



(REVIEW ARTICLE)



## AI ML and cloud computing: exploring models, challenges and opportunities

Harshad Pitkar <sup>1,\*</sup> and Sumedh Ambapkar <sup>2</sup>

<sup>1</sup> Cummins Inc, Columbus, Indiana, USA.

<sup>2</sup> Luddy school of informatics computing and engineering, Indiana University, Bloomington, IN USA.

World Journal of Advanced Research and Reviews, 2025, 25(02), 770-783

Publication history: Received on 29 December 2024; revised on 04 February 2025; accepted on 07 February 2025

Article DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0430>

### Abstract

The collaboration between Artificial Intelligence (AI) and Machine Learning (ML) in the landscape of Cloud Computing (CC) is fundamentally redefining the techniques that enterprises apply in terms of data processing, security, cost-optimization, scalability, and resource oversight. Consequently, cloud platforms provide the essential infrastructure and flexibility necessary for AI and ML algorithms to analyze vast datasets, while automation propelled by AI and ML models enhances cloud services by optimizing performance, reducing latency, and forecasting demand. This intersection not only facilitates dynamic scalability and efficient resource allocation but also unveils revolutionary prospects, including intelligent automation, self-healing systems, and adaptive security frameworks. This state-of-the-art review provides researchers with current trends, challenges in this rapidly growing field and points towards research gaps and unexplored research directions. This literature review explores the synergistic interplay between AI, ML, and Cloud Computing, highlighting significant advancements, intrinsic challenges, and potential opportunities across various sectors. We examine the different techniques that enable AI and ML algorithms to enhance cloud applications, focusing specifically on domains like automated decision-making, optimization, operations, scheduling and security. This review paper aims to provide a comprehensive perspective of the current state of this convergence, the challenges it faces, and the opportunities it presents for the future.

**Keywords:** Cloud Computing; Artificial Intelligence; Machine Learning; Deep Learning; Cybersecurity; Cost Optimization; Scheduling; Automation

### 1. Introduction

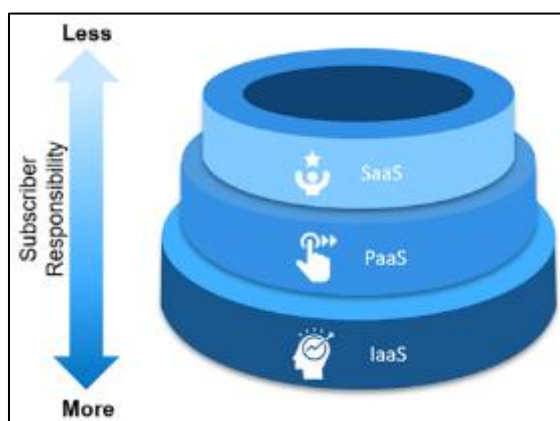
In current scholarly discussions, the rapid progressions in technological domains have facilitated the emergence of a new era characterized by digital transformation, fundamentally influenced by the integration of Artificial Intelligence (AI), Machine Learning (ML), and Cloud Computing. Each of these technological paradigms, in isolation, possesses substantial potential. AI equips systems with the capability to execute tasks that have historically necessitated human cognitive faculties, encompassing decision-making, linguistic analysis, and visual interpretation. ML, which is a subfield of AI, enables systems to learn from data and elevate their performance without the necessity of being programmed. On the flip side, Cloud Computing provides scalable and on-demand access to computing resources which allow organizations to handle massive data collections and complex computation problems in a cost-efficient manner. Together, they have opened doors to never-before-seen possibilities, from personalized recommendations to autonomous systems to scientific discoveries and improved business processes [1]. Cloud Computing serves as the ideal infrastructural ecosystem for the implementation of AI and ML models owing to its almost limitless storage and computing power. This ability enables organizations to train complicated models on large datasets, and deploy them across distributed frameworks, without requiring significant investment in on premise infrastructure. Conversely, AI and ML enrich Cloud Computing by delivering intelligent automation, adaptive learning, and predictive analytics that are used to refine cloud services, reduce operational costs, and improve overall performance. This complementary

\* Corresponding author: Harshad Pitkar

interplay has led to significant breakthroughs across industries, ranging from automated decision-making, personalized recommendations, and self-driving systems to advanced cybersecurity solutions. As organizations progressively transition their operational workloads to cloud environments, the integration of AI and ML technologies within these frameworks is becoming increasingly imperative. This paper endeavors to explore the interdependent nature of AI, ML, and Cloud Computing, scrutinizing their interrelations, primary applications, and the pivotal roles they occupy in shaping the future landscape of technology-driven innovation.

### 1.1. Cloud Computing

The rise of cloud computing denotes a major technological evolution that facilitates the retrieval and storage of information and applications via the internet, as opposed to depending on local hardware or personal computing equipment. Cloud computing facilitates the immediate accessibility of computational resources, particularly in the realms of data storage and processing capabilities, while necessitating no direct oversight or management by the end user [2]. It is like having several data centers available to many users over the internet [3]. Cloud computing service models are typically divided into three categories:



**Figure 1** Three Cloud service models showing shift in subscriber responsibility from Infrastructure as a Service IaaS to Software as a Service SaaS models

**Infrastructure as a Service (IaaS):** Provides virtualized computing resources over the internet. Users can rent IT infrastructure like servers and virtual machines (VMs), storage, networks, and operating systems from a cloud provider [4].

**Platform as a Service (PaaS):** Offers hardware and software tools over the internet, wherein hardware and software run in the Cloud Provider infrastructure [4].

**Software as a Service (SaaS):** Delivers software applications over the internet, that follows a subscription basis. SaaS applications and services can be accessed from anywhere using a device with an internet connection [5]. Figure 1 illustrates the three service models, as subscribers move from top to bottom, their responsibilities around managing resources gradually increase. To explain this further, in the SaaS model, users do not have to be concerned about managing the application or the infrastructure, where in the bottom most layer, which is the IaaS model, users are responsible for managing virtual machines or servers, operating system and the application running on it. PaaS strikes this balance wherein hardware remains the responsibility of the cloud provider while application hosting is users' responsibility. The past decade has seen Cloud Computing evolve into a major technological breakthrough in Information Technology (IT), highlighting a key alteration in the practices surrounding the provision and use of IT services by clients [6] [7].

### 1.2. Artificial Intelligence

AI is the emulation of human intelligence processes by machines, such as learning, reasoning, and self-correction [8]. AI is a broad field that covers several different subfields like Machine Learning (ML) and Deep Learning (DL). AI has vast real-world applications across multiple domains such as medical diagnosis, face recognition, robotics, internet applications, data mining, and industrial applications. AI is utilized in significant industries, including healthcare management, financial research, social networking monitoring, and cloud computing, enhancing performance and efficiency in these fields [9]. AI's integration with cloud computing has led to the development of AI-as-a-service

products, enabling businesses to leverage AI capabilities without substantial infrastructure investments [10]. AI continues to evolve, with ongoing research and development aimed at making predictions, automating complex tasks, and improving decision-making processes across various sectors [10].

