future internet

MDPI

Review

Research on Intelligent Resource Management Solutions for Green Cloud Computing

Amirmohammad Parhizkar [1], Ehsan Arianyan [2,*] and Pejman Goudarzi [2,*]

[1] Department of Industrial Engineering, Tarbiat Modarres University, Tehran 14155-3961, Iran; a_parhizkar@modares.ac.ir
[2] ICT Research Institute, Tehran 14155-3961, Iran
* Correspondence: ehsan_arianyan@itrc.ac.ir (E.A.); pgoudarzi@itrc.ac.ir (P.G.)

Abstract

Cloud computing utilization has experienced progressive expansion over the last decade, which has raised concerns and challenges regarding efficient resource allocation and energy efficiency. The burgeoning increase in the number of cloud computing users and their data exacerbates the difficulty of resolving these challenges using conventional methods. Thus, utilizing intelligent approaches is indispensable. Among the most recent intelligent methods, artificial intelligence-based techniques have gained prominence across numerous research domains, including cloud resource management. Through a literature review aimed at analyzing existing studies addressing the open challenges of cloud computing, we have identified some gaps that are presented in this paper. Moreover, this paper presents a survey on cloud resource management solutions spanning from 2018 to 2025, with a focus on the papers that utilized intelligent methodologies for green computing. More specifically, this study shed light on the prevailing challenges in the field concerning methods, research areas, metrics, tools, and datasets. Furthermore, it provides a clear classification of methods, research areas, and metrics.

Keywords: resource management; cloud computing; artificial intelligence; green computing

1. Introduction

Today, cloud computing is recognized as a relatively new field that aims to transform the traditional network communication architecture into a unique model. It achieves this by leveraging other new sciences and technologies to reduce operational costs and improve agility [1]. The rise in requests to outsource computing and utilize infrastructure resources (such as CPU, RAM, and bandwidth), alongside advancements in virtualization technology, has significantly propelled the progress of cloud computing [2].

There are several definitions for cloud computing, among which the American National Institute of Standards and Technology (NIST) defines cloud computing as a model for having pervasive, easy, and on-demand network access to a set of configurable computing resources (such as networks, servers, storage space, applications, and services) that can be quickly provided or released with minimal work and effort or the need for the intervention of the service provider [3].

Cloud computing provides high scalability and relatively low cost for using heavy computing facilities, but at the same time, the growth in demand for cloud infrastructure has increased energy consumption in data centers. Currently, one of the most important problems in this field is how to optimally use cloud resources in order to reduce energy

Check for updates

Academic Editor: Paolo Bellavista

Received: 1 November 2025
Revised: 23 January 2026
Accepted: 26 January 2026
Published: 1 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

consumption [1]. From a resource management perspective, high energy consumption, in addition to the issue of elevated operating expenses, presents another significant complication. The substantial energy usage of data centers contributes to global warming due to the emission of carbon dioxide and heat. The energy consumed in large data centers, housing thousands of computing machines, is remarkably substantial. Research indicates that by 2030, approximately 3–13% of the world's electricity will be consumed by data centers [4].

Consolidating virtual machines into fewer physical machines is one of the strategies that service providers use to manage energy consumption by using less hardware and thus reducing energy consumption. Among the various strategies that are adopted to increase energy efficiency, the integration of virtual machines is one of the most effective and proven techniques for optimal use of hardware resources and reducing energy consumption. Basically, the topic of virtual machine integration, which is normally the functionality of the hypervisor, is divided into three main steps: (1) workload identification: identifying physical machines that have less than the standard workload or physical machines that have more than the standard workload; (2) choosing virtual machines: choosing suitable virtual machines for integration/migration; and (3) deployment of virtual machines: finding suitable physical machines to deploy virtual machines [5]. The optimal use of resources is not limited to the layer of physical machines, as resource management in cloud computing for online users is another case that can be optimized; in this space, the allocation of each parameter, such as CPU, storage, memory, and bandwidth, should be managed based on requests from users [1].

In cloud computing, the calculations related to the allocation and scheduling of resources due to the wide solution space and heterogeneous resources are in the category of NP-Hard problems, and to solve these problems, innovative, meta-heuristic, and artificial intelligence (AI)-based methods (machine learning, deep learning, etc.) are used [6]. NP-Hard indicates that no polynomial-time algorithm is known for solving the problem optimally in the general case, so practical solutions rely on heuristics, approximations, or metaheuristics. Nowadays, artificial intelligence has permeated numerous scientific disciplines owing to its high computing power and promising outcomes. Bibliographic analyses reveal a significant increase in the utilization of various AI methods in the realm of cloud resource management.

Based on reviews conducted on literature over the last six years, a limited number of review studies specifically addressing the utilization of AI algorithms in cloud computing were identified. None of these papers comprehensively provided a classification for studies conducted in cloud computing utilizing AI methods and focusing on green computing. Also, their review period does not cover the entire span from 2018 to 2025. In this paper, driven by the goal of reducing energy consumption through the proliferation of intelligent methods and addressing related challenges, the literature is reviewed in order to address the following questions:

(RQ1) What are the focusing areas of research in the field of intelligent green computing? What is their classification?

(RQ2) What is the classification of methods used to optimize energy consumption in the field of intelligent green computing?

(RQ3) What is the classification of metrics used to measure and evaluate the efficiency of solutions in the field of intelligent green computing?

(RQ4) Which tools and datasets are used in the field of intelligent green computing for performance evaluation?

(RQ5) What are the challenges that can be addressed as future research directions in the field of intelligent green computing?

The remainder of the paper is organized as follows: First, a review of related review or survey papers is provided. Subsequently, the research methodology employed in this paper is elucidated. The main body of the review encompasses research areas, methods, metrics, tools, and datasets. Finally, Sections 6 and 7 comprise a discussion and conclusion.

2. Related Work

In this paper, we have reviewed about 113 papers, some of which are review/survey papers from different aspects of green cloud computing [1–113]. Here, we will investigate some related ones.

Khan and colleagues [7] have investigated the methods of cloud resource management based on machine learning, considering the three layers of IaaS, PaaS, and SaaS. The main goal in this review is to identify the weak points of the methods, the datasets used, and also how the methods work. In this category, machine learning methods are divided into four main categories, including supervised, unsupervised, semi-supervised, and reinforcement learning methods. Parameters used as optimization components are described, but a specific taxonomy and classification for these parameters is not provided.

Nayeri and colleagues [8] have focused on fog computing. Their main objectives include classification of machine learning methods, evolutionary methods, and hybrid methods in the field of fog computing. In addition to their main goals, the performance metrics that were considered in the optimization methods were also examined. Moreover, the separation of sections in this paper is based on these performance metrics.

Goodarzy and colleagues [9] have separated and reviewed papers based on machine learning methods used to optimize cloud resource management in three general categories: supervised, unsupervised, and reinforcement learning methods. In this paper, for the review of each paper, the methods are explored, as well as other items, including the dataset used as input, the parameters examined from the cloud environment, the outputs, and also the objective functions.

Among the review papers that have covered the topic of unsupervised methods based on machine learning, we can mention the work of Stanley et al. [10], which has separated data centers in the field of cloud computing into two general categories of dynamic and static energy management. In the first category, the clustering methods, and in the second category, the machine learning methods are examined. All the above methods have been checked with CPU, RAM, and disk resources. The research conducted has a narrative aspect; none of the above three categories has been investigated in depth, and no tree or classification table is provided.

Victoria and colleagues [11] have reviewed the machine learning algorithms used in cloud computing from the four aspects of resource provision, resource scheduling, resource allocation, and energy efficiency. In each of the mentioned aspects, the prediction method, the evaluation index, the performance of each method, and the utilized dataset are considered.

Imran and colleagues [12] have investigated the artificial intelligence methods used in this field, emphasizing the migration methods of virtual machines. The metrics analyzed in this paper include workload balance, energy efficiency, service level agreement violation, and network efficiency. The artificial intelligence methods reviewed in this paper have been analyzed separately for four main types of migration based on the optimized index. In addition, in this paper, artificial intelligence methods are not categorized separately, but the types of artificial intelligence methods are analyzed to increase productivity in the four types of migration.

Khallouli and colleagues [13] have addressed the issue of scheduling and resource allocation in cloud space in a clustered manner. In this paper, a wide range of cluster

scheduling solutions at three levels of IaaS, PaaS, and SaaS, and from two aspects of scheduling structure, scheduling goals, and methods are examined.

In the systematic review of the paper by Djigal et al. [14], many papers are reviewed on the topic of using machine learning and deep learning methods in cloud resource management. This paper has focused on edge computing with multiple accesses. The aforementioned methods are deeply investigated in this paper in three main categories of task offloading, task scheduling, and allocation of connected resources. Also, all machine learning and deep learning methods are categorized in terms of advantages, disadvantages, and use cases in cloud resource allocation.

Masdari and colleagues [15] have categorized the load forecasting methods in cloud computing and explained the challenges of load forecasting. In these categories, the methods based on machine learning are divided into three main categories: regression-based methods, classification algorithm-based methods, and hybrid methods.

Aqib and colleagues [16] have investigated the application of machine learning in cloud computing. In this paper, fog computing and machine learning are fully explained. The issues covered in this paper include different applications of machine learning in fog computing, performance evaluation metrics, the best results obtained in each category of machine learning model in fog computing, and challenges.

Singh and Tanwar [17] have reviewed papers in the field of cloud resource management between 2011 and 2020. The main focus of this research is on identifying new research opportunities in the field of green management of cloud resources or green computing. All the methods used in the literature are considered by the researchers and are not specifically focused on a group of methods. Other items reviewed include performance metrics, limitations and challenges, and tools. Bachiega and colleagues [18] have investigated resource management techniques in edge computing environments. The disciplines investigated in this paper include analysis of resource management steps, metrics, resource management techniques, architectures, virtualization techniques, and case studies in the field of fog computing.

Alsharif et al. have investigated research on techniques and recent advances in energy-efficient fog computing [96]. The authors in [103] have performed a systematic review of green-aware management techniques for sustainable data centers. Zhang et al. in [107] have performed a systematic review on zero-carbon development in cloud data centers using waste heat recovery technology.

In summary, examining review papers in the field of intelligent cloud resource management reveals some gaps in present studies pertaining to the utilization of intelligent methods within this domain, which are described in the next sections. Table 1 shows the comparison of this review study and related review papers. The features that are used to compare review papers are as follows:

- Research Sectioning: Have the authors explained the different sections of research focus that can be found in this field of research?
- Green Computing: Have the authors focused on the matter of energy efficiency when searching papers?
- Tools: Have the authors clearly mentioned the tools that are used in papers?
- Datasets: Have the authors explained the datasets that are used in papers?
- Performance Metrics: Have the authors mentioned the performance metrics that are used in papers?
- Context: What is the context and scope of papers that have been reviewed? (Possible options are cloud computing, fog computing, edge computing, etc.)
- Methods: What are the methods that have been used in the papers that have been covered in the scope of the reviewed papers?
- Period: Which period does the review paper cover?

Table 1. Comparison of this study and related papers.

Row	Paper	Year	Research Sectioning	Green Computing	Tools	Datasets	Performance Metrics	Context	Methods	Period
1	[7]	2021				√	√	Cloud Computing	Machine learning [1]	up to 2020
2	[8]	2021					√	Fog and Edge Computing	Deep learning [2] and meta-heuristic algorithms	2017–2020
3	[9]	2020				√		Cloud Computing	Machine learning	up to 2019
4	[10]	2021		√			√	Cloud Computing	Machine learning, meta-heuristics, and deep learning	up to 2020
5	[11]	2021	√			√	√	Cloud Computing	Machine learning and deep learning	2009–2019
6	[12]	2022					√	Issues of Live Virtual Machine Techniques	-	-
7	[13]	2022				√		Cluster Scheduling Problem (Cloud Computing)	Machine learning	up to 2020
8	[14]	2022	√		√	√		Edge Computing	Machine learning and deep learning	up to 2022
9	[15]	2020			√	√	√	Cloud Computing (Workload Forecasting)	Machine learning and deep learning	up to 2019
10	[16]	2022	√				√	Fog Computing	Machine learning	up to 2021
11	[17]	2022		√	√	√	√	Cloud Computing	Various resource allocation approaches	up to 2020
12	[18]	2023	√				√	Fog Computing	Virtualization methods, various optimization methods, including machine learning	2012–2022
13	[96]	2025	√	√			√	Fog Computing	Various optimization methods	up to 2025
14	[103]	2024	√	√		√		Cloud computing	Various optimization heuristics	up to 2024
15	[107]	2025	√	√			√	Cloud computing	Various optimization methods, machine learning	up to 2025
16	This Study	2025	√	√	√	√	√	Cloud computing	Machine learning, deep learning, reinforcement learning, deep reinforcement learning, and meta-heuristics	2018–2025

[1] In this research, the term machine learning is used to represent more classical AI methods, such as decision trees, random forests, regression, SVM, etc. [2] In this research, the term deep learning is used to represent deep neural networks like CNN, RNN, LSTM, etc.

3. Research Methodology

In the paper selection section, the keywords "Resource management" and "Cloud computing" were chosen as the foundation for the search. Papers were retrieved from the Scopus database, with a publication date restricted to 2018–2023. Subsequently, to narrow down the selection to papers focusing on the utilization of intelligent methods, those containing at least one of the keywords "Machine learning," "Deep learning," "Artificial Intelligence," "Learning algorithms," or "Green computing" were extracted. The inclusion of the keyword "Green computing" stems from the research's emphasis on energy consumption reduction.

However, since initially focusing solely on this keyword during the search stage limited the results and excluded other applications of intelligent methods, it was combined with the keywords from the second series using the "or" condition. In the final step, to leverage the expertise of researchers more actively engaged in this domain, the results were further restricted to papers authored by individuals with a minimum of two publications in this field. Ultimately, 396 papers were extracted. In order to avoid any probable errors encountered during the search for papers in scientific databases, mainly to withdraw papers from non-cloud fields and those not primarily focusing on intelligent methods, the abstracts of 396 papers were scrutinized. Subsequently, a total of 74 papers—comprising 62 research papers and 12 review papers—were selected.

To conduct this study, a literature review was carried out. Scientific articles and industrial reports were extracted and reviewed from reputable databases such as IEEE Xplore, Springer, ScienceDirect, MDPI, and other publishers. The selected articles were chosen based on the following criteria:

- Focus on emerging green and intelligent resource allocation/scheduling in cloud computing.
- Contain practical and empirical analyses related to green and intelligent resource allocation/scheduling in cloud computing.

A PRISMA illustration of the research methodology is shown in Figure 1.

