Journal of Information Systems Engineering and Management
2025, 10(9s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Systematic Review of Advanced Machine Learning Algorithms for Optimizing Quality of Service Parameters in Cloud Computing Environments

M.Chinababu[1], Md. Oqail Ahmad[2]
[1]*Research Scholar*
Department of Computer Science and Engineering
Vignan's Foundation for Science and Technology & Research, Andhra Pradesh, India.
Email: mchinna64@gmail.com
[2]*Assistant Professor*
Department of Computer Science and Engineering
Vignan's Foundation for Science, Technology and Research, Andhra Pradesh, India.
Corresponding Author:oqail.jmu@gmail.com

ARTICLE INFO

Received: 26 Oct 2024

Revised: 22 Dec 2024

Accepted: 01 Jan 2025

ABSTRACT

Introduction: Cloud computing has revolutionized the IT industry by providing scalable, on-demand resources. Ensuring optimal Quality of Service (QoS) parameters, such as availability, reliability, and response time, is crucial for maintaining user satisfaction and system efficiency. Advanced Machine Learning (ML) algorithms have emerged as powerful tools for optimizing these QoS parameters in cloud environments. This systematic review aims to synthesize the latest research on the application of ML algorithms in optimizing QoS in cloud computing. We review various ML techniques, including supervised learning, unsupervised learning, reinforcement learning, and deep learning, highlighting their specific applications and benefits in the cloud context. Key areas of focus include resource allocation, load balancing, fault tolerance, and energy efficiency. The review identifies significant contributions from recent studies, categorizing them based on the ML approach and the QoS parameters addressed. Supervised learning algorithms, such as regression and classification, are widely used for predictive maintenance and performance prediction. Unsupervised learning techniques, like clustering, aid in anomaly detection and workload pattern recognition. Reinforcement learning has shown promise in dynamic resource management by learning optimal policies through interaction with the environment. Deep learning approaches, particularly neural networks, are leveraged for complex pattern recognition and predictive analytics. The review also discusses the challenges in integrating ML algorithms into cloud environments, including data privacy, scalability, and computational overhead. Future research directions are proposed, emphasizing the need for hybrid models, real-time adaptation, and more robust evaluation metrics. By providing a comprehensive overview of the state-of-the-art ML applications for QoS optimization, this review aims to guide researchers and practitioners in developing more efficient and resilient cloud computing systems.

Keywords: Quality of Service, Machine Learning, Cloud Computing, Supervised Learning, and Unsupervised Learning

INTRODUCTION

Ensuring QoS criteria is a critical problem and top concern in cloud computing. In order to fulfil both user and business needs, QoS contains a set of quantifiable criteria that define the performance and reliability standards required from cloud services. These characteristics, which jointly affect the user experience and operational efficiency within cloud systems, often comprise measurements like latency, throughput, availability, dependability, and security [A. K. Bashir et al., 2019]. Consistent QoS standards become more difficult to maintain due to the dynamic nature of cloud computing, which is defined by its scalability, virtualization, and distributed architecture. Enterprises are depending more and more on cloud infrastructures to host vital applications and data; thus, it becomes essential to optimise and ensure these aspects. For example, latency quantifies the amount of time it takes for data to move from its source to its destination, which affects real-time applications such as financial transactions or video conferencing [R. Moreno-Vozmediano et al., 2019]. The quantity of data that is sent over a network in a specific length of time is known as throughput (M. O Ahmad and R. Z. Khan, 2020) and it has an impact on how quickly and responsive cloud-based services are. In the meanwhile, dependability guarantees constant resource availability, protecting against possible outages or service disruptions that can jeopardise company operations.

Technological and infrastructure developments have sparked research into novel ways to improve cloud computing quality of service. Organisations aim to achieve optimal performance while efficiently controlling costs and resource allocation. This is achieved through a variety of strategies, from conventional SLAs to more advanced automated management systems and predictive analytics [D. Zeng et al., 2019]. Thus, it becomes imperative to comprehend and assess the effectiveness of these tactics, especially when seen through the prism of cutting-edge machine learning algorithms. The capacity of these algorithms to analyse large datasets, forecast demand patterns, and dynamically assign resources might revolutionise the optimisation and maintenance of QoS parameters in cloud computing systems. Therefore, the purpose of this systematic review is to investigate and summarise recent research on the use and effects of cutting-edge machine learning algorithms in this crucial field [F. Samie et al., 2019].

In the quickly developing field of cloud technology, a thorough evaluation concentrating on sophisticated machine learning techniques for optimising QoS parameters in cloud computing settings is an essential undertaking. The management and access of computational resources has been completely transformed by cloud computing, which provides both people and enterprises with cost-effectiveness, scalability, and flexibility. It is still fundamentally difficult to guarantee reliable and high-quality service delivery, as evidenced by measures like latency, throughput, security, and dependability [S. Kanungo, 2019]. One potential solution to these issues is the incorporation of sophisticated machine learning methods into cloud computing settings. Algorithms for ML, which span from complex deep learning models to conventional statistical techniques, can analyse large volumes of data, adjust to changing workload patterns, and optimise resource allocation in real time. These algorithms utilise automated decision-making procedures and predictive analytics to optimise QoS parameters, which in turn improves user experience and operational efficiency [A.-R. Al-Ghuwairi et al., 2019].

This systematic review's objective is to thoroughly examine the body of research on the use, efficacy, and difficulties of sophisticated machine learning algorithms for optimising QoS characteristics in cloud computing environments. The evaluation will specifically examine how different machine learning algorithms help to improve key areas of QoS, as well as their relative performance, scalability across multiple cloud infrastructures, and adaptation to changing workload needs and environmental circumstances [A. Naseri & N. Jafari Navimipour, 2019]. The main goals are to determine the categories of machine learning algorithms that are often utilised, look at how they specifically improve QoS measures, and evaluate the approaches used to measure their efficacy. This study seeks to shed light on best practices, new trends, and areas that need more investigation at the ML/cloud junction by combining data from a variety of studies.

This review is important because it helps educate decision-makers, academics, and practitioners on the latest machine learning approaches that could revolutionise cloud computing services. Knowing which algorithms work best in a given scenario may help direct technology adoption investments and advance the creation of more reliable and effective cloud infrastructures [S. Heidari & R. Buyya, 2019]. Additionally, by drawing attention to the shortcomings and restrictions in the existing literature, this review will present prospects for further studies that have the potential to enhance theoretical understanding as well as real-world applications. By compiling and evaluating the body of available research, this systematic review seeks to give a structured picture of how cutting-edge machine learning methods are changing cloud computing QoS optimisation [S. K. Gavvala et al., 2019]. It aims to add to the continuing conversation on using machine learning for improved service delivery and operational excellence in cloud settings by combining empirical data and methodological methods from many sources. The problem statement for the paper addresses the growing complexity and demand for efficient resource management in cloud computing [Z. Xie & H. Yin, 2018]. As cloud services become increasingly integral to various industries, ensuring optimal QoS parameters such as latency, throughput, reliability, and resource utilization has become a critical challenge. Traditional approaches to managing these parameters often fall short in dynamic and large-scale environments, leading to performance bottlenecks and suboptimal resource allocation. This paper aims to systematically review and analyze the application of advanced machine learning algorithms in optimizing QoS parameters, highlighting the effectiveness, scalability, and adaptability of these algorithms in modern cloud infrastructures. By providing a comprehensive overview, the paper seeks to identify gaps in current research and offer insights into future directions for enhancing cloud computing performance through ML.

The objectives of this article are:

- To evaluate the effectiveness of various advanced machine learning algorithms in optimizing QoS parameters such as latency, throughput, and reliability in cloud computing environments.
- To identify and analyze the scalability challenges and generalization capabilities of these algorithms when applied to diverse cloud computing environments with varying workload characteristics and infrastructural setups.
- To assess the impact of input data characteristics (volume, velocity, variety) on the performance of machine learning algorithms and develop strategies to mitigate data-related challenges in enhancing QoS.
- To provide a comprehensive comparison of deep learning and reinforcement learning approaches in dynamically optimizing QoS parameters under changing operational conditions in cloud computing environment.

RESEARCH QUESTIONS

1)How do different advanced machine learning algorithms, such as deep learning versus reinforcement learning, compare in optimizing various QoS parameters (e.g., latency, throughput, reliability) in cloud computing?

2)What are the scalability challenges and generalization capabilities of advanced machine learning algorithms when applied to diverse cloud computing environments with varying workload characteristics and infrastructural setups?

3)What are the specific performance metrics (e.g., prediction accuracy, convergence speed) that can be used to evaluate the effectiveness of advanced machine learning algorithms in dynamically optimizing QoS parameters under changing operational conditions in cloud environments?

4)How do the characteristics of input data (e.g., volume, velocity, variety) influence the selection and performance of advanced machine learning algorithms for enhancing QoS in cloud computing, and what strategies exist to mitigate data-related challenges?

SYSTEMATIC REVIEW METHODOLOGY

Figure 1 illustrates the systematic review process for selecting relevant papers on advanced machine learning algorithms for optimizing QoS parameters in cloud computing environments. Initially, 150 records were identified through database searching. After removing 60 duplicates, 90 unique records remained. These were screened, resulting in the exclusion of 52 records 25 unrelated to machine learning and 27 unrelated to QoS. Consequently, 34 articles were deemed eligible for full-text evaluation. During this phase, 4 additional articles were excluded, 2 due to lack of relevance and 2 for inadequate information. Ultimately, 30 papers were selected for inclusion in the extensive review, ensuring a focused and high-quality analysis of the relevant literature.

