Cluster Computing (2023) 26:3209–3235
https://doi.org/10.1007/s10586-023-04024-8

Prediction-based scheduling techniques for cloud data center's workload: a systematic review

Shobhana Kashyap[1] · Avtar Singh[1]

Received: 14 January 2023 / Revised: 27 March 2023 / Accepted: 30 April 2023 / Published online: 18 May 2023

Abstract
A cloud data center provides various facilities such as storage, data accessibility, and running many specific applications on cloud resources. The unpredictable demand for service requests in cloud workloads affects the availability of resources during scheduling. It raises the issues of inaccurate workload prediction, lack of fulfillment in resource demands, load unbalancing, high power consumption due to heavy loads, and problems of under and overutilization of resources. Therefore, an efficient scheduling technique and an accurate forecasting model are needed to overcome these issues. Also, to deal with these challenges and provide optimal solutions, researchers must have a robust knowledge of cloud workloads, their types, issues, existing technologies, their advantages and disadvantages. However, previous research indicates limited systematic review studies exist for cloud workload applications with prediction-based scheduling techniques. Therefore, a survey is required that provides information related to cloud workload. To fulfill this requirement, the current study collects the related articles published in the past years. This paper is a systematic review study of prediction-based scheduling techniques that extract and evaluate data based on five criteria. It includes the datasets of different workload applications, resources, current prediction and scheduling techniques, and their related parameters. The survey is quite useful for academicians who want to select the problem and develop new techniques for issues related to cloud workload applications. It also gives an idea of existing approaches that are already implemented and employed.

Keywords Cloud workload · Data centers · Resources · Prediction-based scheduling · Resource allocation · Load balancing

1 Introduction

Nowadays, many traditional businesses are moving their workloads to Cloud Data Centers (CDCs) [1]. These businesses are experiencing the advantages offered by Cloud Service Providers (CSPs), like cost savings, reliability, performance improvement, scalability, and flexibility. CSPs run the CDC and provide pay-per-use access to computing resources [2, 3] as well as a variety of other services to customers. They operate data centers (DCs) in multiple geographic locations [4], allowing them to provide redundant infrastructure [5] and backup systems [6] to ensure that services remain available even in the event of a failure or outage [7]. Many processes exist to secure data and applications from unauthorized access [6] and other security risks [8] inside the CDCs. These include firewalls [9], access controls [10], encryption [11], and intrusion detection/prevention systems [12]. These advantages have led to a meteoric rise in the number of CSPs and the range of services accessible in the cloud [13].

This rapid growth generates problems like fluctuating demand for resources [14], a lack of Quality of Service (QoS) [15], an uneven distribution of work [16–18], energy efficiency [19, 20], and many others. The dynamism and unpredictability of the resource's demand influence their accessibility during scheduling [21]. Hence, effective resource management is required so that work can be planned according to their execution demands [22, 23]. The QoS is the assessment of an entire service's performance. It

✉ Shobhana Kashyap
shobhanak.cs.19@nitj.ac.in

Avtar Singh
avtars@nitj.ac.in

[1] Department of CSE, Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India

is directly affected by the increase in service requests and the exponential growth of cloud users. To tackle these challenges, efficient prediction-based scheduling is required [24, 25]. Irregular load distribution across Virtual Machines (VMs) leads to inefficient utilization of resources, which affects the scheduling. A prediction-based scheduling scheme that handles the unbalancing of data is required [26, 27].

The scheduling of resources in a Cloud Workload (CW) is not only one of the fundamental difficulties in computing, but it also impacts the primary interests of cloud regulars and facility providers [28]. In fact, this issue is an NP-hard multi-constraint and multi-objective optimization problem [29, 30]. This signifies that finding the optimal solution is very difficult. To overcome these challenges, many researchers are working on integration of prediction methods with heuristic, meta-heuristic and hybrid algorithms [8, 27, 31]. The prediction method helps to estimate the future workload placed on a data center and provides the possible resource consumption patterns. And these patterns provide us appropriate resources and a significant step to building a resource-efficient scheduling method. Therefore, accurate forecasting is essential for preventing performance decline and reducing resource wastage, both of which improve revenue [32]. In CW, managing resources properly allows activities to be scheduled based on execution situations [23].

Many business sectors currently rely on cloud-based workloads [33, 34] and here optimizing resource consumption [33, 35, 36], is an essential component in this context. Hence, it is vital to understand which workloads are appropriate for customers and CSPs to make optimal use of available resources. Next, the delivery of computer resources (such as CPU, memory, network, virtualized servers, etc.) via the internet using cloud computing has become the trend [25, 37, 38]. For the purpose of offering these resources as a service in the cloud market, the CSPs have their own optimization goals. These objectives include limiting maintenance costs and boosting the money produced by underlying computer resources. Furthermore, in order to achieve these objectives, individual resource utilization needs to be predicted and the demanding resources must be deployed properly in the operational environment. However, in this perspective, workload prediction gives more definite or deterministic future knowledge about resource demands. Also, it allows for making appropriate choices to optimize resources in cloud data centers.

The above paragraphs conclude that accurate prediction and optimum scheduling are the most important aspects in achieving the optimization goal. A lot of researches [23, 39–47] have been done to work on these two techniques that help cloud users and providers to make better decisions for accurate load distribution. Also, researchers need to have a complete knowledge of workload and techniques that helps to enhance its performance. In the literature, various intelligent methods and algorithms are reviewed for the selection of best techniques from a large pool of options. Various researchers come up with different ways to solve the problem, and each suggests that their approach is the best. For further clarity, in this survey a Systematic Review Study (SRS) has been conducted. The current research aim is to compares several existing methods for making decisions to choose CW and its suitable techniques. The work also lists the pros and cons of each method and future research directions. Consequently, our research examines the robust specifications of several Cloud Workload Applications (CWAs). This study discusses the theoretical background, current methodologies, and performance metrics of the research topic. This research is anticipated to benefit both scholars and business professionals.

1.1 Motivation

In the present time, many researchers are continuously working to find an optimal solution for their respective fields and coming up with new ideas. For CW, the approaches such as scheduling and prediction plays very important role by ensuring that resources are efficiently allocated to execution requirements of cloud users [25, 40, 48, 49]. Also, these techniques helps to solve the problem of resource allocation [50], inaccurate prediction [51], improper load distribution [52], also improvement in cost [34] and performance optimization of the system [53].

An example is discussed considering a CWDC having multiple virtual machines (VMs) that execute different workloads. The CWDC is equipped with the following resources: 64 GB of RAM, 16 CPU cores, 1 TB of storage, and 1 Gbps network bandwidth. Now, let's suppose there are four VMs running in the DC, and each of them needs the following resources:

VM1: RAM = 16 GB, CORE = 4, STORAGE = 250 GB, 200 Mbps network bandwidth.
VM2: RAM = 8 GB, CORE = 2, STORAGE = 100 GB, 50 Mbps network bandwidth.
VM3: RAM = 32 GB, CORE = 6, STORAGE = 500 GB, 500 Mbps network bandwidth.
VM4: RAM = 8 GB, CORE = 2, STORAGE = 150 GB, 250 Mbps network bandwidth.

To allocate resources to these VMs, CW can use a resource management mechanism such as a hypervisor or container orchestrator. It can be configured to optimize efficiency, reliability, and affordability while allocating resources. Here scheduling technique is required for allocation of resources to the particular task and prediction helps to predict the resources to improve system performance. In the above example during allocation process, if VM1 and VM3 are given the resources they need to perform their workloads, while VM2 and VM4 are given fewer resources as they have lower resource requirements. This allocation also ensures that there is enough network bandwidth available for all VMs to communicate with each other and the outside world. Therefore, resource allocation can be a complex process in a CWDC, especially when there are many VMs running different workloads with varying resource requirements. A good resource management technique such as prediction-based scheduling can help to ensure that resources are allocated in a way that optimizes performance, availability, and cost.

From the previous studies found that limited surveys are conducted for CWAs. Also, there is a lack of information related to its dataset, resources, prediction techniques, scheduling algorithms, and their associated factors. Therefore, this study addresses challenges and motivates us to conduct the survey. The current work covers information related to similar types of CWA and details using various approaches with their appropriate solutions. It also includes the research questions related to the CW. In existing study, prediction models and scheduling techniques have been implemented in different applications of the cloud for proper utilization of resources. Hence the current research aims to address the requirement for a SRS for prediction-based scheduling techniques in CW.

