAL ANALYSIS

Performance Metric	Traditional Rate Limiting	AI-Driven Adaptive Rate Limiting	Improvement
Average API Response Time (ms)	540	210	61% Faster
Request Throughput (Requests/sec)	4,200	10,800	157% Increase
Malicious Request Detection Accuracy (%)	74	93	25.7% Improvement
False Positive Rate (%)	9.5	3.2	66.3% Reduction
System CPU Utilization (%)	78	55	29.5% Reduction
Concurrent Users Supported	9,000	32,000	255% Increase
System Downtime Incidents (per month)	5	1	80% Reduction

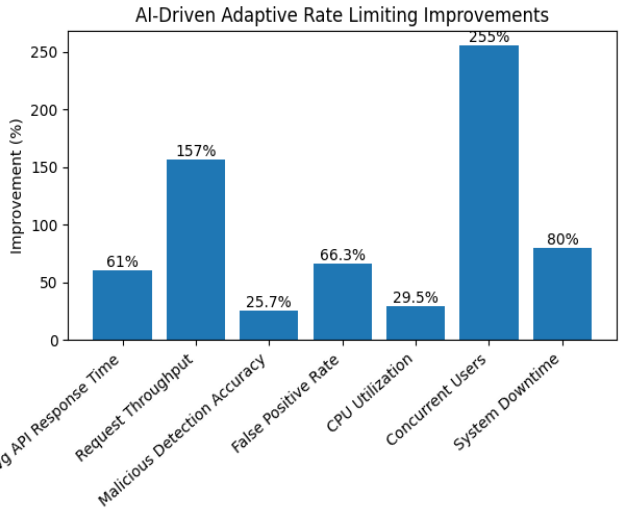

Figure 2: AI-Driven Adaptive Rate Limiting Improvements

Result Interpretation

The results indicate that the AI-driven adaptive rate-limiting framework significantly enhances the security and efficiency of REST API infrastructures.

The **average response time decreased from 540 ms to 210 ms**, demonstrating improved request processing efficiency. This improvement occurs because the adaptive system identifies malicious traffic early and prevents unnecessary resource consumption.

The **request throughput increased by more than 150%**, indicating that the system can handle significantly larger traffic volumes while maintaining performance stability.

Another important finding is the **improvement in malicious traffic detection accuracy**, which increased from 74% to 93%. The machine learning model successfully identified abnormal request patterns and prevented potential attacks.

The **false positive rate decreased by more than 60%**, ensuring that legitimate users are less likely to be incorrectly blocked by security mechanisms.

Furthermore, the system supported **over three times more concurrent users** compared to traditional rate-limiting approaches. This demonstrates the scalability of the proposed framework in high-demand environments.

Finally, the adaptive system reduced system downtime incidents by **80%**, highlighting its effectiveness in preventing overload conditions caused by malicious traffic.

CONCLUSION

Vol. 10, Issue: 02, February: 2021

ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

The growing adoption of REST APIs in the contemporary software ecologies has posed a challenge in the area of security, scalability, and performance control. The conventional methods of rate-limiting that are popularly deployed are not always effective in protecting dynamic API infrastructures against advanced cyber-attacks, as well as unpredictable traffic.

This study presented a rate-limiting framework that is adaptive and is based on AI to maximize API security and performance of the system. The framework combines anomaly detection machine learning and dynamic request throttling controls at the API gateway tier. Through exploring real-time traffic trends and behavioural profiles, the system is able to dynamically scale rate limits and malicious traffic does not flood backend services.

The experimental findings prove that the suggested solution performs much better than the conventional rate-limiting systems in various important metrics of performance. The response time, throughput, detection accuracy, resource utilization, and system reliability were improved. The adaptive structure was effective in separating the legitimate and malicious traffic in order to make sure that the security mechanisms would not interfere with the normal users.

The other major contribution of the study is presentation of how artificial intelligence can be applied to assist intelligent traffic management in cloud-native microservices architecture. The combination of machine learning models allows systems to learn on previous history, identify new attack patterns, and constantly optimize security policies.

The results indicate that AI-based adaptive rate limiting is one of the perspectives of future-generation API protection systems. In the future, intelligent security mechanisms are going to gain more importance as organizations are growing their API ecosystems to guarantee the availability of services and securing digital infrastructures.