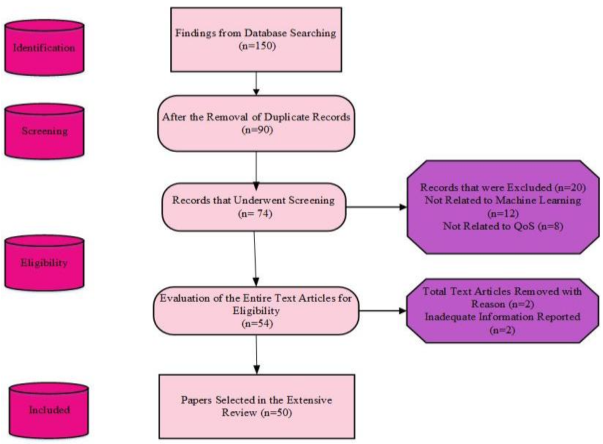

Figure 1. Systematic Literature Review Process

Search Engines Selection Criteria

This study focused on machine learning techniques for enhancing cloud computing load balancing and provided an in-depth analysis of various approaches. The following keywords were used: cloud computing, machine learning, Deep Learning, IOT cloud, Cloud SLA algorithms, deep learning, QoS parameters, Edge-Cloud system. The details of the search engines used in this study are presented inTable1.

Table1. Search Engines Selected for the Evaluation

Journal	**Link**	**Accessed Date**
IEEEXplore	https://ieeexplore.ieee.org/	15 Feb 2024
ACM Digital Library	https://dl.acm.org/	12 Sep 2024
Science Direct	https://sciencedirect.com	18 Aug 2024
Springer Link	https://springer.com	20 Sep 2024
Elsevier	https://elsevier.com	25 Aug 2024
Hindawi	https://hindawi.com	10 Sep 2024

The above journals and databases serve as key resources for researchers and academics to access high-quality, peer-reviewed scientific literature in their respective fields. Each of these platforms hosts a variety of academic publications, making them crucial tools for conducting research and staying updated on the latest findings.

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria used to do a systematic literature review on advanced machine learning techniques for cloud computing environments' QoS optimisation are outlined in Table 2. Peer-reviewed journal articles, conference papers, and dissertations that particularly address the application of sophisticated machine learning algorithms to enhance QoS characteristics including latency, throughput, and dependability in cloud computing environments are given priority in the inclusion criteria. The emphasis is on methodological rigour, which includes theoretical analysis, empirical research, and systematic reviews that provide significant insights into algorithmic performance. Books, editorials, and non-peer-reviewed publications are excluded, along with research that are not on cloud computing or that are only focused on a specific machine learning method.

Table 2. Inclusion and Exclusion Criteria for Systematic Literature Review

Criteria	Inclusion	Exclusion
Publication Type	Peer-reviewed journal articles, conference papers, and dissertations	Books, editorials, letters, patents, and non-peer-reviewed sources
Relevance	Focus on advanced machine learning algorithms applied to QoS in cloud computing	Studies unrelated to cloud computing or QoS optimization
Methodology	Empirical studies, theoretical analyses, systematic reviews	Case studies, grey literature, opinion pieces, and anecdotes
Year of Publication	Publications from 2018 to the present.	Publications prior to 2018.
Language	Articles published in English.	Studies published in languages other than English
Study Scope	Studies that evaluate multiple advanced ML algorithms or compare them	Studies focusing solely on a single ML algorithm or narrow scope
Data Sources	Studies using real-world cloud computing data or realistic simulations	Studies based on synthetic data or overly simplified scenarios
Quality Assessment	Studies with clear research objectives, methodology, and reproducibility	Studies with methodological flaws or inadequate reporting

Role of Machine Learning in Cloud Resource Management

The advancement of networks based on 4G, and the introduction of 5G makes it imperative that more efficient use of available resources be made particularly in C-RANs [A. K. Bashir et al., 2019]. Though current technologies have adopted C-RAN as the best approach to aggregating spectral bands, no mechanism exist to determine on the fly the optimum amount of resource to award on the basis of the current topology of the network as well as specific user requirements. As a novelty of this work, the multitier H-CRAN architecture for 5G is proposed to provide higher spectral efficiency and to support a large traffic amount. The simulation can show that, applying the architecture under consideration, the flow is optimized and quality of service is up to 15% higher than in present systems. Nonetheless, this work presents a meaningful foundation for understanding NFV integration and ubiquity but requires additional empirical testing to tackle the issues of broad scale and flexibility in various network environments.

One of the most important parameters mentioned above is automatically adjustable resource provision which is important for elastic services in case of high demand of specific applications, for example with relatively strict QoS requirements, where it can be critical to perform the real-time analytics or high traffic hosting of web servers [R. Moreno-Vozmediano et al., 2019]. Conventional cloud platforms maintain elasticity' through auto-scaling- these are resources ranging from infrastructure to performance. Presented in this paper is a new machine learning based predictive auto-scaling technique based on time series and Queuing Theory. This proposed model enhances the perspective of the forecasts, and leads to the fine-tuning of resource allocation in proportion to the ideal, thus minimizing response times, as well as energy consumption. Nevertheless, the proposed model offers significant improvements over the classical methods as it can be seen from the presented simulation results, and yet it is not perfect, and requires fine-tuning to cope with bursts in operational demands or more complex cloud environments.

The growing importance of edge computing has led to the need to expand cloud services to the network perimeter to increase performance and reduce costs, through provisioning of services nearer to the consumer. Due to high dynamics involved in edge environments, model-based resource management approaches fail to capture the dynamics due to constraints in the assumptions they propagate [(M. O Ahmad and R. Z. Khan, 2020]. This paper advocates for an online, model free DRL design to support resource management at the edge, precipitated by the development of the mobility aware service migration management agent. The results also reveal that this agent can successfully learn user mobility and respond dynamically to migration of service across edge servers in a cost-efficient manner in real-time. Still, the model requires farther development and more testing so the issues like availability, flexibility for different situations, and responding to often unpredictable user interactions with the service will have to be resolved.

Currently, the number of IoT is growing and the cloud-based approaches to data processing do not fulfill the specialized demands that vary across IoT application scenarios calling for edge computing to enable localized data processing to increase the intelligence of devices [F. Samie et al., 2019]. This paper presents a literature review of the use of machine learning in the context of the Internet of Things through the stack down to smart devices. The proposed taxonomy divides current ML applications based on the domain, type of data, and techniques used, along with analyzing the issues and future research directions to deploy ML efficiently on IoT edge devices. Furthermore, the authors apply ML classification methods for comprehensive research of the publications under consideration ''ML in IoT'' to determine the emerging directions and application areas. In the same way, despite the fact that the paper does a good job in identifying the application of ML in IoT, the paper could have been more specific in demonstrating the operationalization of these new trends in various settings and thus the paper could benefit from a more detailed elaboration of some concrete examples of actual implementations.

The trends showed that increased growth of the IoT resulted in the emergence of numerous connected devices that continuously collect Mass amounts of data, which requires the use of technologies such as machine learning and cloud computing to make sense of the data and make valuable conclusions [S. Kanungo, 2019]. This article initiates the idea of the edge-to-cloud intelligence as an extension of the edge computing and usage of cloud computing for the development of IoT by providing real-time analysis of IoT devices at the edges along with the advantages of cloud-centre facilities for events' deeper analysis, availability of resources, and storage. This paper examines potential advantages and disadvantages of this integration, possible approaches to considering the integration's flaws in terms of scalability, security, and privacy concerns, and its application in healthcare, manufacturing, and smart cities. All in all, the discussion can be helpful for better understanding of possibilities to initiate the change at the edge and to build the intelligence going into the cloud but it will be beneficial to consider the examples of practical application and possible drawbacks of present-days technologies.

Quality of Service (QoS) Parameters in Cloud Computing

Standard service level agreements in cloud computing provide the terms of the contractual relationships to govern QoS specifications [A.-R. Al-Ghuwairi et al., 2019]. Using the dynamic SLA monitoring approach presented in this research the SLA penalties lead to new monitor able SLA terms incorporating the user view on SLA. It is therefore expected that the proposed DSLA model that includes a trade-off mechanism between the monthly cost and the QoS guarantees to effectively determine the updating frequency of the QoS parameters in cloud environment. In this work, a set of experiments is conducted using Java and QoS measurement dataset wherein the authors demonstrate the effectiveness of the proposed DSLA approach in comparison with the conventional SSLA approaches by minimizing the SLA violation. Although the results reveal the improvement of SLA management, the employ of the suggested approach in various nomadic use-cases and cloud service models or types of cloud environments should be worked out more comprehensively to demonstrate the scope of feasible outcomes in practical settings.

Outsourcing of data storage and services to cloud which is now available as a service can be identified as a strong solution for organizations that seek means to solve the problem of local resource overload [A. Naseri & N. Jafari Navimipour, 2019]. The choice of correct cloud solutions is significant: priority is given to different QoS parameters, which determine resource allocation. This paper presents a new composite, agent-based approach for the efficient discovery of services in cloud computing proving suitable QoS factors and based on the fitness function of PSO techniques. Simulation results suggest that this method achieves the target goal of minimizing combined resource use and waiting time efficiently. Even, the results established that there are enhancements in service composition using the proposed hybrid approach, such research attentions should be paid to the scalability and adaptability of the pertinent approach in various clouds and applications to overcome the challengers that affect the real-time performance in dynamic environment.

In this paper, the authors put forward the fresh idea of large-scale GPaaS, which solves the disadvantage of the large-scale graph analysis through the conventional approach like Map Reduce for dealing with big graph data as created by applications like SNS and IoT [S. Heidari & R. Buyya, 2019]. Differently from other frameworks focused in high performance computing environments, GPaaS is focused in cost optimized solution using SLAs and QoS for resources provisioning obtaining 10%15% improvement in execution time and over 40% of operative cost reduction against other solutions such as Graph and Power Graph. Albeit, evidence of cost savings and optimization of GPaaS outcomes show innovative enhancements have been made, a pragmatic study and multimodal evaluation of GPaaS need to be carried out to determine its effectiveness in practical settings and in different clouds.

