A systematic review on resolving service level agreement violation in IT Services

Bungaran Adil Paruhum Siagian[1], Kusworo Adi[2], Adi Wibowo[3]
[1]Doctoral Program of Information Systems, [2,3]Faculty of Science and Mathematics
Diponegoro University, Semarang, Indonesia
[1]adilsiagian@students.undip.ac.id, [2]kusworoadi@lecturer.undip.ac.id,
[3]adiwibowo@lecturer.undip.ac.id.

Abstract—Information Technology (IT) services are crucial for modern organizations, making Service Level Agreements (SLAs) essential for ensuring quality and performance. However, SLA violations remain an unavoidable and significant challenge in IT service management, often leading to financial losses and reputational damage. The increasing complexity of IT landscapes, driven by cloud computing, IoT, and AI, complicates both root cause identification and resolution. This systematic literature review (SLR) identifies, evaluates, and synthesizes existing research on SLA violation resolution, with a particular focus on machine learning (ML) and deep learning (DL) methods. The study aims to identify relevant technology contexts, examine main problems in SLA violations, and determine how ML/DL methods are effectively used in their resolution. Findings indicate a strong research focus on cloud computing, with emerging attention to edge computing and IoT, and highlight the diverse application of ML/DL in mitigating SLA breaches.

Keywords—Deep learning, IT services, Machine learning, Multi-context, Service Level Agreement, Violation.

I. INTRODUCTION

The development of information technology in recent decades has been rapid, marked by the utilization of various new technology services. The simultaneous use of cloud-based services, the Internet of Things, and artificial intelligence provides greater benefits than using each technology separately. These greater benefits have encouraged almost all large and medium-sized enterprises to adopt information technology in running their business operations to increase efficiency and productivity. However, the growing diversity of technology use has resulted in the increasingly complex management of information technology. In addition, the potential for service failure from these complex systems is increasing. Thus, while the integration of diverse information technologies offers significant gains in efficiency and productivity, it simultaneously introduces new complexities and heightened risks of system failure that demand careful management [1].

Most large companies utilize IT services provided by service providers using SLAs to mitigate the risk of service failure and ensure the continuity of IT services. An SLA is not only a service contract between the provider and a customer but also a comprehensive framework that establishes performance standards and service quality benchmarks for customer satisfaction. Failure to meet SLAs has serious consequences, ranging from financial penalties for violations to customer dissatisfaction and reputational damage, which can ultimately lead to the termination of service contracts [2]. Therefore, it is important for IT service providers to maintain a commitment to service quality that meets SLA requirements.

A prior systematic review conducted by Khan et al. examined SLA indicators pertinent to the quality of service within cloud computing services [3]. The quality-of-service (QoS) parameters observed included packet loss, availability, throughput, latency, and response time. SLA violations were detected through various machine learning techniques such as Naive Bayes, Random Forest, and Support Vector Regression. In addition, the study identified QoS remediation techniques that achieve efficient resource utilization, dynamic scalability, SLA compliance (in terms of response time, throughput, and optimal costs) and adaptability to workloads changes using optimization algorithms, integer programming, heuristics, and machine learning. Specifically in the cloud computing, this study observed SLA violations through QoS violation detection as well as modeling and techniques that support the management and handling of these violations. However, it did not evaluate the effectiveness of various techniques in reducing SLA violations.

In a different study, Patel and Chaurasia conducted a systematic review of energy consumption when migrating virtual machines in a dynamic cloud computing environment [4]. Several algorithms, such as a combination of adaptive threshold algorithms, minimum migration time policy, gradient descent regression, and Modified Best Fit Decreasing, were elaborated to obtain efficient VM migration. This study concluded that an energy-aware, adaptive threshold-based VM migration algorithm can effectively optimize VM resource usage, reducing energy consumption while maintaining service performance, thereby reducing the potential for SLA violations. Although it discusses SLA violations, the focus of this study was more on studying energy management than specifically addressing in the violations on themselves.

