A Survey of AI-Based Methods for Cloud Resource Allocation and Optimization

Rim Doukha, Abderrahmane Ez-zahout
Intelligent Processing and Security of Systems Team-Faculty of Sciences, Mohammed V University, Rabat, Morocco

***Abstract*—Cloud computing has become essential for modern digital services, yet efficiently allocating compute, storage, and network resources in large-scale and highly dynamic environments remains a significant challenge. Traditional rule-based approaches often struggle to cope with workload variability, multi-tenancy, and the need for real-time multi-objective optimization. In response, recent research has increasingly explored artificial intelligence techniques to improve prediction, scheduling, and automated resource control in cloud infrastructures. This study presents a comprehensive survey of AI-based methods for cloud resource allocation, including machine learning, deep learning, reinforcement learning, and hybrid approaches. It systematically analyzes selected studies published between 2020 and 2026, examining their learning paradigms, optimization objectives (e.g., performance, cost, energy efficiency), experimental validation strategies, and reported limitations. While classical optimization techniques are briefly discussed to contextualize the evolution of the field, the core analysis is strictly centered on AI-driven approaches. The study concludes by identifying the key challenges that persist in intelligent cloud resource management and outlines promising directions for future research toward more adaptive, reliable, and scalable optimization frameworks.**

***Keywords*—AI techniques; heuristics; metaheuristic; cloud resource management; sustainability; survey**

I. Introduction

Cloud computing has emerged as the foundational infrastructure model of the digital era, offering scalable, on-demand access to computing, storage, and network resources [1]. Its elasticity and cost-efficiency have transformed how enterprises, governments, and end users deploy and manage digital services [2]. However, as workloads have become more dynamic and compute-intensive, driven by trends such as artificial intelligence (AI), big data analytics, and Internet of Things (IoT), the task of efficiently allocating resources has become more complex and critical.

At the heart of this complexity lies the challenge of resource allocation, mainly deciding how to dynamically distribute CPU, memory, bandwidth, and storage across diverse virtualized environments, often in real time. Suboptimal allocation leads to underutilized infrastructure [3], violated service-level agreements (SLAs), increased operational costs, and a growing carbon footprint. According to the International Energy Agency, data centers now consume approximately 1–1.5% of global electricity, a share that is projected to rise due to expanding cloud and AI workloads [4].

Traditional approaches to resource allocation have relied on rule-based heuristics such as First Fit, Round Robin, and Greedy algorithms. While effective in static or predictable environments, these strategies struggle in today's multi-tenant, heterogeneous, and latency-sensitive cloud architectures [5]. They are typically reactive, single-objective, and unable to adapt to workload volatility or evolving performance requirements [6].

Recent advances in AI and the growing availability of telemetry data from cloud environments have enabled a shift toward learning-based resource allocation methods. Machine learning (ML) models are now being used to predict workload demands, detect anomalies, and estimate resource requirements, while deep learning (DL) architectures such as LSTM and Transformer models enable fine-grained time-series forecasting. Reinforcement learning (RL), in particular, has emerged as a promising solution for dynamic and autonomous resource allocation, continuously optimizing placement and scheduling decisions based on environmental feedback. In parallel, metaheuristic techniques like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) continue to evolve as powerful tools for multi-objective optimization [5].

Despite this progress, significant challenges remain. Many AI-based allocation strategies depend on large, clean, and context-specific datasets that are not always available in practice. Moreover, deploying intelligent controllers in production requires satisfying strict latency, reliability, and interpretability constraints. As cloud providers increasingly adopt sustainability goals, resource allocation methods must also account for energy efficiency and carbon awareness, making the optimization task even more multi-dimensional.

While heuristic and metaheuristic approaches have historically played an important role in cloud resource allocation, this study does not aim to provide an exhaustive analysis of these methods as standalone solutions. Instead, they are discussed primarily to contextualize the evolution of cloud resource management and, more importantly, to highlight their role within hybrid frameworks, where they are combined with AI-based techniques to enhance scalability, convergence speed, and decision efficiency. The core focus of this survey is, therefore, on AI-driven resource allocation, including ML, DL, RL, and hybrid AI-based approaches that integrate classical optimization components.

This study presents a comprehensive overview of AI-based methods for optimizing resource allocation in cloud

computing environments. It aims to clarify the current state of research, identify open challenges, and propose future directions for intelligent and sustainable cloud infrastructure management.

This survey is guided by the following research questions:

RQ1: What AI-based techniques are currently used for cloud resource allocation, and how are they applied?

RQ2: What optimization objectives are addressed by AI-driven and hybrid allocation frameworks?

RQ3: To what extent do existing AI-based approaches incorporate sustainability-aware mechanisms, such as energy-efficient scheduling or carbon-aware resource allocation?

RQ4: What methodological limitations persist in current AI-based cloud allocation research, particularly regarding scalability, real-world validation, generalization, and deployment feasibility?

The study is structured as follows: Section II presents the related work and contributions. Section III explores the key methodologies used in cloud resource allocation, including heuristics, metaheuristics, AI-based approaches, and hybrid models. Section IV gives an overview of the existing research on cloud resource allocation. Section V provides insights and research opportunities for the next generation of cloud optimization techniques. Section VI presents future directions. Finally, Section VII concludes the study.

II. Related Work and Contribution

The increasing complexity and scale of cloud environments have driven significant interest in intelligent resource management strategies. Consequently, several survey papers have emerged over the past few years, focusing on different aspects of AI-based cloud resource optimization, including ML, DL, RL and metaheuristics.

Rathee and Dalal [7] conducted a systematic literature review of ML-based resource allocation techniques in cloud environments, focusing primarily on traditional supervised and unsupervised learning paradigms for virtual machine provisioning, scheduling, and load balancing. Their review does not extensively examine RL or DL methods, hybrid AI models, or sustainability-aware strategies, such as carbon-aware scheduling and green autoscaling.

Naidu and Kumar [8] present a survey of AI-based approaches for intelligent cloud resource allocation, with a particular focus on DL, ensemble models, and metaheuristic optimization techniques. Their review highlights predictive resource scaling methods designed to improve cost efficiency, energy consumption, and SLA adherence. While the study provides valuable insights into algorithmic trends, it offers limited cross-comparative analysis and does not deeply explore RL, hybrid AI frameworks, or sustainability-driven scheduling mechanisms.

Pitkar and Ambapkar [9] provide an informative overview of the intersection between AI, ML, and cloud computing, aiming to highlight key models, challenges, and opportunities. Their work brings attention to relevant themes such as scalability, automation, and data-driven optimization, contributing to the conceptual understanding of AI integration in cloud platforms. However, the review adopts a generalist perspective. Advanced AI paradigms such as RL or sustainability-focused strategies are not deeply explored.