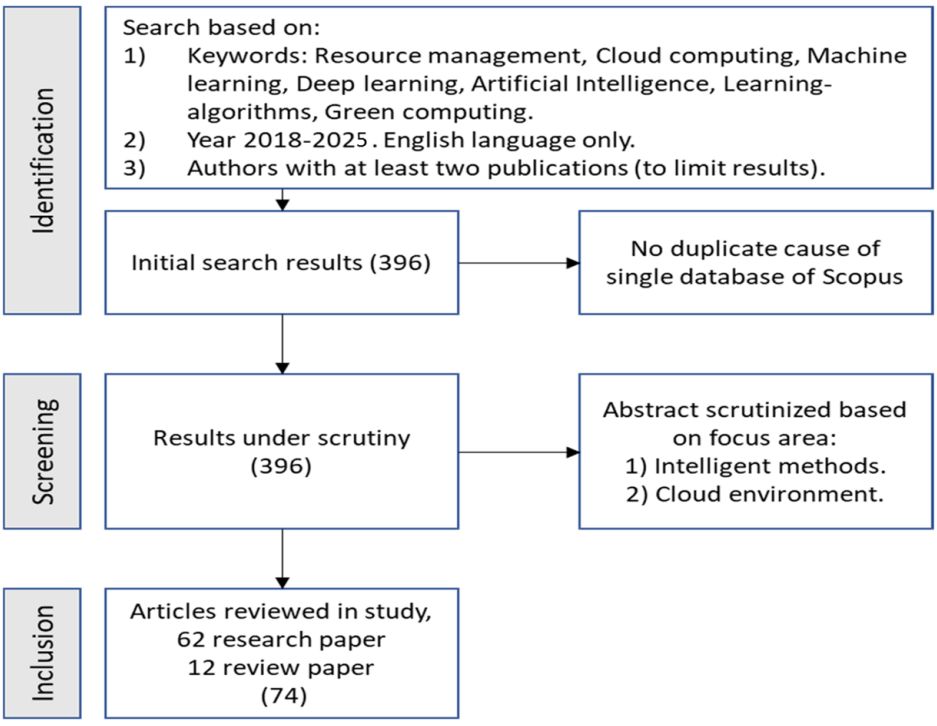


Figure 1. PRISMA chart of research methodology.

Table 2 depicts the conferences and journals from which the selected papers were chosen, with their frequencies and quartiles. Also, Figure 2 provides statistics regarding the number of journals and papers in each quartile. The quartiles of journals are derived from the Scimago website (www.scimagojr.com (accessed on 22 January 2026)), where Q1+ journals have an SJR score exceeding 4, and Q1++ journals have an SJR (SCImago Journal & Country Rank) score surpassing 10. This differentiation is employed to depict the quality of journals more clearly and is not an official classification.

Table 2. Details of reference journals.

Row	Journal	Frequency	Quartile
1	Various Conferences	16	-
2	IEEE Internet of Things Journal	5	Q1+
3	IEEE Access	5	Q1
4	Concurrency and Computation: Practice and Experience	4	Q3
5	Wireless Personal Communications	3	Q2
6	Journal of Network and Computer Applications	2	Q1
7	MDPI energies	2	Q1
8	The Journal of Supercomputing	2	Q2
9	Future Generation Computer Systems	2	Q1
10	Computing (Springer)	2	Q2
11	IEEE Transactions on Vehicular Technology	2	Q1
12	Neurocomputing	2	Q1
13	Journal of Systems and Software	2	Q1
14	IEEE Transactions on Parallel and Distributed Systems	2	Q1
15	Journal of Parallel and Distributed Computing	2	Q1
16	IEEE Transactions on Network and Service Management	2	Q1
17	Computers and Electrical Engineering	1	Q1
18	IEEE Communications Surveys & Tutorials	1	Q1++
19	International Journal of Web Engineering and Technology	1	Q4
20	ACM Computing Surveys	1	Q1+
21	International Journal of Information Technology	1	Q2
22	IEEE Transactions on Services Computing	1	Q1
23	Software: Practice and Experience	1	Q2
24	Computer Networks	1	Q1
25	Computer Communications	1	Q1
26	IEEE Transactions on Industrial Informatics	1	Q1+
27	Computers, Materials and Continua	1	Q2
28	CCF Transactions on High Performance Computing	1	Q3
29	Soft Computing	1	Q2
30	ACM Transactions on Modeling and Performance Evaluation of Computing Systems	1	Q2
31	Technological Forecasting and Social Change	1	Q1

Table 2. *Cont.*

Row	Journal	Frequency	Quartile
32	Applied Nanoscience	1	Q2
33	Sustainable Cities and Society	1	Q1
34	Cluster Computing	1	Q2
35	Simulation Modelling Practice and Theory	1	Q1
36	IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems	1	Q1
37	Energy Reports	1	Q1

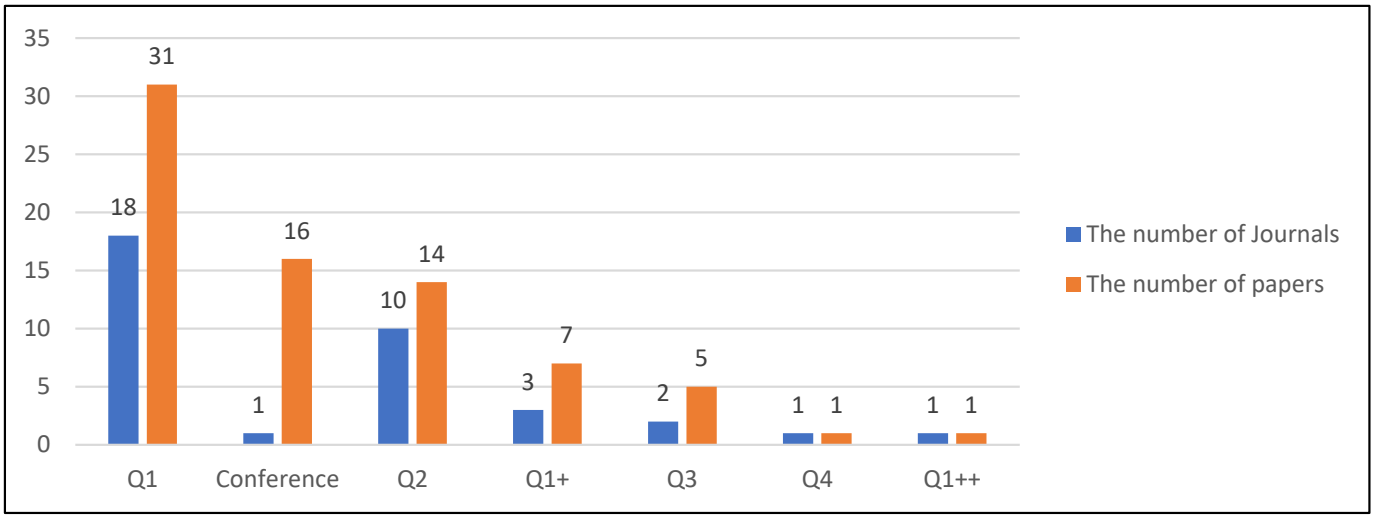


Figure 2. Frequency of papers based on quartile category.

By Q1, Q2, Q3, and Q4, we mean the publications that are in the top 25%, 25–50%, 50–75%, and the lower 25% of high-quality papers published in the field, respectively.

The papers in the journal are more complete and more rigorously reviewed in comparison with conference papers, so we can address more research questions in a journal publication.

4. Intelligent Resource Management for Green Cloud Computing

This section is composed of three subsections. In Section 4.1, the main research areas that we will focus on regarding intelligent resource management for green cloud computing are indicated. Afterwards, in Section 4.2, we will discuss the main methods and metrics used regarding this important topic. Finally, in Section 4.3, we will review the main tools and datasets used in intelligent green cloud computing.

4.1. Research Areas

Our survey conducted on cloud resource management papers shows that various research areas can be defined in this field. Among the main topics of this field, resource scheduling and resource allocation are the primary well-known subjects. The problem of resource allocation seeks to find an optimal allocation of a fixed amount of resources to activities in order to minimize the costs and other optimality factors incurred by the allocation.

The issue of scheduling is the act of allocating resources to do things when the scheduling activity is performed by a process called a scheduler. Schedulers are often designed to keep all computer resources busy, allowing multiple users to share system resources efficiently or achieve a desired quality of service. Therefore, in the literature, when talking

about allocation or scheduling, in general, the goal is to find an optimal allocation of resources to tasks.

In the three environments of cloud computing, edge computing, and fog computing, there are some subtle differences between resource entities and task entities. Despite their architectural differences, cloud, fog, and edge environments all require dynamic resource allocation and scheduling strategies to manage heterogeneous workloads under varying resource constraints.

According to the literature, the difference among these three environments can be explained as follows. The cloud provides centralized elastic computing and long-term analytics, while edge/fog adds geographically distributed micro-data centers close to data sources to reduce round-trip delay, backbone traffic, and potential energy use (depending on workload placement and hardware efficiency) [19,20].

For example, while in data centers, enterprise-grade hardware is considered the main resource, in edge and fog computing environments, these resources can be diminished to user devices, which typically have lower performance compared to hardware in cloud data centers, but the fact of allocation of requests to resources can still be applied. Therefore, the general definition of resource allocation and scheduling can be equally adapted to each environment with some minor conceptual differences.

The elaboration of fog/edge is not the purpose of this article and is out of scope. In the current work, we only focus on centralized cloud data centers.

The definition of this problem is the same in the three environments of cloud computing, edge computing, and fog computing, but there are some subtle differences between resource entities and task entities in these environments that, on the scale of this survey, can be neglected. The papers in this category are introduced as follows.

In [21], the scheduling problem in cloud and edge computing, focusing on the dynamic problem of components in edge computing, was modeled in the form of a Markov decision problem, and optimal results are then achieved via the help of a deep reinforcement learning algorithm and reward mechanism. In [22], the problem of resource allocation in edge computing is studied, focusing on the issue of multi-dimensionality in heterogeneous environments of Internet of Things and edge computing.

Also, the two-stage reinforcement learning algorithm is proposed in order to allocate edge resources to Internet of Things software requests with the aim of increasing the quality of user experience. In [23], a model for automatic long-term decision-making has been designed with the help of learning from the dynamic environment of the network, such as the pattern of incoming requests and energy prices, with the aim of scheduling and allocating cloud resources. In [24], a comprehensive model for managing cloud resources focusing on energy efficiency using artificial intelligence is presented.

This model is designed for private–public hybrid clouds. In this model, the scheduler is designed in a way to reduce thermal hotspots in the data center and thereby achieve energy optimization. In [25], a multi-objective algorithm for scheduling activities in a heterogeneous environment is presented. In this paper, the author has named his research problem a scheduling problem. In [26], a partition-based algorithm for scheduling activities using the proximal policy with the aim of optimizing heterogeneous resources is presented.

In this algorithm, tasks are first placed in partitions based on the current position of heterogeneous partitions and then assigned to appropriate servers in that partition.

In [27], a model in the space of edge computing is presented in order to distribute the computing power required for requests that are in the field of deep artificial intelligence networks. High dynamism and heterogeneity are the characteristics of requests in deep artificial intelligence networks, which are specifically addressed in this paper for the first time.

Another research area in cloud resource management is the issue of load balancing. When the amount of resources requested by the users is greater than the remaining resources in the system, the efficiency of resource allocation decreases, and resource allocation to tasks fails. When the number of requests approaches the processing capacity of the remaining resources in the system, requests are processed slowly. As a result, the tasks are not executed properly, and the load of the cloud data center becomes unbalanced [28]. The papers in this category are introduced as follows.

In [29], the problem of scheduling the allocation and migration of virtual machines in order to optimize cost and energy consumption is the main focus. The parallel allocation of the sequence of devices has been performed with the help of the Q-learning algorithm. The architecture of the solved problem is divided into three main layers: user, scheduler, and infrastructure, and the infrastructure layer itself is divided into three parts: resource management, virtual machine allocation, and virtual machine migration.

The migration of virtual machines is one of the steps that is performed in order to balance the load on resources. In ([30]), the problem of resource allocation and load adjustment in dynamic edge computing is discussed. This problem is modeled as a Markov decision process. In this paper, service caching is also considered in the MEC system, which is one of the innovations of this paper. It should be noted that MEC consists of different types of computing servers, which causes high diversity in this environment.

Migration of virtual machines has also created a line of research called VM consolidation and container consolidation. In [5], a method for merging virtual machines and assigning them to PMs is presented in two stages: choosing virtual machines based on the index of the impact of the virtual machine on the host being too busy, and selecting the host with the aim of reducing the number of active hosts in the network.

According to the objective of optimizing energy consumption and in line with green computing, one of the profitable strategies to reduce energy consumption is the dynamic integration of virtual machines in less physical machines. Most research divides the virtual machine integration problem into three main stages: (1) identification of physical machines that have suffered overloads; (2) choosing the best virtual machines for migration; and (3) choosing the best physical machines to place the virtual machines in.

Another group of researchers has specifically categorized works or resources. Classification of tasks is performed in order to prioritize the handling of tasks in specific conditions, and classification of resources is performed because more appropriate resource allocation is performed based on the general characteristics of resource categories. The papers in this category are introduced as follows.

In [30], branching and enhanced learning algorithms are presented to detect the optimal deployment of activities in the cloud and connected edge. The first algorithm is used to separate the hosts based on their characteristics, and then the second algorithm performs the allocation of activities in the space created by the first algorithm, which has reduced the computational complexity. In [31], by categorizing hosts, different mechanisms for dynamic integration of virtual machines have been proposed for each category of hosts, and in all cases, special attention has been paid to energy consumption.

The paper [32] proposes a model in the application layer that aims to group activities into classes with similar completion times, and by using machine learning algorithms, this model can also consider the dynamics of activity completion times. In [33], a non-invasive host-based method is proposed to describe the workload of virtual machines based on hypervisor tracing. In [33], the effectiveness of clustering virtual machines in reducing their management complexity is shown. Clustering virtual machines is achieved by hypervisor solutions such as VMware (Broadcom) or Microsoft Hyper-V.

Extracting the necessary features for clustering is performed based on hypervisor trace analysis from the kernel virtual machine. Then, Wakeup Reason analysis and Process Ranking Algorithm methods are defined. The first one is to identify different CPU states in virtual machines and collect the analysis metrics of virtual machines, and the second one is to identify different processes and patterns in virtual machines.

Cloud computing has attracted a significant number of users in the last decade; as a result, the number of requests and also the dynamics of requests for cloud computing have also increased. On the other hand, it is customary that in data centers, due to service quality issues, resources are provided more than the amount of requests [34]. For this reason, the problem of predicting requests or workload arises, which is another category of issues raised in this field. The papers in this category are introduced as follows.

In [35], a model based on regression and logistic regression is presented to dynamically identify overactive hosts based on the historical data of various types of workloads entering the data center. In the paper [36], a multi-step approach to forecast workload using machine learning techniques (retraining a hybrid method of clustering and forecasting algorithms) is presented, and then, resource allocation is performed based on these forecasts with the aim of reducing data center energy consumption.