### **1.3. Machine Learning**

Machine Learning (ML) is a subset of Artificial Intelligence (AI) focused on developing algorithms that enable machines to learn from and adapt to data without explicit programming [8]. ML is more about the ability of a machine to learn and to adapt based on experience. There are several different techniques of ML (Supervised learning, Unsupervised learning, Reinforcement learning, etc.), each of which is used for different purposes to analyze data and make predictions. ML has a wide variety of use cases, including but not limited to computer vision, natural language processing, predictive analytics, anomaly detection in industrial equipment, etc. [8], [11]. The continued development of ML techniques and their integration with other technologies like AI and blockchain are expected to drive further innovation and efficiency across various sectors [12], in this review we have examined several models, discussed challenges as well as opportunities that the integration of these three technologies present. The paper is organized as follows: Section 2 discusses AI driven innovations in cloud computing around a few critical use cases, challenges of this integration and future research opportunities. Similarly, section 3 reviews innovations driven by ML on Cloud platforms, the various models proposed and research gaps. Section 4 concludes the paper with our final thoughts.

---

## **2. Artificial Intelligence driven innovations in Cloud Computing**

Artificial Intelligence (AI) driven innovations in cloud computing have significantly transformed the landscape of digital services, enhancing efficiency, scalability, and security across industries. This results in optimizing cloud infrastructure and resource management, which in turn leads to industry-wide improvement. Artificial Intelligence has emerged as a transformative force in the realm of cloud computing, revolutionizing the way we approach data storage, processing, and analysis. The integration of AI and CC offers numerous advantages, addressing the increasing computational demands of AI applications [13]. Leveraging the scalability and flexibility of cloud infrastructure, AI services can solve complex problems faster and more efficiently than ever before. The convergence of big data, machine learning, and cloud super-computing has driven the recent resurgence of AI, paving the way for innovative applications across various industries [14]. In the following sections we have tried to discuss AI related innovations in the Cloud Computing space in detail.

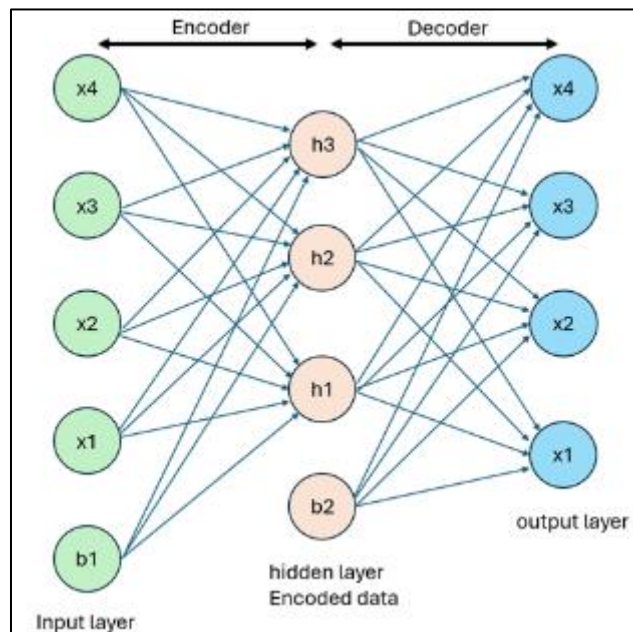
### **2.1. Artificial Intelligence for IT Operations or AI Ops**

AI Ops is a transformative approach that leverages AI and ML to improve the efficiency and reliability of IT operations [15]. By integrating big data analytics, machine learning, and automation, AIOps aims to streamline IT processes, improve system performance, and reduce operational costs [16]. This approach is increasingly being adopted across various industries to address the complexities of modern IT environments in Cloud computing. The most proficient deep learning-oriented paradigms for immediate problem detection and diagnosis within IT operations employing AIOps are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Nevertheless, ultimately, the choice of framework is contingent upon the specific application scenario and the nature of the data involved [17]. AIOps leverages AI to analyze massive volumes of operational data, detect anomalies, predict issues, and facilitate informed decision-making, thereby improving the efficiency, reliability, and resilience of software systems [16]. AIOps significantly enhances anomaly detection accuracy by 15% and increases incident management effectiveness by 50%. This is achieved through a hybrid approach combining supervised and unsupervised learning techniques, which outperform traditional rule-based methods [18]. As more and more organizations, public, private, educational and government included, adopt Cloud Computing, which essentially follows the pay-as-you-go model, one common theme has emerged that is, how do we control Cloud costs? By integrating AI with cloud computing, AIOps help in better resource management, ensuring that even less powerful nodes are effectively utilized [19], [16]. By exploiting big data collected in the form of log, tracing, metric, and network data, AIOps enable detection of faults and issues in services [20]. AIOps facilitates the enhancement of organizational agility and cost efficiency by markedly reducing expenditures associated with infrastructure management, thereby enabling more responsive and flexible operational frameworks. [19], [16]. By combining AI capabilities with DevOps practices, AIOps enhances operations and monitoring in the DevOps lifecycle, promoting collaboration, agility, and faster delivery of high-quality software products [16]. In the next three subsections we have discussed three major use cases, and the models proposed to meet these requirements.

### **2.2. AI-powered fault detection and mitigation in cloud computing infrastructures**

Fault detection is of paramount importance in the realm of cloud computing, given its crucial role in maintaining the reliability and performance of the system and avoiding downtime or service disruptions. As cloud computing services

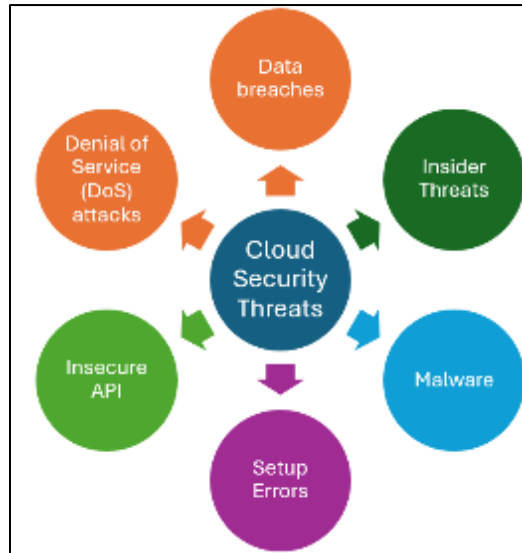
become increasingly essential for both small-scale consumers and large-scale organizations, the assurance of fault tolerance and performance predictability within these systems has emerged as a critical concern [21]. Traditional fault detection methods often rely on static thresholds and predefined rules, making them less adaptable. In contrast, AI-powered approaches can process vast amounts of data in real-time and identify subtle patterns indicative of potential faults [22]. The paper discusses the use of artificial intelligence, specifically through the (Thread level parallelism) TLP-Allocator, to optimize parallel computing in cloud servers [23]. AI can significantly enhance cloud computing by optimizing task scheduling, which reduces energy consumption and improves resource allocation. Intelligent algorithms, such as the Gated Graph Convolution Network (GGCN) and Convolutional Neural Networks (CNNs), can efficiently allocate user-deployed jobs to servers, minimizing performance degradation and operational costs. The application of AI in these areas leads to better job completion rates and lower energy usage, ultimately benefiting cloud providers and users alike. In [24] One of the methods proposed in [25] uses a sparse de-noising auto-encoder to develop a parallel architecture network. The fundamental architecture of the Auto-encoder (AE) as illustrated in Figure 2 closely resembles that of a three-layer Back propagation (BP) neural network. This architecture is divided into two distinct components, the encoding segment and the decoding segment. A salient characteristic of this structure is that the quantity of neurons present in the input layer is equivalent to the quantity of neurons in the output layer; furthermore, the training datasets utilized are unlabeled, with the output values mirroring the input values. The primary objective of employing this model is to extract data characteristics or to achieve a nonlinear dimensionality reduction of the data by analyzing the data structure within the hidden layer. AI techniques can continuously learn and adapt to changing conditions, making them more robust in dynamic environments. This continuous learning ensures that AI models remain effective in detecting and mitigating emerging fault patterns.