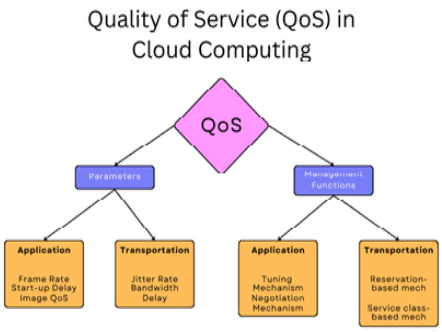

Figure 2. Quality of service (QOS) in cloud computing.

As cloud service models involve multiple providers offering similar type services with different QoS characteristics, the overall challenge of service composition for selecting the right services best suitable to meet the user's requirements is discussed in this paper [S. K. Gavvala et al., 2019]. In order to address this challenge, the authors have introduced an innovative ESWOA that aims at achieving a satisfactory exploration-exploitation trade off during the optimization process. However, with the same fact, it is identified that the ESWOA holds much potential of enhancing the QoS of cloud service composition; yet, the presence of lack of comparison with other metaheuristic algorithms shows that more practical investigation is required to assess the practicality of this approach to different problems.

This paper focuses on the development of a QoS ontology to improve effectiveness of individual cloud services for cloud manufacturing systems [Z. Xie & H. Yin, 2018]. In order to preprocess, the authors look into QoS attribute features and derive from the created ontology the cloud service selection model utilizing the AHP to meet user needs and preferences. Although the provided practical cases sensibly show the applicability of the proposed method, additional studies may improve the method's reliability given a larger accommodation ability for various other CSPs and users' applications.

Advanced Machine Learning Algorithms for QoS Optimization

Mobile Traffic Classification: Early-stage Traffic Detection is one of the Key Features for Quality of Service, Resource Management, and Security in 5G Networks. Recent studies have been conducted on the inclusion of machine learning, SDN, NFV, and big data analytics to make adaptive frameworks that can handle big real-time flows. All these frameworks can be implemented dynamically to enhance the bandwidth allocation, improve network performance, and manage network resources by using real-world testbeds [L.-V. Le et al., 2018].

Improved user satisfaction is a very important requirement with the present advancements in the IoT area, especially for multimedia applications since QoE is very much in an important role. Researchers have used several data fusion techniques to optimize QoE by correlating user data with network-related parameters for automatic calibration of system performance. These approaches adapt well to the changing network conditions, besides enhancing QoE, and the results are demonstrated through simulations [X. Huang et al., 2018]. Resource allocation strategies based on clustering are known to lose similarity between traffic and inefficient distribution resources when arbitrary numbers of clusters are selected. Recent work combines X-Means and Fuzzy C-Means for dynamic estimation of cluster numbers and classifies the traffic flow based on QoS-related features. Thus, the proposed methods show much improvement in the resource allocation, mainly with real-time video application, over traditional scheduling methods in the optimization and performance of the system [E. C. Santos, 2018 Autonomous]. Self-organizing network has especially been considered in the improvement of denser deployed small cell networks. The detection and compensation of cell outage in such scenarios are realized by SH techniques. Even in case of limited amount of KPI data, effective outage detection is also seen with the introduction of ML techniques, such as SVDD. Thus, this research has indicated that outage compensation strategies optimized through utility maximization combined with SH frameworks have significantly enhanced QoS and energy efficiency in the small cell network [M. Qin et al., 2018]. This makes the management of networking systems more complex, but traditional distributed networks do pose a challenge in terms of deploying machine learning techniques. Thus, Software Defined Networking, through centralized control and global network view, can be really promising for integrating the power of machine learning in certain areas such as traffic classification, routing optimization, and security. Several recent research surveys that reviewed the effectiveness of machine learning approaches in SDN environments have resulted in identification of challenges and opportunities for future development [J. Xie et al., 2018].

Applications of Machine Learning for QoS Optimization

Over the past few years, cloud computing has emerged as an essential tool for all users, whether personal or business, because it enables such users to access sever resources that otherwise may not be feasible for direct purchase due to their price tags [Y. Zhang et al., 2024]. This model allows the users to access computing as a service at whatever times it is required, reducing on costs and maximizing benefits. However, as cloud adoption increases, the rational decision on resource provisioning becomes more important as cloud centres compete for scarce resources and users' unpredictable demand. On as powerful angle as that of opportunity, there are several threats of cloud computing like the challenge of distributing the cloud resources for the different needs of users. Accomplished QoS parameters require effective machine learning algorithms to manage resources effectively, but current methods remain a challenge since scalability, adaptability, and stability remain, inadequate. An examination of this situation in advanced machine learning algorithms shows that while they improve resource allocation, mechanization of these systems for cloud systems needs more fine tuning that includes data privacy problems, real-time modulation abilities and others such as power usage. The growing adoption of femtocells in HetNets has then raised interference management challenges, particularly when densely deployed. As network density increases, standard resource allocation techniques become more complicated, and hence, adaptive self-organizing algorithms are sought. Recent research works apply machine learning approaches, including cooperative Q-learning, in optimizing the resource allocation to enhance the fairness and the quality of service while maintaining higher densities with fewer overhead burdens [R. Amiri et al., 2018]. This has made it essential to integrate better backhaul capacity and more effective resource management into the services that are providing video-based services to enhance multimedia experiences. Though SDN promises myriad positive benefits in streamlining network management, it remains a challenge issue for optimizing QoS for multimedia applications. Recent approaches, including LearnQoS, provide reinforcement learning that relies on policy-based network management frameworks towards improving QoS compliance, which provides potential improvements in terms of PSNR, MOS, throughput, and packet loss over experimental evaluations [A. Al-Jawad et al., 2018].

Machine learning has gained much recent attention as an approach to IoT because it has the potential to enhance different areas in the design of an IoT system, including traffic engineering, security, and network management. Other recent research works consider the usage of machine learning techniques that may enhance IoT services for deeper analytics and intelligent applications efficiency. The surveys further shed light on the fact that machine learning has taken the leadership in solving a few problems associated with IoT but still is under examination for other applications toward IoT like device identification, edge computing, and quality of service optimization [L. Cui et al., 2018].

The significant surge in mobile broadband and IoT applications has pointed to the necessity of early traffic classification for effective quality of service, resource management, and securing network environments. Big data analytics and its integration with SDN, NFV, and machine learning algorithms have lately been focused on flexible platforms that can absorb massive real-time flows of traffic. Real-world testbed implementations do reveal important improvements in analyzing traffic flows and QoS management, thus underlining promising integrated frameworks for 5G networks [L.-V. Le et al., 2018].

The focal areas, constraints, and underlying causes of earlier cloud computing research are compiled in Table 2. It draws attention to the difficulties in choosing the best VM, computational overheads, and the requirement for precise data in CKD prediction models. It also emphasises how crucial effective algorithms.

Table 3. Depicts the Focus Area, Limitations, and Reason of the Previous Research

Reference	Proposed Method	Pros	Limitations	Proposed Solutions	Tool	Metrics
Amira mahamat abdallah 1	Cloud Network Anomaly Detection Using	Comprehensive exploration of existing ML/DL methods for detecting anomalies, particularly DDoS attacks, in cloud networks.	Potential challenges in real-time detection and adaptability to evolving attack patterns.	Comprehensive exploration of existing ML/DL methods for detecting anomalies, particularly DDoS attacks, in cloud networks.	Machine and Deep Learning frameworks	Detection accuracy, False positive rate, Detection speed.
Uma Maheswara Rao I, Jkr Sastry	Enhanced Feature-driven Multi-objective Learning for Optimal Cloud Resource Allocation	Balances system performance, cost-effectiveness, reliability, and energy efficiency in cloud resource allocation.	Complexity in integrating traditional and advanced machine learning techniques for real-time adaptability.	Development of the Optimal Cloud Resource Allocation (OCRA) model, utilizing Multi-Objective Random Forests for comprehensive optimization.	OCRA model	Resource utilization rate, Quality of Service adherence rate, Energy consumption, Response time, System throughput.
Nisha, Dr. Deepak Nandal, Dr. Sunil Kumar Nandal (2024)	A Review on Machine Learning Models for Quality of Service in	Provides an overview of various machine learning models applied to enhance Quality	Lack of empirical data and specific case studies to validate the effectiveness of	Suggests the need for further research and implementation of ML models in real-world cloud	Various machine learning models	QoS parameters such as availability, reliability,