In [37], first, with the help of logarithmic operations, time series of resource consumption and load are de-noised, and then an innovative deep Long Short-Term Memory (LSTM) method is proposed that uses the advantages of Bidirectional Long Short-Term Memory (BiLSTM) and GridLSTM simultaneously to accurately predict resource load and time series.

Service availability and quality of services have always been important issues in cloud computing; a report from the COVID-19 pandemic period shows that the number of failures in cloud computing that have not been resolved for a long time has affected many service providers. Also, the most common causes of failures in cloud computing are related to software failures, virtual machines, and servers. Due to the importance of service availability and quality of services, a new research area has emerged under the title of failure prediction in cloud computing. In various studies, cloud security is recommended for future research. The papers in this category are introduced as follows.

In this regard, research [38] has presented a model for maintaining cloud resources in order to reduce costs and increase the availability of services for users based on the LSTM method. In this research, active and reactive methods have been used to prevent service failure at the same time.

In Table 3, the research area and scope of research in different papers are given. Also, in order to better understand the focus of each field of study in the cloud computing environment, Figure 3 is presented. As shown in Figure 3, three columns are provided. The right column classifies entities of the cloud environment into six groups, including physical machines (PMs), virtual machines (VMs), tasks, schedulers, requests, and users. In addition, the graphical representations of the mentioned entities and classification of their research area are shown in the middle and left columns, respectively.

Table 3. Research areas, methods, and metric types.

Row	Paper	Computation Environment	Research Areas							Methods						Metrics					
			RA/RS [1]	Load Balancing	Resource Classification	Task Prioritization	Task Prediction	Failure Prediction	Economic Aspects	Meta-Heuristics	Machine Learning	Deep Learning	Reinforcement Learning	Deep RL	Others	Energy Efficiency	SLA/QoS [2]	Resource Utilization	Model Evaluation	Scalability	Environmental
1	[41]	CDC	√						√	√		√				√	√	√			
2	[27]	EDGE		√								√				√		√	√		
3	[49]	CDC	√							√							√	√			
4	[38]	CDC			√	√		√	√		√	√					√		√		
5	[44]	FOG	√			√			√					√		√	√	√			
6	[52]	EDGE	√						√			√					√				
7	[30]	EDGE	√	√										√			√	√	√		
8	[45]	CDC		√			√					√			√			√	√		
9	[50]	CDC		√				√		√	√					√	√	√	√		
10	[46]	CDC	√			√			√				√		√	√	√	√			
11	[42]	CDC	√						√	√						√	√				
12	[53]	EDGE	√	√										√		√					
13	[54]	CDC					√					√						√	√		
14	[55]	CDC	√	√						√		√					√	√	√		
15	[56]	FOG	√						√			√				√	√				
16	[86]	FOG													√	√					
17	[57]	EDGE	√										√	√					√		
18	[21]	EDGE	√						√			√		√			√		√		
19	[58]	EDGE		√									√								
20	[59]	EDGE	√								√	√				√	√				
21	[29]	CDC	√	√					√				√			√	√		√		
22	[60]	CDC			√	√			√		√						√		√		
23	[5]	CDC		√								√		√		√	√				
24	[61]	CDC					√		√			√					√	√	√		
25	[62]	CDC	√						√					√		√	√	√			
26	[24]	CDC	√						√			√		√		√	√				√
27	[48]	CDC	√	√						√				√		√	√		√		
28	[25]	CDC	√						√		√					√	√	√	√		
29	[63]	CDC	√	√						√			√						√		
30	[64]	CDC	√	√					√	√	√					√	√				
31	[65]	CDC	√			√					√								√		
32	[22]	EDGE	√						√					√			√				
33	[66]	CDC	√								√	√						√	√		
34	[28]	EDGE	√	√	√						√			√			√		√		
35	[26]	CDC	√	√										√	√	√		√	√		
36	[67]	CDC		√					√		√					√	√	√	√		
37	[68]	CDC		√			√							√		√		√	√		

Table 3. *Cont.*

Row	Paper	Computation Environment	Research Areas							Methods					Metrics					
			RA/RS [1]	Load Balancing	Resource Classification	Task Prioritization	Task Prediction	Failure Prediction	Economic Aspects	Meta-Heuristics	Machine Learning	Deep Learning	Reinforcement Learning	Deep RL	Others	Energy Efficiency	SLA/QoS [2]	Resource Utilization	Model Evaluation	Scalability	Environmental
38	[51]	CDC	√							√				√			√				
39	[69]	CDC	√	√					√		√					√	√	√			
40	[70]	CDC	√			√			√					√		√	√				
41	[71]	CDC	√											√		√	√				
42	[36]	CDC	√				√				√				√			√	√		
43	[43]	CDC	√	√						√				√				√	√		
44	[33]	CDC			√						√				√				√		
45	[72]	CDC	√	√					√	√					√	√	√			√	
46	[73]	FOG	√	√						√							√	√	√		
47	[74]	FOG	√						√	√			√				√	√			
48	[75]	CDC	√						√			√		√			√	√	√		
49	[76]	CDC		√							√					√	√	√	√		
50	[32]	CDC				√					√								√		
51	[37]	CDC					√					√						√	√		
52	[77]	CDC	√						√					√		√	√				
53	[78]	CDC	√	√							√	√				√	√		√		
54	[31]	CDC		√	√						√			√			√	√	√		
55	[79]	CDC					√		√			√				√	√	√	√		
56	[80]	CDC	√										√						√	√	
57	[81]	FOG	√	√									√				√				
58	[34]	CDC					√					√							√		
59	[35]	CDC		√			√		√		√					√	√		√		
60	[23]	CDC	√						√				√			√	√				
61	[82]	CDC		√			√		√				√			√	√	√	√		
62	[83]	CDC	√	√									√			√	√		√		
63	[84]	CDC	√	√									√		√	√	√		√		
64	[86]	CDC							√						√						√
65	[87]	CDC							√						√						√
66	[88]	CDC							√						√						√
67	[89]	CDC							√						√						√
68	[90]	CDC							√						√						√
69	[91]	CDC							√						√						√
70	[92]	CDC							√						√						√
71	[93]	CDC							√						√						√
72	[94]	CDC							√						√						√
73	[95]	CDC							√						√						√

[1] Resource Allocation/Resource Scheduling. [2] Service level management/quality of service.

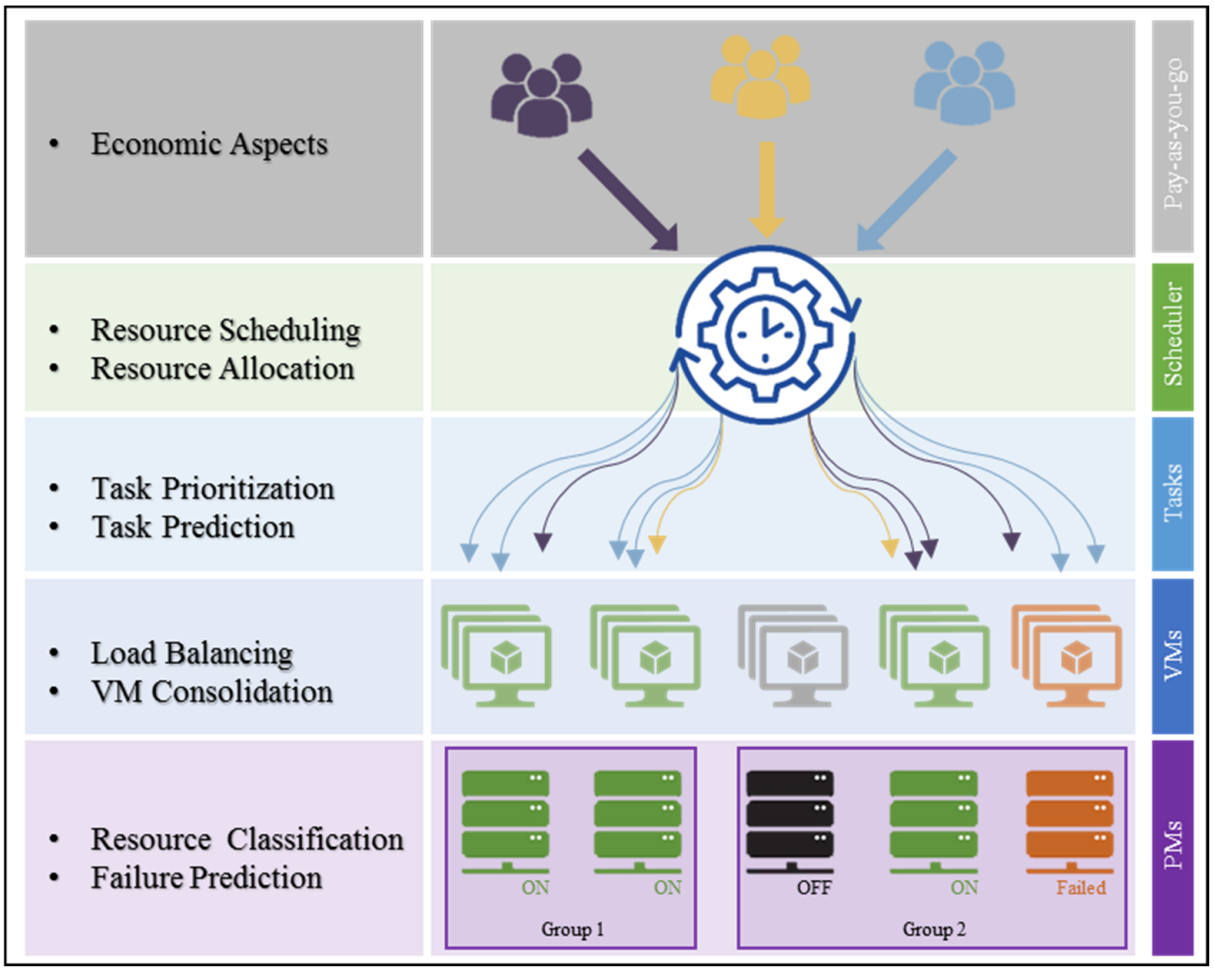


Figure 3. Intelligent resource management research areas in green cloud computing environments.

Other components that affect the data center environment, like the virtualization layers, clustering virtual machines (hypervisor tools like VMware and Microsoft Hyper-V) and cluster management/orchestration layer (tools like vSphere/DRS, System Centre, OpenStack, and Kubernetes for containers), are not directly discussed in this study [39,40].

Hence, the main research areas in the current review are as follows:

- Economic aspects: The role of the pay-as-you-go (PAYG) model of cloud computing is another important concern in green cloud computing. In cloud computing, users select configuration models that meet their specific needs; for cost efficiency, they often opt for minimal configurations, which can also lead to reduced power consumption. Consolidation and right-sizing are driven not only by technical efficiency but also by pricing incentives and billing granularity (e.g., per-second/per-minute instances, reserved/savings plans, and spot/pre-emptible capacity). This is important because "optimal" allocation can differ depending on whether the objective function is provider-side energy minimization, user-side cost minimization, or a multi-objective trade-off between SLA, carbon intensity, and price. Hence, some cost-related metrics (instance-hour cost, overprovisioning ratio, utilization-to-bill ratio) alongside energy metrics (server power models, joules per task) are also important.
- Resource allocation: Resource allocation is a fundamental process in computing systems where available resources (such as CPU cores, memory, storage, bandwidth, or GPUs) are assigned to tasks, applications, or users in a way that meets performance goals while optimizing system efficiency.

- Resource scheduling: Resource scheduling is the decision-making process that orchestrates the execution order and placement of workloads across available resources (CPU, memory, storage, network). The scheduler examines the system's current state and determines the most efficient plan for running tasks so that performance objectives, deadlines, and fairness criteria are met.
- Task prioritization: The process of determining the order in which tasks (or jobs) should be executed based on their importance, urgency, or other criteria. This is a key part of resource scheduling because the scheduler needs to decide which tasks receive access to limited resources first.
- Task prediction: The process of forecasting future tasks or workload characteristics so that resources can be allocated and scheduled more efficiently.
- Load balancing: The process of distributing workloads and computing tasks evenly across available resources to ensure optimal system performance, avoid bottlenecks, and maximize resource utilization. It is closely related to resource allocation and scheduling, but focuses specifically on how tasks are spread across multiple servers, nodes, or clusters. (Load balancing can be considered a subset of RA/RS, but according to the literature, some of them merely focus on load balancing. Therefore, it is defined as a unique research area to highlight the importance.)
- VM consolidation (and VM placement): The process of combining multiple virtual machines onto fewer physical servers in order to optimize resource usage, reduce energy consumption, and improve efficiency. It is closely tied to resource allocation and scheduling. (VM consolidation can be considered a subset of RA/RS, but according to the literature, some of them merely focused on load balancing. Therefore, it is defined as a unique research area to highlight the importance.)
- Resource classification: The process of categorizing computing resources based on their type, characteristics, or capabilities to manage them more effectively for allocation, scheduling, and optimization.
- Failure prediction: The proactive anticipation of potential system failures so that the system can take preventive actions, ensuring reliability, minimizing downtime, and improving overall performance.

4.2. Methods and Metrics

In cloud computing, the calculations related to the allocation and scheduling of resources due to the wide solution space and heterogeneous resources are in the category of NP-Hard problems. Solving these problems needs capable methods including heuristics, meta-heuristic and artificial intelligence-based methods (machine learning, deep learning, etc.) [6] which are examined in this section.

4.2.1. Methods

Meta-heuristic methods are computational intelligence models that are widely used to solve complex and NP-Hard optimization problems. These models do not necessarily find the optimal answer in the feasible space of the problem, but based on a specific pattern, they search the feasible space of the problem and find various possible combinations of model parameters that lead to a reasonable solution for the problem; then, the objective function is calculated and sorted based on the results.

The main reason for the introduction of this class of methods is the impossibility of solving NP-Hard problems in a reasonable time with classical optimization methods. Meta-heuristic models can obtain a relatively optimal solution for this group of problems in a fair time. In general, meta-heuristic methods are included in the category of estimation methods for solving mathematical problems.

There are two main categories for meta-heuristic methods, the first category is inspired by nature, such as genetic algorithm, evolutionary methods, bee colony, swarm intelligence, etc. In the other category, which is not inspired by nature, there are methods such as forbidden search, imperialist competition, harmony, etc. Among the papers that have used this category of methods are the Levy flight firefly algorithm [41], the rock hyrax algorithm [42], and the Coral Reef algorithm [43].

On the other hand, there are methods based on artificial intelligence, for example, machine learning, which is artificial intelligence that uses algorithms and statistical approaches to enable machines to learn from data in a way that can improve their performance in solving problems that are not designed for a specific type of problem. These systems autonomously improve their learning over time using data and information obtained from their interactions with the real world. Among the papers that have used this category of methods are the Double Deep Q-Learning algorithm [44], the hierarchical reinforcement learning algorithm [30], the Bi-LSTM algorithm [45], as well as the SARSA and BWA algorithms [46].