**Figure 2** Architecture of an auto-encoder resembles 3-layer back propagation neural network

### 2.3. AI-powered fault detection and mitigation in cloud computing infrastructures

The facilitation of security and data privacy within cloud platforms has garnered considerable scholarly interest. For instance, sensitive patient information in the cloud necessitates stringent protection [26], and the transmission of such information among various stakeholders must also be safeguarded. Generally, security vulnerabilities in cloud ecosystems have been the subject of thorough investigation. Such vulnerabilities encompass data breaches, data loss, denial of service, service rejection, and threats from malicious insiders arising from challenges such as multi-tenancy [27], diminished control over data, and issues of trust [28]. Figure 3 depicts major types of security threats in Cloud Computing space, and throughout this section we have tried to review the AI based solutions that help in addressing these threats. In [29] authors address Distributed Denial of Service (DDoS) attacks, which are significant threats in cloud environments. These attacks aim to make network services unavailable by overwhelming them with traffic. The



**Figure 3** Common Security Threats in Cloud Computing

proposed model helps in early detection and mitigation of these attacks by generating synthesized malicious samples to improve the accuracy of intrusion detection systems (IDS). In proposed method in [29], IDS uses model CDAAE-KNN which incorporates CDAAE in conjunction with the K-nearest neighbour algorithm enables the generation of malicious borderline samples, which subsequently enhances the overall accuracy of a cloud-based IDS. Gloss (1) is computed as the average log-loss across all generated samples, pushing the generator to create samples that the discriminator cannot easily distinguish from real ones. Minimizing Gloss improves the quality of generated samples. The total loss function of CDAAE is calculated as shown in (2) [29].

$$G_{loss} = \frac{1}{n} \sum_{i=0}^n \log (d_{fake}^i) \dots \dots \dots (1)$$

$$L_{CDAAE} = R_{loss} + D_{loss} - G_{loss} \dots \dots \dots (2)$$

The primary model discussed in [30] is the IDSGT-DNN, which stands for Intrusion Detection System Game Theory-Deep Neural Network. This model is designed to enhance cloud security by incorporating game theory into a deep neural network framework. It focuses on optimizing the classification of normal and attack data within the CICIDS-2017 dataset [35]. The study presents a hybrid CIDS using SCAE for feature extraction and SVM for classification. Experimental results indicate promising performance, with room for classifier optimization. The primary threat mitigated by the proposed model in [32] is Cross-Site Scripting (XSS) attacks. These attacks are a significant concern in cloud computing environments as they can lead to data theft, unauthorized access, and service disruption. The paper highlights the complexity and prevalence of XSS attacks, which are identified as “one of the top ten web security vulnerabilities” by Open Worldwide Application Security Project (OWASP) [32]. Cloud-Based Behaviour Centric Model (CBCM) proposed in [33] is central to the proposed system used to create features from execution traces of suspicious files. The CBCM groups related behaviours according to predefined rules to generate features that are crucial for detecting malware. In [36] authors address the threat of malicious users in cloud computing who may gain access to private data without permission. To tackle this issue, the authors propose the Federated Learning-Driven Malicious User Prediction Model (FedMUP). This innovative model employs federated learning methodologies to scrutinize user behaviors and discern security risk indicators without infringing upon data confidentiality. Table 1 presents a list of models proposed and the threat mitigated by these studies. The integration of AI and cloud computing significantly enhances data security and scalability by leveraging AI’s capabilities to detect, prevent, and respond to security threats while optimizing resource management. AI-driven solutions in cloud environments provide advanced security measures, such as real-time threat detection and adaptive responses, which are crucial for maintaining robust security in the face of evolving cyber threats.

**Table 1** Models proposed, and security threats mitigated in Cloud Computing

Reference article	Model Proposed	Threat Mitigated
deep et. al [29]	CDAEE-KNN Hybrid Model	Low Application layer and low bandwidth DDoS
balamurugan et. al [30]	Game Theory-cloud Security Deep Neural Network (IDSGT-DNN)	Several: account hijacking, malware injection, data breaches,API vulnerabilities
wang et. al [31]	Stacked Contractive Auto-Encoder SCAE-SVM	Network intrusion detection
li et. al [32]	Character-level Bidirectional LSTM with Multi-Head Attention (CMABLSTM)	Cross-site scription attacks
aslan et. al [33]	Cloud-Based Behavior Centric Model (CBCM)	Malware detection
wahab et. al [34]	Stackelberg Security Game Model	Insider attacks, distributed attacks by Malicious VirtualMachines

#### 2.4. Challenges of integrating AI with Cloud Computing

Integration of AI with CC certainly has its own merits, however there are challenges too that need to be overcome, including the distributed nature and massive scale of cloud platforms. The high complexity of cloud systems, combined with the need to leverage vast amounts of runtime and workload data, presents significant hurdles in designing and implementing AI-driven solutions effectively [37]. Some of the challenges include ensuring data security, complexity of AI algorithms, addressing the need for high-performance computation. Cloud service providers bear the obligation of safeguarding the information housed within their infrastructures. Nonetheless, there exist documented instances where cloud service providers have suffered breaches, culminating in the exposure of confidential data [9]. It is essential for organizations to set up measures to protect their information while it resides on cloud platforms [8]. Also, organizations are expected to conform to multiple rules concerning data holding, privacy, and security, which can differ markedly across various countries [9]. Understanding and adhering to these regulations is crucial when deploying AI applications on cloud platforms [8]. The methodologies associated with artificial intelligence that rely on autonomous computing frequently encounter detrimental effects due to sluggish internet connectivity, resulting in latency complications during the transmission of data to the cloud and the subsequent retrieval of responses [19]. The initiation of an AI-integrated edge computing ecosystem requires an upfront financial investment aimed at procuring edge-enabled frameworks that incorporate both software and hardware. This investment often conflicts with company initiatives, making it difficult, particularly in developing countries [19], [16]. Committing to a specific cloud provider can make switching suppliers difficult due to the lack of industry-wide standards. This vendor lock-in can pose significant challenges for businesses looking to adopt or switch AI-integrated cloud solutions [16].