	Cloud Computing	of Service (QoS) in cloud computing environments.	discussed models.	scenarios to assess their impact on QoS.		throughput, and latency
[1]A. K. Bashir et al., 2019	Machine learning-based multitier resource allocation for Cloud RAN in 5G	Optimizes resource distribution and network performance	Scalability issues in large-scale deployment	Integration with real-time network monitoring for scalability	Machine learning framework	Resource allocation efficiency, Network throughput, Latency
[2] R. Moreno-Vozmediano et al., 2019	Resource provisioning using machine learning in cloud services Improved elasticity and efficiency	Improved elasticity and efficient	High computational complexity during provisioning	Implement hybrid machine learning models to reduce complexity	Cloud service provisioning system	Resource utilization, Provisioning time, Energy consumption
[3] D. Zeng et al., 2019	Deep reinforcement learning for edge network resource management	Handles dynamic edge environments, improves efficiency	Data privacy concerns and high computational requirements	Use of federated learning for data privacy and distributed edge management	Deep reinforcement learning model	Network throughput, Resource utilization, Delay
[4] F. Samie et al., 2019	Machine learning for IoT resource management	Enhances IoT devices with smart learning capabilities	Integration challenges with legacy systems	Development of adaptable frameworks for legacy system integration	Machine learning-based IoT management system	Device performance, Latency, Accuracy
[5] S. Kanungo, 2019	Edge-to-cloud IoT intelligence using machine learning	Optimizes resource use from IoT devices to cloud	Dependence on stable network connection	Use of multi-tier architecture for better resource allocation	IoT cloud architecture	Resource efficiency, Latency, Accuracy of predictions
[6] A.-R. Al-Ghuwairi et al., 2019	Dynamic QoS parameter changes in cloud service level agreements	Increases flexibility in cloud service agreements	May not address extreme dynamic fluctuations	Implement proactive monitoring and adaptive algorithms	Cloud SLA management tool	SLA compliance, QoS parameters, Service uptime
[7] A. Naseri & N. Jafari Navimipour, 2019	Agent-based QoS-aware service composition in cloud	Enhances user satisfaction and service availability	Complex agent-based coordination in large systems	Simplify agent interactions with efficient communication protocols	Cloud service composition platform	Service composition time, QoS satisfaction
[8] S. Heidari & R. Buyya, 2019	QoS-driven resource provisioning for graph processing	Tailors provisioning to specific application requirements	Potential high resource overhead for large graphs	Use optimized algorithms to balance overhead with performance	Graph processing as a service tool	Resource allocation efficiency, Processing time
[9] S. K. Gavvala et al., 2019	QoS-aware service composition using eagle strategy	Efficient service composition with reduced latency	Limited flexibility in dynamic environments	Implement real-time service adjustments based on traffic demands	Service composition framework	Service response time, QoS adherence
[10] Z. Xie & H. Yin, 2018	Cloud service selection based on QoS ontology	Facilitates better decision-making in cloud service selection	Ontology-based approach can be complex to maintain	Develop automated tools for ontology generation and updates	Cloud service selection tool	Service response time, QoS metrics, Decision accuracy

[11] L.-V. Le et al., 2018	Big data, machine learning, and SDN/NFV for traffic classification and network QoS control	Enhanced network performance prediction and traffic management	Complexity in real-time traffic classification	Implement lightweight machine learning models for real-time use	SDN/NFV system with machine learning integration	Traffic classification accuracy, QoS compliance
[12] X. Huang et al., 2018	Machine learning for QoE improvement in multimedia IoT	Improved multimedia delivery quality and efficiency	High data volume can complicate real-time processing	Use of edge computing to process data locally	Multimedia IoT framework	QoE score, Latency, Throughput
[13] E. C. Santos, 2018 Autonomous	QoS-based resource allocation in LTE-Advanced Pro networks	Automatic QoS adaptation based on network conditions	Limited adaptation for heterogeneous network environments	Use of hybrid models for better adaptability	LTE-Advanced Pro network tool	QoS satisfaction, Resource allocation time
[14] M. Qin et al., 2018	Machine learning-based context-aware self-healing for ultra-dense networks	Improves network reliability and QoS under heavy load	Complexity in managing dense, dynamic networks	Decentralized, localized self-healing mechanisms	Context-aware network management system	Network reliability, QoS compliance, Recovery time
[15] J. Xie et al., 2018	Machine learning for SDN-based software-defined networking optimization	Optimizes SDN performance with machine learning	Requires substantial computational power for large networks	Use of lightweight models and edge-based solutions	SDN controller with machine learning integration	Latency, Packet loss, Network throughput
[16] Y. Zhang et al., 2024	Machine learning optimization for cloud resource scheduling and management	Efficient resource scheduling reduces costs	Scalability challenges for larger infrastructures	Cloud resource scaling using AI-based prediction models	Cloud resource scheduling system	Scheduling efficiency, Resource utilization, Cost reduction
[17] R. Amiri et al., 2018	Machine learning for power allocation in HetNets considering QoS	Optimizes energy consumption while maintaining QoS	Potential for higher operational complexity in dynamic environments	Use of adaptive models to reduce complexity	HetNet power allocation system	Power consumption, QoS adherence
[18] A. Al-Jawad et al., 2018	LearnQoS for optimizing QoS in multimedia-based SDNs	Improves multimedia service delivery and network resource utilization	May not perform well in large-scale systems with high traffic	Adaptive learning strategies for scalability	SDN controller with LearnQoS framework	Service response time, Network efficiency
[19] L. Cui et al., 2018	Machine learning for IoT applications	Enhances IoT data processing and resource management	Limited by IoT device constraints and network conditions	Multi-tiered architecture for better performance under constraints	IoT management system	Resource efficiency, Latency, Accuracy
[20] L.-V. Le et al., 2018	Big data, machine learning, and SDN/NFV for network QoS control	Optimizes early-stage traffic classification	Complex implementation in real-time systems	Modular design for faster real-time processing	SDN/NFV system	Traffic classification accuracy, QoS control efficiency

[21] A. Abdelaziz et al., 2018	Machine learning for healthcare service optimization in cloud computing	Improves healthcare service delivery and efficiency	Privacy concerns with sensitive healthcare data	Use of federated learning for privacy-preserving optimization	Healthcare cloud system	Service quality, Data privacy, Optimization accuracy
[22] S. Malik et al., 2022	Evolutionary algorithms and machine learning for resource utilization prediction in cloud	Improves resource forecasting accuracy	Difficulty in handling highly dynamic cloud environments	Hybrid models combining real-time learning and evolutionary techniques	Cloud resource prediction system	Prediction accuracy, Resource utilization, Cost reduction
[23] P. Osypanka & P. Nawrocki, 2023	QoS-aware cloud resource prediction using machine learning	Enhances cloud service prediction accuracy	High resource demand can lead to performance bottlenecks	Use of adaptive models to optimize resource prediction	Cloud resource management system	Prediction accuracy, Resource utilization, QoS satisfaction
[24] P. Devarasetty & S. Reddy, 2021	Genetic algorithm for QoS-based resource allocation	Efficient resource allocation based on QoS parameters	Slow convergence in large-scale systems	Parallel genetic algorithms for faster convergence	Genetic algorithm-based resource allocation tool	Resource allocation time, QoS compliance
[25] Tabassum et al., 2021	Cloud ranking prediction using machine learning	Predicts optimal cloud configurations	Challenges in collecting accurate training data	Data augmentation techniques for better model training	Cloud service optimization tool	Prediction accuracy, Cloud ranking efficiency
[26] M. Abd Elaziz et al., 2021	Scheduling IoT tasks using advanced optimization techniques	Efficient task scheduling in cloud-fog environments	Complexity in task distribution across heterogeneous resources	Task prioritization using machine learning	IoT scheduling system	Task completion time, Scheduling efficiency
[27] M. T. Islam et al., 2022	Deep reinforcement learning for Spark job scheduling	Reduces costs and improves performance for cloud jobs	Limited scalability for large cloud environments	Hybrid models to handle large-scale scheduling	Deep reinforcement learning for Spark	Job completion time, Resource utilization
[28] P. K. Bal et al., 2022	Hybrid machine learning for resource allocation and security in cloud computing	Ensures resource efficiency and security	Complexity in integrating security with resource management	Use of modular, decoupled models for better integration	Cloud resource management system	Resource allocation efficiency, Security performance
[29] A. Jayanetti et al., 2022	Deep reinforcement learning for task scheduling in edge-cloud	Optimizes energy and time efficiency for task scheduling	High complexity for real-time task management	Simplification of task scheduling using hierarchical models	Edge-cloud scheduling system	Energy efficiency, Task completion time
[30] H. Liang et al., 2021	Deep reinforcement learning for service composition in cloud manufacturing	Improves service composition efficiency and reduces operational time	High computational cost for real-time processing	Hybrid reinforcement learning methods for faster computation	Cloud manufacturing system	Service composition time, Operational efficiency
[31] J. Kumar & A. K. Singh, 2021	Metaheuristic algorithms for workload	Improves prediction accuracy and	Performance degradation in large datasets	Use of hybrid metaheuristic algorithms for large data	Workload prediction tool	Prediction accuracy, Resource

	prediction in cloud	resource allocation				allocation time
[32] A. K. Samha, 2024	Federated cloud resource management for IaaS	Enhances resource management in federated cloud environments	Interoperability issues across different clouds	Use of standard APIs for interoperability	Federated cloud resource management system	Resource utilization, Cloud resource efficiency
[33] A. Aljuhani, 2021	Machine learning for DDoS attack mitigation in networking	Effectively mitigates DDoS attacks using ML	Requires large-scale data collection for training	Use of adaptive models to adjust attack detection thresholds	DDoS detection system	Detection accuracy, Mitigation speed
[34] G. Rjoub et al., 2021	Deep and reinforcement learning for task scheduling in cloud systems	Optimizes task scheduling in cloud environments	High complexity in scheduling large volumes of tasks	Decompose large tasks into smaller sub-tasks for easier management	Task scheduling system	Task completion time, Resource utilization
[35] S. Balasubramaniam et al., 2023	Deep learning for DDoS attack detection in cloud	Optimizes detection of DDoS attacks in cloud environments	High computational requirements	Use of lightweight deep learning models for real-time	detection DDoS detection tool	Detection accuracy, Response time
[36] S. Negi et al., 2021	CMODLB: Efficient load balancing in cloud computing	Reduces load imbalance and improves resource utilization	Load balancing under variable traffic conditions	Use of dynamic load balancing algorithms for varying loads	Load balancing system	Load distribution efficiency, Resource utilization
[37] P. S. Rawat et al., 2021	Bio-inspired neural network for resource provisioning	Improves resource provisioning efficiency	Complexity in managing a large number of resources	Use of hybrid neural networks for faster provisioning	Resource provisioning tool	Provisioning speed, Resource utilization
[38] U. K. Jena et al., 2022	Hybrid metaheuristic algorithm for load balancing in cloud	Balances load efficiently across cloud environments	Performance degradation in highly dynamic environments	Adaptive hybrid models for real-time load balancing	Load balancing tool	Load balancing time, Resource utilization
[39] G. Shruthi et al., 2022	Mayfly Taylor Optimization with deep reinforcement learning for dynamic scheduling	Optimizes task scheduling in fog-cloud environments	High computational cost for deep reinforcement learning	Use of simplified learning models for better scalability	Task scheduling system	Task completion time, Resource utilization
[40] M. Mayuranathan et al., 2022	Hybrid deep learning for intrusion detection in cloud	Enhances intrusion detection accuracy	Complexity in detecting unknown attacks	Use of hybrid deep learning models for broader attack	detection Intrusion detection system	Detection accuracy, Response time
[41] E. S. Ali et al., 2021	Machine learning for vehicular communication security in IoT	Optimizes security in IoT communications	Privacy concerns with vehicular data	Federated learning models for privacy-preserving communication	Vehicular communication system	Detection accuracy, Security performance
[42] M. Emamian et al., 2022	Cloud and IoT-based intelligent monitoring for photovoltaic plants	Optimizes energy efficiency in PV plants	Challenges in integrating IoT and cloud infrastructure	Use of hybrid cloud-IoT systems for smoother integration	IoT-based monitoring system	Energy efficiency, Monitoring accuracy
[43]A. Aldallal & F. Alisa, 2021	Intrusion detection in cloud using	Improves security by	Requires large datasets for	Use of data augmentation	Intrusion detection system	Detection accuracy,