Pandey and Rawat also conducted a systematic review study on virtual machines [5]. They observed various methods of provisioning VMs in different environments, namely static, dynamic, and metaheuristic, and their impact on various SLA indicators. To prevent SLA violations, the dynamic model is considered the most effective due to its adaptability to runtime conditions and real-time workload changes. While the metaheuristic model also yields favorable results in optimization and violation reduction, it occasionally faces limitations due to its computational complexity. Conversely, the static model struggles to manage variability and adjust to fluctuating loads, which increases the likelihood of SLA violations. This study underscores the importance of using adaptive and automated provisioning methods to deal with load variations and avoid SLA violations.

Although previous SLR studies have explored SLA violation in cloud computing, none have addressed this issue in the context of different IT services contexts (such as cloud, network, and IoT) simultaneously. This study addresses a gap in existing literature regarding multi-context environments, which has not been thoroughly explored. Considering multi-context environments is critical, as actual IT services are often delivered across multi-context environments. The objectives of this study are to examine SLA violations and their causes in a multi-context environment, elaborate on the use of ML and DL to overcome these problems, and identify the effectiveness of ML and DL models in overcoming SLA violations. The results of this study are expected to provide a more comprehensive understanding of SLA violations in IT services in a multi-context and serve as a basis for companies to formulate end-to-end strategies for managing IT service SLAs.

II. METHOD

This study adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. PRISMA is a framework commonly used in SLR because it provides systematic methodological guidance in identifying, screening, and assessing the eligibility of relevant literature for the purposes of the study [6].

A. Identification

This study uses the PICO (Population, Intervention, Comparison, Outcome) framework to limit and clarify the focus of the research. The limitations of the focus of this SLR research are as follows:

- Population: Various main contexts of IT service environments, such as cloud computing, network infrastructure, data centers, the Internet of Things (IoT), and cybersecurity systems, may experience SLA violations.

- Intervention: ML and DL are the methods chosen to detect, predict, and resolve incidents.
- Comparison: This study does not include comparison groups.
- Outcome: Reduction in SLA violations and improvement in overall system performance

Specifically, this research aims to answer the following research questions (RQs):

- RQ1: What is the scope of the technology context in SLA violations in IT services?
- RQ2: What are the main SLA violation problems in IT services?
- RQ3: How are machine learning and deep learning methods effectively used in resolving SLA violations?

  This study uses across academic databases, namely Elsevier Scopus, IEEE, and Science Direct. The search strategy used combined several related terminologies, namely violations, SLA, machine learning, deep learning, and several IT service environments such as cloud computing, data centers, network security, and IoT. The article search is limited to publications from 2020 to June 30, 2025. Duplicates were removed to ensure the uniqueness and relevance of the articles selected for further analysis.

B. Screening & Eligibility

After the identifying phase, the next step is an initial screening of each article found in the search results. The initial screening is conducted on the titles and abstracts of the articles.

The inclusion criteria include:

- Articles categorized as Q1, Q2, Q3, and Q4 in the Scimago Journal Rank.
- Articles categorized as A*, A, B, or C by CORE (Computing Research and Education Association of Australasia).
- Studies that incorporated machine learning and deep learning techniques

The exclusion criteria include:

- Studies that do not use ML and DL techniques.
- Articles that are not publicly accessible in full.
- Articles that do not meet the minimum ranking category.

Articles that have passed the screening criteria are then assessed using eligibility criteria. At this stage, the full text of articles is thoroughly evaluated to ensure their suitability. The eligibility criteria:

- The study population covers IT services domains such as cloud computing, IoT, networks, security, and their derivatives.
- The study includes indicators of SLA violations.
- The research dataset is mentioned or available.
- Quantitative model results related to SLA violations are available.

A study that meets at least the three criteria above is considered eligible, while a study that fails to meet the minimum eligibility criteria will be excluded from this study. This rigorous review aims to ensure that only the most relevant and methodologically sound papers are included.