Srikandabala et al. [10] propose an AI-based framework to enhance data center network performance through flow-aware routing and traffic optimization. Their focus on network-level efficiency offers valuable insights, particularly in reducing congestion and improving throughput. However, the work remains confined to the network layer and does not address broader cloud resource scheduling challenges. Key aspects such as workload prediction, sustainability-aware provisioning, and advanced learning paradigms like reinforcement or hybrid learning are not explored.

Mohammad and Al-Ta'I [11] present a comprehensive comparison of hybrid DL architectures combining CNN, GRU, and LSTM for virtual machine scheduling in cloud environments. While their study effectively demonstrates the predictive capabilities of DL models, it does not address RL, which has become increasingly important for dynamic and feedback-driven resource allocation in modern cloud systems.

Varun et al. [12] present a conceptual overview of AI and ML approaches for cloud resource optimization. While their discussion identifies key challenges and high-level strategies, it does not offer a structured analysis of existing literature or model comparisons.

Alhashimi et al. [13] conducted a structured survey of AI-enabled resource management techniques targeting 6G cloud and edge computing environments. Their work categorizes ML, DL, and RL approaches based on application areas such as task scheduling, energy management, and latency reduction. The study includes a well-defined taxonomy and comparative analysis. However, its scope is primarily centered on 6G-oriented network architectures and edge-cloud integration, with less emphasis on generalized sustainability-aware strategies or hybrid AI models in traditional cloud infrastructures.

Nawrocki and Smendowski [14] provide a structured survey of optimization strategies for cloud resource consumption, reviewing over 70 research articles. Their work introduces a novel taxonomy based on semantic properties, evaluating each method's capabilities, deployment potential, and limitations. While the survey thoroughly addresses cost-awareness and consumption control, DL and RL approaches receive limited attention, and sustainability considerations are only briefly discussed.

Collectively, these surveys provide valuable insights into intelligent cloud resource management. However, several common limitations emerge. Many studies focus on individual paradigms, such as ML or DL, without offering cross-comparative analysis across ML, DL, and RL approaches. Hybrid frameworks that integrate predictive modeling with adaptive control remain underrepresented. Furthermore, sustainability considerations, particularly carbon-aware optimization, are often treated indirectly or receive limited emphasis.

This study addresses these gaps by providing a structured overview of AI-driven resource allocation techniques for cloud computing. It synthesizes main studies published between 2020 and 2026, organizes them into methodological categories spanning ML, DL, RL, and hybrid models, and comparatively analyzes their optimization objectives, including performance, cost, energy efficiency, and sustainability considerations, along with reported limitations. By offering a cross-paradigm synthesis, this survey aims to provide a clear and structured reference for researchers and practitioners in intelligent cloud resource management.

This study makes the following contributions:

- It provides a structured overview of AI-based cloud resource allocation methods, synthesizing selected studies published between 2020 and 2026.
- It offers a comparative synthesis of recent literature, analyzing how AI-driven approaches address key optimization objectives such as performance, cost, energy efficiency, SLA compliance, and sustainability considerations.
- It examines the role of hybrid resource allocation frameworks, particularly those integrating ML-based prediction, RL-based adaptive control, and classical optimization components to enhance adaptability and scalability.
- It identifies open research challenges and outlines future research directions, emphasizing scalability, real-world validation, sustainability-aware optimization, and interpretability in AI-driven cloud resource management systems.

III. Methodologies used in Cloud Resource Allocation

Cloud resource allocation has been approached through a diverse set of methodologies that evolved in response to increasing system complexity and dynamic workload behavior. The earliest solutions were primarily heuristic, relying on static rules that map tasks or virtual machines to available resources. These methods were effective in early cloud environments due to their simplicity, deterministic execution, and low computational overhead. As cloud infrastructures grew in scale and heterogeneity, metaheuristic techniques emerged to overcome the limited search capability of heuristics, offering more flexible and global optimization behavior. The continued increase in workload diversity, real-time constraints, and multi-objective optimization requirements eventually led to the rapid adoption of AI–based methods, including ML, DL, and RL.

These AI techniques provide predictive and adaptive capabilities that allow resource management systems to learn from operational data and continuously improve decision-making. More recently, hybrid models combining classical strategies with AI methods or integrating multiple AI approaches have become increasingly prominent, reflecting the need for systems that balance interpretability, computational efficiency, and intelligent adaptation. The following subsections present these methodological categories:

A. Heuristic Methods

Heuristic methods are among the earliest and most widely applied approaches for managing cloud resources. These techniques rely on predefined rules to determine how virtual machines, containers, or tasks are scheduled across available infrastructure. Their popularity stems from low computational overhead, fast execution, and straightforward implementation, making them suitable for real-time scenarios where speed is prioritized over optimality [15].

Classic heuristics, such as First Fit, Best Fit, Round Robin, and Greedy Scheduling, are commonly implemented in simulation environments like CloudSim and OpenDC [16]. These algorithms follow simple logic [17]. For example, First Fit places tasks on the first available host with sufficient resources, while Best Fit selects the host with the tightest fit without exceeding capacity [18]. Such methods were especially effective in earlier cloud environments characterized by consistent and predictable workloads.

In general, heuristic methods fall short when multiple objectives such as cost, energy, latency, and SLA adherence must be considered simultaneously [19]. Their static rules often lack adaptability and can struggle under fluctuating workloads or failure-prone environments.

Nonetheless, heuristic strategies continue to hold value in modern cloud systems. Their transparent logic, rapid decision-making, and minimal computational demand make them particularly relevant in edge computing and latency-sensitive contexts. When integrated into hybrid optimization frameworks, they often serve as fallback mechanisms or fast initializers within broader decision pipelines. This enduring utility ensures they remain part of the resource allocation toolbox, even as more intelligent and autonomous solutions gain prominence.

B. Metaheuristic Methods

Metaheuristic methods emerged as a response to the shortcomings of classical heuristics. Inspired by natural and physical processes, these algorithms employ stochastic exploration and population-based search strategies to approximate near-optimal solutions in complex and high-dimensional search spaces [20]. Commonly used metaheuristics include GA, PSO, Ant Colony Optimization, Simulated Annealing, and Differential Evolution. Their ability to escape local optima and search globally makes them particularly well-suited for resource allocation problems involving multiple conflicting objectives, such as balancing energy efficiency with performance guarantees or optimizing resource utilization while minimizing operating costs [21].

Unlike heuristics, metaheuristics do not rely on rigid rules but instead adaptively tune their search based on performance feedback. Although more flexible and powerful, metaheuristics introduce their own challenges, including sensitivity to parameter initialization, computational overhead during large search operations, and difficulties associated with scaling these methods to real-time production environments. Nonetheless, they remain a widely adopted solution in research due to their versatility and effectiveness in exploring complex optimization landscapes.