1.2 Contribution

The contribution of current SRS is mentioned as follows:

- Review existing survey papers related to prediction-based scheduling techniques for CW and highlighted their issues.Provides research questions related to the problem and gives appropriate solutions.
- Various criteria for extraction and evaluation of data in CW are discussed and analyzed.
- Selected workload is analyzed for prediction and scheduling-based strategies with associated metrics.
- The research challenges are discussed with future directions in prediction-based scheduling techniques.

The remaining parts of this research are organized into the following categories: The methodology of the study is discussed in Sect. 2. The description of the theoretical background of the topic can be found in Sect. 3. The research questions are answered in Sect. 4 by doing an analysis of the chosen publications, and the material that is necessary for this study is supplied. The research problems and possible future directions are presented in Sect. 5.

2 Research methodology

The purpose of a systematic literature review (SLR) is to search, examine, and draw conclusions from all published works in a certain field of study. The term SRS is a kind of SLR in which latest research material is compiled and organized to provide a comprehensive description of a particular domain through a uniform and efficient research methodology. In this survey, the objective of SRS is to gather and evaluate past research on prediction-based scheduling techniques in CW. Inspired by survey paper [21], the Research Questions (RQ) are formulated at the beginning to serve as a framework for the creation of search strings. An appropriate examination of digital libraries is performed based on a search string to provide an answer to the RQ stated in the first step.

2.1 Research questions (RQ)

In SRS, preparation of RQ is one of the most essential steps, as research findings are examined in the consideration of these questions. This study is intended to answer some RQ. To obtain clarification, a comprehensive literature review has been conducted. This section is devoted to identifying six survey questions. The aim is to find answers to these RQ by examining related papers. The RQs generated with the help of SRS are given below:

RQ1: How to identify dataset and appropriate resources related to CW? (Answered in Sect. 4.1)
RQ2: What are the existing studies that address the issues of CW using predictive models and also mention its parameters? (Answered in Sect. 4.2)
RQ3: What are the current scheduling techniques and parameters that are working to increase performance of the system? (Answered in Sect. 4.3)
RQ4: What are the recent works of prediction-based scheduling related to CWs? (Answered in Sect. 4.4)

2.2 Article search strategy

This section describes the SRS article search strategy. The criteria for articles selection is choosing those studies which are related to CW. Finding relevant studies in digital libraries serves as a preliminary step in conducting a systematic review of the available literature. In SRS, the search technique is essential and has an impact on general performance. This investigation searches published publications from 2015 to the present in two phases: generalized and targeted. As illustrated in Fig. 1, the article search strategy is separated into three stages.

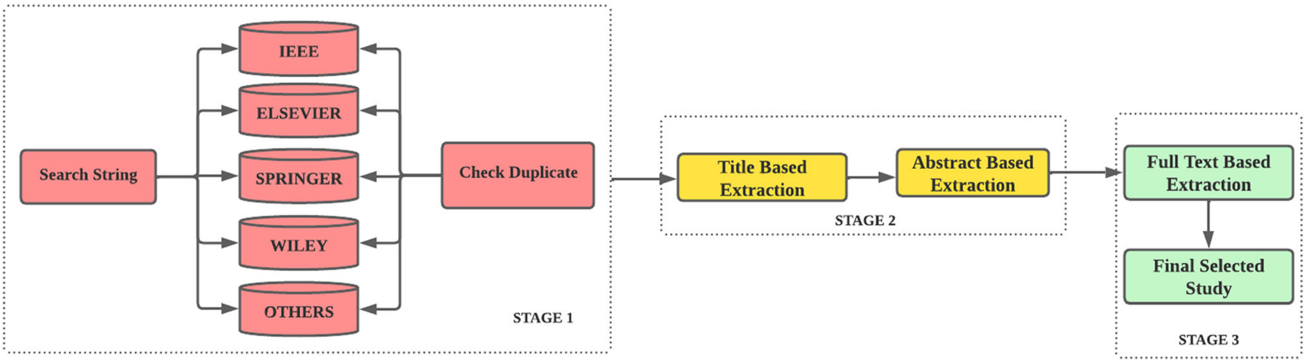

Fig. 1 Stages of article search strategy

2.2.1 STAGE 1: Start searching by entering relevant strings and keywords

The search for strings and keywords is carried out in two stages: generalized and targeted. Google Scholar is used to identifying papers pertinent to the study issue throughout the generalized research phase. The targeted phase identified relevant research articles in four digital libraries (IEEEXplore, SpringerLink, Wiley, and Elsevier). The main actions are carried out during this phase:

Step 1: Building strings and keywords from RQ is the initial stage in the article search strategy. Strings are generated by concatenating the keywords workload prediction, cloud scheduling, resource allocation, resource management, prediction-based scheduling, cloud resources, load balancing, resource overutilization, resource underutilization, and quality of service.

Step 2: Second, we performed both generalized and targeted searches for keywords and strings to identify relevant articles. The search returned 15,639 journal, book, conference paper, note, and chapter, etc. publications that included the previously defined keywords or phrases.

2.2.2 STAGE 2: Eliminating unnecessary and redundant contents

Considering problems related to research, only called high-research articles for our analysis. Also, we place a significant amount of importance on works that have been published in English and have appeared in reputable journals or conferences. Furthermore, unpublished work, reports, publication notes, and book articles are not added in the SRS. Cloud and similar principles include papers selected based on their titles in prestigious journals and conferences that deal with prediction-based scheduling. The following actions are performed during this phase:

Step 3: Third, we filter out results that have already been chosen in the first two steps (generalized and targeted). (6478 papers removed)

Step 4: The fourth step involves sorting the chosen publications by their titles and removing those that don't belong. Finally, it is concluded that 8024 studies have been collected in Stage 2.

2.2.3 STAGE 3: Paper selection based on quality and relevance of data

At this step, the entire texts and abstracts of the chosen papers were reviewed to determine their relevance to the RQ. The participation of each paper was determined by the research problem's relevance and the year of publication. For additional analysis, the abstracts and keywords of the chosen publications were entered into spreadsheets.

Step 5: At this phase, papers are rejected if their abstracts don't have any relevance to the subject under study. (Removed 3997 articles).

Step 6: In this stage, the author studied the whole text of each article. The rejection of articles that failed to address the RQ (described in Section 2.1). (Excluded 3886 articles).

In conclusion, the selected papers are submitted to three phases of review, and only relevant research studies capable of addressing the research questions are selected for further examination. Through this procedure, 15,497 publications were eliminated. In the end, 141 research papers were chosen for the SRS from a total of 15,639 research articles.

3 Theoretical background

This section provides the information related to CW which includes definition, importance, different reasons to adopt CW, types, and challenges.

3.1 Cloud workload (CW)

It is refers to the total amount of work to be performed, whether it in real-time by users interacting with cloud services, or in batches [54]. Also it is a collection of application resources that support a common business goal with multiple services, such as data stores and APIs, functioning together to deliver exact end-to-end functionality [55].

3.2 Importance of cloud workload

Today, a lot of businesses are transferring traditional workload to the cloud [8]. The Cloud Security Alliance (CSA) published its study on January 11, 2019, which examined the condition of cloud adoption and the effect of the cloud on enterprise resource planning applications [56]. Almost 70% of organizations asked questions associated to move, are in the process of enterprise resource planning (ERP) data and analyzing the time of workloads to cloud environments, although most of them are concerned to move [39, 57]. Experts mentioned the main advantages of migration toward ERP schemes to cloud paradigms:

- Scalability: (65% of respondents) the primary advantage of cloud migration for new technologies.
- Lower cost: (61% of respondents) lower ownership costs are a significant benefit of cloud computing.
- Security: (49% of respondents) update and regular service upgrades are a strong reason for migrating to the cloud.

3.3 Reason to adopt cloud workload

In article [58] Gartner suggested some key points why enterprise workloads moving to the public cloud given below:

- Mobility: Mobile technologies and apps make remote work much easier. The adaptable cloud approach is an excellent choice for mobile solutions.
- Disaster recovery: Cloud-based disaster recovery is both cost-effective and safe. It also saves business costs and efforts of maintaining redundant production-quality infrastructure in a different location.
- Web conferencing: Due to the pandemic, this virtual meeting facilitator has become a significant operational role. With the varying networking bandwidth needs, large enterprise cloud providers can reliably supply videoconferencing solutions.
- Collaborative effort and information management: Collaboration is an effective way whereby persons collectively work on a common problem to achieve business benefit. This demonstrates the cloud's usefulness for business productivity software
- Remote workstation management and virtual desktops: A depend-able Virtual Desktop Infrastructure (VDI) is essential for allowing remote work. Cloud-based virtualization and Desktop as a Service (DaaS) are now commonplace, offering a scalable and secure alternative to traditional DC based solutions.