C. Included

The articles that meet all eligibility criteria will be included for data extraction and subsequent analysis. The selected set of 29 articles will undergo a systematic process of data extraction, analysis, and synthesis to answer the research questions defined in the initial stage.

The entire selection process, from identification to inclusion, was transparently documented in a PRISMA flow diagram (Fig.1).

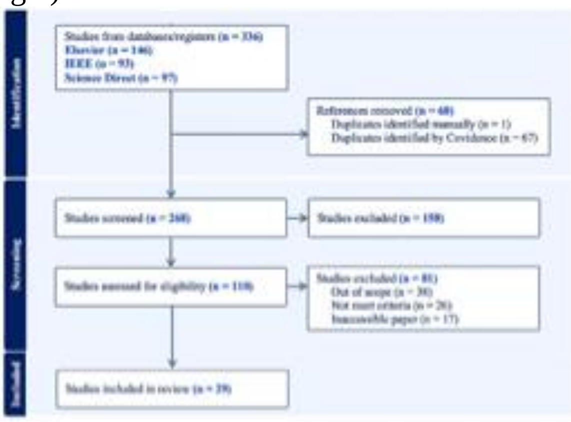

Figure 1. Prisma Flow Diagram on Resolving SLA Violation

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14284

III. RESULTS AND DISCUSSION

A. The scope of the technology context in SLA violations in IT Services

Analysis of the different technology categories and study contexts reveals that most research on resolving SLA violations focuses on resolving SLA violations on cloud computing (Table 1). This is understandable given cloud environments are so complex and constantly changing— they require balancing tasks like autoscaling services, VM consolidation, and container management to keep everything running smoothly. Since cloud systems serve many users at once and need to adjust resources dynamically, ensuring SLAs are a challenging task. Therefore, most research emphasizes on finding smart ways to handle resources and scheduling within clouds to avoid SLA problems.

Edge computing and IoT services are also getting research attention. Even though they're not as extensively studied as cloud computing, these domains present unique challenges from distributing computing power closer to where data is generated—on devices and at the network's edge. Managing workflows and scaling in these decentralized setups, as well as dealing with varied IoT devices, makes keeping SLAs intact trickier. Researchers are starting to focus more on these environments because they require fresh approaches to manage latency, device diversity, and distributed processing, which differ quite a bit from traditional cloud setups. Network services and network virtualization appear less frequently in the literature but present a unique challenge on infrastructure. They manage the fundamental aspects of guaranteeing service quality within the network, especially with technologies like 5G and network slicing. These areas remind us that SLAs don't just depend on the cloud or edge—they also need solid network performance and management. Even though fewer studies cover these topics, they highlight how important it is to combine network-level solutions with the higher-level service management to truly keep SLAs from being broken. Finally, Service-Based Systems (SBSs) receive limited attention but point to issues around managing SLAs at the application or service orchestration layer. This suggests that beyond infrastructure, the quality-of-service composition and orchestration is also crucial. While it's a less-explored this area, it suggests there's room to grow in understanding how to prevent SLA violations from the software composition side.

Table 1 The scope of technology context in IT services

No	Technology Category	Study Context Environment	Ref.
1	Cloud Computing	Autoscaling and scalability	[7], [8], [9], [10], [11], [12]
		VM consolidation and placement	[3], [13], [14],[15], [16],[17], [18]
		Resource scheduling and provisioning	[19], [20]
		Application performance and specific issues	[21], [22], [23]
2	Edge Computing	Scheduling and autoscaling in edge-cloud environments	[24], [25]
3	IoT Services	VM placement, scheduling, and QoS for IoT applications	[26], [27], [28]
4	Network Services	Network Slicing	[29], [30], [31]
		Traffic data and QoS in mobile networks	[32], [33]
5	Network Virtualization	Virtual Network Functions (VNF)	[34]
6	Service-Based Systems	QoS in Service-Based Systems	[35]

In summary, most SLA research focuses heavily on cloud computing because of its complexity and vital role, with edge computing and IoT rising as newer frontiers full of interesting challenges. Network services and service-based systems add important supporting perspectives. This all suggests that handling SLA violations requires understanding the demands of each technology environment.