#### 2.5. Future work or research gap

While there are several challenges around integrating AI with Cloud computing, these challenges also present many opportunities for further research, and this section discusses some of the research gaps. A significant number of literature reviews revolve around a specific topic, such as deep learning anomaly detection and root-cause analysis, rather than offering a comprehensive overview of AIOps in both academia and industry [8]. The data in these applications grows incrementally, and current AIOps techniques may struggle to cope with this rampant increase [18]. There is a need for more advanced techniques in incident detection and failure prediction to reduce mean time to detection (MTTD) and predict potential issues before they occur. This would allow for proactive measures to minimize impact [8]. Developing more sophisticated methods for automated actions and root cause analysis is essential. In [38] authors point out the gaps that exist in healthcare AI due to the medical community's cautious approach to new technologies that hinder AI's full potential. Integrating AIOps into the DevOps lifecycle to address operational challenges is still an emerging area. The research [39] delineates a significant deficiency in contemporary scholarly discourse concerning efficacious methodologies for mitigating DDoS assaults, particularly within cloud computing contexts, thereby underscoring an imperative for more concentrated investigative efforts in this domain to fortify cloud security protocols.

### 3. ML Driven innovations in Cloud Computing

The field of Machine Learning (ML) represents a revolutionary influence across multiple domains, and its amalgamation with Cloud Computing substantially augments the capacity for innovation. The integration of Machine Learning and

Cloud Computing not only enhances the efficacy of data governance and resource distribution but also cultivates an environment in which enterprises can innovate with both speed and effectiveness while keeping infrastructure costs under control. Using machine learning techniques, the redundant data can be identified and compressed efficiently. This minimizes storage space needed, which makes better use of available storage resources. Machine learning algorithms also can classify data based on access patterns and automatically stage less-frequently accessed data into less expensive, slower storage tiers. Machine Learning increases data management efforts by integration and insight-driven decision making. This includes the optimization of how we store, access and manage that data in the cloud [40]. ML can also analyze the patterns based on data access, prediction of data location in the storage infrastructure can be performed to minimize retrieval times. So that most accessed data is accessed fast, added the data [7]. In the subsequent sections we have explored further into the technical aspect of models and their applications.

### 3.1. Leveraging Machine Learning for Big Data in Cloud

Cloud resources can be automatically adjusted based on utilization trends through machine learning algorithms such as linear regression, decision trees, and random forests. By doing so, it helps maximize resource usage by organizing the data that is stored while still providing maximal performance with minimal usage of computational resources [41]. Models of machine learning can be used to analyze big datasets to identify patterns and trends in data management. Such as data storage, retrieval, and effective processing of data to make sure the cloud structure is used to the fullest extent [7]. Cloud security is improved by ML algorithms through dynamic threat detection, automatic access control and fast data encryption. By analyzing historical data, machine learning allows predictive analytics to keep track of future trends. ML algorithms can classify data based on access patterns and automatically move less frequently accessed data to more economical storage tiers. This optimizes storage costs and ensures that high-performance storage is reserved for frequently accessed data. Machine learning algorithms can detect anomalies in data usage patterns, which helps in identifying inefficiencies and potential security issues. This proactive approach ensures that any irregularities are addressed promptly, maintaining the integrity of big data in the cloud [41].

$$a = f(Wx + b) \quad \dots \dots (3)$$

In general, neural networks can be used for file classification, using activation function stated in (3), where  $a$  is the activation,  $f$  is the activation function,  $W$  is the weight matrix,  $x$  is the input vector, and  $b$  is the bias vector. Neural Network-Based Classifiers have been proposed for storage optimization [42], which uses offline classifiers based on neural networks (16-Neuron, 32-Neuron) to optimize storage by comparing with Naive Bayes Classifier and Support Vector Machine, focusing on classifier accuracy and file access latency.

$$c = \sum_{i=1}^n c_i \cdot s_i \quad \dots \dots (4)$$

Data movement costs can be calculated using (4), where  $C$  is the total cost,  $c_i$  is the cost per unit size for moving file  $i$ , and  $s_i$  is the size of file  $i$ . Data selection extracts relevant features and classifiers determine storage, proposes a comparative study of some feature selection algorithms (Fisher score, F-score and L1L21) and classifiers (SVM, K-NN, Neural Network) [43] and how they found the best combination between them for correct predictions. With these techniques, they also guarantee that only the best relevant data features are used, hence storing space can be allocated effectively [44]. The approach shown in [45] applies to the XGBoost algorithm, an enhanced gradient boosting tree model, to the optimization of data storage at a system level in distributed file systems. This machine learning framework is specifically designed to accurately predict file access patterns and subsequently move data across different storage tiers accordingly. The ARM System uses reinforcement learning with TensorFlow to learn and optimize storage based on analysis of system-level performance metrics of Ceph Storage [46]. Random forest algorithm giving best result in predicting execution need of MapReduce jobs, research output [47]. When compared against alternative models tested, including linear regression; decision trees; and gradient-boosted regression trees, it showed an unprecedented degree of predictive accuracy and effectiveness. The investigation highlighted that aspects like data volume and resource allocation become crucial in determining the effectiveness of the job. Such features were significantly useful to improve the prediction accuracy of the model and, in turn, increase resource management efficiency in the cloud-based Hadoop environments. By accurately predicting job execution times, the proposed methodology aims to improve the efficiency of resource utilization. Such an improvement is essential for the effective use of cloud-based Hadoop clusters that are used for large-scale data processing tasks [47]. Through dynamic resource allocation, better data management, increased security, predictive analytics, automated data tiering, anomaly detection, data integration, and scalability, these points collectively prove the role of machine learning in the management and enhancement of big data in the cloud.