3.4 Challenges in cloud workload

With the advancements and development of cloud computing services, businesses confront a variety of cloud computing issues. The cloud provides businesses with advantages, but resource management is not a simple procedure. The dynamic workload may lead following issues with associated to cloud-hosted services. The points mentioned below are the hurdles that must be overcome to use the cloud benefits efficiently.

- Dynamic environment: The business environment is quickly changing that impact the market [23, 59, 60], businesses have to incorporate these changes and provide new ideas, solutions, and services to retain technology and new developments.
- Understanding of resource demands: It may be dangerous to distribute resources arbitrarily without first analyzing incoming requirements, identifying priorities, or considering the organization's goals [61, 62]. Before distributing resources, it is essential to have full awareness of coming demand and resource availability according to matching skill sets and responsibilities.
- Over-provisioning: It can be a host or computational node that has assigned unused computing resources namely CPU, memory, I/O, disk, or network at peak time [25, 63, 64].
- Under-provisioning: The issue of resource under-provisioning occurs when service gives out fewer resources than are needed, it won't be able to provide a good service to its users [51, 63, 64]. If there aren't enough resources for the website, it might seem slow or out of reach. People stop using the Web, which means the service provider loses customers.
- Oscillation: When oscillation (auto-scaling) is used, it leads to both over-provisioning and under-provisioning issues [65]. It is a combination of under-provisioning and over-provisioning.

- Inefficient prediction: This type of prediction leads to unnecessary outcomes. Inaccurate prediction in the cloud raises issues of resource allocation, load balancing, energy efficiency, and many more difficulties. When training models formulate predictions, the sequence establishes an order on the data that must be maintained. In general, prediction issues involve a sequence of data which raises prediction problems [26, 27].
- Efficient Scheduling: This type of scheduling working on these factors such as to maximize the quality [18, 32, 66], service provider productivity [67], and efficiency of the system [24, 68]. The objective of these three factors helps to explore many businesses with alternative methods. Inefficient scheduling can affect each of the three goals.

3.5 Types of cloud workload

Depending on the purpose of the application, various types of workloads are classified. They are:

3.5.1 Business-critical workload applications

These applications are required for survival and long-term activities, although its failure does not necessarily result in an instant disaster [61, 69]. During normal operations, enterprises are using a broad range of apps, but not all of them are necessary for immediate survival during power failures and other disasters. Failure of a business-critical application suffers from decreased user productivity and a worse overall experience for those using the application [39]. Although, the organization is required to perform basic operation for limited hours, without causing serious harm to operations and profits. In general, the organization may continue the task with current resources or through alternative ways to the unsuccessful system. Most of the time, the applications on the list below are considered business-critical, but this can be changed.

- *Financial applications* It gives a business a way to handle money and information [8, 70]. Banks offer a wide range of financial services, each of which is made to meet a specific need. Each organization chooses the financial application that best fits its needs. Then, each application is categorized and given a priority based on how it affects the organization. Most organizations need a financial app to make sure they have a steady flow of money coming in, but how the organization gets paid is a key factor in how important these apps are. For instance, a short interruption of service might not have a big effect on a company that processes subscriptions once a month at a certain time. If an ecommerce site that needs to process purchases during the holidays goes down, even for a short time, it could lose a lot of money. Also, organizations must follow the rules when using financial applications to make sure that the transactions are safe and that sensitive and personal information stays private.
- *Messaging systems* It provides information between workers, business partners, customers, and other stakeholders. Organizations make use of a broad range of communications technologies, such as email programs, text messaging, and cross-functional platforms [71]. Messages often include critical information required for typical corporate operations. Private and sensitive information, proprietary data, and trade secrets may also be included in messages. All of these interactions and information exchanges are often necessary to keep routine corporate activities running. These systems also have security issues, for example, an Email system can put the firm at risk of security. If threat actors acquire access to email accounts, they may use them to steal information, trick people into disclosing information, use them as a gateway into the business network, and engage in other criminal actions.
- *Legacy systems* A legacy system is usually found for a long time within an organization's ecosystem [72]. It is a known methodology that has possibly been modified to meet the company's specific needs. These technologies come with a cost—the initial setup cost, followed by the cost of continuous maintenance. Because many legacy systems have not been designed to work in the cloud, they must be significantly updated when transferred. When legacy systems are designated as business-critical, they must be managed carefully to minimize interruptions. As a result, many businesses are still relying on old systems and are unwilling to shift to the cloud.

3.5.2 Mission-critical workload applications

In this workload gathering of application resources are described, which need to be extremely dependable on the platform [73]. The workload is consecutively available, quick recovery of failures, and quick operations. The organizations depend on instance operations based on mission-critical models and machines. With the condition mission critical resource experience downtime, further in brief, maybe a reason for the huge disruption and negative impacts are shown at once in the short and long term. Mission-critical machines and workloads deal with a maximum priority that needs to be continued to guarantee operations stay viable.

3.5.3 Low-priority workload applications

Organizations categorize applications as low-priority or non-critical, when they proceed with similar operations for a large duration without making use of the application [74]. This application highlights the organization suffering minimal consequences but alternatively performing the entire necessary work. These are frequently adopted because the system development simplifies operations and is highly productive.

3.5.4 Scientific computing workload applications

Scientific computing workload applications are software programs meant to do complex scientific computations and simulations using advanced mathematical models and methods [75–77]. These applications are often used in domains such as physics, chemistry, biology, engineering, finance, and weather forecasting, and need a huge amount of computing power. Solving systems of equations, numerical optimization, data analysis, and data visualization are often involved. They also need access to massive datasets and be able to effectively manage and modify enormous volumes of data.

- *High-performance computing applications* Generally, high-performance computing refers to the practice of grouping computing power in a single technology [78]. It provides considerably higher performance than other systems such as desktop computers or workstations to resolve large issues in the field of engineering, science or business [79]. It is the capacity to rapidly assimilate data and do complex computations.

In the above paragraphs, various types of applications are mentioned whose computing criteria are different from each other. The applications such as databases and web servers are referred to as business-critical workloads [31] and can be recovered from any losses. A site transfer is possible because the data for the business-critical job is mirrored. When determining what should be prioritized, it is common practice to use a relative measure that is established and relevant to the specific demands of the industry. For instance, one corporation may consider a communications system to be mission crucial, but another may consider it to be business-critical or even low priority.

4 Reporting of literature

This section provides the solution of RQ by analyzing the selected papers with their discussions to resolve the issues.

4.1 Identify dataset and appropriate resources for CW (RQ1)

In this section, datasets for CW from selected literature are analyzed. Next, the resources used in previous studies are examined based on their maximum utilization. The dataset and resources both are analyzed; this process will help the researcher for selection of dataset and choose appropriate resources for their study.

Earlier research has demonstrated that the dataset is generated either synthetic approach or real time approach. Workload generators are used to produce synthetic workloads, though real workloads are collected via benchmark datasets namely NASA dataset [80], Google Cluster Trace [33, 60], etc., or recovered from existent cloud platforms. To provide a better understanding, Table 1 lists the various datasets in relationship to the CW categories also provides its description.

Based on the data in the table, we can conclude that category of business-critical workload has many traces which comprise the same type of features. For example, Materna, Bitbrain, Alibaba, and Google Cluster have cloud resources-related features. Dynamic prediction and scheduling can be computed by integrating these datasets. The next category is a scientific workflow which is widely used in a variety of fields, including physics, chemistry, biology, and engineering. Last, the mission-critical workload is essential to the operation of a business, organization, or system.

Also, the description given by this section helps the researchers to have a clear picture of the problem area. The information provided here helps them in the initial phase of investigation and also gives direction to the designing phase. Table 1 helps the researchers to know which dataset falls under the category of CW.

Next, the survey analyzes cloud resources which are also an essential part in CW. The DC provides internet-based supply of computational resources, such as Artificial Intelligence (AI), processing power, storage, networking, databases, analytic, and software applications over the internet. By outsourcing these services, businesses may obtain computing assets on-demand basis without having to acquire and maintain an on-premise IT infrastructure. This allows and permits for adaptable resources, quick innovation, and fast scaling of economies. Table 2, presents the list of cloud resources this analysis helps the researcher finds the most appropriate resources for resolving the various issues in CW.