Using the combination of meta-heuristic methods and artificial intelligence is another approach that has attracted the attention of researchers in recent years [47]. Due to the complexity of the computing environment and high computation time of single-handed models, hybrid models were introduced to address various performance challenges of single algorithms, including increasing convergence rate of models, avoiding local optimum, etc. [48]. For example, an algorithm called whale-based convolution neural framework has been introduced in [49], which is a convolutional network based on the whale optimization algorithm. In [50], krill herd, whale optimization algorithms are used along with a deep learning method. In [51], the greedy adaptive butterfly algorithm is used based on the deep reinforcement learning (DRL) method.

According to the literature, artificial intelligence algorithms are divided into three general groups: machine learning algorithms, deep learning, and reinforcement learning. In the lower layer, the machine learning algorithms used in the literature are divided into two categories: unsupervised and supervised algorithms, where supervised algorithms are used for prediction and unsupervised algorithms are used for clustering purposes.

Table 3 depicts the result of analysis of the selected papers regarding (1) computation environment including cloud data center (CDC), edge, and fog; (2) research areas including failure prediction, task prediction, task prioritization, resource classification, and load balancing; (3) methods including meta-heuristics, machine learning, deep learning, reinforcement learning, and DRL; and (4) metrics including energy efficiency, SLA/QoS, resource utilization, model evaluation, scalability, and environment. Also, in Table 3, the types of algorithms used in each paper, as well as the more complete details about the algorithms used in each paper, are provided.

4.2.2. Metrics

Energy consumption is the most important metric in green computing. Considering the direct relationship that cloud has with the end users, quality of service is another important metric category, including execution time [23], response time, delay time [52], and completion time [48]. In addition, to evaluate the efficiency of resource allocation and scheduling, the number of migration index are used.

The next category is metrics that measure the optimal use of resources, such as the number of active machines and resource usage [36]. Other metrics that are considered in various papers are derivatives of the two mentioned metric categories. For instance, the number of inactive machines is related to the concept of the energy consumption metric [53–82].

The metrics introduced so far investigate the effects of the utilized solution on the computing environment. In order to evaluate the efficiency of the utilized method or algorithm, other specific metrics are utilized. For instance, precision, recall, and accuracy are used to evaluate the efficiency of prediction algorithms [65]. In addition, silhouette score indices [36], intra-cluster and inter-cluster similarity indices [33] are used to evaluate the efficiency of clustering algorithms. Furthermore, error measurement indices such as mean absolute error, mean absolute percentage error, and root mean square error [54] are used to evaluate the efficiency of statistical methods.

The environmental factors and scalability metrics are rarely used in the literature. The temperature metric is one example of an environmental factor considered in [82]. The scalability metric is used in [72,80]. The scalability metric evaluates the efficiency of the algorithm by expanding the scale of the environment. In other words, if an algorithm provides admissible results in both average and large-sized data centers, the solution has scalability. Table 4 summarizes the evaluation methods and metrics utilized in selected papers.

4.3. Tools and Datasets

4.3.1. Tools

As shown in Table 5, the most widely utilized tool to simulate a cloud environment is Cloudsim which is a powerful tool for simulating a cloud environment with predefined packages. The second prevalent tool is the ones developed by Python v 3.13.0. Other utilized tools are MATLAB R2024a, Java v21, Rapid-Miner v 10.3.1, iFogSim v2, R v4.4.0, and C++, based on their frequency of usage in the literature.

4.3.2. Datasets

The gathered statistics in Table 6 show that among public datasets that can be used for different purposes in cloud computing, Google traces is the most widely used one. This dataset is provided by the well-known company, Google, and is updated regularly. While this dataset provides a large database, researchers select a subset of this dataset according to their research objective. The second rank is custom real datasets that are provided by researchers. Bitbrains, SPECpower, and PlanetLab are the other most utilized public datasets.

Table 4. Method and metric details.

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
1	[41]	RQ2, RQ3, RQ4	The authors used a combination of Levy flight firefly (LFF) and a modified genetic algorithm (GA) to improve task arrangement in VM Migration. This hybrid approach enhances exploration speed and avoids local optimums. They also employed E-ANFIS for task mapping, achieving at least 8% performance improvement compared to other methods.	Computing capacity, task execution time, request response time, mean request waiting time, and energy/CPU consumption.
2	[27]	RQ2, RQ4	Authors propose the Road-Runner method to manage Edge and CDC tasks, with offline and online versions using complex CNNs. This method reduces energy consumption by 88% and speeds up processes compared to the Neurosurgeon baseline, using VGG and AlexNet models for DNN benchmarks.	Performance, energy consumption, and prediction accuracy.
3	[49]	RQ2, RQ3, RQ4	The Whale-based convolutional neural framework schedules jobs based on deadlines to reduce resource use and costs. It predicts task deadlines, prioritizes assignments, and shows 8% higher accuracy and 50% faster execution than other models. The authors did not clarify the convolutional neural framework.	Task scheduling rate, prediction accuracy, time to complete the entire process, task scheduling time, resource consumption, and task completion time.
4	[38]	RQ2, RQ4	The author presents the SRE-HM model for task and resource prioritization, which enhances service availability by 19.56%, reduces active servers and power consumption by 26.67% and 19.1%, respectively, and predicts failures using LSTM. The model is tested with real-world data.	Mean time between failures, mean time to repair, resource consumption, energy consumption, failure prediction rate (accuracy), and resource availability.
5	[44]	RQ2, RQ4	The text presents a hybrid model improving cloud resource management, enhancing energy efficiency by 10% and task completion by 1%.	Task execution time, task execution rate, resource consumption, energy consumption.
6	[52]	RQ1, RQ5	To address power limitations in mobile edge computing, the authors introduced a DNN model using policy reinforcement learning for task management and resource allocation. While lacking a suitable comparison baseline, the study evaluates various scenarios and real-world constraints like user device delay.	Rate of task execution.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
7	[30]	RQ3, RQ5	The authors proposed two policy optimization DQNs to reduce resource wastage in remote clouds, servers, and devices, focusing on workload offloading and load balancing. They introduced a cost function to enhance resource use, achieving better results than three baselines. Future work should improve fault tolerance and use real-world data.	System cost, resource usage rate, and delay in task execution.
8	[45]	RQ3, RQ4	The authors address overloads in CDCs due to cloud computing's dynamic nature by proposing a Proactive Hybrid Pod Auto Scaling method. It monitors workload and uses an attention-based bidirectional LSTM for resource pattern extraction, improving CPU usage by 23.39% and memory by 42.52%.	CPU consumption, Memory consumption, RMSE [1], MAE [2], and R-squared.
9	[50]	RQ3, RQ4	The proposed hybrid model combines Krill herd and Whale-based techniques for load balancing in CDCs, improving performance over existing methods, but lacks dataset details.	Energy consumption, resource consumption, memory efficiency, task execution time, task distribution balance, and time complexity index.
10	[46]	RQ2, RQ3	Cloud computing challenges include task prioritization and energy use; uRank-TOPSIS improves scheduling and resource allocation through intelligent agents.	Energy consumption, task execution time, request response time, resource efficiency, and throughput.
11	[42]	RQ2, RQ3, RQ4	The authors proposed a dynamic resource allocation method using the rock hyrax meta-heuristic algorithm for improving QoS in cloud environments. The model, tested in CloudSim and compared with other algorithms, achieved a 3–8% QoS improvement but lacked real-world dataset evaluation.	Time to do tasks and energy consumption.
12	[53]	RQ2, RQ3, RQ5	To minimize power usage during network caching and request processing, the authors created a DRL-based algorithm for optimizing task offloading and resource allocation. The TORA-DRL model uses a Deep Q Network and epsilon greedy for exploration. In tests, it showed less power consumption than baseline models, although only one metric was used for evaluation.	Energy consumption.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
13	[54]	RQ2, RQ3, RQ4	The authors propose a multivariate time-series forecasting model, BiLSTM, to understand the relationship between provisioned and utilized resources in cloud data centers (CDCs). The model uses various attributes like CPU and memory metrics. While it improves service level agreements, its complexity creates a computational burden.	Mean absolute error (MAE); root mean square error (RMSE); and mean absolute percentage error (MAPE) on CPU, memory, disk, and network consumption.
14	[55]	RQ2, RQ3	A proposed migration method enhances resource utilization in cloud computing through workload forecasting, requirement identification, and optimized resource allocation.	Normalized MSE, CPU consumption, resource wastage, and ratio of rejected activities.
15	[56]	RQ2, RQ3, RQ5	Cloud-fog expansion necessitates resource optimization to reduce energy consumption. A CNN-based approach was proposed, showing 15% improvement over traditional models, but it needs real dataset validation.	Energy consumption, time of activities.
16	[57]	RQ2, RQ3, RQ5	A novel deep reinforcement learning deadline-driven allocation algorithm optimally addresses RMSA and AR problems, outperforming traditional methods in cloud-edge environments.	Model runtime and loss.
17	[21]	RQ2, RQ3, RQ5	The proposed DRL algorithm improves end-edge-cloud orchestrated computing (EECOC) performance by incorporating movement prediction and sensitivity analysis.	Computing delay, convergence, and accuracy.
18	[58]	RQ5	Quantum-inspired reinforcement learning improves task offloading in vehicles.	Service delay, energy consumption, and cost metrics.
19	[59]	RQ5	A scheduling model using spectral clustering and LSTM improves response time in IoT Edge environments.	Model convergence, execution time, delay, response time, and the utilization of network resources.
20	[29]	RQ2, RQ3, RQ5	Proposed RL-based VM scheduling reduces energy, cost, and SLA violations but increases migration latency.	Energy consumption, cost, number of performed migrations, and service level agreement.
21	[60]	RQ2, RQ3, RQ5	Dynamic interference-aware cloud scheduling improves performance by utilizing online Bayesian change point detection.	Response time, software interference index in resource allocation, and classification time.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
22	[5]	RQ2, RQ3, RQ4	Proposed VM consolidation method improves load balance using DRL, LSTM for resource allocation, and real-world data.	Energy consumption, service level agreement.
23	[61]	RQ4, RQ5	A generative adversarial network generates realistic workloads with specific patterns.	F1-Score, RMSE, resource consumption, cost, and time.
24	[62]	RQ2, RQ4, RQ5	The authors proposed an Actor–Critic DRL method to tackle cloud data center challenges, improving efficiency over basic A2C methods.	Delay in activities, activity loss rate, and energy consumption.
25	[24]	RQ2, RQ4	The HUNTER model improves energy efficiency in cloud computing compared to several baseline models.	Service level agreement, energy consumption, cost, and temperature.
26	[48]	RQ2, RQ3, RQ4, RQ5	The authors propose DRLPPSO, a model utilizing deep reinforcement learning and Parallel Particle Swarm Optimization to address complexity and performance issues in high-dimensional data, focusing on load balancing, reducing make span time, and energy consumption.	Model accuracy, energy consumption, and completion time.
27	[25]	RQ2, RQ3, RQ4	A multi-objective task workflow scheduling model using decision trees aims to minimize make span, enhance load balance, and maximize resource utilization. It involves task priority calculation, resource matrix creation, and allocation, showing a 5% efficiency gain but increasing energy consumption by 30%.	Resource consumption index, load balance, busy time, idle time, total execution time, and energy consumption.
28	[63]	RQ2, RQ3, RQ4	The authors propose a hybrid multi-objective job scheduling method in cloud computing, combining Artificial Bee Colony and Q-learning algorithms to enhance load balancing and minimize make span and costs. Compared to existing algorithms, it achieves over 10% improvement in make span, though simulations show a low number of tasks.	Cost and make span.
29	[64]	RQ1, RQ3, RQ4	The Energy-Aware–ABC algorithm minimizes VM migrations, using a Modified Best Fit Decrease algorithm for PM allocation. SVM identifies overload patterns, reducing migrations compared to state-of-the-art models.	Energy consumption and service level agreement.
30	[65]	RQ2, RQ3, RQ4, RQ5	The proposed model enhances admission control in fog-cloud environments by clustering requests, using K-means for clustering, and a decision tree classifier for labeling, outperforming baselines in performance and execution time.	Model accuracy, execution time, precision, and recall.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
31	[22]	RQ2, RQ3, RQ5	DeepEdge addresses resource allocation challenges in heterogeneous applications by utilizing a QoE model and a two-stage DRL. It outperforms AD and DR-learning, achieving 15% less latency and a 10% higher task completion rate in edge-cloud environments.	Operational delay and the number of completed tasks.
32	[66]	RQ1, RQ3, RQ4, RQ5	The proposed Kubernetes scaling engine utilizes various machine learning methods, including AR, HTM, LSTM, and RL, to improve user request management in dynamic environments. It adapts Horizontal Pod Auto scaling for better performance and resource usage.	Excess of pod usage, loss, lost requests, and resource usage.
33	[28]	RQ2, RQ3, RQ5	The proposed joint cloud-edge task deployment strategy uses a pruning algorithm to select suitable edge-cloud resource pairs and the Deterministic Strategy Gradient Algorithm (DDPG) for effective task deployment, achieving improved load balancing and reduced computation compared to FIFO and FERPTS methods.	Activity time, response time, and workload balance in the network.
34	[26]	RQ2, RQ3, RQ4	Cloud computing faces high energy consumption, leading to increased CO_2 emissions. To promote green computing, the authors propose VM consolidation software focusing on VM selection and placement. Using a VM utilization-based overload detection algorithm (VUOD) and selection policy, the model shows slightly improved energy efficiency over default CloudSim baselines.	Energy consumption, the number of migrations, service level agreement, and resource wastage.
35	[67]	RQ2, RQ3, RQ4, RQ5	The proposed fuzzy SARSA method enhances VM placement during consolidation, reducing VM migrations and improving energy efficiency by up to 50% compared to FFD. Incorporating FIS increases DRL speed and adaptability in fluctuating workloads.	Energy consumption, resource wastage, and the number of migrations.
36	[68]	RQ2, RQ3, RQ5	Addressing the resource management challenge of wastage reduction, the authors have proposed a novel energy-aware technique that considers both VM and container consolidation, and a DQL algorithm chooses which of them or both of them should be migrated. Unfortunately, no evaluation is available in this paper.	Energy consumption, the number of migrations, and the amount of resource wastage.
37	[51]	RQ2, RQ3	Authors propose a multi-agent system for resource allocation and scheduling, addressing historical resource wastage. The system features local agents reformulating cloud requests and a global agent handling VM management, using a greedy adaptive firefly algorithm. Simulation results outperform existing methods.	Activity execution time, delay time, response time, accuracy, and efficiency.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
38	[69]	RQ2, RQ3, RQ4	Approximately 70% of peak power is wasted by idle servers, increasing costs. A proposed method using VM scheduling and migration aims to address energy consumption. It involves PM clustering via Bayes Theorem, allocating VMs based on load and energy consumption, and migrating VMs from underloaded PMs for optimal resource utilization.	Number of active machines, energy consumption, and execution time.
39	[70]	RQ2, RQ3, RQ4, RQ5	The authors propose a partition-based dynamic task scheduling model leveraging proximal policy optimization-clip and an auto-encoder for high-dimensional processing. Evaluation against H_2O-Cloud and Tetris shows reduced energy use, but waiting time remains unchanged, with undefined partitioning strategy drawbacks.	Energy consumption and waiting time.
40	[71]	RQ2, RQ3, RQ4, RQ5	A complex mathematical model for VM placement has been proposed, utilizing a Rainbow DQN to sequentially select VMs based on 20 constraints, including security. Six heuristic methods improve deployment, but the model's complexity results in intensive calculations.	Quality of service experience, consumed energy, and durability of quality with respect to required energy.
41	[38]	RQ2, RQ3, RQ4, RQ5	A hybrid retrain-based prediction approach combines k-means clustering with forecasting techniques (SVM, DT, LR, K-NN, and Gradient Boosting) to predict VM resource consumption. The Best Fit Decrease algorithm facilitates VM placement. Evaluation shows the proposed method competes well with LSTM and GRUED, offering long-term forecasts and diverse algorithm use.	Silhouette score, RMSE, RAE, CPU and Memory consumption, as well as the average number of active physical machines that the data center needs in one hour.
42	[43]	RQ2, RQ3, RQ5	The proposed hybrid model combines multi-objective Coral Reef Optimization and Q-learning to enhance cloud resource management through improved allocation, migration decisions, and load balancing, although it struggles with large-scale data applicability.	Productivity index and migration index.
43	[33]	RQ2, RQ3, RQ4	The authors propose an agent-less method for clustering VMs to manage cloud resources efficiently while maintaining privacy. Using WRA for feature extraction and K-means clustering, results show reduced overhead to 10% compared to agent-based methods.	Silhouette, intra- and inter-cluster similarity metrics.
44	[72]	RQ1, RQ3, RQ4	An ant colony optimization (ACO) combined with a deterministic spanning tree protocol (STP) improves task scheduling in large-scale data centers, enhancing energy efficiency and performance metrics compared to existing methods like ICGA, ICSA, and RACO.	Energy consumption, make span time calculations, customer experience, and scalability.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
45	[73]	RQ2, RQ3, RQ5	Addressing conventional issues of fog-cloud environment as high latency caused by delayed response from devices, authors have proposed a load balancer approach using the PSO meta-heuristic algorithm. In comparison with Round Robin, the proposed approach achieves up to 20% lower response time.	Response time, network consumption, calculation time.
46	[74]	RQ2, RQ3, RQ4, RQ5	The proposed context-aware approach in fog computing optimizes energy consumption through intelligent sleep cycles, genetic algorithms for node allocation, and RL algorithms for optimal duty cycles, achieving 11–21% less energy use.	Service delay, service level agreement violation, and the number of busy nodes.
47	[75]	RQ2, RQ3, RQ5	A cloud automation framework for task scheduling using RL, DQN, RNN-LSTM, and DRL-LSTM is proposed, achieving optimal accuracy and minimal resource utilization, as previous research lacks this combination.	Execution delay, increase resource utilization (CPU, RAM consumption), and model accuracy.
48	[76]	RQ2, RQ3, RQ4	Container-based virtualization faces challenges, particularly service inter-dependency during consolidation. The proposed Energy Aware Service consolidation using Bayesian optimization (EASY) enhances container placement but has a 10% slower response time than BFD and FFD, lacking clear overload detection strategies.	Energy consumption, number of active hosts, and resource wastage.
49	[32]	RQ2, RQ3, RQ4	Resource allocation in cloud environments can cause interference, degrading performance, and QoS. An interference-aware classifier using SVM and K-means quantifies interference, achieving 97% accuracy and a Rand index of 0.82, surpassing previous studies.	Accuracy, F1-Score, and Rand Relax.
50	[37]	RQ2, RQ3, RQ4, RQ5	The authors propose a novel prediction model combining BiLSTM and GridLSTM to forecast workload time series, addressing high variance and accuracy issues. They utilize logarithmic transformation and Savitzky–Golay filtering, achieving superior results compared to traditional methods using a large-scale dataset.	CPU consumption, Ram, Cumulative Activity Entry Index, MSE, and RMSLE.
51	[77]	RQ2, RQ3, RQ4	A DQN algorithm optimizes energy consumption and task make span in cloud resource management, reducing energy use compared to random, RR, and MoPSO for over 150 tasks.	Energy consumption and average activity make span.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
52	[78]	RQ2, RQ3, RQ4	The authors propose an individual prediction model for CPU usage of each VM, accompanied by a VM migration method with three policies: overload/underload detection, VM selection, and destination host selection using a modified bin packing algorithm. Multiple machine learning methods enhance forecast accuracy.	Cross-validation, MSE, energy consumption, service level agreement, and number of migrations, as well as ESV, which is obtained by multiplying the number of violations of service level agreements by the amount of energy consumption.
53	[31]	RQ2, RQ3, RQ4, RQ5	The proposed Q-learning method addresses green computing by defining migration strategies for overloaded and underloaded hosts, using ON/OFF strategies for light and heavy underloads, and optimizing VM selection and allocation, outperforming MAD, IQR, LR, and LRR baselines.	Number of migrations, violation of the service level agreement, energy consumption, and resource consumption.
54	[79]	RQ1, RQ3, RQ4	The PACPA algorithm, addressing resource forecasting, utilizes a multi-layer perceptron with sliding windows for overload detection and VM selection, reducing energy consumption by 18% and SLAV by over 34%.	RMSE, MAE, MSE, MAPE, and metrics related to data center performance, including CPU/bandwidth/energy consumption, number of service level agreement violations, number of active hosts (relative to total), and total number of migrations.
55	[80]	RQ2, RQ3, RQ5	A Q-learning reinforcement learning framework, implemented in the SmartYARN prototype, optimizes multi-resource configuration for CPU/Memory in Apache YARN, achieving up to 98% cost reduction compared to the optimal scenario.	Search steps, cost, and scalability.
56	[81]	RQ2, RQ3, RQ5	Authors propose a joint optimization method using actor-critic reinforcement learning and DNN to address fog-enabled IoT challenges, minimizing latency and backhaul requirements; however, evaluations lack real datasets.	Average end-to-end delay time.