B. The main SLA Violation problems in IT Services

The main problems associated with SLA violations in IT services (summarized in Table 2) can be categorized as follows:

1. Inefficient resource management and elasticity control.
   The inability of systems to manage resources and perform adaptive scaling in response to workload dynamics is a common cause of problems in cloud computing. Delays in scaling lead to under-provisioning and over-provisioning of resources. Meanwhile, static threshold-based autoscaling has proven inadequate in responding to temporal changes, thereby increasing SLA violations [8]. To overcome this, advanced methods such as Deep Reinforcement Learning (DRL) with LSTM and Actor-Critic, as well as Q-learning have been applied in studying historical data and real-time conditions to make optimal scaling and resource allocation decisions [8], [9], [20]. Studies [29] and [31] have applied deep learning and machine learning approaches to predict resource requirements proactively in 5G networks. These approaches improve resource utilization, reduce operational costs, and significantly reduce SLA violations.
2. Complex Inter-Service Dependency and Interaction Failures.
   This problem often occurs in the management of microservices-based applications, where complexity and dependencies between services are very high. Resource scheduling processes that only considers one service without accounting for the cascading effects on other services often triggers SLA violations [7], [10], [21]. To overcome this problem, dynamic modeling and consideration of service dependencies using Graph Neural Networks and adaptive graph learning have been able to mitigate the occurence of SLA violations [7]. Similarly, for predicting SLA violations, supervised learning methods based on GNN and Bayesian Regularization Neural Network have been used, which have been proven to improve the accuracy of early detection of violations in systems with complex dependencies [21], [23].
3. Suboptimal Workload Orchestration and Placement.
   Inefficiencies in task scheduling, workload orchestration, and VM placement often occur in cloud and edge infrastructure. Previous studies [17], [19], [25], [26] have shown that un-orchestrated workloads can lead to uneven workloads, excessive energy consumption, and escalated risk of host overload. To address this, cutting-edge methods such as deep learning and reinforcement learning were used to optimize resource orchestration and placement. For example, Deep Q-Network (DQN) has been applied in dynamic task scheduling, while MetaNet utilized surrogate neural networks for QoS prediction and dynamic policy selection [19], [26]. Meanwhile, in edge-cloud workflows, the MCDS algorithm (combining MCTS, deep surrogate, and graph convolution) resulted in smarter and more adaptive scheduling [25].
4. Inadequate Anomaly Detection and Predictive Capabilities.
   This problem is triggered by the system's limitations in detecting operational anomalies and proactively predicting potential SLA violations due to changes in workload and uncertain conditions. Previous studies found that traditional detection methods often fail to proactively anticipate changes in traffic, overload, or other anomalies and typically only function after an SLA violation occurs [15], [22], [27], [32], [34]. Local Outlier Factor (LOF), Isolation Forest, and One-Class SVM were applied for adaptive traffic change detection and model

retraining timing [32]. Deep learning methods, such as LSTM, GAN, and RNN, have been developed for long-term workload and anomaly behavior prediction [15], [22], [27].

5. Unpredictable External Dynamics and Network Conditions.
These challenges are triggered by external dynamics and unpredictable network conditions, such as mobile traffic fluctuations, performance interference, and sudden changes in load and resource availability [11], [12], [24], [30], [33]. These conditions impact resource scaling or orchestration to be delayed or inaccurate, resulting in SLA violations due to response times exceeding thresholds or unexpected resource exhaustion. To overcome this problem, time series-based prediction methods (LSTM, ARIMA) were used for traffic and resource demand forecasting [12], [33]. In another study, adaptive online machine learning and Markov Decision Process were applied for real-time auto-scaling decision-making in edge-cloud systems [24]. Constrained reinforcement learning methods (PPO, Three-way Decision RL) were also developed to ensure scaling remains adaptive to external dynamics and maintains near-zero SLA violations during online learning [11], [30].