Table 2 concludes that both CPU and memory are highly occupied resources in previous studies. In future work, the researcher may include other resources to test and check the effect on the performance of the system.

Table 1 Cloud workload dataset description

References	Dataset	Workload type	Description
[20, 33, 81]	Bitbrain trace	Business-critical workload	• It is a trace of job submissions and resource usage in a large-scale distributed computing system • It was obtained from the Bitbrains grid over a period of many months • It is a commercial grid computing system that provides resources for computing and data processing applications
[66, 82]	Materna trace	Business-critical workload	• This dataset is a collection of job submissions and resource usage data from a large-scale computing system • Materna GmbH is German IT consulting company • The dataset was collected over a period of several months • Includes information about various types of jobs and workloads, like web, application, and database servers
[17, 33, 83, 84]	PlanetLab trace	Business-critical workload	• It is a network traffic dataset collected from the PlanetLab network • It includes traffic data from over 1000 PlanetLab nodes located in more than 500 institutions worldwide • The trace contains traffic data for a period of several years, from 2004 to 2009 • The data in the trace is captured using the NetFlow protocol, which allows for the collection of traffic statistics such as packet counts, byte counts, and protocol types • The trace includes both IPv4 and IPv6 traffic also including HTTP, BitTorrent, Skype, and DNS
[33, 60, 79]	Google cluster trace	Business-critical workload	• It is a trace of job submissions and resource usage in a large-scale computing cluster operated by Google • Data is related with huge amount of jobs submitted over a period of several months • Includes information about various aspects of the jobs and the resources used by them
[43, 49, 85–87]	Alibaba trace	Business-critical workload	• Dataset released by Alibaba in 2017 • Contains a one-month trace of the operations and resource usage of a large-scale cluster running on Alibaba's production environment • Dataset includes information about job submission, job scheduling, task execution, and resource allocation • Support research in the area of cluster computing
[74, 80, 83, 88]	HTTP trace	Business-critical workload	• It is a collection of HTTP traffic logs captured from various web servers and clients • The dataset includes information about HTTP requests and responses, as well as information about the timing and size of the requests and responses
[80, 88–90]	NASA trace	Business-critical workload	• Dataset is a set of logs of the job submissions and resource usages in a large computational cluster at the NASA Advanced Supercomputing (NAS) facility • Collected over a period of several months and contains information about approximately 2 million jobs • The dataset is widely used in research on scheduling, resource allocation, and workload characterization in large-scale distributed systems
[89]	Wikipedia clickstream	Business-critical workload	• This is a publicly available dataset • Contains information about user clickstreams on Wikipedia • It includes the sequence of requests made by users as they navigate through Wikipedia, and the corresponding Wikipedia pages that they visit
[5, 32, 76, 91, 92]	Cyber shake	Scientific computing workload	• It is a scientific application used to study the effects of earthquakes on critical infrastructure, such as buildings, bridges, and pipelines • The application uses high-performance computing resources to simulate how earthquakes propagate through the Earth's crust and how the resulting ground motions affect the built environment • Offers on-demand access to work out on resources that permit researchers to rapidly scale up needed simulations

Table 1 (continued)

References	Dataset	Workload type	Description
[44]	Floodplain	Scientific computing workload	• Floodplain modeling is an important scientific application that is used to simulate and predict the behavior of water in floodplains • It allows customers to use computational resources online. The combination of these two technologies can be used to create powerful floodplain modeling workflows that can help to improve flood management and response
[92]	Montage	Scientific computing workload	• Used in various cloud scientific applications to study different aspects of the universe • Used to train ML methods to classify different types of galaxies based on their morphology • Used to study the distribution of dark matter in the universe, to create high-resolution maps of the sky, and to analyze the structure and evolution of galaxies
[70, 93, 94]	Financial system	Mission-critical workload	• It is a key application area for mission critical workloads • Financial institutions such as banks, investment firms, and insurance companies rely heavily on their financial systems to process transactions, manage customer accounts, and analyze market data • Any disruption or downtime in these systems can result in significant financial losses, regulatory violations, and reputational damage
[95, 96]	Electronic health records (EHRs)	Mission-critical workload	• EHR systems are used by healthcare providers • Helps to manage patient information, medical histories, prescriptions, and other critical data • Interruptions in the functioning of these systems can result in patient safety issues, regulatory compliance violations, and legal liabilities
[97]	Industrial control systems (ICS)	Mission-critical workload	• Controlling and monitoring industrial processes that are essential for the functioning of critical infrastructure • These systems are used to manage processes such as power generation, water treatment, transportation systems, and chemical processing, among others • ICS systems are designed to be highly reliable and resilient, with redundancy built in to ensure that processes continue to function even if there is a hardware or software failure • To ensure that ICS systems are able to meet the demands of mission-critical workloads, they need to be designed with high availability and fault tolerance

4.2 Predictive models and their effective parameters (RQ2)

Workload prediction is an important for intelligent resource scaling and load balancing that maximizes cloud service provider's economic development and user's Quality of Experience (QoE) [18]. Predicting workload helps cloud-end cluster maintainers determine whether the resource allocation method is appropriate or not [85]. To properly manage cloud resources, it's essential to accurately forecast VM workload for resource provisioning [17]. Scheduling of resources is a set of rules and policies used to distribute jobs to appropriate resources (bandwidth, memory, and CPU) to maximize performance and resource usage. The proper planning of resources is a critical issue in cloud computing to improve the overall performance of the system. The present study describes the benefit of algorithms that are used most frequently in the past and suggests new algorithms as per the demand.

In a CDC load prediction is required for the allocation of resources that depend on demand applications [107]. According to [60] for efficient resource utilization in the cloud, ML approaches are used in the development that produces reliable prediction solutions. Many scholars focus on workload prediction and use different methodologies. Generally, there are two common methods of prediction: history-based and homeostatic. History-based models are easy and popular. Models examine past workloads to forecast future needs this technique involves prior workload patterns, whereas the homeostatic model uses the mean. The homeostatic prediction, estimate the future workload by adding or removing the present workload from the mean of prior workloads here values can be static or dynamic.

Table 2 Cloud resources description

References	Resources	Description
[39, 90, 98]	Cost	• It is a critical feature of CWs • It refers to the amount of money that an organization pays for the cloud resources they use • CSPs offer different pricing models and cost features to help organizations optimize their cloud costs • Cost features includes Pay-as-you-go pricing, Reserved instances, Spot instances, etc
[16, 19, 25, 60, 67, 99–103]	CPU	• It is primary computing resource in a VM or container • It is responsible for executing the instructions of a computer program and performing mathematical and logical operations • CSPs offer different CPU configurations to meet the varying needs of applications
[104]	Disk I/O	• It refers to the data transfer between an application and the disk storage • It can be a bottleneck for applications that rely heavily on disk access, such as databases or file storage applications • CSPs offer different types of disk storage with varying I/O performance characteristics, such as solid-state drives (SSDs) or magnetic disks
[46]	GPU	• GPUs are specialized processors that are designed to perform complex mathematical calculations and render graphics • They are used in applications such as gaming, scientific simulations, and machine learning • CSPs offer different GPU configurations to meet the varying needs of applications
[104]	Input/ output	• It refers to the number of read and writes operations that can be performed on a disk storage resource per second • It has a significant impact on the performance of applications that require frequent disk access, such as databases or file storage applications • CSPs offer different levels of I/O for their storage resources
[67]	No. of requests	• It refers to the number of times an application sends a request to a cloud resource or service • The number of requests can have a significant impact on the performance and cost of CWs
[17, 42, 49, 99, 102, 105]	Memory	• It is also known as RAM, is a temporary storage area that a computer's CPU can access quickly • It is used to store data and instructions that the CPU needs to access frequently • CSPs offer different memory configurations to meet the varying needs of applications
[63]	Bandwidth	• It refers to the amount of stuff that may be sent via a network connection in a particular amount of time • In the cloud, bandwidth is used to transfer data between VMs, storage, and other resources • CSPs offer different bandwidth configurations to meet the varying needs of applications
[60, 80, 101, 106]	Network	• Network interfaces provide a virtualized network connection to VMs • They are used to connect VMs and containers • CSPs offer different network interface configurations to meet the varying needs of applications

Previous studies in different areas such as global solar radiation [108], the stock market [109], the sports industry [110], and rainfall forecasting [111] state that prediction with time series data is not an easy task. In time series group data points are studied by consecutive points at constant intervals of time. Classical methods used in time series forecasting [88] include Auto Regression (AR) model, Exponential Smoothing (ES) model, Autoregressive Integrated Moving Average (ARIMA) model, Moving Average (MA) model, etc.