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
57	[34]	RQ2, RQ3, RQ5	A novel model, LSTMtsw, forecasts data center workloads in 24 h cycles, comprising future workload prediction and noise reduction phases, achieving 20% more accuracy than LSTM and BPNN through error moderation.	Number of iterations in the prediction process and the error rate of the prediction models.
58	[35]	RQ2, RQ3	Authors propose a model for overload host detection using linear and logistic regression to reduce energy consumption, SLAV, and VM migrations through VM consolidation, employing various VM selection algorithms for testing.	Service level agreement, energy consumption, and the number of migrations.
59	[23]	RQ2, RQ3, RQ5	A DRL-Cloud framework addresses scalability in large-scale data centers, optimizing resource provisioning and task scheduling without task dependencies. Utilizing directed acyclic graphs and deep Q-learning, it significantly outperforms traditional methods by up to 200% with extensive datasets.	Energy consumption, execution time, and number of completed tasks.
60	[82]	RQ2, RQ3, RQ5	The authors propose a reinforcement learning-based approach called Random Learning Automata Overload Detection (LAOD) to optimize cloud resource management by addressing SLA and energy efficiency. The model includes a global manager, local managers, and four steps for VM consolidation, outperforming NPA and DVFS metrics.	Energy consumption, number of physical machines off, number of migrations, and service level agreement.
61	[83]	RQ2, RQ3, RQ4, RQ5	Hosts operating at 60% capacity achieve optimal energy efficiency and performance. A self-optimizing reinforcement learning model, ARLCA, enhances VM allocation while ensuring SLA compliance, suggesting neural networks for improved learning.	Energy consumption, number of migrations, and service level agreement.
62	[84]	RQ2, RQ3, RQ4	The authors propose a two-step model for VM consolidation addressing drawbacks like CPU-load correlation, inter-VM communication, and large-scale data centers. Both methods involve VM clustering, classification, and allocation, with a hyper-heuristic selecting the optimal approach, outperforming baseline models.	Energy consumption, data exchange, service level agreement violation rate, number of migrations, and execution time.
63	[86]	RQ1, RQ3	The seminal paper defines cloud economics, including PAYG, as a core tenet. It establishes the economic argument for cloud adoption.	IRR/vCPU-hour, IRR/GPU-hour, IRR/GB-hour, IRR/TPU-hour, $kgCO_2e$, and gCO_2e/kWh.
64	[87]	RQ2, RQ3	This paper breaks down the true TCO of cloud data centers, showing how cost components (servers, power, and networking) drive pricing models like PAYG.	instance-hour cost, IRR/vCPU-hour, utilization-to-bill ratio, and Power Usage Effectiveness (PUE).

Table 4. *Cont.*

Row	Paper	RQMapping	Detailed Description, Including Methods	Metrics
65	[88]	RQ1, RQ3, RQ4	Comprehensive survey linking resource allocation (directly tied to PAYG metrics) to energy efficiency. Discusses scheduling, virtualization, and pricing.	IRR/vCPU-hour, IRR/GPU-hour, IRR/GB-hour, IRR/GB-month storage, IRR/GB data transfer, and PUE.
66	[89]	RQ1, RQ3	Presents algorithms for temporal and spatial workload shifting based on carbon intensity, using PAYG flexibility.	IRR/GB data transfer, kgCO_2e, and gCO_2e/kWh.
67	[90]	RQ1, RQ3	Empirical study showing how cloud service layers (IaaS, PaaS, and SaaS) differ in energy proportionality—critical for understanding which PAYG services yield greenest outcomes.	IRR/vCPU-hour, IRR/GPU-hour, IRR/GB-hour, IRR/GB-month storage, IRR/GB data transfer, and PUE.
68	[91]	RQ2, RQ3	Analyzes various cloud pricing models (spot, reserved, on-demand) and their implications for cost and resource efficiency.	instance-hour cost, spot, reserved, and on-demand models.
69	[92]	RQ2, RQ3	The foundational text connecting cloud financial operations with technical optimization. Chapter 8 explicitly discusses sustainability.	PUE, overprovisioning ratio, and utilization-to-bill ratio.
70	[93]	RQ1, RQ3	Details about a decision support framework to explore pricing policies of cloud services.	CloudPricingOps, SLA, and PUE.
71	[94]	RQ1, RQ2	Contains detailed chapters on cloud economics, resource management, and energy-efficient computing.	Instance-hour cost, overprovisioning ratio, and utilization-to-bill ratio.
72	[95]	RQ3, RQ4	Details how Google converts PAYG usage data (compute, storage, and network) into carbon emissions using region-specific grid carbon intensity. Explains the data models and calculations behind AWS's carbon reporting tool, which uses detailed billing records.	kgCO_2e, gCO_2e/kWh, and PUE.

[1] Root mean square error. [2] Mean absolute error.

Table 5. Tools usage statistics.

Tool	Papers	Usage
CloudSim v7.0.0	[5,24–26,31,42–44,46,60,63,67,69,78,79,82–84]	18
Python v3.13.0	[27,28,38,48–50,52,62,66,71,75–77]	13
MATLAB R2024a	[34,44,61,64,65]	5
Java platform v21	[29,41,51]	3
Other	[72,73]	2
Rapid-miner v10.3.1	[64]	1
iFogSim v2	[74]	1
R v4.4.0	[32]	1
C++	[84]	1

Table 6. Datasets usage statistics.

Dataset	Description	Paper	Usage
Google cluster traces	This dataset is used in two versions, 2011 and 2019, each of which contains cluster usage of thousands of resources in terms of CPU and Memory usage. This data is used for usage prediction, simulation of task entry, and clustering purposes.	[5,23,37,45,48,61–64,70,74,75,77]	13
Custom real dataset	These datasets have been generated by using static distributions or have been collected from real servers.	[33,34,71,72,79,83]	6
Bitbrains	GWA-T12 BitBrains dataset containing CPU, Memory, network, and bandwidth for a total of 8000 tasks through 1750 VMs, sampled every 5 min. The dataset is used for resource allocation, task scheduling, usage prediction, and clustering.	[36,46,54,84,85]	4
SPECpower benchmark	SPECpower benchmark is used for the configuration of hosts and energy consumption modeling.	[28,36,70]	3
PlanetLab	Real Cloud trace log that has been released by the CoMon project (Planet Lab) is used for CPU usage prediction. The dataset contains 11,746 VMs' CPU utilization for one day, and the CPU utilization is sampled every 5 min.	[61,78,82]	3
GCP pricing	GCP pricing is used for the cost model.	[75]	1
HiBench benchmark	HiBench is a benchmarking suite designed for big data, allowing the evaluation of various big data frameworks based on performance, throughput, and system efficiency.	[80]	1

Table 6. *Cont.*

Dataset	Description	Paper	Usage
Server Power	Server Power is a dataset containing the specifications of servers that is used for servers' power consumption.	[81]	1
Amazon EC2	Amazon EC2 data is used for real hardware configuration.	[5]	1
Berkeley Research Lab dataset	The Intel Berkeley research lab, consisting of 2.3 million records, is used, which is suitable for cloud–edge resource management and clustering purposes. The dataset contains attributes of humidity, light, voltage, and temperature.	[59]	1
Alibaba Trace	This dataset has different versions from 2017 up to 2023 for now. Overall, this dataset can be used for prediction and simulation in cloud computing, comprising CPU, GPU, Disk Usage, and other attributes that are available via a simple Google search.	[61]	1

5. Discussion

Our review shows that there are challenges and lines for future research regarding the aforementioned taxonomies provided in previous sections, some of which are general and have the aspect of improving the results, and some of them specifically seek to provide new innovations in relevant fields of study. In this section, based on our studies, various cases are presented.

Before stating the new research challenges, it should be noted that there is a lack of information in some of the reviewed papers about the way of evaluating the output results, including the tools, data sets, and metrics utilized.

Figure 4 depicts the research area in the reviewed papers. As shown in Figure 4, the highest frequency is related to the category of resource allocation and scheduling, then load balancing. Also, since most researchers have assumed that there is no failure in the cloud environment, failure prediction is the least frequent.

Furthermore, it can also be seen in Figure 5 that even without considering the area of Resource Allocation/Resource Scheduling (RA/RS), the contribution of resource classification, task prioritization, and task forecasting has not taken significant focus in the studies of recent years. Therefore, focusing on less studied research areas as well as considering the dynamics of cloud environments in terms of the possibility of failure are proposed as the new challenges that researchers should consider. Probable failures in cloud environments can be due to network failures such as a lack of bandwidth, packet delays, congestion, or physical failures in network links.