6. Overhead and Disruption from Management Operations.
This issue undertakes overhead and disruptions during infrastructure management operations such as consolidation, migration, and VM rescheduling. Previous studies presented that overly frequent or inefficient VM migration processes can cause performance degradation [13], [16], [18], [28]. In addition, energy savings through VM consolidation often result in a trade-off between reduced energy consumption and increased risk of SLA violations. To mitigate these impacts, Deep Reinforcement Learning (DRL) was used in conjunction with Influence Coefficient-based VM selection to choose the most appropriate VMs for migration, thereby minimizing overload and SLA violations [13].

Table 2 The main SLA problems in IT Services

No	Main Problem	Technology Category	Method	Ref.
1	Inefficient resource management and elasticity control.	Cloud Computing, Network Service	Deep Reinforcement Learning (DRL), Q-learning, LSTM, Actor-Critic, Hybrid RL-ML	[8], [9], [14], [20], [29], [31]
2	Complex inter-service dependency and interaction failures.	Cloud Computing	Graph Neural Network (GNN), Adaptive Graph Learning, DeepScaler, DRL, Bayesian Regularisation.	[7], [10], [21], [23]
3	Suboptimal workload orchestration and placement.	Cloud Computing, Edge Computing	Deep Q-Network (DQN), DRLBTSA, MetaNet, A3C, MCDS, Monte Carlo Tree Search, Surrogate Models	[17], [19], [25], [26]
4	Inadequate anomaly detection and predictive capabilities.	Cloud Computing, Network Virtualization IoT Service, Service-Based System.	Unsupervised Learning (LOF, OC-SVM, Isolation Forest), LSTM, GAN, RNN, Supervised ML	[15], [22], [27], [32], [34], [35]
5	Unpredictable external dynamics and network conditions	Cloud Computing, Edge Computing, Network Service	LSTM, ARIMA, Online ML, Markov Decision Process, PPO, TWD-RL	[11], [12], [24], [30], [33]
6	Overhead and disruption from management operations	Cloud Computing, IoT Service	DRL, SVR, Kalman Filter, Q-learning	[13], [16], [18], [28]

Table 3. The best model used

No	Best Model	Main Problem Resolved	Dataset	Violation Reduction	Ref.
1	Deep Q-Network (DQN) based model.	Inefficient resource management and elasticity control	Simulated 5G network slicing datasets, synthetic load scenarios.	50%	[29]
2	Q-Learning (Reinforcement Learning).		Real-world workload logs used include ClarkNet and Google Traces	78%	[20]
3	DeepScaler—GNN with EM adaptive graph learning.	Complex inter-service dependency and interaction failures	Real-time telemetry data	41%	[7]
4	DL with surrogate.	Suboptimal workload orchestration and placement.	Real-life AI-based benchmark applications	13%	[26]
5	Monte Carlo Deep Surrogate (MCDS) Scheduler.		WFCommons workflow generator dataset; real-world workloads such as BLAST, Montage scientific workflows	45%	[25]
6	VTGAN combined with power-aware best fit decreasing placement.	Inadequate anomaly detection and predictive capabilities.	Real workload traces from PlanetLab	85%	[15]
7	TWD-RL.	Unpredictable external dynamics and network conditions	Google Cloud real workload data; synthetic workload patterns	25%	[11]
8	Algorithm Q-learning with ε-greedy.	Overhead and disruption from management operations	Realistic workflow traces and actual platforms for experiments	91%	[28]

dominance of machine learning and deep learning technologies for real-time prediction, detection, and automated decision-making. These methods have proven to not only improve operational efficiency but also provide an adaptive foundation for the complexity, dynamics, and uncertainty that characterize today's digital IT services.

C. The best model used in resolving SLA violations

Based on Table 3, the SLA violation reduction varies considerably from a low of 13% to a high of 91%. For instance, the Deep Learning (DL) model with surrogate achieved only a 13% reduction, while the Three-Way Decision-Based Reinforcement Learning (TWD-RL) and DeepScaler GNN with EM adaptive graph learning models demonstrated moderate reductions of approximately 25% and 41%, respectively [7], [11].

