AUTOMATED TESTING METHODOLOGIES FOR DISTRIBUTED WEB SYSTEMS: A MACHINE LEARNING-BASED APPROACH TO ANOMALY DETECTION

METODOLOGII DE TESTARE AUTOMATIZATĂ PENTRU SISTEME WEB DISTRIBUITE: O ABORDARE BAZATĂ PE ÎNVĂȚAREA MAȘINĂ A DETECȚIEI ANOMALIILOR

CZU: 004.738.52:004.8 **https://doi.org/10.59295/spd2024n.102**

Marian ILEANA,
ORCID: 0009-0008-9624-2202
National University of Science and Technology Politehnica, Bucharest, Romania;
Pitesti University Center, Pitesti, Romania

Summary. *In the era of modern technologies, distributed web systems play an important role in global IT infrastructures, providing scalability and resilience. However, verification, validation, and testing of these increasingly complex systems remain a major challenge for both specialists working in the field and researchers who want to explore this field. Due to their heterogeneous nature and dynamic interdependencies between existing system components, this important stage of system testing and validation has high complexity. In this context, the paper explores automated testing based on machine learning techniques for anomaly detection in distributed web systems. The methodology used uses advanced machine learning algorithms such as Convolutional Neural Networks, Recurrent Neural Networks and Long-Short Term Memory, K-means Clustering, and Random Forest to analyze data flows and interactions between microservices in real time. It aims to develop a robust framework that includes modules for data collection and preprocessing, model training and validation, as well as an automatic alert system for detected anomalies. Experimental results demonstrate that the proposed approach outperforms traditional testing methods by increasing accuracy and reducing fault identification time.*

Keywords*: Anomaly Detection, Automated Testing, Distributed Web Systems, Machine Learning, Test Automation.*

INTRODUCTION

Distributed web systems are essential to modern digital infrastructures [1], supporting critical services such as healthcare [2] and financial applications [3]. These systems have experienced significant growth in recent years, being a fundamental paradigm that supports complex and scalable applications. These systems can also meet the highest standards for low electricity consumption [4], so they can be said to be environmentally friendly. These systems are characterized by interconnected components distributed across various platforms and geographic locations that provide scalability and resilience, but they also present significant challenges in terms of verification, validation, and testing.

However, automated testing of distributed web systems remains a significant challenge due to their distributed nature, heterogeneity, and intrinsic dynamics. Traditional rule-based automated testing approaches struggle to keep up with the complexity and dynamism of modern distributed web systems, often resulting in insufficient test coverage and a high false positive rate [5].

This article proposes a machine learning (ML)-based approach for anomaly detection within these systems, providing an efficient solution for automated testing of these complex systems [6]. Using supervised and unsupervised machine learning techniques, the proposed model aims to be able to learn from previously provided test data and identify patterns of normal behavior. Deviations from these methods are then classified as potential anomalies, providing valuable insight into potential problems and failure encountered within distributed web systems [7]. The paper includes a description of the proposed architecture. The architecture explains the operation of the distributed web system. The collection of data from the logs and how they will be processed by the chosen machine learning algorithm. And finally, a comparison between normal data and anomalies will be made. It also presents a detailed methodology for implementing the machine-learning-based approach. In addition, we discuss future challenges and opportunities in automated testing of distributed web systems based on machine learning [8].

The proposed ML-based approach contributes significantly to the field of automatic testing of distributed web systems, providing an efficient and adaptable method for testing these systems that have a high degree of complexity [9]. This method can significantly support their testing, improving the quality and reliability of these vital systems. Through this work, we want to open new perspectives for the integration of ML technologies in software quality assurance [10].

Review of previous work in the field

An article published by Wang, Poovendran, and Manokaran [11] highlights the use of machine learning technologies for fault detection in the power system. The system includes electricity, gas, heat, and other various energy sources. It plays a crucial role in power supply and fault detection during dynamic processing. This paper addresses several key challenges, including collaborative planning, tracking optimization, threat assessment, condition assessment, situational awareness, and overall demand management. In addition, the research highlights the significant practical and strategic importance of the integrated energy system using machine learning technology for studies and practices related to China’s energy system development. The experimental results indicate that the integrated energy system - machine-learning achieves a high accuracy of 98.67% in the detection and control of faults within the integrated energy system.

Research by Zhang et al. [12], indicates that by predicting failures in critical IT services is essential to prevent downtime and ensure reliability. Traditional methods using console logs for post-incident analysis, while recent term frequency - inverse document frequency (TF-IDF) text mining techniques create high-dimensional features that are

not easily generalizable. Their proposed system automatically processes streamed documents, groups similar patterns, and treats them as words in documents to reduce the dimensionality of features. Using Long Short-Term Memory (LSTM) networks, real data from several enterprise logs demonstrate the superior performance of our system in predicting IT failures, pioneering the use of LSTM in this context.