ML approach helps organizations to generate precise predictions of a query based on past data to the desired solutions. It can be about anything from a user point of view to suspected dishonest conduct. These give insights to the business that result in genuine financial revenues. For instance, if a model predicts client is likely to switch business, user personal chat can be targeted to avoid losing that customer. The prediction-based selection makes it possible to prevent the upcoming inactive virtual machine (VMs) foremost to decrease the inactive request with retransmission from inactive VMs to active ones [67].

The fast advancement of AI has drawn significant attention to DL techniques [91]. Latest years have seen a grow-up attention in short-term time-series prediction using deep learning [112].When analyzing complicated nonlinear patterns in data, deep neural networks have an advantage over traditional ML structures because DL [113] can examine hierarchical and distributed characteristics. In the paper [79], the authors designed BG-LSTM an integrated method for time series data using a DL approach. The proposed method helps to increase the accuracy of DC predictions and examine patterns in the workload and resource consumption data.

Cloud computing has scalable and flexible sharing of resource services through resource management. In a cloud environment the foundation is laid on a resource that depends on monitoring and prediction to obtain resource automation and manage high performance. The issues associated in [114] is that with monitoring and prediction of resources are addressed in the cloud computing model, execution and design of flexible resource monitoring schemes for cloud computing, and current resource prediction technique founded on VAR through the relationship among different resources.

Table 3 displays the available and newly generated prediction algorithms selected from the recent year's studies. It also gives idea related to its respective parameters.

To accurately predict the workload in the CDC the regression technique is considered the most favorable and significant method. Time feature is associated with data therefore time series forecasting methods are applicable for this type of workload. The previous studies show that various prediction techniques such as statistical methods (i.e. AR, MA, ARIMA, ES, Holt's method, etc.), ML methods (i.e. LR, SVM, NN, etc.), DL methods (i.e. LSTM, DBN, etc.), ensemble techniques (i.e. stacking, voting, boosting, etc.) nature inspired algorithms (i.e. ANN, DE, etc.) are used to solve various cloud workload issues. Many authors tried to combine two or more techniques to resolve the issues.

The above table shows that the parameters such as accuracy and error metrics have been the maximum pick-up ratio. And error metrics include MAE, MSE, MAPE, and RMSE measure values. The researcher can be considering these parameters for their future work.

4.3 Scheduling techniques and its optimize parameters (RQ 3)

In the cloud, each user task uses many virtualized resources and scheduling plays an important role to manage computing resources [121]. The task scheduling method developed in the study [23] aims to minimize service-level agreement (SLA) violations, overall execution time, and costs while distributing m number of tasks of a particular application among a collection of diverse VMs. As a result, an effective scheduling algorithm is required that can balance competing priorities.

In a cloud paradigm, tasks are sorted into two types: they are computing intensity and data intensity. Though the task scheduling requires computing intensity, the data is migrated to the scheduler having high output resources; hence it minimizes the implementation time of the tasks. Alternatively, task scheduling requires data intensity, which helps to decrease the number of data migrations which results to reduce data transfer time [29].

Currently, cloud service numbers are increasing, which in exchange increases the load on cloud nodes for processing. Hence, needs an efficient method to schedule tasks and resources are managed in the cloud environments [122]. Many researchers have enhanced heuristic techniques to achieve better performance in scheduling and others are working on the meta-heuristic algorithm as well as hybrid approaches too. Table 4 gives the recent scheduling techniques and their description.

Scheduling problem is an NP-hard; means it cannot be solved in a polynomial amount of time. It require finding best scheduling methods that relay on various factors such as work characteristics, different goals, and multiple machine conditions. Therefore, optimization techniques are used to provide efficient scheduling. From the Table 4 we have conclude that the scheduling algorithms categorize in such a way i.e. classical optimization algorithms (i.e. FCFS, Round Robin, Min–Max, Min–Min, etc.), nature-inspired optimization algorithms (i.e. GA, PSO, ACO, etc.), Fuzzy theory based optimization algorithms, and many more are used to schedule the task to the VM.

In scheduling, evaluation parameters refer to the criteria or metrics used to assess the performance or efficacy of a timetable. These metrics are used to analyze how effectively a schedule meets its goals and objectives, as well as to suggest opportunities for improvement. Metrics are significant because it allow schedulers to evaluate a current schedule performance and make accurate decisions about schedule modifications. Schedulers may increase schedule

Table 3 Comparison between prediction models and its parameters

References	Approach	Problem	Algorithms	Dataset and parameters	Best one (s)
[115]	Integration of statistical and ML models	Improve prediction accuracy	AME-WPC, RF, KNN	AuverGrid trace MSE (AME-WPC = 42, RF = 175, KNN = 104), NMSE (AME-WPC (2,5), RF(10,3), KNN (6,2))	AME-WPC
[57]	Time series forecasting models	Dynamic resource provisioning	AR(1), MA(1) SES, DES, ETS, ARIMA, NN	Google Cluster MAPE (AR(1) = 51.44, MA(1) = 50.128 SES = 85.89, DES = 75.33, ETS = 72.26, ARIMA = 85.89, NN = 52.35) Intel Netbatch MAPE (AR(1) = 27.70604, MA(1) = 40.856 SES = 29.45, DES = 29.18, ETS = 31.43, ARIMA = 29.72, NN = 29.75)	Google Cluster = MA(1) Intel Netbatch = AR(1)
[79]	Integration DL with time series models	Achieving optimal resource provision	ARIMA, SVM, LSTM, BiLSTM, GridLSTM, SG-LSTM, SG-BiLSTM, SG-GridLSTM, BG-LSTM	Google Cluster RMSLE = (ARIMA = 0.93, SVM = 0.86, LSTM = 0.83, BiLSTM = 0.77, GridLSTM = 0.80, SG-LSTM = 0.74, SG-BiLSTM = 0.19, SG-GridLSTM = 0.17, BG-LSTM = 0.15)	BG-LSTM
[116]	Time series and statistical model	Automatic scaling for elasticity mechanism	MA, AR, ARIMA, DM, MM, Kalman Filter, Pattern matching model	Aliyun MAPE = (MA = 0.2820, AR = 0.2958, ARIMA = 0.2690, DM = 0.4504, MM = 0.2744, Kalman Filter = 0.2367, Pattern matching model = 0.2501)	Kalman Filter
[117]	Time series model	Achieving optimal resource provision	PRESS, AGILE, ARIMA, NARNN, LSTM-U, BLSTM-U, LSTM-M, BLSTM-M	Google Cluster RMSE = (PRESS = 0.2620, AGILE = 0.0159, ARIMA = 0.0198, NARNN = 0.0135, LSTM-U = 0.0123, BLSTM-U = 0.0115, LSTM-M = 0.0105, BLSTM-M = 0.0095)	BLSTM-M
[84]	DL model	Accurate prediction	NN, DBN, Proposed DL based on the canonical polyadic decomposition	PlanetLab MAPE = (NN = 0.26, DBN = 0.22, Proposed DL based on the canonical polyadic decomposition = 0.21) RMSE = (NN = 10.06, DBN = 10.26, Proposed DL based on the canonical polyadic decomposition = 9.17)	Proposed DL based on the canonical polyadic decomposition
[118]	ML model	Auto Scaling	NN, LR, RepTree, M5P	Wikipedia server traffic data Time = (NN = 3.568, LR = 0.01, RepTree = 0.02, M5P = 0.06)	LR
[17]	ML model	Load Balancing	FFT, improved FFT, LSTM	Accuracy = (FFT = 78%, improved FFT = 85%, LSTM = 90%)	LSTM
[119]	ML model	Workload and energy estimation	AP, TCLA, TK-means, TP-teda, TSSAP	Bitbrain Accuracy = (AP = 9%, TCLA = 12%, TK-means = 50%, TP-teda = 50%, TSSAP = 67%)	TSSAP
[120]	Time series forecasting model	Workload prediction	ARIMA(1,1,1), ARIMA (1,0,1), ARIMA (5,0,2), ARIMA (5, 0, 0)	Standard Error = (ARIMA(1,1,1) = 0.094074943, ARIMA (1,0,1) = 0.089732489, ARIMA (5,0,2) = 0.148992588, ARIMA (5, 0, 0) = 0.258320039)	ARIMA(1,0,1)

Table 3 (continued)