### 3.2. Task Scheduling in Cloud Environments Using Machine Learning

ML has seen a dramatic increase in its use for solving the problem of efficient task scheduling in cloud computing [48]. This section highlights the main aspects about the role of ML on task scheduling in cloud computing. A revolutionary phenomenon of introducing ML in task scheduling frameworks has been reported in the literature with promise of significant betterment of several metrics of interest, such as makespan, energy consumption, resource utilization levels and so on. Authors have proposed a deep reinforcement learning approach, LSTM-Double Deep Q-Network (LSTM-DoubleDQN), to overcome the difficulties involved in scheduling problems [49]. This innovative strategy is specifically tailored to more adeptly navigate the dynamic characteristics inherent in cloud manufacturing as compared to traditional approaches. The LSTM-DoubleDQN algorithm shows a fast convergence speed compared to the performance of other algorithms (like DQN, LSTM-DQN, and DoubleDQN). This is an indication of its effectiveness at finding optimal solutions in a shorter time. The scheduling framework developed based on LSTM-DoubleDQN improves the QoS, making it better suited to cloud manufacturing scenarios. This indicates that the algorithm not only completes the task efficiently but delivers high-quality service as well [49]. In combination with MLRHE, the PSO-BATS bandwidth aware task allocation technique introduced in this work shows a completely new scheduling algorithm as proposed in the research [50]. The proposed methodology aims to distribute the tasks well, while minimizing the allocation errors in assigning tasks to Virtual Machines (VMs). Within the proposed methodology, this component known as PSO-BATS employs Particle Swarm Optimization (PSO) that allows for bandwidth awareness in task scheduling. It does an efficient job of laying task boundaries for optimal resource assignment and load balancing. To enhance the scheduling method, Multi-Layered Regression Host Employment (MLRHE) is integrated with PSO-BATS. The adoption of the proposed scheduling technique enhances the performance in general by minimizing cost, increasing Performance Improvement Rate (PIR), and minimizing makespan, which is the overall time needed to do a collection of jobs [50]. The present research shows the successful possible solution of task scheduling via hybridization of Moth Flame Optimization (MFO) with an Artificial Neural Network using Back-Propagation Algorithm (ANN-BPA). Compared with other swarm intelligence methods such as Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) and Cuckoo Search Optimization (CSA) [51], this hybrid method outperforms them in terms of task division, task finishing, time used and energy consumption. The findings reveal that the MFO and ANN-BPA framework delivers exceptional performance, enhancing productivity, resource efficiency, and overall operational effectiveness within cloud computing environments. The proposed Multiple Controlled Toffoli driven Adaptive Quantum Neural Network (MCT-AQNN) model works to improve workload prediction by further improving exploration, adaptation, and exploitation through quantum learning and overall system performance [54]. The approach proposed in [55] attempts to address the task scheduling challenge through a model that combines the Integral-Valued Pythagorean Fuzzy Set (IVPFS) with the Dyna Q+ algorithm. The IVPFS framework is employed to effectively address the uncertainties associated with task scheduling parameters. This integration seeks to facilitate judicious task scheduling determinations, resulting in considerable enhancements in execution duration, makespan duration, operational expenditure, and resource utilization efficiency. These points collectively illustrate the significant role of machine learning in optimizing cloud environments through dynamic resource allocation, predictive analytics, enhanced security, automated task scheduling, anomaly detection, efficient data management, cost optimization, and improved user experience. Table 2 lists the different approaches proposed and models used in the space of process optimization in cloud computing. In conclusion, the findings augment the existing body of knowledge by demonstrating the potential of machine learning-enhanced optimization methodologies in improving task scheduling within cloud computing environments.

### 3.3. Resource Provisioning and Optimization using ML in Cloud Computing

Turning the attention now to role of ML in resource provisioning and optimization, both of these are fundamental components in the domain of CC, essential for ensuring the effective utilization of resources, minimizing expenditures, and sustaining elevated performance levels. The inherently dynamic characteristics of cloud environments, variable workloads and diverse user requirements, demand sophisticated strategies for the allocation of resources. The incorporation of ML methodologies into cloud resource management systems facilitates more accurate projections of resource requirements and optimization of resource allocation, consequently resulting in enhanced system performance and diminished latency. In this paper [56] the authors propose a reinforcement learning-based proactive resource allocation framework designed for optimal resource provisioning in cloud environments. It has been shown that the RLPRAF framework can reduce overall costs by 30% and SLA (Service Level Agreement) violations by 77.7%, which is a testament to its ability to proactively and efficiently manage resources [56]. Model-as-a-Service (MaaS) is an evolving cloud computing paradigm in which machine learning models for Generative AI are deployed and accessed as on-demand services [57]. An interesting technique called Machine Learning Optimized Systems (MLOS) tackles the difficulty of efficiently benchmarking, experimenting with, and optimizing software systems in the cloud. It automates the traditional manual process of setting up experiments, collecting metrics, and analyzing results, especially in complex multi-VM setups [58]. The main algorithm that is proposed in [59] is the Depth-First-Search Coalition Reinforcement Learning (DFSCRL) provisioning policy. It includes combining these physical machines to create a coalition that may be



utilized to facilitate the provisioning of resources. It is a reinforcement learning approach for dynamic generation of the best bundle of multi-type VM instances. CILP (Co-simulation based Imitation Learner for Dynamic Resource Provisioning in Cloud Computing Environments), a type of intelligent VMs provisioning, minimizes execution cost and improves resource utilization efficiency by estimating workload demands and dynamically modifying the provisioning plans. By leveraging neural networks (NNs) and imitation learning, CILP shows better performance than state-of-the-art methods to make better decisions in the cloud environments, leading to improved overall Quality of Service QoS scores and better resource efficiency and utilization [60].

**Table 2** Algorithms proposed for task scheduling optimization in Cloud Computing

Reference article	Model Proposed	Process Optimized
shaheen et al. [50]	MLRHE (Multi-Layered Regression Host Employment)	Resolving scheduling dilemmas and promoting operational efficiency
sajjad et al. [52]	MCSO (Metaheuristic Cuckoo Search Optimization Algorithm)	Minimizing makespan and financial expenditure associated with task execution
kaur et al. [53]	MFO-ANN-BPA (Moth Flame Optimization with Artificial Neural Network using Back-Propagation Algorithm)	Resource efficiency, and overall operational effectiveness.
cao et al. [49]	LSTM-DoubleDQN	Complexities associated with scheduling issues
shamkhi et al. [52]	Neural Network-Based Classifiers	Optimize storage by comparing with Naive Bayes Classifier and Support Vector Machine

### 3.4. Challenges of integrating ML with Cloud Computing

The integration of machine learning (ML) with cloud computing introduces several critical challenges that organizations are compelled to address. A foremost obstacle is the deficiency of skilled personnel; there exists a considerable scarcity of qualified individuals who demonstrate proficiency in both ML and cloud computing technologies, thereby hindering effective deployment. The main challenge in integrating ML with cloud computing, as highlighted in the paper is to satisfy response-time Service-Level Objectives (SLOs) for inference workloads whilst minimizing serving costs. Additionally, handling occasional unpredictable load spikes poses a challenge, necessitating the use of flexible serverless instances to ensure efficient and cost-effective ML inference serving on public cloud platforms [61]. The constraints imposed by data latency and bandwidth can impede the effective transmission of extensive datasets to and from cloud environments, thereby affecting the efficacy and rapidity of data processing operations. The pay-as-you-go model of cloud computing [27] necessitates diligent oversight and enhancement of resource utilization to prevent unforeseen financial liabilities. Machine learning methodologies predicated on autonomous computing are often impeded by sluggish internet connections, resulting in latency complications associated with the transmission of data to cloud infrastructures and the subsequent receipt of responses.