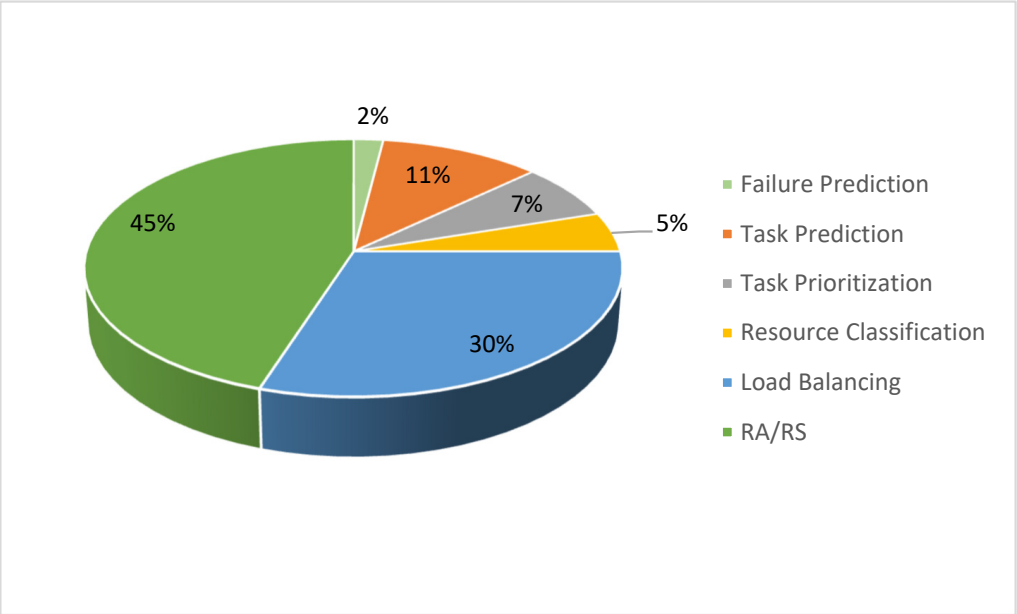


Figure 4. Research area in reviewed papers.

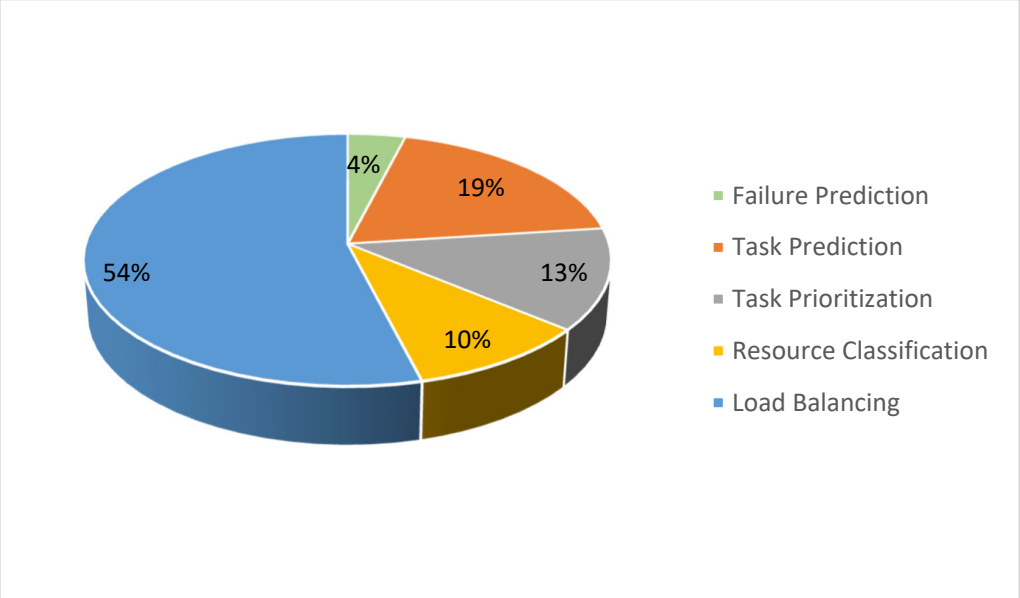


Figure 5. Research area in reviewed papers without RA/RS.

Moreover, the most important open research areas reported in the reviewed papers are sorted according to their occurrences as follows.

- When using learning models to predict system load, most of the research fails to predict the future load of the system, and their solution is based on the current load of the system. However, utilizing load prediction methods in scheduling models can provide adaptation with complex and dynamic cloud computing environments [74,77].
- By reducing the workload in the system by providing isolation while separating activities, in general, it is assumed that one activity is assigned to one physical host. However, in order to increase the resource utilization, it is recommended to share the workload of one activity to more than one host if possible [21,29].

- Load prediction using machine learning methods with a dynamic scheduling architecture can forecast the pattern of task entry, which can help scheduler agents to make more efficient decisions for resource allocation [60].
- Considering the priority in the processing of tasks, creating a parameter based on the deep reinforcement learning method can automatically detect and optimize the connection between databases [62].
- Virtual machines were grouped based on the consumption of different resources, providing a machine learning model for each group independently; grouping would lessen the computational burden of a scheduler because decisions are more based on a group of resources [78].
- Reducing computational complexity by reducing data overlap with clustering methods can be seen as grouping activities [59].

Figure 6 depicts the frequency of metric consideration in reviewed papers. As shown in Figure 6, a limited number of papers have considered the scalability and environmental metrics, including cooling management techniques and thermal parameters of the hosts, in their studies. Consequently, another challenge that is proposed to take more focus in future research is the issues related to the scalability of the solutions, as well as considering environmental metrics.

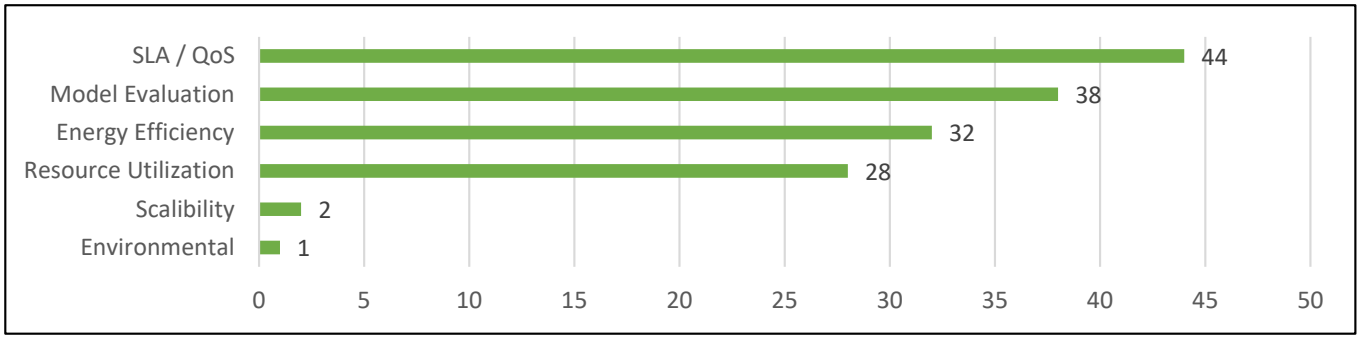


Figure 6. Metric-type frequency in reviewed papers.

In other words, bearing in mind the fact that the designed solutions should possess the generalization characteristic in various situations, such as preserving applicability when facing different data sizes, especially big data, or different request entry patterns, the scalability metric is one of the operational challenges that should be addressed with more emphasis in future studies. In this regard, generalization of the models using an automatic data augmentation method and a multilayer architecture for learning methods in the face of large-scale data sets is recommended for further research in the future.

Figure 7 shows that even though the reviewed papers focused on the topic of green computing, only 52% of papers (32 papers) have considered the energy efficiency metric as their evaluation metric. According to the justifications provided for energy efficiency importance in the Section 1, it is necessary for future research to consider energy consumption metric as their research objective, and also to use multi-objective models to increase the energy productivity in cloud environments.

In recent years, the use of artificial intelligence-based methods has made significant progress, as can be seen in Figure 7; among the artificial intelligence-based methods, deep reinforcement learning and deep learning are the most frequent ones.

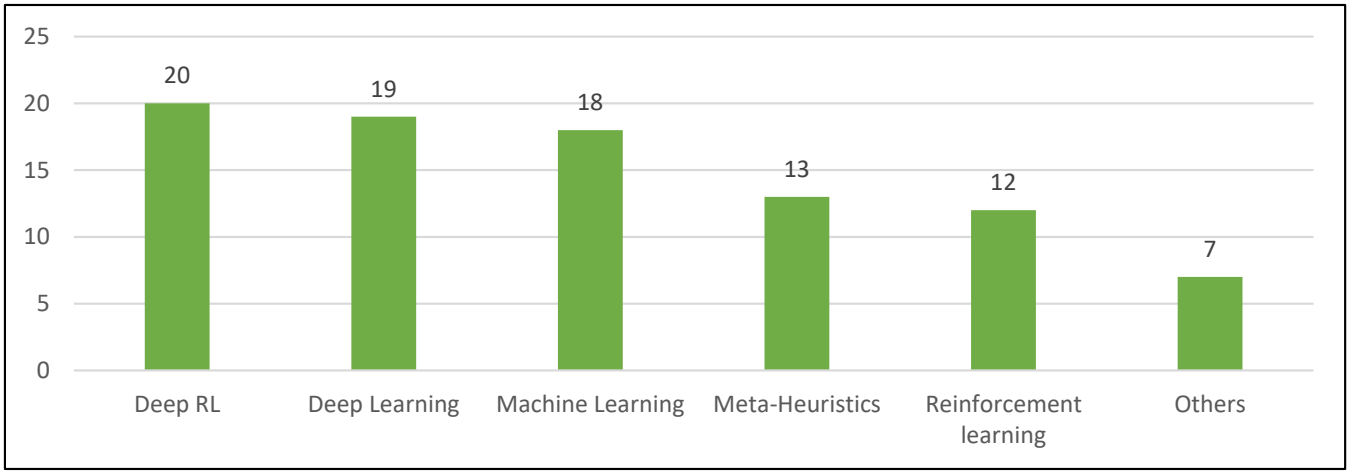


Figure 7. Method frequency.

6. Investigating Challenges and Opportunities

The challenges and opportunities that have been indicated in the reviewed papers are listed in Table 7.

Table 7. Challenges and opportunities.

Reference	Research Challenge	Research Opportunity
[5]	Computational burden of virtual machine consolidation in energy-efficient cloud data center	Optimization of LSTM algorithms and DRLs
[41]	Secure resource allocation based on deep reinforcement learning in serverless multi-cloud edge computing	Using feature selection algorithms before training machine learning models, and to consider remained capacity of energy as a limitation
[52]	Resource allocation optimization for energy-constrained hierarchical edge–cloud systems	Combining CNNs and GCNNs with RNNs
[43,44,55,59]	Computational complexity in intelligent green cloud computing	The use of hybrid algorithms to reduce the computational complexity
[66]	Best method selection for machine learning-based scaling management for Kubernetes edge clusters	Designing architectures in which several intelligent methods compete with each other, and the best method is selected
[82]	Energy and SLA efficient consolidation of virtual machines in cloud data centers	Providing an algorithm to identify machines that have less load than their capacity, taking into account the topology of the servers, temperature, and cooling systems
[86–95]	Ignoring the economic aspects of intelligent resource management in green cloud computing	Applying PAYG metrics for green intelligent resource management of cloud computing
[96–99]	Guaranteeing energy efficiency in fog computing	Energy-efficient fog computing by investigating techniques and recent advances
[100–102]	Lack of energy efficiency in cloud-based IoT networks	Improving energy efficiency through green cloud computing in IoT networks
[103–107]	Green sustainability in cloud data centers	Green-aware management techniques for sustainable data centers

Table 7. *Cont.*

Reference	Research Challenge	Research Opportunity
[108–111]	High computational cost in green intelligent RA/RS for cloud computing	Cost-related green intelligent RA/RS techniques for cloud computing
[112,113]	Reducing the computational cost of RA/RS in large-scale cloud computing	Green quantum-based RA/RS cloud computing

7. Conclusions

The emission of carbon dioxide is a critical environmental concern, exacerbated by the booming growth of cloud computing users. This surge necessitates utilizing more resources, which consequently produces higher energy consumption. Although the number of studies on energy efficiency in cloud data centers has increased in recent years, further research to mitigate energy consumption and carbon dioxide emissions is inevitable. Our statistical analysis revealed that only 52 percent of the reviewed papers directly address energy as their main research objective. It is important to note that other types of research objectives, such as resource utilization is related to energy consumption, but there is still a need for a distinct focus on green computing.

Another significant open research challenge overlooked in the literature is scalability. Our review results indicated that only two papers have evaluated their models regarding scalability. Therefore, we recommended addressing scalability as a future research area. Also, various aspects, including the variability of data centers, their structural aspects, their operational contexts, the request patterns, data center's geographical differences, and data center size, should be considered in proposing new solutions.

In addition, in this paper, we showed that there has been an increase in the use of deep models. Another deduction of our paper was the imperative utilization of deep techniques in AI-based methods for green resource management in cloud environments due to the exponential growth of data. Consequently, there should be a future emphasis on developing hybrid deep models rather than solely relying on machine learning models, which may not always be scalable enough to handle big data challenges.

This review selected a subset of papers published between 2018 and 2025, focusing on green computing and utilizing intelligent methods, including AI-based methods and meta-heuristics. Furthermore, this paper scrutinized the research areas considered regarding state-of-the-art green cloud computing and categorized them into six areas. Additionally, detailed descriptions and statistics of methods, metrics, tools, and datasets were extracted and presented. It is important to note that all categorizations and provided statistics in this research are based on the subset of papers selected for review in this study.

Author Contributions: Conceptualization, A.P. and E.A.; methodology, A.P. and E.A.; formal analysis, E.A.; investigation, A.P.; resources, A.P. and E.A.; data curation, A.P.; writing—original draft preparation, A.P. and P.G.; writing—review and editing, P.G. and E.A.; visualization, A.P.; supervision, E.A.; project administration, A.P.; funding acquisition, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No data were used in this manuscript.