References	Approach	Problem	Algorithms	Dataset and parameters	Best one (s)
[27]	ML model	Accurate prediction	{LR, SVM, EN, LASSO, RR, NNLS} + Proposed Window Technique (WT)	Alibaba NMSE = (LR = 0.40, SVM = 0.38, EN = 0.36, LASSO = 0.39, RR = 0.37, NNLS = 0.41) Materna NMSE = (LR = 0.49, SVM = 0.49, EN = 0.46, LASSO = 0.47, RR = 0.48, NNLS = 0.42) Bitbrains NMSE = (LR = 0.20, SVM = 0.22, EN = 0.17, LASSO = 0.20, RR = 0.19, NNLS = 0.20)	Alibaba = EN + WT Materna = NNLS + WT Bitbrains = EN + WT
[26]	ML model	Accurate prediction	GBT, LR, SVM, Krining, Liu, proposed model	Bitbrain RMSE = (GBT = 2.31, LR = 2.40, SVM = 2.35, Krining = 2.28, Liu = 2.26, Proposed = 2.22) MAE = (GBT = 1.24, LR = 1.32, SVM = 1.28, Krining = 1.24, Liu = 1.24, Proposed = 1.14)	Proposed model

efficiency, accuracy, flexibility, and overall quality by identifying areas where a schedule may be running low. Table 5 represents the list of effective parameters used in scheduling techniques.

From Table 5 we can see that cost, energy consumption, times are the most useful parameters in the previous study. Researchers may choose other metric values for their future work.

4.4 Prediction based scheduling techniques (RQ 4)

In this section several articles are presented in the domain of workload prediction and resource scheduling. Also it discusses the research gaps, importance and need of the current work.

Prediction based scheduling is an important technique to optimize the allocation of resources based on predictions about future demand [40]. This technique involves predicting future workloads and scheduling resources accordingly to ensure efficient and effective utilization. It is important because it can help improve system performance and reduce costs. The CSP can optimize resource utilization, reduce idle time, and avoid over provisioning resources, which can result in cost savings for the provider and its customers. In addition, prediction-based scheduling can help improve user experience by ensuring that resources are available when needed.

For instance, if a website is expecting a sudden surge in traffic, prediction-based scheduling can allocate more resources to the website to ensure that it remains responsive and does not crash due to overload. Overall, prediction-based scheduling is important because it can help improve system performance, reduce costs, and improve user experience by efficiently allocating resources based on predictions about future demand.

The research gaps are discusses here that are found from the previous studies. It helps to design the relevant questions for doing this survey. According to the studies [25, 57, 60, 86] there is a need for more accurate prediction models that can effectively predict the workload of cloud applications. Existing models may not account for certain variables that can affect the performance of the system. Many prediction-based scheduling techniques rely on heuristics [17, 18, 34, 90] and rule-based approaches [42, 133–135]. There is a need for more sophisticated techniques, such as DL, hybrid and ensemble to improve the accuracy of predictions. Cloud applications are often composed of multiple heterogeneous workloads that have different resource requirements [46, 61, 63, 98, 136]. There is a need for prediction-based scheduling techniques that can effectively manage and schedule heterogeneous workloads. Prediction-based scheduling techniques must be scalable to handle large numbers of requests and data [137]. There is a need for techniques that can handle large-scale prediction and scheduling tasks in real-time.

Table 4 Scheduling techniques

References	Algorithm name	Descriptions
[67]	Autonomic scheduling	• The process of automatically scheduling tasks or resources based on the changing needs of a system, without the need for human intervention • Tasks are scheduled automatically based on factors such as system load, resource availability, and user priorities • The goal of autonomic computing is to create systems that can self-manage and self-optimize, reducing the need for human intervention • Used to manage the allocation of resources such as virtual machines and storage to meet changing demands
[123]	Centralized Scheduling	• The process of scheduling tasks or resources in a centralized manner, typically by a single scheduler or resource manager • Used to manage the allocation of resources such as virtual machines, storage, and network bandwidth to meet the demands of multiple users or applications • The scheduler receives requests for resources from multiple users or applications, and then allocates those resources based on a set of predetermined rules or policies • It provide better visibility and control over resource usage, as all requests for resources are managed by a single entity
[3, 124]	FCFS/FIFO	• It is widely used scheduling algorithm in computer systems, including cloud workload management • The first task or request that arrives is executed first, and subsequent tasks or requests are executed in the order in which they arrived • Used to manage the allocation of resources such as virtual machines, storage, and network bandwidth to multiple users or applications • Tasks are executed in the order in which they arrive, without any priority given to certain tasks or users. This can help to ensure that all users or applications receive a fair share of resources over time
[125]	Greedy Algorithm (GA)	• It is an algorithm that makes locally optimal decisions at each step in order to achieve a global optimum • Used to optimize the allocation of resources such as virtual machines, storage, and network bandwidth to multiple users or applications • Allocate resources to the user or application with the highest priority or the greatest need • It prioritizes shorter tasks or requests, in order to optimize the overall throughput of the system
[2, 54]	Deep reinforcement learning	• Using DRL for scheduling in cloud workload management is that it can learn to adapt to changing conditions over time • It can learn to allocate more resources to a user or application that is experiencing a sudden surge in workload, or to allocate fewer resources to a user or application that is experiencing low utilization • The agent may require a large amount of training data in order to learn effective policies for resource allocation
[67]	Honey-Bee	• It is a nature-inspired optimization algorithm that has been applied to various problems, including scheduling in cloud workload management • Used to optimize the allocation of resources such as virtual machines, storage, and network bandwidth to multiple users or applications • It can explore a large search space efficiently and effectively • The challenges of using honey bee scheduling for CW management is the complexity of the algorithm and the potential for the algorithm to converge on suboptimal solutions
[46]	Horus scheduling	• It is a heuristic algorithm for scheduling tasks in CW management • The algorithm is based on a hierarchical structure of clusters and nodes that represents the available resources in the cloud environment • It can efficiently allocate resources in a hierarchical structure, which can be more scalable and easier to manage than a flat structure • It prioritize tasks based on their resource requirements and priorities, which can improve the overall performance of the system • Challenge of using Horus scheduling for CW management is that it requires prior knowledge of the available resources in the cloud environment and the task requirements

Table 4 (continued)

References	Algorithm name	Descriptions
[54, 67, 124]	Load balancing strategy (LBS)	• It is a popular approach for managing cloud workloads and improving the utilization of resources in cloud environments • It distributes the workload evenly across multiple computing resources to avoid overloading some resources while others remain underutilized • Improve the utilization of resources in cloud environments, which can lead to improved performance and cost-effectiveness • The challenge of LBS is that it requires real-time monitoring of the workload and computing resources to adjust the workload distribution
[3, 24, 92]	Max Min	• The algorithm works by allocating resources to tasks based on their needs, and then reallocating the remaining resources to the next task in a greedy manner • It begins by assigning a minimum amount of resources to each task, which ensures that every task has a fair share of resources • Then proceeds to allocate resources to the tasks in a greedy manner, starting with the task that has the highest resource requirement • It continues to allocate resources to the next task in a similar manner until all resources are allocated or no more tasks can be scheduled • It may not be suitable for highly dynamic environments where the resource availability and task requirements can change rapidly
[3, 24, 92]	Min Min	• It is a popular scheduling algorithm used in cloud computing for allocating tasks to available computing resources • This algorithm works by selecting the task with the minimum execution time from the set of available tasks and assigning it to the resource with the minimum completion time • This algorithm improves the overall performance of the system by minimizing the completion time of all tasks • The main disadvantage of the Min-Min algorithm is that it may not always produce the optimal solution
[67]	Round Robin	• It is a simple and fair scheduling algorithm that assigns tasks to computing resources in a cyclic manner, based on a time slice or quantum • Each task is allocated a fixed time slice, and the tasks are scheduled in a circular order • It ensures fair resource allocation among tasks • It requires minimal overhead and can be implemented easily in most cloud computing environments • The main disadvantage of RR scheduling is that it may not be suitable for tasks with varying resource requirements

Table 6 shows that comparison between previous studies related to the prediction and scheduling approaches which contain eight cells fine the information which includes: the references, problems in workload, what are requirements for the problem, the proposed technique, the solution given by the proposed approach, the related environment in which the experiment performed, experiment setup and future works.