Conversely, models such as Value and Trend Generative Adversarial Network (VTGAN) and the Q-learning with ϵ-greedy algorithm showcased considerably higher effectiveness in reducing SLA violations, at 85% and 91%, respectively [15], [28]. This wide range indicates that there is no "one-size-fits-all" solution for addressing SLA violations; rather, outcomes are highly contingent upon the specific problem context and details.

Differences in outcomes can be attributed to some factors. First, the varying complexity of the problems in each study contributes to the different outcomes. Models built on specific and controlled cases such as overloaded hosts or VM performance degradation during migration (e.g., VTGAN and the Q-learning with ϵ-greedy algorithm) tend to provide better SLA violation reduction results [15], [28]. Meanwhile, models that handle multi-faceted problems such as complex service dependencies and cascading effects in microservices (e.g., DeepScaler and Monte Carlo Deep Surrogate Scheduler) produce more moderate outcomes [7], [25]. Second, the definitions and different metrics used to assess SLA violation reductions in each study also influence the variability in outcomes.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14284

Based on the results obtained, ML and DL methods have proven to be very effective in solving main SLA problems. These methods prevent IT systems identify potential violations and facilitate proactive measures, including dynamic resource allocation and intelligent scaling mechanisms, which prevent violations from occurring in the first place. In addition, various types of reinforcement learning (RL) facilitate better decision-making in allocating and scheduling resources. As a result, ML and DL methods enhance SLA compliance and operational reliability.

IV. CONCLUSION

This systematic literature review has comprehensively addressed the research questions related to Service Level Agreement (SLA) violations in IT services. This study exposes that research on SLA violations frequently focused on the cloud computing, where issues such as autoscaling, VM consolidation, and resource management are central. Furthermore, edge computing and IoT services are gaining attention, where the issues studied include distributed computing, latency, and device diversity. Network services (including network slicing and mobile networks) and service-based systems (focused on application-level quality and orchestration) are less discussed but introduce important perspectives.

The main SLA violation problems are classified into six distinct areas: inefficient resource management and elasticity control; complex inter-service dependency and interaction failures; suboptimal workload orchestration and placement; inadequate anomaly detection and predictive capabilities; unpredictable external dynamics and network conditions; and overhead and disruption from management operations. Machine Learning (ML) and Deep Learning (DL) methods are predominantly used for proactive prediction and intelligent resource management to resolve SLA violations. These techniques accommodate the means to anticipate breaches by forecasting workloads and resource demands, leading to proactive interventions like dynamic resource allocation and smart scaling. However, the effectiveness of these models varies significantly depending on the specific problem context and details.

Despite the multi-contextual analysis, significant research gaps remain, especially regarding standardized datasets spanning multiple IT service domains and in-depth study of security-related SLA violations. Future research should aim to address these gaps to further improve SLA management across diverse IT environments.

REFERENCES

[1] A. A. Nacer, O. Perrin, and F. Charoy, “Identification of Comparison Key Elements and Their Relationships for Cloud Service Selection,” in *Service-Oriented and Cloud Computing*, vol. 12054, A. Brogi, W. Zimmermann, and K. Kritikos, Eds., in Lecture Notes in Computer Science, vol. 12054. , Cham: Springer International Publishing, pp. 74–82,2020.

[2] X. Zeng *et al.*, “Detection of SLA Violation for Big Data Analytics Applications in Cloud,” *IEEE Transactions on Computers*, vol. 70, no. 5, pp. 746–758, May 2021.

[3] H. M. Khan, F. F. Chua, and T. T. V. Yap, “A Review on Quality of Service Monitoring, Violation and Remediation for the Cloud,” *Journal of System and Management Sciences*, vol. 13, no. 5, pp. 107–126, 2023.

[4] A. Patel and N. Chaurasia, “A Systematic Review of Energy Consumption and SLA Violation Conscious Adaptive Threshold based Virtual Machine Migration,” in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, Jalandhar, India: IEEE, pp. 39–44, May 2021.