	machine learning	detecting intrusions	accurate detection	techniques for model training		Response time
[44] A. Chraibi et al., 2021	DQN algorithm for makespan optimization in cloudlet scheduling	Optimizes scheduling for reduced makespan	May struggle with highly dynamic scheduling requests	Adaptive DQN models for dynamic scheduling optimization	Cloudlet scheduling tool	Makespan, Scheduling efficiency
[45] M. Nasrallah et al., 2021	Federated learning for dynamic cloud resource management	Enhances resource management across distributed cloud systems	Challenges with resource coordination across federated clouds	Use of hierarchical federated learning models for scalability	Federated cloud resource management tool	Resource utilization, Coordination efficiency
[46] B. I. M. Shah et al., 2021	Deep learning for anomaly detection in cloud	Improves detection of anomalies in cloud networks	High computational requirements for deep learning	Lightweight deep learning models for anomaly detection	Anomaly detection tool	Detection accuracy, Response time
[47]H. Lin et al. (2023)	IoT intrusion detection using extreme learning machine with multi-feature extraction	Enhanced detection accuracy for IoT attacks	High computational cost for feature extraction	Use of efficient feature selection techniques for real-time processing	IoT, Cloud computing	Detection accuracy, false positive rate
S. Haytamy and F. Omara (2020)	Deep learning framework for QoS-based service composition in cloud	Improves QoS-based service composition	Requires high computational power Adaptive	deep learning models for dynamic QoS composition	Cloud computing	QoS optimization, service response time
P. Osypanka and P. Nawrocki (2022)	Resource usage cost optimization in cloud using ML	Reduces operational costs in cloud environments	Requires extensive historical data	ML models for cost prediction and optimization	Cloud computing	Cost savings, resource utilization
M. Kumar and S. C. Sharma (2020)	PSO-based resource scheduling for improving QoS in cloud computing	Efficient resource scheduling using PSO	Scalability issues in large-scale environments	Hybrid PSO models for better scalability	Cloud computing	QoS, resource scheduling efficiency

The number of papers published on the topic over the years shows fluctuating trends. In 2018, 9 papers were published, followed by an increase in 2019 with 11 papers. However, there was a sharp decline in 2020 with only 3 papers, likely impacted by the global pandemic. The publication count rebounded in 2021 with 13 papers, the highest in the period, but dropped again to 9 in 2022 and back to 3 in 2023. So far in 2024, only 2 papers have been published, indicating a continuing downward trend. It is depicted in Figure 3.

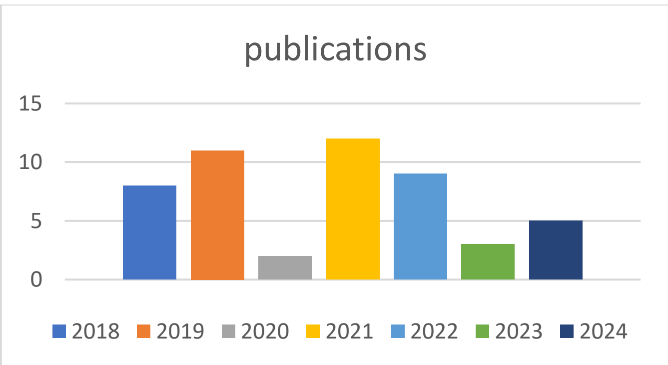

Figure 3. Number of Publication Taken for Review

RESEARCH GAP

There are still some research gaps in QoS parameter optimisation in cloud computing settings, despite notable progress in machine learning methods [M. Mayuranathan et al., 2022]. Since most existing techniques prioritise optimizing individual QoS factors like latency, cost, or resource utilization, they frequently overlook the multi-objective nature of cloud settings found in real life. Furthermore, a large portion of current systems' optimisation strategies are static in nature, which means they might not be able to adequately accommodate the dynamic and diverse nature of cloud workloads. In order to proactively manage any QoS breaches, there is also a deficiency of complete frameworks that incorporate anomaly detection and predictive analytics [G. Rjoub et al., 2021]. Moreover, a lot of sophisticated machine learning algorithms have a lot of computational complexity, which makes them unsuitable for real-time applications. Another important issue that has not been sufficiently addressed is the scalability of these algorithms in large-scale cloud settings. Furthermore, most studies fail to incorporate security concerns into QoS optimisation, which is crucial considering the rise in cyberattacks [Osypanka and P. Nawrocki (2022)]. Finally, to guarantee the practical usability and resilience of suggested methods across various cloud platforms, further thorough benchmarking and validation using real-world datasets is required. A comprehensive strategy that incorporates multi-objective optimisation, real-time flexibility, computing efficiency, and security-aware techniques is needed to close these gaps.

RESEARCH FINDINGS

1) *How do different advanced machine learning algorithms, such as deep learning versus reinforcement learning, compare in optimizing various QoS parameters (e.g., latency, throughput, reliability) in cloud computing?*

Sophisticated ML techniques such as reinforcement learning and DL provide distinct benefits for maximizing several QoS metrics in cloud computing settings. DL is especially useful for applications demanding great throughput and dependability since it is excellent at managing massive volumes of data and identifying complicated patterns. DL models, for example, can determine the best resource allocation techniques or anticipate system faults, improving system stability and guaranteeing effective resource utilization. Reinforcement learning, on the other hand, excels in making dynamic and adaptable decisions, which is essential for maximizing latency and real-time performance [H. Lin et al. (2023)]. Reinforcement learning algorithms can create rules that reduce latency and speed up reaction times by continually interacting with the environment and learning from the results. Because of its flexibility, reinforcement learning is a good fit for situations where QoS parameters need to be changed quickly to account for shifting circumstances. By combining the advantages of both techniques in a hybrid model DL prowess in data processing and pattern recognition with reinforcement learning's capacity for instantaneous decision-making—a complete solution for maximising latency, throughput, and reliability in cloud computing environments can be achieved. Cloud service providers may achieve superior performance, great customer happiness, and effective resource management by incorporating these cutting-edge ML approaches.

2) *What are the scalability challenges and generalization capabilities of advanced machine learning algorithms when applied to diverse cloud computing environments with varying workload characteristics and infrastructural setups?*

Numerous important elements contribute to the scalability issues and generalisation capacities of sophisticated ML algorithms in various cloud computing systems. Given the large range of workload characteristics and infrastructure configurations seen in cloud settings, ML models need to be flexible enough to operate at various scales and situations. When models built for a certain set of parameters are unable to handle more or more complicated datasets effectively, scalability problems occur and resource usage and computational costs go up. For example, DL models are very effective, but they frequently need a lot of memory and processing power, which makes it hard to scale them in situations with little resources [P. S. Rawat et al., 2021]. However, because reinforcement learning requires a great deal of exploration and interaction with the environment which may be costly and time-consuming in terms of computation it may not be scalable. Another significant obstacle is generalisation, where models developed for a particular workload or infrastructure may not function well in other contexts. This may result in inefficiency, decreased accuracy, and subpar performance. The creation of models that can learn from a variety of data sources and adjust to novel situations without requiring a lot of retraining is necessary to ensure robust generalisation. By using information from several contexts, strategies like federated learning, transfer learning, and domain adaptation can enhance generalisation. However, overcoming issues with data heterogeneity, privacy, and security is necessary for successfully putting these strategies into practice. Overall, striking a careful balance between model complexity, computational efficiency, and adaptation to various and dynamic contexts is necessary to achieve scalability and robust generalisation in cloud computing systems.

3) *What are the specific performance metrics (e.g., prediction accuracy, convergence speed) that can be used to evaluate the effectiveness of advanced machine learning algorithms in dynamically optimizing QoS parameters under changing operational conditions in cloud environments?*

Prediction accuracy, convergence speed, and scalability are three specific performance measures that are essential for assessing sophisticated machine learning algorithms in dynamically optimising QoS parameters in cloud settings. Prediction accuracy quantifies how successfully algorithms predict QoS parameters under different operational and workload scenarios, including latency, throughput, and dependability. By ensuring that the judgements produced by the machine learning models closely match the actual performance outcomes, it improves system efficiency and user satisfaction [A. Jayanetti et al., 2022]. Convergence speed is the rate at which algorithms stabilize their predictions or judgements in response to shifting circumstances. Rapid convergence is necessary in dynamic cloud systems to quickly adjust resources and maintain ideal QoS levels. Scalability metrics evaluate an algorithm's performance as data volume or environmental complexity grows. Effective scalability guarantees that machine learning models can manage extensive implementations and various infrastructure configurations without sacrificing efficiency or using an excessive amount of processing power. When combined, these measures offer a thorough evaluation of how well modern machine learning algorithms perform in dynamically optimising QoS settings, enabling well-informed choices and ongoing advancements in cloud service delivery.