A dynamic analysis was carried out on the subjects of distributed web systems and machine learning, with the specialized literature offering a multitude of specialized works and having a large number of authors. In this research, an empirical investigation was conducted using the database provided by Dimensions.ai, and the platform's search engine was used to search for relevant keywords. The results, returned as a CSV file, were uploaded and processed using the Vos Application Viewer desktop application to facilitate analysis by authors researching these topics. Figure 1 illustrates that the groups obtained from the data include a total of 10478 authors, which are groups of authors from an author network.

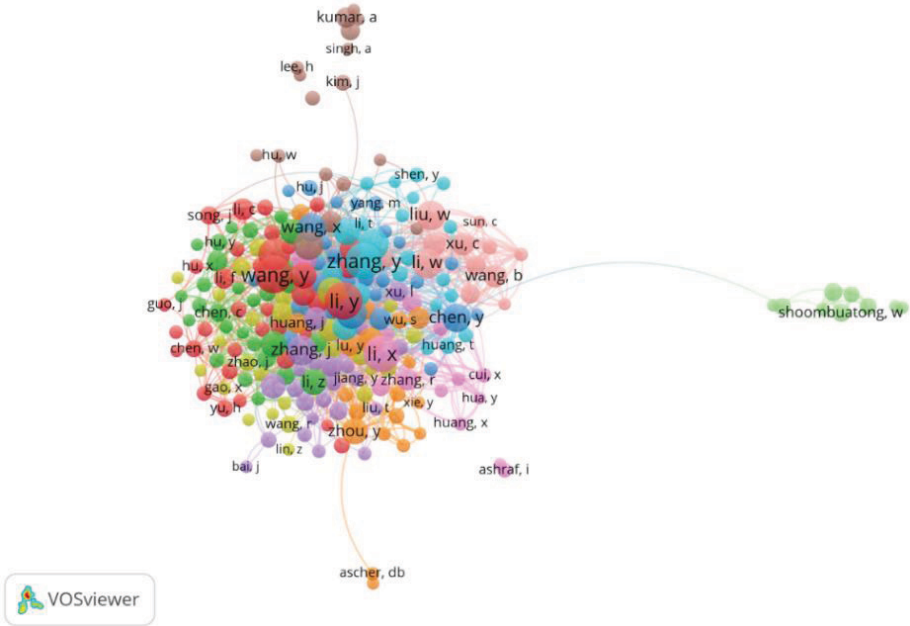


Fig. 1. A graphic map of authors for the search term "distributed web system and machine learning"

The processed data includes a total of 10478 authors. To be considered relevant, an author had to have written at least five papers and had at least five citations. A total of 244 authors met these criteria (Fig. 1).

Using the same data set, an analysis was made of the countries where the authors who publish most frequently on the topic of distributed web systems and machine lear-

ning come from. Thus, a total of 108 authors was obtained (Fig. 2). For a country to be considered relevant, it must have at least five published papers and at least five citations. Thus, 69 countries met the criteria.