C. AI-Based Methods

AI refers to the computational capability of machines to perform tasks that typically require human intelligence, including learning, reasoning, decision-making, and problem-solving [22]. In the context of cloud computing, AI enables the development of systems that can autonomously manage, adapt, and optimize infrastructure behavior based on real-time data and environmental feedback. As cloud infrastructure becomes increasingly dynamic and complex, with fluctuating workloads, multi-tenant architectures, and stringent service-level objectives, conventional optimization techniques such as heuristic or metaheuristic methods fall short in their ability to generalize, adapt, and respond in real time. This limitation has catalyzed growing interest in AI-based methods, which introduce the potential for systems to make data-driven, context-aware decisions that continuously improve over time [23].

Within the domain of cloud resource allocation, AI-based methods are typically organized into three major paradigms: ML, DL, and RL.

ML involves predictive modeling based on historical or streaming data, commonly applied for forecasting resource demand, detecting anomalies, or estimating execution times. DL, a subfield of ML, leverages multi-layer neural network architectures to model complex temporal and multivariate patterns such as workload bursts or correlated latency anomalies. RL, on the other hand, focuses on sequential decision-making in uncertain environments, enabling systems to learn optimal resource-control policies through trial-and-error interactions with their environment. Each of these paradigms contributes unique strengths: ML offers fast and interpretable prediction, DL excels at modeling non-linear high-dimensional data, and RL supports adaptive, autonomous control in dynamic settings.

These approaches have gained significant traction due to the growing availability of cloud telemetry data, such as system logs, performance counters, and application traces, along with large-scale public datasets, including Google Cluster Data, Alibaba traces, and Microsoft Azure VM traces. Additionally, simulation frameworks such as CloudSim and OpenDC have provided testbeds for evaluating AI-driven strategies in controlled environments. As the complexity of resource allocation continues to increase, AI-based methods are rapidly emerging as central to the design of intelligent and sustainable cloud management systems.

D. Hybrid Model

Hybrid models have gained increasing attention as a practical response to the limitations of standalone optimization approaches in cloud resource allocation. These models combine complementary techniques to improve scalability, adaptability, and efficiency in dynamic cloud environments.

In practice, hybrid approaches may integrate multiple AI paradigms, such as using predictive models for workload estimation alongside adaptive control mechanisms, or combine AI-based methods with classical heuristics or metaheuristics to balance learning capabilities with fast and stable decision-making. By leveraging the strengths of different methodologies, hybrid models enable flexible trade-offs between performance, computational overhead, and sustainability objectives, making them well-suited for modern, large-scale cloud infrastructures.

IV. Literature Selection Methodology

This survey adopts a structured literature analysis approach to examine AI-based cloud resource allocation techniques published between 2020 and 2026. Relevant peer-reviewed journal and conference papers were identified through major scientific databases, including IEEE Xplore, ACM Digital Library, Scopus, and ScienceDirect, using keywords related to cloud resource allocation, resource scheduling, AI, ML, DL, RL, sustainability-aware scheduling, and hybrid optimization.

The initial set of retrieved studies was filtered using explicit inclusion and exclusion criteria. The inclusion criteria comprised: 1) peer-reviewed publications, 2) studies focusing on resource allocation or scheduling in cloud or cloud-edge environments, and 3) approaches based on AI techniques or hybrid optimization frameworks. The exclusion criteria included: 1) studies not directly addressing resource allocation, 2) works lacking experimental validation, and 3) duplicate or non-peer-reviewed sources.

After screening titles, abstracts, and full texts, a final set of representative studies was selected for detailed analysis. These studies were then organized into methodological and thematic categories, including RL-based dynamic allocation, predictive ML and DL approaches, hybrid AI and metaheuristic frameworks, network- and multi-data-center-aware allocation, and sustainability- and carbon-aware resource allocation. The selected works were comparatively analyzed based on their optimization objectives, evaluation strategies, and reported limitations.

The analysis reveals a clear transition from static scheduling policies toward adaptive and learning-driven optimization frameworks in recent cloud resource allocation research.

V. Existing Research on Cloud Resource Allocation

Table I provides a structured comparative overview of the analyzed studies, highlighting their cloud context, AI techniques, optimization objectives, sustainability considerations, strengths, and limitations.

The thematic categories presented in this section were derived inductively from the comparative analysis summarized in Table I. The reviewed studies were examined to identify recurring methodological patterns, architectural strategies, and optimization priorities in recent AI-based cloud resource allocation research.

TABLE I. COMPARATIVE ANALYSIS OF AI-BASED AND HYBRID CLOUD RESOURCE ALLOCATION APPROACHES (2020–2026)