Acknowledgments: The authors must thank the ICT Research Institute for its financial support during the research. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AWS	Amazon Web Service
CDC	Cloud Data Center
CPU	Central Processing Unit
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
EC2	Elastic Compute Cloud
GCP	Google Cloud Platform
GPU	Graphics Processing Unit
IaaS	Infrastructure-as-a-Service
IoT	Internet of Things
LSTM	Long Short-Term Memory
MEC	Mobile Edge Computing
PaaS	Platform-as-a-Service
PAYG	Pay-as-You-Go
PM	Physical Machine
PUE	Power Usage Effectiveness
QoS	Quality of Service
RA/RS	Resource Allocation/Resource Scheduling
SaaS	Software-as-a-Service
SLA	Service-Level Agreement
SLAV	Service-Level Agreement Violation
TCO	Total Cost of Ownership
TPU	Tensor Processing Unit
vCPU	Virtual CPU
VM	Virtual Machine

References

1. Mohamed, H.; Alkabani, Y.; Selmy, H. Energy Efficient Resource Management for Cloud Computing Environment. In *Proceedings of the 2014 9th International Conference on Computer Engineering & Systems (ICCES), Cairo, 22–23 December 2014*; IEEE: New York, NY, USA, 2014. [CrossRef]
2. Helali, L.; Omri, M.N. A survey of data center consolidation in cloud computing systems. *Comput. Sci. Rev.* **2021**, *39*, 100366. [CrossRef]
3. Gholipour, N.; Arianyan, E.; Buyya, R. Recent Advances in Energy-Efficient Resource Management Techniques in Cloud Computing Environments. In *New Frontiers in Cloud Computing and Internet of Things*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 31–68. [CrossRef]
4. Helali, L.; Omri, M.N. Intelligent and compliant dynamic software license consolidation in cloud environment. *Computing* **2022**, *104*, 2749–2783. [CrossRef]
5. Zeng, J.; Ding, D.; Kang, K.; Xie, H.; Yin, Q. Adaptive DRL-Based Virtual Machine Consolidation in Energy-Efficient Cloud Data Center. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 2991–3002. [CrossRef]
6. Zubair, A.A.; Razak, S.A.; Ngadi, A.; Al-Dhaqm, A.; Yafooz, W.M.S.; Emara, A.-H.M.; Saad, A.; Al-Aqrabi, H. A Cloud Computing-Based Modified Symbiotic Organisms Search Algorithm (AI) for Optimal Task Scheduling. *Sensors* **2022**, *22*, 1674. [CrossRef] [PubMed]
7. Khan, T.; Tian, W.; Zhou, G.; Ilager, S.; Gong, M.; Buyya, R. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *J. Netw. Comput. Appl.* **2022**, *204*, 103405. [CrossRef]
8. Nayeri, Z.M.; Ghafarian, T.; Javadi, B. Application placement in Fog computing with AI approach: Taxonomy and a state of the art survey. *J. Netw. Comput. Appl.* **2021**, *185*, 103078. [CrossRef]
9. Goodarzy, S.; Nazari, M.; Han, R.; Keller, E.; Rozner, E. Resource Management in Cloud Computing Using Machine Learning: A Survey. In *Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020*; IEEE: New York, NY, USA, 2021; pp. 811–816.

10. Jayaprakash, S.; Nagarajan, M.D.; de Prado, R.P.; Subramanian, S.; Divakarachari, P.B. A Systematic Review of Energy Management Strategies for Resource Allocation in the Cloud: Clustering, Optimization and Machine Learning. *Energies* **2021**, *14*, 5322. [CrossRef]
11. Tsakalidou, V.N.; Mitsou, P.; Papakostas, G.A. Machine learning for cloud resources management—An overview. In *Computer Networks and Inventive Communication Technologies*; Springer: Berlin/Heidelberg, Germany, 2021.
12. Imran, M.; Ibrahim, M.; Din, M.S.U.; Rehman, M.A.U.; Kim, B.S. Live virtual machine migration: A survey, research challenges, and future directions. *Comput. Electr. Eng.* **2022**, *103*, 108297. [CrossRef]
13. Khallouli, W.; Huang, J. Cluster resource scheduling in cloud computing: Literature review and research challenges. *J. Supercomput.* **2022**, *78*, 6898–6943. [CrossRef]
14. Djigal, H.; Xu, J.; Liu, L.; Zhang, Y. Machine and Deep Learning for Resource Allocation in Multi-Access Edge Computing: A Survey. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 2449–2494. [CrossRef]
15. Masdari, M.; Khoshnevis, A. A survey and classification of the workload forecasting methods in cloud computing. *Clust. Comput.* **2020**, *23*, 2399–2424. [CrossRef]
16. Aqib, M.; Kumar, D.; Tripathi, S. Machine Learning for Fog Computing: Review, Opportunities and a Fog Application Classifier and Scheduler. *Wirel. Pers. Commun.* **2022**, *129*, 853–880. [CrossRef]
17. Singh, S.; Singh, P.; Tanwar, S. A comprehensive review and open issues on energy aware resource allocation in cloud. *Int. J. Web Eng. Technol.* **2022**, *17*, 353. [CrossRef]
18. Bachiega, J.; Costa, B.; Carvalho, L.R.; Rosa, M.J.F.; Araujo, A. Computational Resource Allocation in Fog Computing: A Comprehensive Survey. *ACM Comput. Surv.* **2023**, *55*, 181. [CrossRef]
19. Hong, C.H.; Varghese, B. Resource management in fog/edge computing: A survey on architectures, infrastructure, and algorithms. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–37. [CrossRef]
20. Mutluturk, M.; Kor, B.; Metin, B. The role of edge/fog computing security in IoT and industry 4.0 infrastructures. In *Research Anthology on Edge Computing Protocols, Applications, and Integration*; IGI Global: Hershey, PA, USA, 2022; pp. 468–479.
21. Dai, B.; Niu, J.; Ren, T.; Atiquzzaman, M. Toward Mobility-Aware Computation Offloading and Resource Allocation in End–Edge–Cloud Orchestrated Computing. *IEEE Internet Things J.* **2022**, *9*, 19450–19462. [CrossRef]
22. Alqerm, I.; Pan, J. DeepEdge: A New QoE-Based Resource Allocation Framework Using Deep Reinforcement Learning for Future Heterogeneous Edge-IoT Applications. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 3942–3954. [CrossRef]
23. Cheng, M.; Li, J.; Nazarian, S. DRL-cloud: Deep reinforcement learning-based resoaurce provisioning and task scheduling for cloud service providers. In *Proceedings of the 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jeju, Republic of Korea, 22–25 January 2018*; IEEE: New York, NY, USA, 2018; pp. 129–134. [CrossRef]
24. Tuli, S.; Gill, S.S.; Xu, M.; Garraghan, P.; Bahsoon, R.; Dustdar, S.; Sakellariou, R.; Rana, O.; Buyya, R.; Casale, G.; et al. HUNTER: AI based holistic resource management for sustainable cloud computing. *J. Syst. Softw.* **2022**, *184*, 111124. [CrossRef]
25. Mahmoud, H.; Thabet, M.; Khafagy, M.H.; Omara, F.A. Multi-objective Task Scheduling in Cloud Environment Using Decision Tree Algorithm. *IEEE Access* **2022**, *10*, 36140–36151. [CrossRef]
26. Jena, S.; Sahu, L.K.; Mishra, S.K.; Sahoo, B. VM Consolidation based on Overload Detection and VM Selection Policy. In *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021*; IEEE: New York, NY, USA, 2021; pp. 252–256. [CrossRef]
27. Kakolyris, A.K.; Katsaragakis, M.; Masouros, D.; Soudris, D. RoaD-RuNNer: Collaborative DNN partitioning and offloading on heterogeneous edge systems. In *Proceedings of the 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 17–19 April 2023*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
28. Yunmeng, D.; Xu, G.; Zhang, M.; Meng, X. A High-Efficient Joint 'Cloud-Edge' Aware Strategy for Task Deployment and Load Balancing. *IEEE Access* **2021**, *9*, 12791–12802. [CrossRef]
29. Ma, X.; Xu, H.; Gao, H.; Bian, M.; Hussain, W. Real-Time Virtual Machine Scheduling in Industry IoT Network: A Reinforcement Learning Method. *IEEE Trans. Ind. Inform.* **2022**, *19*, 2129–2139. [CrossRef]
30. Zhang, W.; Zhang, G.; Mao, S. Deep Reinforcement Learning Based Joint Caching and Resources Allocation for Cooperative MEC. *IEEE Internet Things J.* **2023**, *11*, 12203–12215. [CrossRef]
31. Hu, Y.; Ding, D.; Kang, K.; Li, T. Adaptive Multi-Threshold Energy-Aware Virtual Machine Consolidation in Cloud Data Center. In *Proceedings of the 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), Beijing, China, 28–30 October 2019*; IEEE: New York, NY, USA, 2020; pp. 1–6. [CrossRef]
32. Meyer, V.; Kirchoff, D.F.; da Silva, M.L.; De Rose, C.A. An Interference-Aware Application Classifier Based on Machine Learning to Improve Scheduling in Clouds. In *Proceedings of the 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), Västerås, Sweden, 11–13 March 2020*; IEEE: New York, NY, USA, 2020; pp. 80–87. [CrossRef]
33. Nemati, H.; Azhari, S.V.; Shakeri, M.; Dagenais, M. Host-Based Virtual Machine Workload Characterization Using Hypervisor Trace Mining. *ACM Trans. Model. Perform. Eval. Comput. Syst.* **2021**, *6*, 197. [CrossRef]

34. Wang, H.; Pannereselvam, J.; Liu, L.; Lu, Y.; Zhai, X.; Ali, H. Cloud Workload Analytics for Real-Time Prediction of User Request Patterns. In *Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, UK, 28–30 June 2018*; IEEE: New York, NY, USA, 2019; pp. 1677–1684. [CrossRef]
35. Daraghmeh, M.; Melhem, S.B.; Agarwal, A.; Goel, N.; Zaman, M. Linear and Logistic Regression Based Monitoring for Resource Management in Cloud Networks. In *Proceedings of the 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), Barcelona, Spain, 6–8 August 2018*; IEEE: New York, NY, USA, 2018; pp. 259–266. [CrossRef]
36. Banerjee, S.; Roy, S.; Khatua, S. Efficient resource utilization using multi-step-ahead workload prediction technique in cloud. *J. Supercomput.* **2021**, *77*, 10636–10663. [CrossRef]
37. Bi, J.; Li, S.; Yuan, H.; Zhou, M. Integrated deep learning method for workload and resource prediction in cloud systems. *Neurocomputing* **2020**, *424*, 35–48. [CrossRef]
38. Saxena, D.; Singh, A.K. A High Availability Management Model Based on VM Significance Ranking and Resource Estimation for Cloud Applications. *IEEE Trans. Serv. Comput.* **2023**, *16*, 1604–1615. [CrossRef]
39. Baktir, A.C.; Kulahoglu, Y.C.; Erbay, O.; Metin, B. Server virtualization in ICT infrastructure in Turkey. In *Proceedings of the 2013 21st Telecommunications Forum Telfor (TELFOR)*; IEEE: New York, NY, USA, 2013; pp. 13–16.
40. Taiepisi, J. Exploring the Effects of the Implementation of Server Virtualization in Small to Medium-Sized Organizations: A Qualitative Study. Ph.D. Thesis, Colorado Technical University, Colorado Springs, CO, USA, 2019.
41. Zhang, H.; Wang, J.; Zhang, H.; Bu, C. Security computing resource allocation based on deep reinforcement learning in serverless multi-cloud edge computing. *Futur. Gener. Comput. Syst.* **2024**, *151*, 152–161. [CrossRef]
42. Singhal, S.; Gupta, N.; Berwal, P.; Naveed, Q.N.H.; Lasisi, A.; Wodajo, A.W. Energy Efficient Resource Allocation in Cloud Environment Using Metaheuristic Algorithm. *IEEE Access* **2023**, *11*, 126135–126146. [CrossRef]
43. Asghari, A.; Sohrabi, M.K. Combined use of coral reefs optimization and reinforcement learning for improving resource utilization and load balancing in cloud environments. *Computing* **2021**, *103*, 1545–1567. [CrossRef]
44. Javaheri, S.; Ghaemi, R.; Naeen, H.M. An autonomous architecture based on reinforcement deep neural network for resource allocation in cloud computing. *Computing* **2023**, *103*, 1545–1567. [CrossRef]
45. Jeong, B.; Baek, S.; Park, S.; Jeon, J.; Jeong, Y.-S. Stable and efficient resource management using deep neural network on cloud computing. *Neurocomputing* **2023**, *521*, 99–112. [CrossRef]
46. Arasan, K.K.; Anandhakumar, P. Energy-efficient task scheduling and resource management in a cloud environment using optimized hybrid technology. *Softw. Pract. Exp.* **2023**, *53*, 1572–1593. [CrossRef]
47. Karimi-Mamaghan, M.; Mohammadi, M.; Meyer, P.; Karimi-Mamaghan, A.M.; Talbi, E.-G. Machine learning at the service of meta-heuristics for solving combinatorial optimization problems: A state-of-the-art. *Eur. J. Oper. Res.* **2022**, *296*, 393–422. [CrossRef]
48. Pradhan, A.; Bisoy, S.K.; Kautish, S.; Jasser, M.B.; Mohamed, A.W. Intelligent Decision-Making of Load Balancing Using Deep Reinforcement Learning and Parallel PSO in Cloud Environment. *IEEE Access* **2022**, *10*, 76939–76952. [CrossRef]
49. Neelakantan, P.; Yadav, N.S. Proficient job scheduling in cloud computation using an optimized machine learning strategy. *Int. J. Inf. Technol.* **2023**, *15*, 2409–2421. [CrossRef]
50. Neelakantan, P.; Yadav, N.S. An Optimized Load Balancing Strategy for an Enhancement of Cloud Computing Environment. *Wirel. Pers. Commun.* **2023**, *131*, 1745–1765. [CrossRef]
51. Karat, C.; Senthilkumar, R. Optimal resource allocation with deep reinforcement learning and greedy adaptive firefly algorithm in cloud computing. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6657. [CrossRef]
52. Su, Y.; Fan, W.; Gao, L.; Qiao, L.; Liu, Y.; Wu, F. Joint DNN Partition and Resource Allocation Optimization for Energy-Constrained Hierarchical Edge-Cloud Systems. *IEEE Trans. Veh. Technol.* **2023**, *72*, 3930–3944. [CrossRef]
53. Fang, C.; Hu, Z.; Meng, X.; Tu, S.; Wang, Z.; Zeng, D.; Ni, W.; Guo, S.; Han, Z. DRL-Driven Joint Task Offloading and Resource Allocation for Energy-Efficient Content Delivery in Cloud-Edge Cooperation Networks. *IEEE Trans. Veh. Technol.* **2023**, *72*, 16195–16207. [CrossRef]
54. Ullah, F.; Bilal, M.; Yoon, S.-K. Intelligent time-series forecasting framework for non-linear dynamic workload and resource prediction in cloud. *Comput. Netw.* **2023**, *225*, 109653. [CrossRef]
55. Sangaiah, A.K.; Javadpour, A.; Pinto, P.; Rezaei, S.; Zhang, W. Enhanced resource allocation in distributed cloud using fuzzy meta-heuristics optimization. *Comput. Commun.* **2023**, *209*, 14–25. [CrossRef]
56. Iftikhar, S.; Ahmad, M.M.M.; Tuli, S.; Chowdhury, D.; Xu, M.; Gill, S.S.; Uhlig, S. HunterPlus: AI based energy-efficient task scheduling for cloud–fog computing environments. *Internet Things* **2023**, *21*, 100667. [CrossRef]
57. Zhu, R.; Li, G.; Wang, P.; Xu, M.; Yu, S. DRL-Based Deadline-Driven Advance Reservation Allocation in EONs for Cloud–Edge Computing. *IEEE Internet Things J.* **2022**, *9*, 21444–21457. [CrossRef]
58. Wang, D.; Song, B.; Lin, P.; Yu, F.R.; Du, X.; Guizani, M. Resource Management for Edge Intelligence (EI)-Assisted IoV Using Quantum-Inspired Reinforcement Learning. *IEEE Internet Things J.* **2022**, *9*, 12588–12600. [CrossRef]