### 3.5. Research Gaps in Machine Learning for Cloud Computing

Identifying the impact of workload fluctuations in one component on another is essential for anticipatory resource allocation, however, this domain remains insufficiently examined in contemporary scholarly literature [62]. In edge-cloud environments there exists a necessity for scholarly inquiry aimed at the development of platforms or protocols that enhance the exchange and aggregation of component status information, thereby ensuring dependable service provision across a multitude of components [62]. Geo-distributed systems underscore the necessity for more resilient solutions that are capable of adjusting to the intricacies inherent in the management of distributed cloud infrastructures [63]. The current literature largely prioritizes static measurements and anticipatory scaling approaches that do not respond adequately to dynamically shifting tasks, jobs, or service calls present in cloud computing frameworks. This signifies a shortcoming in the evolution of adaptive autoscaling approaches that can accommodate real-time variations in workload needs [64]. In adaptive autoscaling the existing methods in cloud resource allocation face limitations such as low convergence rates, which can hinder the effectiveness of energy consumption reduction strategies and overall performance optimization [65]. While substantial research exists, a radical examination of the performance of different

ML strategies for mitigating security vulnerabilities in cloud platforms is still needed [7]. Applying a variety of algorithms and transferring a high volume of data needs further research to improve cloud performance. Research is needed to develop ML algorithms that can efficiently scale with the increasing data and computational demands in cloud environments. Developing foolproof systems that are resistant to both external and internal attacks remains a field requiring extensive research [8]. High energy consumption and computational overheads are significant concerns in several applications including cybersecurity. Research is needed to develop more energy-efficient ML algorithms for cloud computing [8]. These points highlight the key research gaps in machine learning for cloud computing, emphasizing the need for comprehensive evaluations, addressing communication overheads, latency, scalability, security, system complexity, and energy consumption.

---

#### 4. Conclusion

The convergence of Artificial Intelligence (AI), Machine Learning (ML), and Cloud Computing have become a driving force in reshaping the technological landscape. Together, these technologies offer unprecedented opportunities to enhance scalability, efficiency, and innovation across various industries. Cloud Computing provides the flexible infrastructure required to support the vast computational and storage demands of AI and ML models, enabling organizations to process and analyze large volumes of data in real time. In return, AI and ML augment cloud services by introducing automation, intelligent decision-making, and predictive analytics, thereby optimizing resource management, improving performance, and lowering operational costs. This synergistic relationship has already demonstrated transformative impacts, from accelerating innovation in sectors like healthcare and finance to enabling smarter, more responsive digital systems in retail, manufacturing, and beyond. On the downside, as with any technologies, there are challenges that arise with the integration of these technologies including data privacy issues, ethical considerations in AI decision-making, and the complexity of managing continually evolving cloud environments. Solving these challenges will be at the heart of unlocking the true potential of AI, ML, and Cloud Computing. The future will witness even deeper integration of these technologies through advances such as edge computing, federated learning, and quantum computing, which are set to take their synergy to new heights. In conclusion, the synergy between AI, ML, and Cloud Computing is not only shaping today's digital transformation but also setting the foundation for a smarter, more agile technological future. Their complementary strengths offer a compelling pathway to optimizing business processes, solving complex problems, and enabling scalable solutions across industries.

---

#### Compliance with ethical standards

##### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

---

#### References

- [1] Feifei Shi, Huansheng Ning, Wei Huangfu, Fan Zhang, Dawei Wei, Tao Hong, and Mahmoud Daneshmand. Recent progress on the convergence of the internet of things and artificial intelligence. *IEEE Network*, 34(5):8–15, 2020.
- [2] Chuan Zhang, Liehuang Zhu, Chang Xu, and Rongxing Lu. PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-healthcare system. *Future Generation Computer Systems*, 79:16–25, 2018.
- [3] Ahmed Saleh Bataineh, Jamal Bentahar, Rabeb Mizouni, Omar Abdel Wahab, Gaith Rjoub, and May El Barachi. Cloud Computing as a Platform for Monetizing Data Services: A Two-Sided Game Business Model. *IEEE Transactions on Network and Service Management*, 19(2):1336–1350, June 2022.
- [4] Peter Mell and Timothy Grance. The NIST definition of cloud computing. *The NIST Definition of Cloud Computing*, 2012.
- [5] Mehmet Sahinoglu and Luis Cueva-Parra. CLOUD computing. *WIREs Computational Statistics*, 3(1):47–68, 2011-01.
- [6] Aptin Babaei, Parham M. Kebria, Mohsen Moradi Dalvand, and Saeid Nahavandi. A Review of Machine Learning-based Security in Cloud Computing, September 2023. arXiv:2309.04911 [cs].
- [7] Wassim Safi, Sameh Ghwanmeh, Mahmoud Mahfuri, and Waleed T. Al-Sit. Enhancing Cloud Security: A Comprehensive Review of Machine Learning Approaches. In *2024 2nd International Conference on Cyber Resilience (ICCR)*, pages 1–10, Dubai, United Arab Emirates, February 2024. IEEE.