Ref	Cloud Context	AI Technique	Hybrid Components	Optimization Objectives	Sustainability Aspect	Key Strengths	Limitations
[24]	Mobile Edge Computing	DRL	Hybrid DRL + multi-objective Pareto-optimal selection	Task latency reduction, task failure rate reduction, load balancing	Energy considered in reward function, but no explicit green or sustainability optimization	Hierarchical and distributed DRL improves scalability and significantly reduces latency and failures	High training and communication overhead; increased network traffic due to multi-level coordination
[25]	Fog-Cloud environments	Deep Q-Network (DQN) RL + Fuzzy logic	Hybrid fuzzy logic + DRL	Minimize delay, energy consumption, and makespan	Explicit energy reduction objective (energy consumption is measured)	Prioritizes tasks via fuzzy inference and improves scheduling performance across multiple metrics	Simulation only; real-world deployment not validated
[26]	Cloud-Edge computing environment	DRL (DQN-based)	None	Energy consumption reduction, cost minimization, makespan reduction, QoS satisfaction	Energy-aware, sustainability mentioned indirectly; no explicit carbon or green cloud optimization	Multi-objective RL framework with strong performance across energy, cost, and QoS metrics; validated on realistic traces	High training complexity; evaluated only in simulation; carbon and thermal aspects not addressed
[27]	Kubernetes-based cloud environment	ML	None	Proactive autoscaling, improved responsiveness, cost efficiency	Not explicitly addressed (resource efficiency only)	Practical cloud-native integration of ML with HPA; proactive scaling reduces delays and resource waste	Trained on synthetic traffic data; limited validation on complex multi-service workloads
[28]	Cloud computing environment	DL (LSTM for demand prediction) + RL (DQN for scheduling)	Hybrid AI (DL + RL)	Resource utilization maximization, response time reduction, cost minimization, SLA compliance	Energy considered as part of operational cost; no explicit green or carbon-aware optimization	Strong practical validation; significant gains in utilization, latency, and cost; scalable real-world deployment	High model complexity; substantial training and monitoring overhead; sustainability not explicitly addressed
[29]	Large-scale cloud–edge environment (Content Delivery Networks)	Statistical & ML models (S-ARIMA, LSTM, Bi-LSTM, OS-ELM)	Hybrid statistical + ML + online learning framework	Accurate workload prediction, proactive autoscaling, low rejection rates, SLA preservation	Resource efficiency addressed indirectly; no explicit energy or carbon-aware optimization	Strong empirical validation on real production data; adaptive model switching balances accuracy and overhead; directly supports proactive resource allocation	Focuses on prediction rather than allocation policy design; energy and sustainability not explicitly modeled
[30]	Cloud computing environment for smart manufacturing	Swarm Intelligence (Multi-Objective Particle Swarm Optimization - MOPSO)	Hybrid multi-objective swarm optimization with enhanced operators (mutation/crossover)	VM placement optimization, resource utilization efficiency, inter-VM communication cost reduction	Not addressed	Joint optimization of computation and communication during VM placement; addresses a practical cloud placement problem	Limited benchmarking scope and absence of statistical significance analysis
[31]	IaaS cloud computing environment with multiple data centers and VMs	Multi-Agent System (agent-based intelligent resource management)	Hybrid approach combining dynamic resource allocation + agent-based scheduling + power minimization heuristic	Minimize VM cost, minimize makespan, reduce response time, reduce task rejection rate, reduce energy consumption	Energy-aware (explicit power minimization and VM sleep/active modes); sustainability addressed indirectly (no carbon modeling)	Clear agent-based architecture; joint optimization of cost, makespan, and energy; significant improvements over FCFS and PBTS in simulation	Simulation-only validation; limited scale (few hosts); no real-world cloud deployment; no learning-based adaptation (rules/heuristics dominate)
[32]	Edge–Cloud computing	Multi-Objective RL (MORL) with Proximal Policy Optimization (PPO)	None	Minimize response time, maximize task success rate, improve energy efficiency	Energy-aware multi-objective optimization	Up to 25% higher reward, 20% lower response time, 18% higher success rate, 30–35% energy savings vs baselines	Reward tuning sensitivity and significant training cost; limited large-scale production validation
[33]	Cloud computing environment (IaaS)	QT-DNN	DNN-based prediction + BBSO	Minimize energy consumption, makespan, and response time; improve QoS	Energy-aware optimization; no carbon-aware modeling	98.1% prediction accuracy, response time 1.598, makespan 12, ~20% energy savings vs baselines	simulation-only

[34]	Edge computing networks	Multi-Objective RL (MORL) with PPO, modeled as an MOMDP	Entropy-weighted static-dynamic resource modeling + MOMDP	Minimize task transmission latency and energy consumption simultaneously	Explicit energy consumption modeling (offloading + execution energy)	Strong Pareto optimization; real testbed validation)	Centralized design; scalability limits; no carbon modeling
[35]	Cloud data center (VM-based workload scheduling)	DRL (DQN-based MDP scheduling)	None	Energy consumption minimization, latency reduction, throughput improvement, QoS maximization, load balancing	Explicitly energy-aware; focuses on energy efficiency but no carbon-aware or thermal-aware modeling	Strong multi-metric performance gains (energy reduction, 92% load balancing, 95% utilization, 97% QoS); scalable to large task sets	Simulation-only validation; high training overhead; no real production cloud deployment
[36]	Serverless cloud environment	RL (tabular Q-learning)	None	Optimize concurrency level to maximize throughput and reduce latency	Not explicitly addressed	Demonstrates adaptive learning of optimal concurrency per workload; outperforms default Knative autoscaling; real cluster experimentation	Synthetic workloads; simple reward function; tabular RL limits scalability; no energy or sustainability modeling
[37]	Multi-data-centre cloud environment (AI healthcare workloads)	DL (LSTM workload prediction)	Hybrid predictive ML + adaptive scheduling + multi-data-centre orchestration	Minimize SLA violations, reduce task completion time, optimize resource utilization, reduce operational cost	Not explicitly addressed	Strong formal problem formulation; integrates workload forecasting with proactive scheduling; significant SLA reduction (<5%) and utilization improvement (85%)	Simulation-only validation; domain-specific (healthcare); no RL; no sustainability or energy modeling
[38]	Cloud environment supporting real-time financial analytics (containerized workloads)	LSTM workload forecasting + RL-based dynamic allocation	Hybrid predictive ML + RL adaptive scaling	Reduce processing latency, improve resource utilization, maintain SLA compliance, optimize cost	Not explicitly addressed	Strong hybrid predictive-RL integration; significant latency reduction (31%) and utilization gains (26%); SLA-aware design	Simulation-based evaluation; domain-specific (financial analytics); no energy or carbon modeling
[39]	Large-scale cloud environment (5,000 nodes, Azure & Google traces)	BiLSTM + Multi-Agent RL (MARL) + Ape-X	Hybrid predictive (BiLSTM) + decentralized MARL + distributed prioritized replay	SLA compliance, latency minimization, utilization optimization, carbon emission reduction, cost minimization	Explicit carbon-aware VM placement; 22% energy reduction; carbon masking in reward	End-to-end unified architecture; strong scalability; multi-objective reward; economic validation; distributed training efficiency	Assumes relatively homogeneous workloads; retraining required for structural workload shifts; complex architecture
[40]	Kubernetes-based cloud environment	LSTM workload prediction + DQN reinforcement learning	Hybrid predictive ML + RL-based orchestration	Improve resource utilization (31%), reduce latency (26%), minimize SLA violations	Energy efficiency discussed but no explicit carbon-aware modeling	Hybrid predictive + adaptive framework; integrates forecasting and RL; shows SLA improvement vs reactive scaling	Simulation on limited-scale VM setup; single-node validation
[41]	Distributed multi-cloud environments	ML + RL-based optimization framework	Forecasting (LSTM/ARIMA) + RL rightsizing + anomaly detection + hybrid AI schedulers	Cost forecasting, rightsizing, workload distribution, pricing optimization	Mentions carbon-aware scheduling as future trend; not experimentally evaluated	Comprehensive AI cost governance framework	Conceptual framework; no detailed algorithm formulation
[42]	VM optimization in virtualized cloud (Proxmox-based environment)	Random Forest, LSTM	None	CPU usage prediction, dynamic resource allocation, performance improvement	Not explicitly addressed	High prediction accuracy (MAPE 2.65% with RF), real-time monitoring integration	Evaluated in simulated environment; limited workload diversity; no multi-cloud validation
[43]	VM resource forecasting in AWS cloud	ARIMA, Linear Regression, LSTM	Statistical + ML hybrid	CPU, memory, disk usage prediction; cost-efficient provisioning	Not explicitly addressed	Hybrid model improves forecasting accuracy over individual models	Focused on forecasting; allocation impact not experimentally validated; limited real-world deployment

Based on this synthesis, the literature is discussed according to dominant research paradigms and emerging thematic trends, including RL-based dynamic resource allocation, predictive ML and DL approaches, hybrid AI and metaheuristic integration, network- and multi-data-center-aware allocation, and sustainability- and carbon-aware resource allocation.