[5] M. C. Pandey and P. S. Rawat, “Efficient Virtual Machine Provisioning Techniques in Cloud Computing: A Systematic Review,” in *2023 16th International Conference on Security of*

Information and Networks (SIN), Jaipur, India: IEEE, pp. 1–11, Nov. 2023.
[6] Matthew J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *Syst Rev*, vol. 10, no. 1, p. 89, Dec. 2021.
[7] C. Meng, S. Song, H. Tong, M. Pan, and Y. Yu, "DeepScaler: Holistic Autoscaling for Microservices Based on Spatiotemporal GNN with Adaptive Graph Learning," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Luxembourg, Luxembourg: IEEE, pp. 53–65, Sept. 2023.
[8] P. S. Shruthi and D. R. Umesh, "Hybrid Scaling in Cloud Computing for Optimizing Resource Utilization: An LSTM-Enhanced Actor-Critic Learning Model," *Journal of Internet Services and Information Security*, vol. 15, no. 1, pp. 468–485, Feb. 2025.
[9] H. Yang, L. Pan, and S. Liu, "Faster or Cheaper: A Q-learning based cost-effective mixed cluster scaling method for achieving low tail latencies," *Future Generation Computer Systems*, vol. 157, pp. 264–274, Aug. 2024.
[10] C. Meng, J. Tong, M. Pan, and Y. Yu, "HRA: An Intelligent Holistic Resource Autoscaling Framework for Multi-service Applications," in *2022 IEEE International Conference on Web Services (ICWS)*, Barcelona, Spain: IEEE, pp. 129–139, July 2022.
[11] C. Jiang, G. Mao, and B. Xie, "Three-way decision-based reinforcement learning for container vertical scaling," *Information Sciences*, vol. 708, 2025.
[12] P. Souza *et al.*, "Predicting and Avoiding SLA Violations of Containerized Applications using Machine Learning and Elasticity," 2022. [Online]. Available: https://orcid.org/0000-0003-4945-3329
[13] J. Zeng, D. Ding, K. Kang, H. M. Xie, and Q. Yin, "Adaptive DRL-Based Virtual Machine Consolidation in Energy-Efficient Cloud Data Center," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2991–3002, Nov. 2022.
[14] B. Wang, F. Liu, and W. Lin, "Energy-efficient VM scheduling based on deep reinforcement learning," *Future Generation Computer Systems*, vol. 125, pp. 616–628, Dec. 2021.
[15] A. I. Maiyza, H. A. Hassan, W. M. Sheta, K. Banawan, and N. O. Korany, "VTGAN based proactive VM consolidation in cloud data centers using value and trend approaches," *Sci Rep*, vol. 15, no. 1, p. 20133, June 2025.
[16] U. Arshad, M. Aleem, G. Srivastava, and J. C.-W. Lin, "Utilizing power consumption and SLA violations using dynamic VM consolidation in cloud data centers," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112782, Oct. 2022.
[17] P. Wei, Y. Zeng, B. Yan, J. Zhou, and E. Nikougoftar, "VMP-A3C: Virtual machines placement in cloud computing based on asynchronous advantage actor-critic algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 5, p. 101549, May 2023.
[18] M. Awad, N. Kara, and A. Leivadeas, "Utilization prediction-based VM consolidation approach," *Journal of Parallel and Distributed Computing*, vol. 170, pp. 24–38, Dec. 2022.
[19] S. Mangalampalli, G. R. Karri, M. Kumar, O. I. Khalaf, C. A. T. Romero, and G. A. Sahib, "DRLBTSA: Deep reinforcement learning based task-scheduling algorithm in cloud computing," *Multimed Tools Appl*, vol. 83, no. 3, pp. 8359–8387, Jan. 2024.
[20] R. Panwar and M. Supriya, "RLPRAF: Reinforcement Learning-Based Proactive Resource Allocation Framework for Resource Provisioning in Cloud Environment," *IEEE Access*, vol. 12, pp. 95986–96007, 2024.
[21] A.-C. Maroudis, T. Theodoropoulos, J. Violos, A. Leivadeas, and K. Tserpes, "Leveraging Graph Neural Networks for SLA Violation Prediction in Cloud

Computing," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 605–620, Feb. 2024.