- [8] Meghna Manoj Nair and Amit Kumar Tyagi. AI, IoT, blockchain, and cloud computing: The necessity of the future. In *Distributed Computing to Blockchain*, pages 189–206. Elsevier, 2023.
- [9] Nachaat Mohamed, L. Sridhara Rao, Manju Sharma, SureshBabuRajasekaranl, BadriaSulaimanAlfurhood, and Surendra Kumar Shukla. In-depth review of integration of AI in cloud computing. In *2023 3<sup>rd</sup> International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1431–1434, Greater Noida, India, May 2023. IEEE.
- [10] Yuan Zheng and Xiangbin Wen. The Application of Artificial Intelligence Technology in Cloud Computing Environment Resources. *Journal of Web Engineering*, October 2021.
- [11] Mamata Rath, Jyotirmaya Satpathy, and George S. Oreku. Artificial Intelligence and Machine Learning Applications in Cloud Computing and Internet of Things. In *Artificial Intelligence to Solve Pervasive Internet of Things Issues*, pages 103–123. Elsevier, 2021.
- [12] Beniamino Di Martino, Antonio Esposito, and Ernesto Damiani. Towards AI-Powered Multiple Cloud Management. *IEEE Internet Computing*, 23(1):64–71, Jan-uary 2019.
- [13] Neelesh Mungoli. Scalable, distributed AI frameworks: Leveraging cloud computing for enhanced deep learning performance and efficiency. Version Number: 1.
- [14] Nicolas Mialhe and Cyrus Hodes. The third age of artificial intelligence. *Field Actions Science Reports*, pages 6–11, 2017. Publisher: Institut Veolia Environnement.
- [15] Shijun Shen, Jiuling Zhang, Daochao Huang, and Jun Xiao. Evolving from traditional systems to AIOps: Design, implementation and measurements. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications( AEECA)*, pages 276–280. IEEE, 2020.
- [16] Qian Cheng, Doyen Sahoo, Amrita Saha, Wenzhuo Yang, Chenghao Liu, Gerald Woo, Manpreet Singh, Silvio Saverese, and Steven C. H. Hoi. AI for IT Operations (AIOps) on Cloud Platforms: Reviews, Opportunities and Challenges, April 2023. arXiv:2304.04661[cs].
- [17] Latha Narayanan Valli, N. Sujatha, and E. Joice Rathinam. A Study on Deep Learning Frameworks to Understand the Real Time Fault Detection and Diagnosis in IT Operations with AIOps. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, pages 1–6, Bengaluru, India, October 2023. IEEE.
- [18] Syed Imran Abbas and Ankit Garg. AIOps in DevOps: Leveraging Artificial Intelligence for Operations and Monitoring. In *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pages 64–70, Bhimdatta, Nepal, March 2024. IEEE.
- [19] Sukhpal Singh Gill, Minxian Xu, Carlo Ottaviani, Panos Patros, Rami Bahsoon, Arash Shaghaghi, Muhammed Golec, Vlado Stankovski, Huaming Wu, Ajith Abraham, Manmeet Singh, Harshit Mehta, Soumya K. Ghosh, Thar Baker, Ajith Kumar Parlikad, Hanan Lutfiyya, Salil S. Kanhere, Rizos Sakellariou, Schahram Dustdar, Omer Rana, Ivona Brandic, and Steve Uhlig. AI for next generation computing: Emerging trends and future directions. *Internet of Things*, 19:100514, August 2022.
- [20] Sasho Nedelkoski, Jorge Cardoso, and Odej Kao. Anomaly detection and classification using distributed tracing and deep learning. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 241–250. IEEE, 2019.
- [21] Chetankumar Kalaskar and S. Thangam. Fault Tolerance of Cloud Infrastructure with Machine Learning. *Cybernetics and Information Technologies*, 23(4):26–50, November 2023.
- [22] Mandeep Kaur, Vishal Jain, Parma Nand, and Nitin Rakesh. *Software-Defined Network Frameworks: Security Issues and Use Cases*. CRC Press, 1 edition, 2024.
- [23] Everton C. De Lima, F´abio D. Rossi, Marcelo C. Luizelli, Rodrigo N. Calheiros, and Arthur F. Lorenzon. A neural network framework for optimizing parallel computing in cloud servers. *Journal of Systems Architecture*, 150:103131, May 2024.
- [24] Sundas Iftikhar, Mirza Mohammad Mufleh Ahmad, Shreshth Tuli, Deepraj Chowdhury, Minxian Xu, Sukhpal Singh Gill, and Steve Uhlig. HunterPlus: AI based energy-efficient task scheduling for cloud-fog computing environments. *Internet of Things*, 21:100667, April 2023.
- [25] Weipeng Gao and Youchan Zhu. A Cloud Computing Fault Detection Method Based on Deep Learning. *Journal of Computer and Communications*, 05(12):24–34, 2017.


- [26] Nagamany Abirami and M. S. Anbarasi. An efficient multilayer approach for securing e-healthcare data in cloud using crypto – stego technique. *Engineering World*, 6:128–135, 2024.
- [27] Andrew McDole, Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal, and Mamoun Alazab. Deep learning techniques for behavioral malware analysis in cloud IaaS. In Mark Stamp, Mamoun Alazab, and Andrii Shalaginov, editors, *Malware Analysis Using Artificial Intelligence and Deep Learning*, pages 269–285. Springer International Publishing, 2021.
- [28] Hadeel T. El-Kassabi, Mohamed Adel Serhani, Mohammad M. Masud, Khaled Shuaib, and Khaled Khalil. Deep learning approach to security enforcement in cloud workflow orchestration. *Journal of Cloud Computing*, 12(1):10, January 2023.
- [29] Ly Vu, Quang Uy Nguyen, Diep N. Nguyen, Dinh Thai Hoang, and Eryk Dutkiewicz. Deep Generative Learning Models for Cloud Intrusion Detection Systems. *IEEE Transactions on Cybernetics*, 53(1):565–577, January 2023.
- [30] E Balamurugan, Abolfazl Mehbodniya, Elham Kariri, Kusum Yadav, Anil Kumar, and Mohd Anul Haq. Network optimization using defender system in cloud computing security based intrusion detection system with game theory deep neural network (IDSGT-DNN). *Pattern Recognition Letters*, 156:142–151, April 2022.
- [31] Wenjuan Wang, Xuehui Du, Dibin Shan, Ruoxi Qin, and Na Wang. Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine. *IEEE Transactions on Cloud Computing*, 10(3):1634–1646, July 2022.
- [32] Xiaolong Li, Tingting Wang, Wei Zhang, Xu Niu, Tingyu Zhang, Tengting Zhao, Yongji Wang, and Yufei Wang. An LSTM based cross-site scripting attack detection scheme for Cloud Computing environments. *Journal of Cloud Computing*, 12(1):118, August 2023.
- [33] Omer Aslan, Merve Ozkan-Okay, and Deepti Gupta. Intelligent Behavior-Based Malware Detection System on Cloud Computing Environment. *IEEE Access*, 9:83252–83271, 2021.
- [34] Omar Abdel Wahab, Jamal Bentahar, Hadi Otrok, and Azzam Mourad. I Know You Are Watching Me: Stackelberg-Based Adaptive Intrusion Detection Strategy for Insider Attacks in the Cloud. In 2017 IEEE International Conference on Web Services (ICWS), pages 728–735, Honolulu, HI, USA, June 2017. IEEE.
- [35] Noushin Pervez. noushinpervez/intrusion-detection CICIDS2017. original-date: 2023-10-19T21:59:11Z.
- [36] Kishu Gupta, Deepika Saxena, Rishabh Gupta, Jatinder Kumar, and Ashutosh Kumar Singh. FedMUP: Federated learning driven malicious user prediction model for secure data distribution in cloud environments. *Applied Soft Computing*, 157:111519, 2024.
- [37] Hsiao-Wuen Hon. AI for System - Infusing AI into Cloud Computing Systems. *ACM SIGMETRICS Performance Evaluation Review*, 49(1):39–40, June 2021.
- [38] Sarina Aminizadeh, Arash Heidari, Mahshid Dehghan, Shiva Toumaj, Mahsa Rezaei, Nima Jafari Navimipour, Fabio Stroppa, and Mehmet Unal. Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. *Artificial Intelligence in Medicine*, 149:102779, 2024.
- [39] Mohamed Ouhssini, Karim Afdel, Mohamed Akouhar, Elhafed Agherrabi, and Abdallah Abarda. Advancements in detecting, preventing, and mitigating DDoS attacks in cloud environments: A comprehensive systematic review of state-of-the-art approaches. *Egyptian Informatics Journal*, 27:100517, 2024.
- [40] Abdul Salam Mohammad and Manas Ranjan Pradhan. Machine learning with big data analytics for cloud security. *Computers & Electrical Engineering*, 96:107527, December 2021.
- [41] Rui Huang and Shucheng Fang. Machine learning and big data analytics in the cloud environment. *International Journal of Cloud Computing and Database Management*, 5(1):06–08, January 2024.
- [42] Jinting Ren, Xianzhang Chen, Yujuan Tan, Duo Liu, Moming Duan, Liang Liang, and Lei Qiao. Archivist: A Machine Learning Assisted Data Placement Mechanism for Hybrid Storage Systems. In 2019 IEEE 37<sup>th</sup> International Conference on Computer Design (ICCD), pages 676–679, Abu Dhabi, United Arab Emirates, November 2019. IEEE.
- [43] Karamjit Kaur and Rinkle Rani. Managing data in healthcare information systems: Many models, one solution. *Computer*, 48(3):52–59, 2015.
- [44] Anindita Sarkar Mondal, Anirban Mukhopadhyay, and Samiran Chattopadhyay. Machine learning-driven automatic storage space recommendation for object-based cloud storage system. *Complex & Intelligent Systems*, 8(1):489–505, February 2022.