5 Research challenges and future directions

The research covered the challenges in CW solved by prediction models and scheduling techniques. The survey [39, 41, 47, 49], stated that accurately predicting workload demands is vital for effective scheduling. But the prediction of resource allocation in the CW is difficult due to unpredictable demand of resources. Also, this changing workload patterns in real time makes difficult to scheduler to schedule tasks. In this survey, the possible solutions related to these challenges have been discussed. Apart from these, there are another challenges are required to outlook. Hence, future directions addressed that challenges driven by the research communities can be further investigated are given below:

- It is necessary to enhance data quality before utilizing ML and DL methods. Irrelevant features lower model performance. Preprocessing technique will help the researcher find the right domain by removing unnecessary features.
- Compliance requirements can vary based on industry, geography, and other factors, making it challenging to

Table 5 Scheduling techniques effective parameters

References	Measure	Description
[23, 44, 53, 126]	Cost	• It refers to the financial cost or the economic value associated with the allocation and utilization of cloud computing resources for executing a particular workload or task • It determines the financial cost of executing a particular workload and helps optimize resource allocation to minimize the overall cost • The appropriate value of the cost parameter depends on the specific requirements and characteristics of the cloud computing environment, the workload characteristics, and the available resources
[53]	Deadline violation	• This measures help evaluate the efficiency and effectiveness of scheduling algorithms in meeting task deadlines • They also aid in identifying the reasons for missed deadlines, such as insufficient resources, unexpected changes in workload, or inaccurate workload characterization • It is often critical to ensure timely delivery of services and maintain user satisfaction
[1, 19, 44, 101, 127]	Energy consumption	• This is the total energy consumed by the data center to execute the workload. A lower EC indicates better energy efficiency • This measure helps to evaluate the efficiency and effectiveness of scheduling algorithms in optimizing resource allocation to minimize energy consumption while meeting performance requirements • They also aid in identifying the reasons for high energy consumption, such as inefficient resource utilization, over-provisioning, or under-provisioning of resources
[1, 53, 54, 68, 128]	Execution time	• It is an important performance metric used to evaluate the efficiency and effectiveness of scheduling algorithms • Execution time measures the time taken by a task to complete its execution from the start of its allocation to the release of its results • This measure help evaluate the efficiency and effectiveness of scheduling algorithms in meeting performance requirements and improving the overall execution time of the workload • It also aid in identifying the reasons for long execution times, such as resource contention, inefficient resource allocation, or insufficient resources
[28, 46, 67]	Latency	• It is an important performance metric used to evaluate the efficiency and effectiveness of scheduling algorithms • Latency measures the delay between a task request and its response from the system • It helps to evaluate the efficiency and effectiveness of scheduling algorithms in meeting latency requirements and improving the overall latency of the workload • It also aid in identifying the reasons for high latency, such as network congestion, resource contention, or inefficient resource allocation
[46, 67, 129]	Makespan	• This is the time taken to complete all the tasks • A shorter makespan indicates better performance • In cloud environments, where multiple tasks are executed concurrently, the makespan can be influenced by various factors such as resource availability, task dependencies, and workload characteristics • The makespan can be calculated using the following formula: Makespan = Finish time of last task—Start time of first task • Scheduling algorithms aim to minimize the makespan by optimizing the allocation of resources and scheduling of tasks
[67]	Resource utilization	• It is an important metric in cloud workload scheduling that measures the degree to which available resources are being utilized for executing tasks • The goal of scheduling algorithms is to achieve high resource utilization by efficiently allocating resources to tasks while minimizing resource wastage • Resource utilization can be calculated using the following formula: Resource Utilization = (Total time resources used by tasks) / (Total time resources available) • Higher resource utilization indicates better performance as it implies that available resources are being utilized effectively to execute tasks

Table 5 (continued)

References	Measure	Description
[44, 67, 127, 130, 131]	Response time	• This is the time taken by a task to receive a response from the system after submitting its request • A lower response time indicates better performance • It is the time interval between the submission of a task request and the receipt of its response from the system • The goal of scheduling algorithms is to minimize response time by efficiently allocating resources to tasks and ensuring timely execution of tasks • Response time can be measured using various metrics such as average response time, median response time, and percentile response time
[23, 34, 132]	SLA rate	• It is an important metric in cloud workload scheduling that measures the percentage of tasks that meet their SLA requirements • It defines the expectations and guarantees between the cloud service provider and the user, specifying the level of service that is expected to be provided • The goal of scheduling algorithms is to optimize the allocation of resources and scheduling of tasks to meet SLA requirements and maximize SLA rate • SLA rate is affected by various factors such as resource availability, network latency, and task dependencies • Scheduling algorithms aim to maximize SLA rate by efficiently allocating resources to tasks and ensuring timely execution of tasks
[46, 52, 92, 123]	Throughput	• It is an important performance metric in cloud workload scheduling that measures the rate at which tasks are completed by the system • It represents the number of tasks that can be completed within a given time period and is often expressed in terms of tasks per unit time • The goal of scheduling algorithms is to optimize the allocation of resources and scheduling of tasks to maximize throughput, i.e., to complete as many tasks as possible within a given time period • This is achieved by minimizing resource contention, ensuring timely execution of tasks, and avoiding idle time for resources
[123]	Transmission delay	• It refers to the time it takes for data to travel from one point to another in a network • In the context of scheduling for cloud workload, transmission delay can be an important factor to consider when deciding how to allocate resources for a particular workload • When a cloud workload is scheduled, the data associated with that workload needs to be transmitted from the user's computer or device to the cloud server where it will be processed • The time it takes for that data to be transmitted can have an impact on the overall performance of the workload

maintain compliance in the cloud. CSPs must ensure that their services comply with regulations and standards to avoid penalties and legal issues.

- Moving workloads from one cloud provider to another can be challenging due to the differences in technology and infrastructure. This can lead to vendor lock-in, where businesses find it challenging to switch providers, limiting their options and potentially increasing costs.
- High computation capabilities are required due to the large size of datasets in the cloud environment.
- Researchers could explore new prediction techniques that leverage ML and AI to analyze large datasets and identify patterns in workload demands. This could involve the use of DL methods to automatically extract features and learn complex relationships between different variables, or the use of ensemble methods that combine multiple prediction models to improve accuracy.
- Increased focus on the use of prediction-based scheduling for edge computing, which involves processing data closer to where it is generated, rather than in a centralized data center.
- By using prediction algorithms to anticipate workload demands at the edge, cloud providers can optimize the allocation of resources and ensure that workloads are processed efficiently and cost-effectively.
- CWs store and process sensitive data, making them a prime target for cyber attacks. Ensuring data security and privacy in the cloud is a critical issue that needs to be addressed in future.

Table 6 Existing prediction based scheduling approaches

References	Problem	Need	Proposed technique	Environment	Experiment setup	Pros	Limitations
[46]	It has been discovered that when DL workloads are effectively co-located with the same GPU, the machine decelerates and this results in system interference	Optimizing GPU consumption and productivity requires	Horus: an interference aware and prediction-based resource manager	DL Framework	CUDA Toolkit	Horus eliminates the requirement for online monitoring and partitions reserved GPUs by anticipating heterogeneous deep learning applications' GPU utilization	It needs a significant quantity of data in order to outperform other strategies
[67]	The dynamic load balancing increases inter-VM communication overheads	Communication overhead is one of the key limitations of dynamic load balancing solutions; a dynamic scheduling mechanism is required for optimum distribution of requests on VMs	Autonomous Load Balancing Method	Virtualized Cloud Environment	CloudSim	The suggested technique enhances resource usage and minimizes turnaround time and computation time by distributing requests proportionally across VMs	The proposed technique limits its work to generalize the proposed technology to geographic clouds with distributed DCs. Also, this approach can be used to optimize the autonomous multiobjective scheduling method
[23]	The dynamic behavior of cloud resource burden affects resource availability during scheduling	. Effective resource allocation is required so that resources may be dispersed in accordance with the requirements of their respective executions	OPSA	Virtualized Cloud Environment	Experimentation platforms consist of Java SDK 8, CloudSim 3.0, WorkflowSim 1.0, Netbeans IDE 8.2, and Microsoft Azure	Applications are assigned to appropriate VMs	The work limits its work expanded to include epigenomics, montage, anomaly prediction and weather prediction
[25]	As a popular IT service, more companies have shifted to cloud DCs. Cloud service providers (CSPs) should offer flexibility, cost effectiveness, and QoS for their clientele. Achieving QoS while minimizing costs is a challenging task	To effectively manage cloud resources, it is essential to offer an accurate approach for predicting the workload of VMs	Workload Forecasting Using m-Gaps and Cluster Analysis	Cloud Trace	Tensorflow	In initial stage technique forecasts time before projected time point for task scheduling. And then groups comparable workloads	This study will improve DL algorithm design using a clustering based approach
[53]	Existing resource schedulers lack Big Data analytical applications' efficiency and deadline meeting	Platforms for streaming large data must enhance their scheduling efficiency and deadline meeting probability	Deadline-Aware-Scheduling Method	Virtualized Cloud Environment	CloudSim	The proposed approach addresses public cloud cost, efficiency variance, deadline compliance, and delay reduction for streaming Big Data analytical applications	It is planned to expand the proposed work to multi-cloud resources in the future. Next, streaming Big Data fluctuation prediction can be use to improve resource provisioning. Research can also explore Big Data health concepts