[22] S. Tuli, S. S. Gill, P. Garraghan, R. Buyya, G. Casale, and N. Jennings, "START: Straggler Prediction and Mitigation for Cloud Computing Environments using Encoder LSTM Networks," *IEEE Transactions on Services Computing*, pp. 1–1, 2021.

[23] A. Pandita, P. K. Upadhyay, and N. Joshi, "Prediction of Service-Level Agreement Violation in Cloud Computing Using Bayesian Regularisation", pp. 231–242, 2021.

[24] T. P. da Silva, A. R. Neto, T. V. Batista, F. C. Delicato, P. F. Pires, and F. Lopes, "Online machine learning for auto-scaling in the edge computing," *Pervasive and Mobile Computing*, vol. 87, p. 101722, Dec. 2022.

[25] S. Tuli, G. Casale, and N. R. Jennings, "MCDS: AI Augmented Workflow Scheduling in Mobile Edge Cloud Computing Systems," Dec. 14, 2021, *arXiv*: arXiv:2112.07269.

[26] S. Tuli, G. Casale, and N. R. Jennings, "MetaNet: Automated Dynamic Selection of Scheduling Policies in Cloud Environments," in *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*, Barcelona, Spain: IEEE, pp. 331–341, July 2022.

[27] N. Staifi and M. Belguidoum, "ViolationPredictor: a Solution for Predicting SLA Violations of IoT Applications," 2022. [Online]. Available: http://ceur-ws.org

[28] X. Ma, H. Xu, H. Gao, M. Bian, and W. Hussain, "Real-Time Virtual Machine Scheduling in Industry IoT Network: A Reinforcement Learning Method," *IEEE Trans. Ind. Inf.*, vol. 19, no. 2, pp. 2129–2139, Feb. 2023.

[29] H. Chergui and C. Verikoukis, "Offline SLA-Constrained Deep Learning for 5G Networks Reliable and Dynamic End-to-End Slicing," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 2, pp. 350–360, Feb. 2020.

[30] Q. Liu, N. Choi, and T. Han, "OnSlicing: Online End-to-End Network Slicing with Reinforcement Learning," in *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*, pp. 141–153, Dec. 2021.

[31] N. Yarkina, A. Gaydamaka, D. Moltchanov, and Y. Koucheryavy, "Performance Assessment of an ITU-T Compliant Machine Learning Enhancements for 5G RAN Network Slicing," *IEEE Trans. on Mobile Comput.*, vol. 23, no. 1, pp. 719–736, Jan. 2024.

[32] V. Gudepu, V. R. Chintapalli, P. Castoldi, L. Valcarenghi, B. R. Tamma, and K. Kondepu, "Adaptive Retraining of AI/ML Model for Beyond 5G Networks: A Predictive Approach," in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, Madrid, Spain: IEEE, pp. 282–286, .June 2023.

[33] E. Tuna, A. Baykal, and A. Soysal, "Multivariate and multistep mobile traffic prediction with SLA constraints: A comparative study," *Ad Hoc Networks*, vol. 163, p. 103594, Oct. 2024.

[34] J. Hong, S. Park, J.-H. Yoo, and J. W.-K. Hong, "Machine Learning based SLA-Aware VNF Anomaly Detection for Virtual Network Management," in *2020 16th International Conference on Network and Service Management (CNSM)*, IEEE, pp. 1–7, Nov. 2020.

[35] S. Xie, "Optimizing Self-Adaptation in Service-Based Systems: Leveraging Ensemble Prediction with DNN-ILSTM Models," presented at the Proceedings - 2024 IEEE International Conference on Software Maintenance and Evolution, ICSME 2024, Institute of Electrical and Electronics Engineers Inc., pp. 843–847, 2024.