- [45] Herodotos Herodotou and Elena Kakoulli. Automating distributed tiered storage management in cluster computing. *Proceedings of the VLDB Endowment*, 13(1):43–56, September 2019.
- [46] Faisal S. Alsubaei, Ahmed Y. Hamed, Moatamad R. Hassan, M. Mohery, and M. Kh. Elnahary. Machine learning approach to optimal task scheduling in cloud communication. *Alexandria Engineering Journal*, 89:1–30, February 2024.
- [47] Mohammed Bergui, Soufiane Hourri, Said Najah, and Nikola S. Nikolov. Predictive modelling of MapReduce job performance in cloud environments using machine learning techniques. *Journal of Big Data*, 11(1):98, 2024.
- [48] Mohamed Lahby, Al-Sakib Khan Pathan, and Yassine Maleh. *Combatting Cyberbullying in Digital Media with Artificial Intelligence*. Chapman and Hall/CRC, 1 edition, 2023.
- [49] Yu Cao, Ming Lv, Xingbo Qiu, Yongkui Liu, and Xubin Ping. Cloud Manufacturing Task Scheduling Under Machinery Breakdown Based on Deep Reinforcement Learning. In *2023 China Automation Congress (CAC)*, pages 3742–3747, Chongqing, China, November 2023. IEEE.
- [50] Anwar R Shaheen and Sundar Santhosh Kumar. Tasks Scheduling in Cloud Environment Using PSO-BATS with MLRHE. *Intelligent Automation & Soft Computing*, 35(3):2963–2978, 2023.
- [51] Xin-She Yang, editor. *Cuckoo Search and Firefly Algorithm: Theory and Applications*, volume 516 of *Studies in Computational Intelligence*. Springer International Publishing, Cham, 2014.
- [52] sajjad shamkhi jaber, Yossra Ali, and Nuha Ibrahim. Task Scheduling in Cloud Computing Based on The Cuckoo Search Algorithm. *Iraqi Journal of Computer, Communication, Control and System Engineering*, pages 86–96, March 2022.
- [53] Surinder Kaur, Jaspreet Singh, and Vishal Bharti. A Comparative Study of Optimization Based Task Scheduling in Cloud Computing Environments Using Machine Learning. In *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 731–740, Tirunelveli, India, March 2024. IEEE.
- [54] Ishu Gupta, Deepika Saxena, Ashutosh Kumar Singh, and Chung-Nan Lee. A Multiple Controlled Toffoli Driven Adaptive Quantum Neural Network Model for Dynamic Workload Prediction in Cloud Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7574–7588, December 2024.
- [55] Bhargavi Krishnamurthy and Sajjan G. Shiva. Integral valued pythagorean fuzzy-set-based dyna q+ framework for task scheduling in cloud computing. *Sensors*, 24(16):5272, 2024.
- [56] Reena Panwar and M. Supriya. RLPRAF: Reinforcement Learning-Based Proactive Resource Allocation Framework for Resource Provisioning in Cloud Environment. *IEEE Access*, 12:95986–96007, 2024.
- [57] Harshad Pitkar, Sanjay Bauskar, Devendra Singh Parmar, and Hemlatha Kaur Saran. Exploring model-as-a-service for generative ai on cloud platforms. *Review of Computer Engineering Research*, 11(4):140–154, Dec. 2024.
- [58] Brian Kroth, Sergiy Matushevych, Rana Alotaibi, Yiwen Zhu, Anja Gruenheid, and Yuanyuan Tian. MLOS in action: Bridging the gap between experimentation and auto-tuning in the cloud. *Proceedings of the VLDB Endowment*, 17(12):4269–4272, 2024.
- [59] Waheed Iqbal, Abdelkarim Erradi, Muhammad Abdullah, and Arif Mahmood. Predictive Auto-Scaling of Multi-Tier Applications Using Performance Varying Cloud Resources. *IEEE Transactions on Cloud Computing*, 10(1):595–607, January 2022.
- [60] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. CILP: Co-simulation based Imitation Learner for Dynamic Resource Provisioning in Cloud Computing Environments, April 2023. arXiv:2302.05630 [eess].
- [61] Chengliang Zhang, Minchen Yu, Wei Wang, and Feng Yan. Enabling Cost-Effective, SLO-Aware Machine Learning Inference Serving on Public Cloud. *IEEE Transactions on Cloud Computing*, 10(3):1765–1779, July 2022.
- [62] Thang Le Duc, Rafael Garcí'a Leiva, Paolo Casari, and Per-Olov Östberg. Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. *ACM Computing Surveys*, 52(5):1–39, September 2020.
- [63] Ninad Hogade and Sudeep Pasricha. A Survey on Machine Learning for Geo-Distributed Cloud Data Center Management. *IEEE Transactions on Sustainable Computing*, 8(1):15–31, January 2023.

- [64] Istv'an Pintye, J'ozsef Kov'acs, and R'obert Lovas. Enhancing machine learning-based autoscaling for cloud resource orchestration. *Journal of Grid Computing*, 22(4):68, 2024.
- [65] Stanly Jayaprakash, Manikanda Devarajan Nagarajan, Roc'io P'erez De Prado, Sugumaran Subramanian, and Parameshachari Bidare Divakarachari. A Systematic Review of Energy Management Strategies for Resource Allocation in the Cloud: Clustering, Optimization and Machine Learning. *Energies*, 14(17):5322, August 2021.

---

### Author's short biography

<p>Authors Name: Harshad Pitkar</p> <p>Harshad Pitkar holds a Master's degree in Data Science from Indiana University, Bloomington, Indiana. He has over 20 years of experience in the field of Information Technology. He is currently working as an Engineering Leader at Cummins Inc. and likes to stay at the forefront of emerging technologies. He is an independent researcher, with a focus on fields like Cloud computing, automation, machine learning and artificial intelligence.</p>	
<p>Authors Name: Sumedh Ambapkar</p> <p>Sumedh Ambapkar is pursuing master's in data science from Indian University of Bloomington. He has over 13 years of experience in Information Technology and worked with organization like IBM and Amazon Web Services Inc. Currently he is working as Senior Analyst at Cummins Inc. He has expertise over serverless technologies, AI-driven analytics, and Zero-emission Initiatives. He drives innovation in cloud Services, workload migrations and sustainable IT strategies ensuring impactful solutions aligned with business and environmental goals.</p>	