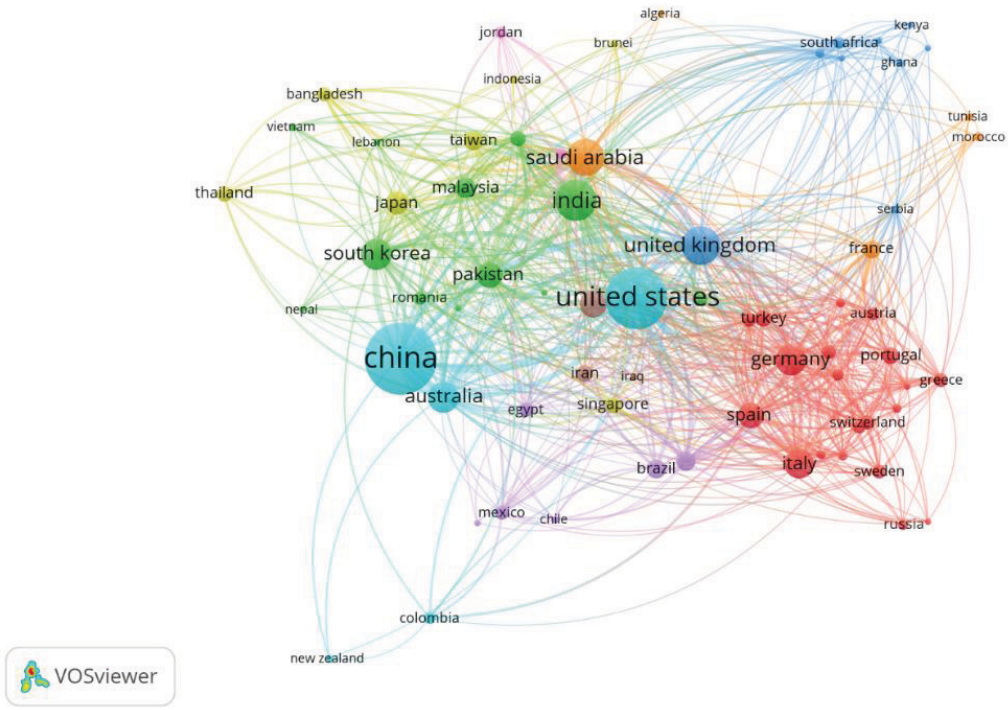


Fig. 2. A country map of authors for the search term “distributed web system and machine learning”

Material and method

Automated testing of distributed web systems is a complex and challenging task due to their dynamic, scalable, and heterogeneous nature [13]. Traditional automated testing approaches such as unit testing, integration testing, and functional testing can be expensive in terms of time and resources. Also, traditional types of testing can pose scalability issues within large and complex distributed web systems [14].

Fig. III illustrates a distributed web system that integrates machine learning for anomaly detection. A client sends an HTTPS request to www.example.com, which is initially resolved by a local DNS server, yielding the IP address 181.12.34.2 [15].

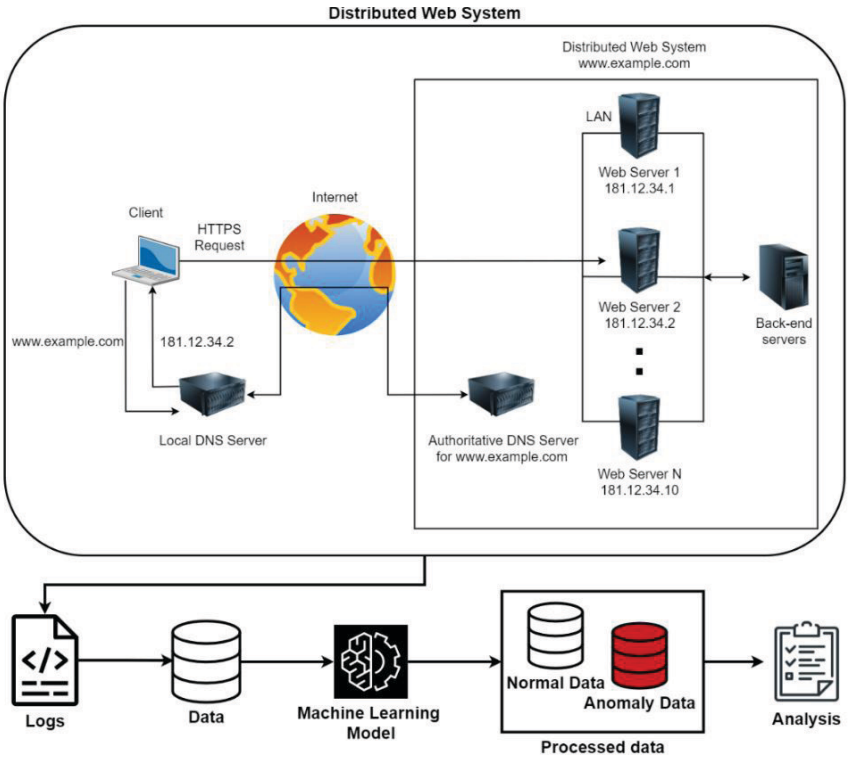


Fig. 3. Distributed Web System with Machine Learning Analysis for Anomaly Detection

The request is then sent over the Internet to the distributed web system, where the authoritative DNS server for www.example.com directs the request to one of the internal web servers (Web Server 1, Web Server 2, ..., Web Server N, with IP addresses of 181.12.34.1 to 181.12.34.10) connected to a local area network (LAN) [16]. These web servers process the request and communicate with the back-end servers to provide the requested data. The logs generated by these interactions are collected and stored, then analyzed using a machine learning model. This machine learning model classifies data into abnormal and normal, and then analysis is used to evaluate system performance and security, facilitating the detection and prevention of potential problems [17].