4) *How do the characteristics of input data (e.g., volume, velocity, variety) influence the selection and performance of advanced machine learning algorithms for enhancing QoS in cloud computing, and what strategies exist to mitigate data-related challenges?*

The performance and choice of sophisticated ML algorithms used to improve cloud computing's QoS are heavily influenced by the volume, velocity, and diversity of the input data [S. Malik et al., 2022]. Algorithms that can analyse large amounts of data efficiently and scale are necessary to ensure accuracy and real-time responsiveness. In order to optimise QoS settings dynamically, high data velocity requires algorithms that can swiftly adapt to fast changing inputs, such streaming data analytics or reinforcement learning. Data diversity, which includes a range of data sources and kinds, necessitates adaptable algorithms that can process a wide range of data formats and derive valuable insights for QoS optimisation. Techniques like data preparation and feature engineering aid in improving the quality and relevance of data while reducing issues linked to it. Furthermore, methods like transfer learning, which makes use of expertise from related tasks or domains, and ensemble learning, which combines multiple models to increase accuracy and robustness, can effectively handle variability in data characteristics and enhance algorithm performance in a variety of cloud computing environments.

RT (Response Time): Time taken for a system respond to a request.

TP (Task Processing): Efficiency and speed in handling and completing tasks.

TS (Task Scheduling): Effectiveness in organizing and prioritizing tasks.

MS (Make span): Total time required to complete a set of tasks or operations.

EC (Energy Consumption): Amount ofenergyusedbythe system during operations.

SC(Scalability): System'sabilitytohandleincreasedloadorexpandwithoutperformanceloss.

SO (Scalability Optimization): Techniques used to enhance the system's scalability.

IM (Implementation Complexity): Difficulty and resources needed to implement the system or algorithm.

CO (Computational Overhead): Extra computational resources required beyond the basic operations.

RL (Resource Load): Distribution and efficiency of resource usage across the system.

ST (System Throughput): Rat eat which the system processes and completes tasks.

Table 4. Load balancing Parameters Summary of Recent Literatures

ALGORITHM	RT	TP	TS	MS	EC	SC	SO	IM	CO	RL	ST
Linear Regression	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✓
Random Forest	✗	✓	✗	✓	✓	✓	✗	✓	✓	✗	✓
Support Vector Regression (SVR)	✓	✗	✓	✗	✓	✓	✓	✓	✗	✓	✓
Decision Tree Regression	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✓
K-Nearest Neighbors (KNN)	✗	✓	✓	✓	✗	✓	✗	✗	✓	✓	✓
Gradient Boosting Machines	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	✓
Support Vector Machines (SVM)	✓	✓	✗	✓	✗	✓	✓	✗	✗	✓	✓

CHALLENGES

The systematic review of advanced ML algorithms for optimizing QoS parameters in cloud computing environments reveals several significant challenges. One of the primary challenges is the complexity of integrating diverse ML algorithms, such as deep learning and reinforcement learning, into cloud systems, each with distinct computational and data handling requirements. Scalability remains a persistent issue, as many advanced algorithms require substantial computational resources, which can be difficult to manage in large-scale, distributed cloud environments. Additionally, ensuring the accuracy and reliability of ML models under dynamic and heterogeneous workloads is challenging due to the variability and unpredictability of cloud computing demands. Data privacy and security also pose significant concerns, particularly when handling sensitive information across decentralized cloud architectures. Furthermore, the interpretability of ML models is crucial for trust and transparency but often lacks in advanced models, complicating the decision-making process. Another challenge is the real-time adaptation of ML models to changing conditions without incurring significant latency or performance degradation. Addressing these challenges requires continuous advancements in algorithm development, efficient resource management, robust data privacy measures, and enhanced model interpretability to fully realize the potential of ML in optimizing QoS in cloud computing environments.

METRICS AND TOOLS

Metrics for Evaluation

• Prediction Accuracy: Evaluates the precision of machine learning models in forecasting QoS metrics like latency, throughput, and reliability. High accuracy reflects successful optimization.

• Convergence Speed: Gauges the rapidity with which ML algorithms achieve optimal or near-optimal solutions, crucial for adaptive and real-time cloud environments.

• Scalability: Measures the capability of algorithms to maintain performance across different dataset sizes and workloads, ensuring they are suitable for extensive and complex cloud infrastructures.

• Latency: Tracks the response time of the cloud system, with lower latency signifying improved real-time performance and enhanced user experience.

• Throughput: Assesses the volume of data processed by the system within a specific time frame, indicating the efficiency and capacity of cloud services.

• Reliability: Evaluates the consistency of the system in maintaining QoS parameters across varying conditions and workloads.

• Resource Utilization: Analyzes the effectiveness of cloud resource usage (CPU, memory, storage), aiming for high utilization rates without overburdening the system.

• Energy Efficiency: Measures the energy consumption of the cloud infrastructure, focusing on reducing energy use while maintaining optimal performance.

Implementation and Analytical Tools

The graphical representations of the implementation tools and analytical tools with respect to their metrics are depicted in Figure 4 and Figure 5.

Implementation Tools

• TensorFlow: An open-source platform for machine learning that offers comprehensive tools for developing and deploying ML models. It supports both deep learning and reinforcement learning algorithms, making it versatile for various QoS optimization tasks in cloud computing.

• PyTorch: A flexible and easy-to-use deep learning framework that allows for dynamic computation graphs, which are particularly useful for complex and adaptive ML models. PyTorch is well-suited for research and production environments.

• Scikit-learn: A Python library that provides simple and efficient tools for data mining and data analysis. It is suitable for implementing classical machine learning algorithms and includes modules for model selection, evaluation, and preprocessing.

• Apache Spark: A unified analytics engine for large-scale data processing. It supports distributed computing, which is essential for handling big data in cloud environments, and includes MLlib for scalable machine learning.

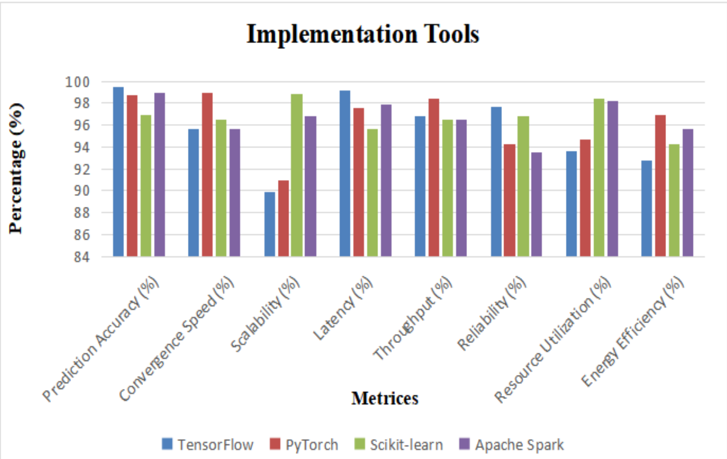

Figure 4. Implementation Tools

Analytical Tools

• Jupyter Notebooks: An open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. It is ideal for data cleaning, transformation, and visualizing the performance of ML models.

• MATLAB: A high-level language and interactive environment for numerical computation, visualization, and programming. MATLAB offers extensive tools for data analysis, machine learning, and model evaluation, making it a powerful tool for research and development.

• Tableau: A powerful data visualization tool that helps in transforming raw data into understandable and interactive visualizations. It is useful for analyzing and presenting the performance metrics of ML models.

• Power BI: A business analytics service by Microsoft that provides interactive visualizations and business intelligence capabilities. Power BI can be used to create detailed reports and dashboards to monitor the QoS metrics and ML model performance.

• TensorBoard: A suite of visualization tools provided with TensorFlow that helps in understanding, debugging, and optimizing machine learning models. TensorBoard offers insights into the model's performance, training process, and computational graph.