A. Reinforcement Learning for Dynamic Resource Allocation

RL has emerged as one of the most prominent paradigms for adaptive cloud resource control. Unlike predictive-only approaches, RL-based frameworks directly optimize allocation decisions through sequential interaction with the environment.

Mohammadi Ghaleh [24] proposes a two-level distributed PPO scheduler for mobile edge computing, combining hierarchical policy learning with multi-objective optimization.

While the framework improves latency and load balancing, it introduces non-trivial communication overhead.

Similarly, Moazzami et al. [25] integrate fuzzy logic with DRL to enhance scheduling adaptability in fog–cloud systems, explicitly minimizing energy consumption alongside delay. Yu et al. [26] further extend RL into multi-objective optimization, jointly minimizing cost, makespan, and energy consumption.

In serverless environments, Schuler et al. [36] demonstrate that even tabular Q-learning can outperform reactive autoscaling policies by learning optimal concurrency levels.

More recently, advanced multi-agent RL (MARL) systems have been introduced. Manhary et al. [39] propose a scalable MARL-ApeX framework combining BiLSTM forecasting with decentralized RL, explicitly incorporating carbon-aware VM placement and achieving measurable energy reductions. EdgeSched-DQN [32] and Liu et al. [34] extend RL into edge-cloud settings, optimizing latency-energy trade-offs using multi-objective formulations.

These works confirm that RL is particularly effective for dynamic, feedback-driven environments. However, most frameworks remain simulation-based and introduce high training overhead, raising concerns regarding deployment feasibility and reproducibility.

B. Predictive Machine Learning and Deep Learning Approaches

Predictive models play a central role in proactive resource allocation. Instead of reacting to overload conditions, ML-based systems forecast workload trends and adjust resource provisioning in advance.

Abdelhamid et al. [27] integrate XGBoost-based traffic prediction with Kubernetes autoscaling, demonstrating improved responsiveness in cloud-native environments. Wang and Yang [28] combine LSTM demand prediction with DQN-based scheduling to achieve significant utilization and latency gains in real cloud deployments.

Duc et al. [29] propose a hybrid forecasting ensemble (LSTM, Bi-LSTM, ARIMA, OS-ELM) for proactive autoscaling in cloud-edge CDNs, showing strong empirical validation on production data.

In our previous work, Doukha and Ez-zahout [42] proposed a real-time monitoring and predictive modeling framework integrating Random Forest and LSTM models for virtual machine optimization. The study demonstrated that ensemble-based learning, particularly Random Forest, can achieve high prediction accuracy for CPU utilization forecasting. However, the evaluation was conducted in a controlled environment and did not address large-scale heterogeneous or multi-cloud deployment scenarios.

Singh [37] and Wei [38] further illustrate hybrid predictive-control systems for healthcare and financial workloads, respectively, where SLA adherence and cost efficiency are primary objectives.

While predictive models enhance foresight and reduce SLA violations, most studies focus on performance and cost objectives, with limited integration of sustainability-aware metrics.

C. Hybrid AI and Metaheuristic Integration

Hybridization has emerged as a dominant design paradigm in cloud resource allocation research. Rather than relying exclusively on predictive learning or standalone reinforcement learning, many recent frameworks integrate learning components with classical optimization techniques or heuristic search strategies. This integration aims to combine the adaptability of AI models with the stability and convergence guarantees of structured optimization methods.

One hybridization pattern involves combining deep learning–based workload prediction with metaheuristic schedulers. Sharma et al. [33] integrate QT-DNN-based task–VM prediction with Binary Bird Swarm Optimization (BBSO) to enhance energy-efficient scheduling. In this design, the neural predictor estimates workload-resource mappings, while the swarm-based optimizer refines allocation decisions to minimize energy consumption and makespan. This separation of prediction and optimization helps reduce convergence instability often observed in pure RL approaches.

Similarly, Xaba et al. [44] employ Particle Swarm Optimization (PSO) for network-aware VM placement across multiple data center topologies. Although the approach does not incorporate adaptive learning, it represents a hybrid form of automation where heuristic optimization is guided by network stress metrics. The study highlights the importance of topology-aware placement but lacks learning-driven adaptability and sustainability quantification.

Another hybrid direction integrates predictive ML with reinforcement learning control. Chileshe [40] combines LSTM workload forecasting with DQN-based orchestration in Kubernetes environments. The predictive layer anticipates resource demand, while the RL agent dynamically adjusts scaling decisions to minimize SLA violations and latency. This predictive-control hybrid structure reflects a growing trend toward proactive rather than purely reactive allocation.

Rabaaoui et al. [31] propose an agent-based hybrid architecture combining dynamic allocation mechanisms with heuristic energy minimization strategies. Their approach emphasizes distributed coordination among agents,

demonstrating that hybrid agent–heuristic systems can improve cost, makespan, and energy metrics in simulated cloud environments.

In another previous study, Doukha et al. [43] introduced a hybrid forecasting framework combining ARIMA, linear regression, and LSTM models for virtual machine resource prediction. The integration of statistical and deep learning techniques aimed to capture both linear temporal structures and nonlinear workload dynamics, improving forecasting accuracy compared to standalone models. Nevertheless, the study primarily evaluated predictive performance and did not experimentally validate large-scale dynamic allocation mechanisms.

Collectively, these works illustrate three main hybridization patterns:

- DL prediction with metaheuristic optimization
- ML forecasting with RL control
- Agent-based coordination with heuristic energy minimization
- Statistical time-series models integrated with machine learning and deep learning for enhanced workload prediction

Hybrid approaches frequently report improved multi-objective performance compared to single-paradigm systems, particularly in balancing latency, utilization, and energy consumption. However, most of them remain validated in simulation environments, with limited deployment-scale benchmarking or real cloud trace validation. Moreover, hybrid systems often introduce architectural complexity and parameter tuning challenges, which may hinder practical adoption.

D. Network- and Multi-Data-Center-Aware Allocation

While early cloud allocation research primarily focused on compute utilization and VM efficiency, recent studies increasingly consider communication overhead, topology constraints, and geographic distribution as key optimization factors. In large-scale cloud and edge infrastructures, network congestion, inter-rack latency, and cross-data center traffic significantly affect application performance and operational cost.