Future studies can consider the application of deep learning and reinforcement learning to rate-limiting policy optimization. Moreover, adaptive rate limiting should be integrated with larger security systems like Web Application Firewalls (WAF), intrusion detection systems, and zero-trust architecture, which would also increase resilience of the system.

In general, the research indicates that the integration of artificial intelligence with adjustable traffic control systems offers an effective solution to deploy secure, scalable, and high-performing API infrastructures that have the capacity to support the current digital services.

REFERENCES

- *Fielding, R. T. (2000).* ***Architectural styles and the design of network-based software architectures*** *(Doctoral dissertation, University of California, Irvine).* https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
- *Jacobson, V., Nichols, K., & Poduri, K. (1999).* ***An expedited forwarding PHB****. Internet Engineering Task Force (IETF RFC 2598). https://doi.org/10.17487/RFC2598*
- *Al-Rawi, M., Wahsheh, H., & Binsalleeh, H. (2019).* ***Detecting distributed denial-of-service attacks using machine learning techniques****. Journal of Information Security and Applications, 46, 136–148. https://doi.org/10.1016/j.jisa.2019.03.003*
- *Sommer, R., & Paxson, V. (2010).* ***Outside the closed world: On using machine learning for network intrusion detection****. IEEE Symposium on Security and Privacy, 305–316.* https://doi.org/10.1109/SP.2010.25
- *Buczak, A. L., & Guven, E. (2016).* ***A survey of data mining and machine learning methods for cyber security intrusion detection****. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.* https://doi.org/10.1109/COMST.2015.2494502
- *Yu, S., Zhou, W., Doss, R., & Jia, W. (2013).* ***Traceback of DDoS attacks using entropy variations****. IEEE Transactions on Parallel and Distributed Systems, 22(3), 412–425. https://doi.org/10.1109/TPDS.2010.169*
- *Chandola, V., Banerjee, A., & Kumar, V. (2009).* ***Anomaly detection: A survey****. ACM Computing Surveys, 41(3), 1–58.* https://doi.org/10.1145/1541880.1541882
- *Ahmed, M., Mahmood, A. N., & Hu, J. (2016).* ***A survey of network anomaly detection techniques****. Journal of Network and Computer Applications, 60, 19–31.* https://doi.org/10.1016/j.jnca.2015.11.016
- *Breiman, L. (2001).* ***Random forests****. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324*
- *Chen, T., & Guestrin, C. (2016).* ***XGBoost: A scalable tree boosting system****. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.* https://doi.org/10.1145/2939672.2939785
- *Kreutz, D., Ramos, F. M., Verissimo, P., Rothenberg, C., Azodolmolky, S., & Uhlig, S. (2015).* ***Software-defined networking: A comprehensive survey****. Proceedings of the IEEE, 103(1), 14–76.* https://doi.org/10.1109/JPROC.2014.2371999
- *Pahl, C. (2015).* ***Containerization and the PaaS cloud****. IEEE Cloud Computing, 2(3), 24–31. https://doi.org/10.1109/MCC.2015.51*
- *Newman, S. (2015).* ***Building microservices: Designing fine-grained systems****. O'Reilly Media.*
- *Richardson, L., & Ruby, S. (2007).* ***RESTful web services****. O'Reilly Media.*
- *Goodfellow, I., Bengio, Y., & Courville, A. (2016).* ***Deep learning****. MIT Press.* https://www.deeplearningbook.org
- *Kim, G., Lee, S., & Kim, S. (2014).* ***A novel hybrid intrusion detection method integrating anomaly detection with misuse detection****. Expert Systems with Applications, 41(4), 1690–1700. https://doi.org/10.1016/j.eswa.2013.08.066*
- *Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018).* ***A deep learning approach to network intrusion detection****. IEEE Transactions on*

Vol. 10, Issue: 02, February: 2021

ISSN: (P) 2347-5412 ISSN: (O) 2320-091X

Emerging Topics in Computational Intelligence, 2(1), 41–50. https://doi.org/10.1109/TETCI.2017.2772792

- *Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Young, M. (2015).* ***Hidden technical debt in machine learning systems****. Advances in Neural Information Processing Systems (NeurIPS), 2503–2511. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html*