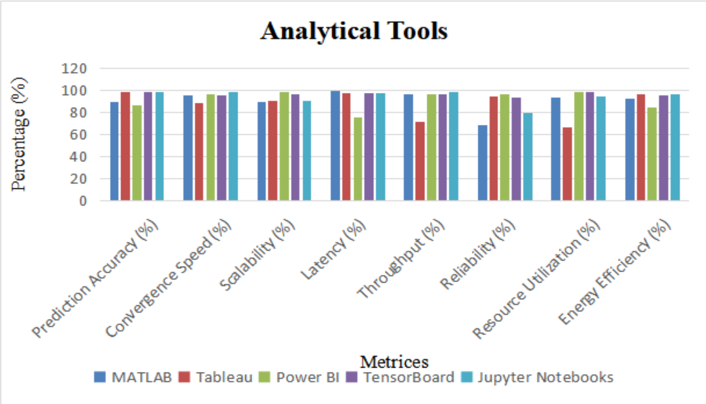

Figure 5. Analytical Tools

CONCLUSION AND FUTURE WORKS

This systematic research concludes by highlighting the noteworthy advancements made in using cutting-edge ML algorithms to optimise QoS parameters in cloud computing systems. The examined literature shows how several ML approaches, including deep learning, reinforcement learning, and ensemble methods, may be used to improve latency, throughput, dependability, and other crucial QoS measures. DL models have demonstrated promise in managing intricate patterns and massive amounts of data, improving system efficiency and prediction accuracy. Reinforcement learning, on the other hand, has shown promise in real-time decision-making and dynamic resource allocation, both of which are essential for adjusting to shifting cloud operational conditions. Looking ahead, a number of important areas should be the focus of future study. First, by leveraging the advantages of various strategies, the creation of hybrid machine learning models incorporating numerous methods might improve QoS optimisation by enhancing scalability and flexibility. Second, especially in crucial cloud contexts, addressing the interpretability of ML models is still crucial to guaranteeing openness and confidence in decision-making procedures. Further developments in edge computing and federated learning may also provide QoS enhancements to decentralised cloud systems, guaranteeing reliable performance in a variety of infrastructure configurations. Furthermore, investigating the integration of cutting-edge technologies like blockchain and quantum computing may completely alter the way QoS parameters are optimised in cloud settings, opening up new security and efficiency opportunities. In addition, responsible implementation and acceptance of ML-driven QoS optimisation depend heavily on addressing ethical issues like bias mitigation and data protection. In conclusion, even though sophisticated machine learning algorithms have shown a great deal of promise for improving cloud computing quality of service, further research should concentrate on interdisciplinary partnerships, methodological advancements, and ethical issues in order to fully realise the transformative potential of these algorithms in terms of future cloud architectures and services.

REFRENCES

[1] K. Bashir, R. Arul, S. Basheer, G. Raja, R. Jayaraman, and N. M. F. Qureshi, "An optimal multitier resource allocation of cloud RAN in 5G using machine learning," Transactions on emerging telecommunications technologies, vol. 30, no. 8, p. e3627, 2019.

[2] R. Moreno-Vozmediano, R. S. Montero, E. Huedo, and I. M. Llorente, "Efficient resource provisioning for elastic cloud services based on machine learning techniques," Journal of Cloud Computing, vol. 8, no. 1, pp. 1–18, 2019.

[3] D. Zeng, L. Gu, S. Pan, J. Cai, and S. Guo, "Resource management at the network edge: A deep reinforcement learning approach," IEEE Network, vol. 33, no. 3, pp. 26–33, 2019.

[4] F. Samie, L. Bauer, and J. Henkel, "From cloud down to things: An overview of machine learning in internet of things," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4921–4934, 2019.

[5] S. Kanungo, "Edge-to-Cloud Intelligence: Enhancing IoT Devices with Machine Learning and Cloud Computing," International Peer-Reviewed Journal, vol. 2, no. 12, pp. 238–245, 2019.

[6] A.-R. Al-Ghuwairi, M. Khalaf, Z. Salah, and A. Alsarhan, "Dynamic changes of QoS parameters in cloud computing service level agreement," International Journal of Business Information Systems, vol. 32, no. 1, pp. 73–90, 2019.

[7] A. Naseri and N. Jafari Navimipour, "A new agent-based method for QoS-aware cloud service composition using particle swarm optimization algorithm," Journal of ambient intelligence and humanized computing, vol. 10, pp. 1851–1864, 2019.

[8] M. O Ahmad and R. Z. Khan, "An efficient load balancing scheduling strategy for cloud computing based on hybrid approach", International Journal of Cloud Computing, Vol. 9, Issue 4, 2020. InderScience Publication.

[9] S. Heidari and R. Buyya, "Quality of Service (QoS)-driven resource provisioning for large-scale graph processing in cloud computing environments: Graph Processing-as-a-Service (GPaaS)," Future Generation Computer Systems, vol. 96, pp. 490–501, 2019.

[10] Amira Mahamat Abdallah, Aysha Saif Rashed Obaid Alkaabi (2024). "Cloud Network Anomaly Detection Using Machine and Deep Learning Techniques—Recent Research Advancements." Journal of Cloud Computing, 2024.

[11] Uma Maheswara Rao I, Jkr Sastry (2024). "Enhanced Feature-driven Multi-Objective Learning for Optimal Cloud Resource Allocation." International Journal of Cloud Computing and Big Data, 2024.

[12] Nisha, Dr. Deepak Nandal, Dr. Sunil Kumar Nandal (2024). "A Review on Machine Learning Models for Quality of Service in Cloud Computing." Cloud Computing and Machine Learning Journal, 2024.

[13] A. K. Bashir, R. Arul, S. Basheer, G. Raja, R. Jayaraman, and N. M. F. Qureshi, "An optimal multitier resource allocation of cloud RAN in 5G using machine learning," Transactions on emerging telecommunications technologies, vol. 30, no. 8, p. e3627, 2019.

[14] R. Moreno-Vozmediano, R. S. Montero, E. Huedo, and I. M. Llorente, "Efficient resource provisioning for elastic cloud services based on machine learning techniques," Journal of Cloud Computing, vol. 8, no. 1, pp. 1–18, 2019.

[15] D. Zeng, L. Gu, S. Pan, J. Cai, and S. Guo, "Resource management at the network edge: A deep reinforcement learning approach," IEEE Network, vol. 33, no. 3, pp. 26–33, 2019.

[16] F. Samie, L. Bauer, and J. Henkel, "From cloud down to things: An overview of machine learning in internet of things," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4921–4934, 2019.

[17] S. Kanungo, "Edge-to-Cloud Intelligence: Enhancing IoT Devices with Machine Learning and Cloud Computing," International Peer-Reviewed Journal, vol. 2, no. 12, pp. 238–245, 2019.

[18] A.-R. Al-Ghuwairi, M. Khalaf, Z. Salah, and A. Alsarhan, "Dynamic changes of QoS parameters in cloud computing service level agreement," International Journal of Business Information Systems, vol. 32, no. 1, pp. 73–90, 2019.

[19] A. Naseri and N. Jafari Navimipour, "A new agent-based method for QoS-aware cloud service composition using particle swarm optimization algorithm," Journal of ambient intelligence and humanized computing, vol. 10, pp. 1851–1864, 2019.

[20] S. Heidari and R. Buyya, "Quality of Service (QoS)-driven resource provisioning for large-scale graph processing in cloud computing environments: Graph Processing-as-a-Service (GPaaS)," Future Generation Computer Systems, vol. 96, pp. 490–501, 2019.

[21] S. K. Gavvala, C. Jatoth, G. Gangadharan, and R. Buyya, "QoS-aware cloud service composition using eagle strategy," Future Generation Computer Systems, vol. 90, pp. 273–290, 2019.

[22] Z. Xie and H. Yin, "Selection of optimal cloud services based on quality of service ontology," ISI, vol. 23, no. 6, 2018.

[23] L.-V. Le, B.-S. Lin, and S. Do, "Applying big data, machine learning, and SDN/NFV for 5G early-stage traffic classification and network QoS control," Transactions on Networks and Communications, vol. 6, no. 2, p. 36, 2018.

[24] X. Huang, K. Xie, S. Leng, T. Yuan, and M. Ma, "Improving Quality of Experience in multimedia Internet of Things leveraging machine learning on big data," Future Generation Computer Systems, vol. 86, pp. 1413–1423, 2018.

[25] E. C. Santos, "Autonomous QoS-based mechanism for resource allocation in LTE-Advanced Pro networks," in 2018 IEEE Colombian Conference on Communications and Computing (COLCOM), IEEE, 2018, pp. 1–6.

[26] M. Qin, Q. Yang, N. Cheng, H. Zhou, R. R. Rao, and X. Shen, “Machine learning aided context-aware self-healing management for ultra dense networks with QoS provisions,” IEEE Transactions on Vehicular Technology, vol. 67, no. 12, pp. 12339–12351, 2018.

[27] J. Xie et al., “A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 393–430, 2018.

[28] Y. Zhang, B. Liu, Y. Gong, J. Huang, J. Xu, and W. Wan, “Application of machine learning optimization in cloud computing resource scheduling and management,” in Proceedings of the 5th International Conference on Computer Information and Big Data Applications, 2024, pp. 171–175.

[29] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, “A machine learning approach for power allocation in HetNets considering QoS,” in 2018 IEEE international conference on communications (ICC), IEEE, 2018, pp. 1–7.

[30] A. Al-Jawad, P. Shah, O. Gemikonakli, and R. Trestian, “LearnQoS: A learning approach for optimizing QoS over multimedia-based SDNs,” in 2018 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB), IEEE, 2018, pp. 1–6.

[31] L. Cui, S. Yang, F. Chen, Z. Ming, N. Lu, and J. Qin, “A survey on application of machine learning for Internet of Things,” International Journal of Machine Learning and Cybernetics, vol. 9, pp. 1399–1417, 2018.

[32] L.-V. Le, B.-S. Lin, and S. Do, “Applying big data, machine learning, and SDN/NFV for 5G early-stage traffic classification and network QoS control,” Transactions on Networks and Communications, vol. 6, no. 2, p. 36, 2018.

[33] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. M. Riad, “A machine learning model for improving healthcare services on cloud computing environment,” Measurement, vol. 119, pp. 117–128, Apr. 2018, doi: 10.1016/j.measurement.2018.01.022.

[34] S. Malik, M. Tahir, M. Sardaraz, and A. Alourani, “A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques,” Applied Sciences, vol. 12, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/app12042160.

[35] P. Osypanka and P. Nawrocki, “QoS-Aware Cloud Resource Prediction for Computing Services,” IEEE Transactions on Services Computing, vol. 16, no. 2, pp. 1346–1357, Mar. 2023, doi: 10.1109/TSC.2022.3164256.

[36] P. Devarasetty and S. Reddy, “Genetic algorithm for quality of service-based resource allocation in cloud computing,” Evol. Intel., vol. 14, no. 2, pp. 381–387, Jun. 2021, doi: 10.1007/s12065-019-00233-6.