Xaba et al. [44] explicitly incorporate network stress metrics into VM placement decisions using a PSO-based NetDEO algorithm. Their evaluation across FatTree, BCube, and Tree topologies demonstrates that topology-aware allocation can reduce traffic imbalance and congestion compared to baseline placement strategies. However, the approach remains heuristic-driven and does not incorporate adaptive learning or carbon-aware routing.

Beyond topology-aware placement within a single data center, multi-data-center orchestration introduces additional complexity. Singh [37] proposes predictive orchestration for healthcare workloads across geographically distributed cloud environments, where SLA adherence and resource utilization must be maintained under cross-site migration scenarios.

The framework integrates workload forecasting with adaptive scheduling to mitigate SLA violations during migration and scaling events.

Wei [38] extends predictive allocation to real-time financial analytics, targeting containerized deployments across distributed infrastructures. By combining LSTM forecasting with RL-based dynamic allocation, the framework seeks to optimize latency and resource efficiency in geographically distributed environments.

Collectively, these studies reflect a broader shift from centralized, compute-centric allocation toward network-aware and geographically distributed orchestration. However, several limitations persist. First, most works optimize network or SLA metrics independently rather than integrating them into unified multi-objective carbon-aware formulations. Second, cross-region energy pricing and carbon-intensity variations are rarely incorporated into allocation decisions. Third, empirical validation across real multi-region cloud deployments remains limited.

As cloud-native applications increasingly span hybrid and multi-cloud infrastructures, network-aware, as well as geography-aware scheduling, is likely to become a core requirement for future AI-driven resource management systems.

E. Sustainability and Carbon-Aware Resource Allocation

Sustainability has become an increasingly important objective in cloud resource allocation. As data center energy demand continues to grow [3], recent studies have begun incorporating energy-related metrics into scheduling and allocation policies.

Several works integrate energy consumption directly into optimization objectives. Moazzami et al. [23] and Yu et al. [27], for example, include energy terms within their RL reward functions, treating power consumption as part of a multi-objective trade-off alongside latency and cost. However, in most cases, energy is considered primarily as an operational efficiency metric rather than an explicit environmental indicator.

A more explicit sustainability perspective appears in Manhary et al. [47], which models carbon-aware VM placement using emission-based reward shaping within a scalable MARL framework. This approach moves beyond energy minimization toward emission-sensitive allocation. Harun also explores carbon-aware scheduling combined with constrained reinforcement learning and digital twin validation, representing an early attempt to integrate environmental considerations into hybrid AI frameworks.

Despite these advances, explicit carbon-intensity modeling remains limited across the broader literature. Many allocation systems continue to treat energy consumption as a proxy for cost rather than directly optimizing environmental impact, indicating that fully carbon-aware cloud resource management remains an open research direction.

F. Synthesis and Observations

Across the surveyed literature, several patterns emerge:

- RL dominates dynamic allocation research, but suffers from training complexity.
- Metaheuristics remain relevant, particularly for placement optimization.
- Sustainability is increasingly considered, but often indirectly.
- Most frameworks rely on simulation environments rather than production validation.
- Carbon-aware and thermal-aware scheduling remain underexplored.
- Explainability and interpretability are rarely addressed.

A key observation across the surveyed literature is the presence of trade-offs between competing optimization objectives. For instance, minimizing latency often requires increased resource provisioning, which leads to higher operational cost and energy consumption. Conversely, strategies focused on energy efficiency may negatively impact performance and SLA compliance. These trade-offs highlight the importance of multi-objective optimization frameworks capable of balancing latency, cost, energy efficiency, and reliability within unified allocation strategies.

In addition, the effectiveness of resource allocation approaches varies depending on workload characteristics. Batch workloads, which are delay-tolerant, can benefit from global optimization and energy-aware scheduling strategies. In contrast, interactive and latency-sensitive workloads require fast, real-time decision-making, favoring lightweight heuristics or low-latency inference models. This distinction is critical when evaluating the practical applicability of AI-based allocation methods in real-world cloud environments.

Table I highlight that only a minority of works explicitly integrate environmental metrics beyond energy consumption, and even fewer validate at a large scale using real cloud traces.

VI. Future Directions

The analysis of existing AI-based cloud resource allocation research reveals clear progress toward adaptive, learning-driven orchestration. However, several structural limitations remain. Most surveyed approaches prioritize performance, latency, and cost minimization, while sustainability is often treated as a secondary objective. Although energy-aware reward functions are increasingly incorporated into RL frameworks [25], [29], and more explicitly modeled in carbon-aware formulations such as [39], explicit carbon-intensity-aware scheduling remains rare. Future research should, therefore, move beyond energy minimization as a cost proxy and integrate real-time carbon intensity signals as an additional objective or constraint within resource allocation decisions. Emerging carbon-aware computing initiatives demonstrate the feasibility of such approaches. For example, Radovanović et al. [45] show how carbon-aware workload shifting can reduce emissions without compromising reliability.

Another recurring limitation identified in the literature is the heavy reliance on simulation-based validation. Many RL and hybrid systems demonstrate promising results in controlled environments but lack large-scale real-cloud benchmarking. Bridging this gap requires reproducible evaluation pipelines, standardized workload traces, and safe deployment mechanisms within production cloud clusters. Strengthening empirical validation will be essential to ensure robustness, scalability, and operational reliability of AI-driven allocation systems.

Scalability also remains a challenge, particularly for multi-agent RL systems in geographically distributed cloud infrastructures. While decentralized and distributed training paradigms [46] have demonstrated scalability in other domains, their adaptation to carbon-aware, SLA-constrained cloud environments remains underexplored. Future work should investigate hierarchical and communication-efficient MARL designs tailored specifically for multi-data-center orchestration.

From a predictive standpoint, most surveyed studies rely on LSTM- or GRU-based workload forecasting. Recent advances in transformer-based time-series modeling suggest improved long-term dependency capture and generalization capacity [47]. Integrating transformer-based forecasting into hybrid predictive–control allocation frameworks may enhance robustness under bursty or highly non-stationary workloads.

Finally, explainability and operational transparency remain largely absent from current AI-driven allocation systems. As cloud infrastructures increasingly support critical applications, allocation policies must be interpretable and auditable. Foundational discussions on explainable AI in high-stakes systems [48] highlight the risks of opaque decision-making models. Incorporating interpretable reward structures, SLA-risk visualization, and policy explainability into resource orchestration engines represents an important direction for trustworthy AI-driven cloud management.

In summary, future AI-based cloud resource allocation research should prioritize: 1) explicit carbon-intensity–aware optimization, 2) real-world validation beyond simulation, 3) scalable and stable multi-agent learning, 4) advanced workload forecasting architectures, and 5) explainable and trustworthy allocation mechanisms. Addressing these dimensions will move the field from performance-centric optimization toward sustainable, production-ready, and environmentally responsible cloud orchestration.