59. Vijayasekaran, G.; Duraipandian, M. An Efficient Clustering and Deep Learning Based Resource Scheduling for Edge Computing to Integrate Cloud-IoT. *Wirel. Pers. Commun.* **2022**, *124*, 2029–2044. [CrossRef]
60. Meyer, V.; da Silva, M.L.; Kirchoff, D.F.; De Rose, C.A. IADA: A dynamic interference-aware cloud scheduling architecture for latency-sensitive workloads. *J. Syst. Softw.* **2022**, *194*, 111491. [CrossRef]
61. Lin, W.; Yao, K.; Zeng, L.; Liu, F.; Shan, C.; Hong, X. A GAN-based method for time-dependent cloud workload generation. *J. Parallel Distrib. Comput.* **2022**, *168*, 33–44. [CrossRef]
62. Chen, Z.; Hu, J.; Min, G.; Luo, C.; El-Ghazawi, T. Adaptive and Efficient Resource Allocation in Cloud Datacenters Using Actor-Critic Deep Reinforcement Learning. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 1911–1923. [CrossRef]
63. Kruekaew, B.; Kimpan, W. Multi-Objective Task Scheduling Optimization for Load Balancing in Cloud Computing Environment Using Hybrid Artificial Bee Colony Algorithm with Reinforcement Learning. *IEEE Access* **2022**, *10*, 17803–17818. [CrossRef]
64. Talwani, S.; Alhazmi, K.; Singla, J.; Alyamani, H.J.; Bashir, A.K. Allocation and Migration of Virtual Machines Using Machine Learning. *Comput. Mater. Contin.* **2022**, *70*, 3349–3364. [CrossRef]
65. Sham, E.; Vidyarthi, D.P. Intelligent admission control manager for fog-integrated cloud: A hybrid machine learning approach. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6687. [CrossRef]
66. Toka, L.; Dobreff, G.; Fodor, B.; Sonkoly, B. Machine Learning-Based Scaling Management for Kubernetes Edge Clusters. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 958–972. [CrossRef]
67. Jumnal, A.; Kumar, S.M.D. Optimal VM Placement Approach Using Fuzzy Reinforcement Learning for Cloud Data Centers. In *Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021*; IEEE: New York, NY, USA, 2021; pp. 29–35. [CrossRef]
68. Gholipour, N.; Shoeibi, N.; Arianyan, E. An Energy-Aware Dynamic Resource Management Technique Using Deep Q-Learning Algorithm and Joint VM and Container Consolidation Approach for Green Computing in Cloud Data Centers. In *Proceedings of the Distributed Computing and Artificial Intelligence, Special Sessions, 17th International Conference*; ResearchGate: Rome, Italy, 2021; pp. 227–233.
69. Swathy, R.; Vinayagasundaram, B. Bayes Theorem Based Virtual Machine Scheduling for Optimal Energy Consumption. *Comput. Syst. Sci. Eng.* **2022**, *43*, 159–174. [CrossRef]
70. Li, J.; Zhang, X.; Wei, Z.; Wei, J.; Ji, Z. Energy-aware task scheduling optimization with deep reinforcement learning for large-scale heterogeneous systems. *CCF Trans. High Perform. Comput.* **2021**, *3*, 383–392. [CrossRef]
71. Caviglione, L.; Gaggero, M.; Paolucci, M.; Ronco, R. Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters. *Soft Comput.* **2021**, *25*, 12569–12588. [CrossRef]
72. Aliyu, M.; Murali, M.; Zhang, Z.J.; Gital, A.; Boukari, S.; Huang, Y.; Yakubu, I.Z. Management of cloud resources and social change in a multi-tier environment: A novel finite automata using ant colony optimization with spanning tree. *Technol. Forecast. Soc. Change* **2021**, *166*, 120591. [CrossRef]
73. Baburao, D.; Pavankumar, T.; Prabhu, C.S.R. Load balancing in the fog nodes using particle swarm optimization-based enhanced dynamic resource allocation method. *Appl. Nanosci.* **2021**, *13*, 1045–1054. [CrossRef]
74. Reddy, K.H.K.; Luhach, A.K.; Pradhan, B.; Dash, J.K.; Roy, D.S. A genetic algorithm for energy efficient fog layer resource management in context-aware smart cities. *Sustain. Cities Soc.* **2020**, *63*, 102428. [CrossRef]
75. Rjoub, G.; Bentahar, J.; Wahab, O.A.; Bataineh, A.S. Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems. *Concurr. Comput. Pract. Exp.* **2020**, *33*, e5919. [CrossRef]
76. Nath, S.B.; Addya, S.K.; Chakraborty, S.; Ghosh, S.K. Green Containerized Service Consolidation in Cloud. In *Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020*; IEEE: New York, NY, USA, 2020; pp. 1–6. [CrossRef]
77. Peng, Z.; Lin, J.; Cui, D.; Li, Q.; He, J. A multi-objective trade-off framework for cloud resource scheduling based on the Deep Q-network algorithm. *Clust. Comput.* **2020**, *23*, 2753–2767. [CrossRef]
78. Moghaddam, S.M.; O'sUllivan, M.; Walker, C.; Piraghaj, S.F.; Unsworth, C.P. Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers. *Futur. Gener. Comput. Syst.* **2020**, *106*, 221–233. [CrossRef]
79. Shaw, R.; Howley, E.; Barrett, E. An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. *Simul. Model. Pract. Theory* **2019**, *93*, 322–342. [CrossRef]
80. Xu, Y.; Yao, J.; Jacobsen, H.-A.; Guan, H. Enabling Cloud Applications to Negotiate Multiple Resources in a Cost-Efficient Manner. *IEEE Trans. Serv. Comput.* **2018**, *14*, 413–425. [CrossRef]
81. Wei, Y.; Yu, F.R.; Song, M.; Han, Z. Joint Optimization of Caching, Computing, and Radio Resources for Fog-Enabled IoT Using Natural Actor-Critic Deep Reinforcement Learning. *IEEE Internet Things J.* **2018**, *6*, 2061–2073. [CrossRef]
82. Ranjbari, M.; Torkestani, J.A. A learning automata-based algorithm for energy and SLA efficient consolidation of virtual machines in cloud data centers. *J. Parallel Distrib. Comput.* **2018**, *113*, 55–62. [CrossRef]

83. Shaw, R.; Howley, E.; Barrett, E. An Advanced Reinforcement Learning Approach for Energy-Aware Virtual Machine Consolidation in Cloud Data Centers. In *Proceedings of the 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), Cambridge, UK, 11–14 December 2017*; IEEE: New York, NY, USA, 2018. [CrossRef]
84. Pahlevan, A.; Qu, X.; Zapater, M.; Atienza, D. Integrating Heuristic and Machine-Learning Methods for Efficient Virtual Machine Allocation in Data Centers. *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **2017**, *37*, 1667–1680. [CrossRef]
85. Abu Zohair, L.M. Prediction of Student's performance by modelling small dataset size. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 27. [CrossRef]
86. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.H.; Konwinski, A.; Lee, G.; Patterson, D.A.; Rabkin, A.; Stoica, I.; et al. *Above the Clouds: A Berkeley View of Cloud Computing*; UC Berkeley: Berkeley, CA, USA, 2009.
87. Greenberg, A.; Hamilton, J.; Maltz, D.A.; Patel, P. *The Cost of a Cloud: Research Problems in Data Center Networks*; ACM SIGCOMM: Barcelona, Spain, 2009.
88. Mehor, Y.; Rebbah, M.; Smail, O. *Energy-Aware Task Scheduling and Resource Allocation in Cloud Computing*; Atlantis Press: Paris, France, 2025.
89. Breukelman, E.; Hall, S.; Belgioioso, G.; Dörfler, F. Carbon-Aware Computing in a Network of Data Centers. In *Proceedings of the 2024 European Control Conference (ECC), Stockholm, Sweden, 25–28 June 2024*; IEEE: New York, NY, USA, 2024.
90. Amahrouch, A.; Saadi, Y.; El Kafhali, S. Optimizing Energy Efficiency in Cloud Data Centers: A Reinforcement Learning-Based Virtual Machine Placement Strategy. *Network* **2025**, *5*, 17. [CrossRef]
91. EdinatA. Cloud Computing Pricing Models: A Survey. *Int. J. Sci. Eng. Res.* **2019**, *6*, 22–26. [CrossRef]
92. APPTIO. FinOps: A New Approach to Cloud Financial Management. 2019. Available online: https://www.apptio.com/resources/ebooks/finops-new-approach-cloud-financial-management/ (accessed on 22 January 2026).
93. Fragiadakis, G.; Tsadimas, A.; Filiopoulou, E.; Kousiouris, G.; Michalakelis, C.; Nikolaidou, M. Cloud PricingOps: A Decision Support Framework to Explore Pricing Policies of Cloud Services. *Appl. Sci.* **2024**, *14*, 11946. [CrossRef]
94. Dan, C. *Marinescu, Cloud Computing: Theory and Practice*, 3rd ed.; Newnes: Newton, MA, USA, 2023.
95. Amazon Web Services. *AWS Customer Carbon Footprint Tool: Methodology*; Amazon Web Services: Seattle, WA, USA, 2023.
96. Alsharif, M.H.; Jahid, A.; Kannadasan, R.; Singla, M.K.; Gupta, J.; Nisar, K.S.; Abdel-Aty, A.-H.; Kim, M.-K. Survey of energy-efficient fog computing: Techniques and recent advances. *Energy Rep.* **2025**, *13*, 1739–1763. [CrossRef]
97. Mijuskovic, A.; Chiumento, A.; Bemthuis, R.; Aldea, A.; Havinga, P. Resource Management Techniques for Cloud/Fog and Edge Computing: An Evaluation Framework and Classification. *Sensors* **2021**, *21*, 1832. [CrossRef] [PubMed]
98. Alomari, A.; Subramaniam, S.K.; Samian, N.; Latip, R.; Zukarnain, Z. Resource Management in SDN-Based Cloud and SDN-Based Fog Computing: Taxonomy Study. *Symmetry* **2021**, *13*, 734. [CrossRef]
99. Mustafa, S.; Nazir, B.; Hayat, A.; Khan, A.U.R.; Madani, S.A. Resource management in cloud computing: Taxonomy, prospects, and challenges. *Comput. Electr. Eng.* **2015**, *47*, 186–203. [CrossRef]
100. Shekhar, A.; Aleem, A. Improving Energy Efficiency through Green Cloud Computing in IoT Networks. In *Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 9–10 February 2024*; IEEE: New York, NY, USA, 2024; pp. 929–933. [CrossRef]
101. Prasad, V.K.; Dansana, D.; Bhavsar, M.D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A. Efficient Resource Utilization in IoT and Cloud Computing. *Information* **2023**, *14*, 619. [CrossRef]
102. Bhandari, K.S.; Cho, G.H. An Energy Efficient Routing Approach for Cloud-Assisted Green Industrial IoT Networks. *Sustainability* **2020**, *12*, 7358. [CrossRef]
103. Lin, W.; Lin, J.; Peng, Z.; Huang, H.; Lin, W.; Li, K. A systematic review of green-aware management techniques for sustainable data center. *Sustain. Comput. Inform. Syst.* **2024**, *42*, 100989. [CrossRef]
104. Mondal, S.; Faruk, F.B.; Rajbongshi, D.; Efaz, M.M.K.; Islam, M.M. GEECO: Green Data Centers for Energy Optimization and Carbon Footprint Reduction. *Sustainability* **2023**, *15*, 15249. [CrossRef]
105. Murino, T.; Monaco, R.; Nielsen, P.S.; Liu, X.; Esposito, G.; Scognamiglio, C. Sustainable Energy Data Centres: A Holistic Conceptual Framework for Design and Operations. *Energies* **2023**, *16*, 5764. [CrossRef]
106. Luo, J.; Li, H.; Liu, J. How Green Data Center Establishment Drives Carbon Emission Reduction: Double-Edged Sword or Equilibrium Effect? *Sustainability* **2025**, *17*, 6598. [CrossRef]
107. Zhang, L.; Zhao, Z.; Chen, B.; Zhao, M.; Chen, Y. Zero-Carbon Development in Data Centers Using Waste Heat Recovery Technology: A Systematic Review. *Sustainability* **2025**, *17*, 10101. [CrossRef]
108. Asadi, S.K.M. Green Computing and Its Role in Reducing Cost of the Modren Industrial Product. *Acad. Account. Financ. Stud. J.* **2021**, *25*, 1–13.
109. Yu, X.; Mi, J.; Tang, L.; Long, L.; Qin, X. Dynamic multi objective task scheduling in cloud computing using reinforcement learning for energy and cost optimization. *Sci. Rep.* **2025**, *15*, 45387. [CrossRef] [PubMed]
110. Ji, K.; Zhang, F.; Chi, C.; Song, P.; Zhou, B.; Marahatta, A.; Liu, Z. A joint energy efficiency optimization scheme based on marginal cost and workload prediction in data centers. *Sustain. Comput. Inform. Syst.* **2021**, *32*, 100596. [CrossRef]

111. Zhang, S.; Yuan, D.; Pan, L.; Liu, S.; Cui, L.; Meng, X. Selling Reserved Instances through Pay-as-You-Go Model in Cloud Computing. In *Proceedings of the 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, USA, 25–30 June 2017*; IEEE: New York, NY, USA, 2017; pp. 130–137. [CrossRef]
112. Bandyopadhyay, A.; Choudhary, U.; Tiwari, V.; Mukherjee, K.; Turjya, S.M.; Ahmad, N.; Haleem, A.; Mallik, S. Quantum Game Theory-Based Cloud Resource Allocation: A Novel Approach. *Mathematics* **2025**, *13*, 1392. [CrossRef]
113. Brahmi, Z.; Derouiche, R. EQGSA-DPW: A Quantum-GSA Algorithm-Based Data Placement for Scientific Workflow in Cloud Computing Environment. *J. Grid. Comput.* **2024**, *22*, 57. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.