[37] Tabassum et al. - 2021 - Prediction of Cloud Ranking in a Hyperconverged Cl.pdf.” Accessed: Jul. 15, 2024. [Online]. Available: https://cdn.techscience.cn/ueditor/files/TSP_CMC_67-3/TSP_CMC_14729/TSP_CMC_14729.pdf

[38] M. Abd Elaziz, L. Abualigah, and I. Attiya, “Advanced optimization technique for scheduling IoT tasks in cloud-fog computing environments,” Future Generation Computer Systems, vol. 124, pp. 142–154, Nov. 2021, doi: 10.1016/j.future.2021.05.026.

[39] M. T. Islam, S. Karunasekera, and R. Buyya, “Performance and Cost-Efficient Spark Job Scheduling Based on Deep Reinforcement Learning in Cloud Computing Environments,” IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 7, pp. 1695–1710, Jul. 2022, doi: 10.1109/TPDS.2021.3124670.

[40] P. K. Bal, S. K. Mohapatra, T. K. Das, K. Srinivasan, and Y.-C. Hu, “A Joint Resource Allocation, Security with Efficient Task Scheduling in Cloud Computing Using Hybrid Machine Learning Techniques,” Sensors, vol. 22, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/s22031242.

[41] A. Jayanetti, S. Halgamuge, and R. Buyya, “Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge–cloud computing environments,” Future Generation Computer Systems, vol. 137, pp. 14–30, Dec. 2022, doi: 10.1016/j.future.2022.06.012.

[42] H. Liang, X. Wen, Y. Liu, H. Zhang, L. Zhang, and L. Wang, “Logistics-involved QoS-aware service composition in cloud manufacturing with deep reinforcement learning,” Robotics and Computer-Integrated Manufacturing, vol. 67, p. 101991, Feb. 2021, doi: 10.1016/j.rcim.2020.101991.

[43] J. Kumar and A. K. Singh, “Performance evaluation of metaheuristics algorithms for workload prediction in cloud environment,” Applied Soft Computing, vol. 113, p. 107895, Dec. 2021, doi: 10.1016/j.asoc.2021.107895.

[44] A. K. Samha, "Strategies for efficient resource management in federated cloud environments supporting Infrastructure as a Service (IaaS)," Journal of Engineering Research, vol. 12, no. 2, pp. 101–114, Jun. 2024, doi: 10.1016/j.jer.2023.10.031.
[45] Aljuhani, "Machine Learning Approaches for Combating Distributed Denial of Service Attacks in Modern Networking Environments," IEEE Access, vol. 9, pp. 42236–42264, 2021, doi: 10.1109/ACCESS.2021.3062909.
[46] G. Rjoub, J. Bentahar, O. Abdel Wahab, and A. Saleh Bataineh, "Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems," Concurrency and Computation: Practice and Experience, vol. 33, no. 23, p. e5919, 2021, doi: 10.1002/cpe.5919.
[47] S. Balasubramaniam et al., "Optimization Enabled Deep Learning-Based DDoS Attack Detection in Cloud Computing," International Journal of Intelligent Systems, vol. 2023, no. 1, p. 2039217, 2023, doi: 10.1155/2023/2039217.
[48] S. Negi, M. M. S. Rauthan, K. S. Vaisla, and N. Panwar, "CMODLB: an efficient load balancing approach in cloud computing environment," J Supercomput, vol. 77, no. 8, pp. 8787–8839, Aug. 2021, doi: 10.1007/s11227-020-036017.
[49] P. S. Rawat, P. Dimri, P. Gupta, and G. P. Saroha, "Resource provisioning in scalable cloud using bio-inspired artificial neural network model," Applied Soft Computing, vol. 99, p. 106876, Feb. 2021, doi: 10.1016/j.asoc.2020.106876.
[50] U. K. Jena, P. K. Das, and M. R. Kabat, "Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, Part A, pp. 2332–2342, Jun. 2022, doi: 10.1016/j.jksuci.2020.01.012.
[51] G. Shruthi, M. R. Mundada, B. J. Sowmya, and S. Supreeth, "Mayfly Taylor Optimisation-Based Scheduling Algorithm with Deep Reinforcement Learning for Dynamic Scheduling in Fog-Cloud Computing," Applied Computational Intelligence and Soft Computing, vol. 2022, no. 1, p. 2131699, 2022, doi: 10.1155/2022/2131699.
[52] M. Mayuranathan, S. K. Saravanan, B. Muthusenthil, and A. Samydurai, "An efficient optimal security system for intrusion detection in cloud computing environment using hybrid deep learning technique," Advances in Engineering Software, vol. 173, p. 103236, Nov. 2022, doi: 10.1016/j.advengsoft.2022.103236.
[53] E. S. Ali et al., "Machine Learning Technologies for Secure Vehicular Communication in Internet of Vehicles: Recent Advances and Applications," Security and Communication Networks, vol. 2021, no. 1, p. 8868355, 2021, doi: 10.1155/2021/8868355.
[54] M. Emamian, A. Eskandari, M. Aghaei, A. Nedaei, A. M. Sizkouhi, and J. Milimonfared, "Cloud Computing and IoT Based Intelligent Monitoring System for Photovoltaic Plants Using Machine Learning Techniques," Energies, vol. 15, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/en15093014.
[55] A. Aldallal and F. Alisa, "Effective Intrusion Detection System to Secure Data in Cloud Using Machine Learning," Symmetry, vol. 13, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/sym13122306.
[56] Chraibi, S. Ben Alla, and A. Ezzati, "Makespan Optimisation in Cloudlet Scheduling with Improved DQN Algorithm in Cloud Computing," Scientific Programming, vol. 2021, no. 1, p. 7216795, 2021, doi: 10.1155/2021/7216795.
[57] D. G. Rosero, N. L. Díaz, and C. L. Trujillo, "Cloud and machine learning experiments applied to the energy management in a microgrid cluster," Applied Energy, vol. 304, p. 117770, Dec. 2021, doi: 10.1016/j.apenergy.2021.117770.
[58] R. SaiSindhuTheja and G. K. Shyam, "An efficient metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment," Applied Soft Computing, vol. 100, p. 106997, Mar. 2021, doi: 10.1016/j.asoc.2020.106997.
[59] H. Lin, Q. Xue, J. Feng, and D. Bai, "Internet of things intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine," Digital Communications and Networks, vol. 9, no. 1, pp. 111–124, Feb. 2023, doi: 10.1016/j.dcan.2022.09.021.
[60] S. Haytamy and F. Omara, "A deep learning based framework for optimizing cloud consumer QoS-based service composition," Computing, vol. 102, no. 5, pp. 1117–1137, May 2020, doi: 10.1007/s00607-019-00784-7.

[61] P. Osypanka and P. Nawrocki, “Resource Usage Cost Optimization in Cloud Computing Using Machine Learning,” IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 2079–2089, Jul. 2022, doi: 10.1109/TCC.2020.3015769.
[62] M. Kumar and S. C. Sharma, “PSO-based novel resource scheduling technique to improve QoS parameters in cloud computing,” Neural Comput & Applic, vol. 32, no. 16, pp. 12103–12126, Aug. 2020, doi: 10.1007/s00521-019-04266-x.
[63] S. Zhang, S. Zhang, X. Chen, and X. Huo, “Cloud computing research and development trend,” in 2010 Second international conference on future networks, IEEE, 2010, pp. 93–97. A.
[64] Abid, M. F. Manzoor, M. S. Farooq, U. Farooq, and M. Hussain, “Challenges and Issues of Resource Allocation Techniques in Cloud Computing,” KSII Transactions on Internet & Information Systems, 14, no. 7, 2020.
[65] X. Zhang, T. Wu, M. Chen, T. Wei, J. Zhou, S. Hu, and R. Buyya, “Energy-aware virtual machine allocation for cloud with resource reservation,” Journal of Systems and Software, 147 (2019), pp. 147–161.
[66] L.-Y. Zuo, “Multi-objective integrated ant colony optimization scheduling algorithm based on cloud resource,” Journal of Computer Applications, 32, no. 07, 2012, pp. 1916.
[67] L. Tai, “Multi-objective dynamic scheduling of manufacturing resource to cloud manufacturing services,” China Mechanical Engineering, 24, no. 12, 2013, pp. 1616.
[68] H. Jiang, J. Yi, S. Chen, and X. Zhu, “A multi-objective algorithm for task scheduling and resource allocation in cloud-based disassembly,” Journal of Manufacturing Systems, 41 (2016), pp. 239–255.
[69] S. Yuan, L. Chun-guo, H. Yong-ming, and Y. Lu-xi, “Multi-Objective Resource Allocation Design Algorithm in CloudRadio Access Network,” fififi, 33, no. 3, 2017, pp. 294–303.
[70] P. Devarasetty and Ch. S. Reddy, “Multi objective Ant colony Optimization Algorithm for Resource Allocation in Cloud.
[71] Computing,” International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-2S2, December 2018, pp. 68–73.
[72] M. B. Gawali and S. K. Shinde, “Task scheduling and resource allocation in cloud computing using a heuristic approach,” Journal of Cloud Computing, 7, no. 1, 2018, pp. 1–16.
[73] M. Afrin, J. Jin, A. Rahman, Y.-C. Tian, and A. Kulkarni, “Multi-objective resource allocation for edge cloud based robotic workflow in smart factory,” Future Generation Computer Systems, 97 (2019), pp. 119–130.
[74] Mohan, B. R. ., M, D. ., Bhuria, V. ., Gadde, S. S. ., M, K. ., & N, A. P. . (2023). Potable Water Identification with Machine Learning: An Exploration of Water Quality Parameters. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 178–185. https://doi.org/10.17762/ijritcc.v11i3.6333.