VII. Conclusion

Cloud resource allocation has evolved from static heuristic scheduling toward adaptive, data-driven optimization frameworks powered by artificial intelligence. This survey analyzed recent AI-based approaches published between 2020 and 2026, highlighting the growing dominance of reinforcement learning for dynamic control, the continued importance of predictive modeling for proactive provisioning, and the increasing adoption of hybrid architectures that combine forecasting, learning-based control, and classical optimization techniques.

While significant performance gains have been reported across latency, cost, and resource utilization metrics, most frameworks remain validated primarily in simulation environments. Sustainability considerations are increasingly integrated through energy-aware objectives, yet explicit carbon-intensity modeling and large-scale real-world validation remain limited. Future research should prioritize scalable deployment, carbon-aware optimization, and interpretable decision mechanisms to advance trustworthy and production-ready intelligent cloud resource management systems.

REFERENCES

[1] S. Alzide, "Cloud Computing: Evolution, Challenges, and Future Prospects," Journal of Information Technology, Cybersecurity, and Artificial Intelligence, vol. 1, pp. 52–63, Dec. 2024, doi: 10.70715/jitcai.2024.v1.i1.007.

[2] T. R. Merlo, F. Fard, and S. Hawamdeh, "Cloud Computing's Impact on the Digital Transformation of the Enterprise: A Mixed-Methods Approach," Sustainability, vol. 17, no. 13, p. 5755, Jan. 2025, doi: 10.3390/su17135755.

[3] N. Kumawat, N. Handa, and A. Kharbanda, Cloud Computing Resources Utilization and Cost Optimization for Processing Cloud Assets. 2020, p. 48. doi: 10.1109/SmartCloud49737.2020.00017.

[4] "Energy demand from AI – Energy and AI – Analysis," IEA. Accessed: Nov. 12, 2025. [Online]. Available: https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

[5] R. Aron and A. Abraham, "Resource scheduling methods for cloud computing environment: The role of meta-heuristics and artificial intelligence," Engineering Applications of Artificial Intelligence, vol. 116, p. 105345, Nov. 2022, doi: 10.1016/j.engappai.2022.105345.

[6] H. Madni, A. L. Shafie, M. Abdullahi, S. Abdulhamid, and M. Usman, "Performance comparison of heuristic algorithms for task scheduling in IaaS cloud computing environment," PLOS ONE, vol. 12, p. e0176321, May 2017, doi: 10.1371/journal.pone.0176321.

[7] A. Rathee and S. Dalal, "A systematic literature review of machine learning-based resource allocation techniques in cloud computing," Computing, vol. 107, no. 9, p. 179, Aug. 2025, doi: 10.1007/s00607-025-01526-8.

[8] R. N. K. Bhanu and K. K. J, "Deep Learning and Optimization Algorithms for Intelligent Cloud Resource Allocation: A Survey," IEEE Conference Proceedings, vol. 2025, no. ICTMIM, pp. 1487–1494, 2025.

[9] H. Pitkar and S. Ambapkar, "AI ML and cloud computing: exploring models, challenges and opportunities," World Journal of Advanced Research and Reviews, vol. 25, pp. 770–783, Feb. 2025, doi: 10.30574/wjarr.2025.25.2.0430.

[10] K. Srikandabala et al., "Optimization and Management of Data Center Networks: A Scoping Review on Key Themes, Challenges, and Artificial Intelligence and Machine Learning Approaches," IEEE Access, pp. 1–1, 2025, doi: 10.1109/ACCESS.2025.3593523.

[11] S. I. Mohammad, "Enhancing Cloud Resource Allocation with a Hybrid Deep Learning-Based Framework: A Comparative Study," Journal of Information Systems Engineering and Management, vol. 10, no. 41s, pp. 1019–1034, May 2025, doi: 10.52783/jisem.v10i41s.8106.

[12] V. N R, S. D P, T. S. Agarwal, V. H M, and S. D, "Optimizing Cloud Resource Allocation with AI and Machine Learning," 2025 International Conference on Computing Technologies (ICOCT), pp. 1–5, Jun. 2025, doi: 10.1109/ICOCT64433.2025.11118766.

[13] H. F. Alhashimi et al., "Survey on AI-Enabled Resource Management for 6G Heterogeneous Networks: Recent Research, Challenges, and Future Trends," CMC, vol. 83, no. 3, pp. 3585–3622, 2025, doi: 10.32604/cmc.2025.062867.

[14] P. Nawrocki and M. Smendowski, "A Survey of Cloud Resource Consumption Optimization Methods," J Grid Computing, vol. 23, no. 1, p. 5, Jan. 2025, doi: 10.1007/s10723-024-09792-0.

[15] N. Soltani, B. Soleimani Neysiani, and B. Barekatain, "Heuristic Algorithms for Task Scheduling in Cloud Computing: A Survey," International Journal of Computer Network and Information Security, vol. 9, pp. 16–22, Aug. 2017, doi: 10.5815/ijcnis.2017.08.03.

[16] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," in 2009 International Conference on High Performance Computing & Simulation, Jun. 2009, pp. 1–11. doi: 10.1109/HPCSIM.2009.5192685.

[17] S. A. A. M. Zarif et al., "GLOPS: A Hybrid Approach for Enhanced Scheduling in Cloud Computing Environments via Machine Learning-Based Process Prediction," IEEE Access, vol. 13, pp. 134385–134402, 2025, doi: 10.1109/ACCESS.2025.3592211.

[18] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers," Concurrency and Computation: Practice and Experience, vol. 24, no. 13, pp. 1397–1420, 2012, doi: 10.1002/cpe.1867.

[19] S. H. H. Madni, M. Faheem, M. Younas, M. H. Masum, and S. Shah, "Critical review on resource scheduling in IaaS clouds: Taxonomy, issues, challenges, and future directions," The Journal of Engineering, vol. 2024, no. 8, p. e12420, 2024, doi: 10.1049/tje2.12420.

[20] P. Singh, M. Dutta, and N. Aggarwal, "A review of task scheduling based on meta-heuristics approach in cloud computing," Knowl Inf Syst, vol. 52, no. 1, pp. 1–51, Jul. 2017, doi: 10.1007/s10115-017-1044-2.

[21] Y. Lang, Y. Gao, T. Chen, and H. Wang, "Centered Collision Optimizer: A novel and efficient physics-based metaheuristic optimization algorithm for solving complex real-world engineering optimization problems," Computer Methods in Applied Mechanics and Engineering, vol. 448, p. 118491, Jan. 2026, doi: 10.1016/j.cma.2025.118491.