Choosing a machine learning algorithm according to the system requirements we have. You can choose between Convolutional Neural Networks, Recurrent Neural Networks and Long-Short Term Memory, K-means Clustering, and Random Forest [18]. Each of these has advantages and disadvantages. Convolutional Neural Networks can be successfully applied to log data, they will treat logs as data sequences and learn relevant features through 1D convolutions [19]. Using Recurrent Neural Networks and LSTMs, they are trained to detect normal access patterns and anomalies based on temporal sequences in logs, helping to identify unusual behaviors in data streams [18]. K-means

clustering: K-means clustering will vectorize logs according to their characteristics, then K-means will divide the data into clusters, and logs that do not fit into any cluster will be considered anomalies. Using the Random Forest algorithm, the algorithm uses multiple decision trees trained on different subsets of the log data to achieve robust classification, and abnormal logs are identified by the majority of decision trees [20].

Results and discussion

The machine learning-based approach significantly improved the accuracy and efficiency of anomaly detection in distributed web systems, demonstrating a robust ability to identify both known anomalies and unknown anomalies from logs.

CONCLUSIONS

Automated testing methodologies based on machine learning provide a powerful tool to improve the reliability and security of distributed web systems through effective anomaly detection. This approach not only optimizes testing and maintenance processes but also helps reduce incident response time and minimize operational risks, demonstrating the significant value of integrating emerging technologies into distributed systems management.

REFERENCE

1. TALAVER, V.; VAKALIUK, Tetiana A. Reliable distributed systems: review of modern approaches. *Journal of Edge Computing*, 2023, 2.1: 84-101.
2. ILEANA, Marian. Elevating Medical Efficiency and Personalized Care Through the Integration of Artificial Intelligence and Distributed Web Systems. In: *International Conference on Intelligent Tutoring Systems*. Cham: Springer Nature Switzerland, 2024. p. 3-11.
3. HUANG, Huawei, et al. When blockchain meets distributed file systems: An overview, challenges, and open issues. *IEEE Access*, 2020, 8: 50574-50586.
4. ILEANA, Marian. Optimizing Energy Efficiency in Distributed Web Systems. In: *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*. IEEE, 2023. p. 1-5.
5. SÆTHER, Didrik. *Security in Docker swarm: Orchestration service for distributed software systems*. 2018. Master's Thesis. The University of Bergen.
6. PANDEY, Sushant Kumar; MISHRA, Ravi Bhushan; TRIPATHI, Anil Kumar. Machine learning based methods for software fault prediction: A survey. *Expert Systems with Applications*, 2021, 172: 114595.
7. KONDA, Sreedhar Reddy; SHAH, Varun. Machine Learning-Enhanced Software Development: State of the Art and Future Directions. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, 2022, 6.4: 136-149.

8. VERBRAEKEN, Joost, et al. A survey on distributed machine learning. *Acm computing surveys (csur)*, 2020, 53.2: 1-33.
9. BRAIEK, Houssem Ben; KHOMH, Foutse. On testing machine learning programs. *Journal of Systems and Software*, 2020, 164: 110542.
10. MAYER, Ruben; JACOBSEN, Hans-Arno. Scalable deep learning on distributed infrastructures: Challenges, techniques, and tools. *ACM Computing Surveys (CSUR)*, 2020, 53.1: 1-37.
11. WANG, Peng; POOVENDRAN, Parthasarathy; MANOKARAN, Karthik Bala. Fault detection and control in integrated energy system using machine learning. *Sustainable Energy Technologies and Assessments*, 2021, 47: 101366.
12. ZHANG, Ke, et al. Automated IT system failure prediction: A deep learning approach. In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016. p. 1291-1300.
13. KILINCER, Ilhan Firat; ERTAM, Fatih; SENGUR, Abdulkadir. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks*, 2021, 188: 107840.
14. LIU, Kaijun, et al. A review of android malware detection approaches based on machine learning. *IEEE access*, 2020, 8: 124579-124607.
15. ILEANA, Marian; OPROIU, Maria Ioana; MARIAN, Constantin Viorel. Using Docker Swarm to Improve Performance in Distributed Web Systems. In: *2024 International Conference on Development and Application Systems (DAS)*. IEEE, 2024. p. 1-6.
16. KULESZA, Raoni, et al. Evolution of web systems architectures: a roadmap. *Special Topics in Multimedia, IoT and Web Technologies*, 2020, 3-21.
17. DOBRICA, Liliana. ADDRESSING CHALLENGES IN MODELING REAL-TIME APPLICATIONS FOR COMPLEX DISTRIBUTED SYSTEMS.
18. [MAHESH, Batta. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 2020, 9.1: 381-386.
19. SARKER, Iqbal H. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2021, 2.3: 160.
20. PANESAR, Arjun; PANESAR, Arjun. Machine learning algorithms. *Machine learning and AI for healthcare: Big data for improved health outcomes*, 2021, 85-144.