[22] Y. Xu et al., "Artificial intelligence: A powerful paradigm for scientific research," The Innovation, vol. 2, no. 4, p. 100179, Nov. 2021, doi: 10.1016/j.xinn.2021.100179.

[23] B. Barua and M. S. Kaiser, "AI-Driven Resource Allocation Framework for Microservices in Hybrid Cloud Platforms," Dec. 03, 2024, arXiv: arXiv:2412.02610. doi: 10.48550/arXiv.2412.02610.

[24] A. Mohammadi Ghaleh, A 2-Level Distributed PPO Scheduling Approach For Real-time Heterogeneous Mobile Edge Computing. 2025.

[25] S. Moazzami, A. Mirzaei, M. Aminian, R. Karimi, and N. Mikaeilvand, "A Hybrid Fuzzy Logic and Deep Reinforcement Learning Algorithm for Adaptive Task Scheduling and Resource Allocation in Heterogeneous Fog–Cloud Environments," Sustainable Computing: Informatics and Systems, p. 101260, Nov. 2025, doi: 10.1016/j.suscom.2025.101260.

[26] X. Yu, J. Mi, L. Tang, L. Long, and X. Qin, "Dynamic multi objective task scheduling in cloud computing using reinforcement learning for energy and cost optimization," Sci Rep, Nov. 2025, doi: 10.1038/s41598-025-29280-z.

[27] Y. Abdelhamid, W. Anis Aziz, and J. Soliman, Kubernetes Autoscaling with Machine Learning based on traffic load prediction. 2025.

[28] Y. Wang and X. Yang, "Intelligent Resource Allocation Optimization for Cloud Computing via Machine Learning," acss, vol. 9, no. 1, 2025, doi: 10.23977/acss.2025.090109.

[29] T. L. Duc, C. Nguyen, and P.-O. Östberg, "Workload Prediction for Proactive Resource Allocation in Large-Scale Cloud-Edge Applications," Electronics, vol. 14, no. 16, p. 3333, Jan. 2025, doi: 10.3390/electronics14163333.

[30] "AI-Driven Resource and Communication-Aware Virtual Machine Placement Using Multi-Objective Swarm Optimization for Enhanced Efficiency in Cloud-Based Smart Manufacturing," Computers, Materials and Continua, vol. 81, no. 3, pp. 4743–4756, Dec. 2024, doi: 10.32604/cmc.2024.058266.

[31] S. Rabaaoui, H. Hachicha, and E. Zagrouba, "An efficient and autonomous dynamic resource allocation in cloud computing with optimized task scheduling," Procedia Computer Science, vol. 246, pp. 3654–3663, Jan. 2024, doi: 10.1016/j.procs.2024.09.191.

[32] N. Yamsani and P. Chenna Reddy, "EdgeSched-DQN: An intelligent deep reinforcement learning-based framework for optimized task scheduling in edge-cloud environments," Array, vol. 29, p. 100645, Mar. 2026, doi: 10.1016/j.array.2025.100645.

[33] P. Sharma, D. Yadav, B. Sharma, S. Khan, and A. Almusharraf, "Energy Efficient and Resource Allocation in Cloud Computing Using QT-DNN and Binary Bird Swarm Optimization," CMC, vol. 85, no. 1, pp. 2179–2193, 2025, doi: 10.32604/cmc.2025.063190.

[34] W. Liu et al., "Resource Scheduling Algorithm for Edge Computing Networks Based on Multi-Objective Optimization," Applied Sciences, vol. 15, no. 19, p. 10837, Jan. 2025, doi: 10.3390/app151910837.

[35] A. R. Malipatil et al., "Energy-Efficient Cloud Computing Through Reinforcement Learning-Based Workload Scheduling," International Journal of Advanced Computer Science and Applications, vol. 16, no. 4, 2025.

[36] L. Schuler, S. Jamil, and N. Kuhl, "AI-based Resource Allocation: Reinforcement Learning for Adaptive Auto-scaling in Serverless Environments," 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pp. 804–811, May 2021, doi: 10.1109/CCGrid51090.2021.00098.

[37] S. Singh, "Predictive Resource Orchestration for AI-Driven Healthcare Workloads in Multi-Data-Centre Cloud Migrations," IJAIBDCMS, vol. 7, pp. 54–61, 2026, doi: 10.63282/3050-9416.IJAIBDCMS-V7I1P109.

[38] D. L. Wei, "Predictive AI-Based Resource Management for Real-Time Financial Analytics in Cloud Environments".

[39] F. N. Manhary, M. H. Mohamed, and M. Farouk, "A scalable machine learning strategy for resource allocation in database," Sci Rep, vol. 15, no. 1, p. 30567, Aug. 2025, doi: 10.1038/s41598-025-14962-5.

[40] L. Chileshe, "Intelligent Resource Orchestration Using AI-Driven Predictive Algorithms for Scalable Cloud Systems," AIJCST, vol. 5, 2023, doi: 10.63282/3117-5481/AIJCST-V5I6P102.

[41] K. Rathore, K. Bansal, and D. N. Saraswat, "Smart Cloud Economics: AI-Driven Optimization In Distributed Cloud Environments," vol. 9, no. 6, 2025.

[42] R. Doukha and A. Ez-zahout, "Enhanced Virtual Machine Resource Optimization in Cloud Computing Using Real-Time Monitoring and Predictive Modeling.," International Journal of Advanced Computer Science & Applications, vol. 16, no. 2, p. 658, Feb. 2025, doi: 10.14569/ijacsa.2025.0160267.

[43] R. Doukha, A. Ez-Zahout, and A. Ndayikengurukiye, "Forecasting virtual machine resource utilization in cloud computing: a hybrid artificial intelligence approach," vol. 37, no. 3.

[44] X. C. Xaba, K. Ogudo, and S. Chabalala, "Optimizing Cloud Computing and Network Resource Allocation Through Cloud Automation," in Proceedings of the 2025 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, Cape Town South Africa: ACM, Nov. 2025, pp. 1–6. doi: 10.1145/3759023.3759102.

[45] A. Radovanović et al., "Carbon-Aware Computing for Datacenters," IEEE Transactions on Power Systems, vol. 38, no. 2, pp. 1270–1280, Mar. 2023, doi: 10.1109/TPWRS.2022.3173250.

[46] L. Espeholt et al., "IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures," in Proceedings of the 35th International Conference on Machine Learning, PMLR, Jul. 2018, pp. 1407–1416. Accessed: Feb. 17, 2026. [Online]. Available: https://proceedings.mlr.press/v80/espeholt18a.html

[47] H. Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," Mar. 28, 2021, arXiv: arXiv:2012.07436. doi: 10.48550/arXiv.2012.07436.

[48] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat Mach Intell, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.