Table 6 (continued)

References	Problem	Need	Proposed technique	Environment	Experiment setup	Pros	Limitations
[28]	Recent rise in cloud DC capacity and customer quality demands have affected the cloud system's structure	Resource scheduling management is required for complex cloud environments	Online RSF Based on the DQN	Simulation Environment	Python environment with Tensor-Flow	Successfully optimizing complicated multi-objective optimization problems. By altering the weight of incentives, an effective resource allocation and job scheduling technique balances makespan and energy usage	Effective load forecasting helps manage cloud resources and scheduling tactics, hence a multi-learning model is needed
[30]	Cloud growth leads to more users, which increases the load on DCs and increases power consumption. However, many idle network resources can be used as cloud resource providers. Distributed resources with limited capacity may support resource intensive applications by collaborating and sharing	New scheduling method required to improve job execution stability and minimize power usage	Crowed Funding Model	Cloud Environment	CloudSim	The proposed method of scheduling aim is to boost task execution stability and minimize power usage	To encourage unused resources to engage with the resource pool
[60]	Utilization of available resources is one of the most significant considerations for a cloud service provider. In dynamic resource utilization, accurate prediction is a challenging task	Predicting workload improves resource utilization. Machine learning helps create reliable prediction models. In machine learning, Ensemble processes improve prediction accuracy by using multiple learners	Ensemble based workload prediction mechanism	Cloud Trace	Over Produce and Choose (OPC) approach	Ensemble mechanisms help in enhancing the accuracy of predictions based on stack generalization, which employs a collection of learners instead of one learner	In future, it can be feasible to provide an automated resource allocation strategy based on workload forecasting. A resource allocator that has a prediction module that is as accurate as possible is essential to getting the most out of your available resources. In addition, a module for ensemble prediction might be built for users of the cloud, which would assist them in making decisions on resource scaling

Table 6 (continued)

References	Problem	Need	Proposed technique	Environment	Experiment setup	Pros	Limitations
[103]	The extreme variation of workloads with SLA requirements and lower QoS may affect production application performance	There is a need for a strategy that can schedule hybrid workloads. This technique also included forecasting resource availability before scheduling a new job, and then adapting scheduling choices based on the relative importance of the tasks being scheduled	ARMA Prediction Model, Feedback-Based-Online-AR Model, and MPHW) Scheduling	Cloud Trace	IBM SPSS Statistics toolkit	The multi-prediction model calculates, for each distinct kind of activity, the total number of resources that are free and can be applied to new endeavors	In the future, this research may investigate the interference that can occur between different hybrid jobs running on the same host, as well as the upper limit of resources that can be allocated to new tasks while minimizing the risk of interference
[88]	Cloud computing presents a number of challenges, two of the most significant of which are the scalability of dynamic resources and the use of electricity. Because of these traits, a cloud system is inefficient and costly	Predicting the amount of work to be done by users is one of the aspects that may be utilized to improve a cloud's efficiency and lower its running expenses	NN And SADE Algorithm	Simulation Environment	MATLAB	The model can learn the most suited mutation approach as well as the optimal crossover rate	The goal of future work is to increase forecast accuracy while also implementing a VM placement and relocation process to prevent SLA breaches. Furthermore, a prediction model based on multiple-workload characteristics may be investigated further
[138]	It is challenging to maximize the benefits of a cloud cluster while minimizing computing expenses	An accurate forecast of cloud workload is essential for maximizing resource usage	Recurrent Neural Networks (RNN)	Cloud Environment	Orthogonal Experimental Design (OED)	It's possible that successfully representing data volatility requires using RNN to develop mathematical models of curve fitting and parameter estimation	Although the RNN base technique predicts time sequences accurately, it can only solve short-term time sequences. RNN-based methods are unsuitable for long-term temporal sequence prediction tasks. To tackle this issue, this study may attempt different approaches such as LSTM or CW-RNNs
[24]	Cloud computing is widely utilized; however, job and resource scheduling need improvements	The optimal allocation of resources to tasks decreases completion time and optimizes resource usage	DAG based prediction using PTCT	Simulation Environment	MATLAB	Utilizing PCA and minimizing ETC matrix, the suggested technique delivers a considerable improvement in the makespan and minimizes computation and complexity	Benchmarking and dynamic scheduling for real-world application graphs will be done in the future. The priority will be on optimizing task scheduling overall energy use and can with state of-the-art cloud energy methods

Table 6 (continued)

References	Problem	Need	Proposed technique	Environment	Experiment setup	Pros	Limitations
[123]	Existing data distribution systems depend excessively on real-time channel updates, causing excessive message overhead and broadcast latency	VANETs require a channel predictionbased scheduling approach for data dissemination to reduce transmission cost and improve system efficiency	Channel Prediction-Based Scheduling (CPDS)	Simulation Environment	Not Given	Apply CPDS to help the CS in gathering scheduling in sequence. The approach achieves large-scale channel prediction with minimal computing cost, making it appropriate for real-time processing	Dynamic scheduling not done in this channel
[139]	The provisioning of resources for computational workloads is a most important question	To build a complete resource provisioning model, it is important to estimate the future resource consumption of upcoming computing processes	Machine Learning-Based Prediction Models	Cloud Environment	opensource Java library Deeplearning4j (DL4J)	Basic linear regression or fixed values are used to anticipate resource utilization to refine the provisioning strategy	In the future, performance will be enhanced by utilizing more complex models, such as recurrent ANN. This strategy can also apply to online learning environments
[116]	Cloud computing can automatically decrease and increase resources to fit user needs	To achieve elasticity, the automatic scaling mechanism has to be initiated. Analyzing workload is a common solution	Kalman Filter Model, Time Series Model, Pattern Matching	Cloud Environment	Aliyun	For monitoring data, three models are used to forecast workload. First, examine monitoring data using time series. Then to anticipate the cloud workload filter is used. Next, matching method is used to forecast workload	Future studies will improve the pattern matching model and trigger approach. Also, this research focused on elasticity evaluation to enhance the prediction model
[126]	To increase cloud computing quality, not only should cost and bandwidth is met, but system friendliness should also be emphasized	Dynamic Resource Scheduling(DRS) is required to improve the efficiency and QoS	Fuzzy Based Scheduling (FCTRS)	Cloud Environment	Cloud Sim	The system improves cloud computing QoS by splits user needs and obtainable resources into numerous fuzzy levels	For future work, the method needs to be tested in real-world cloud computing platforms to see whether it works to improve service quality

6 Conclusion

Currently, traditional businesses are rapidly migrating to cloud environments and the increasing demand for multiple resources to perform a single task is the main challenges for cloud service providers. This paper deals with different CWs and their applications where virtual resources are provided with specific features. To maximize the management of diverse resources, efficient scheduling is needed. From the literature, it is also observed that prediction plays an important role to improve scheduling performance. Therefore, an efficient prediction-based scheduling framework is needed to address the challenges in CWs. The present study examined various research articles with their issues, research gaps, and research challenges. Also it gives solutions of related RQ and future directions for the upcoming problem. The research paper is reviewed from the year 2015 to the current and presented in tabular forms with their comparative analysis. The information provided by each table helps the researchers to reduce the time to resolve the issues related to the various scenarios. Here, the data is classified into various categories including workload datasets, resources used in the cloud, prediction models, scheduling algorithms used in previous studies, and their respective evaluation parameters. Finally, the survey paper concludes that designing a framework for prediction-based scheduling with a hybrid deep learning method is the best suite for workload prediction, and integrating with this for scheduling a nature-inspired optimization algorithm can perform excellent work.

Author contributions The authors confirm their contribution to the paper as follows: SK: Study conception and design, Data collection, Analysis and interpretation of results, Draft manuscript preparation. AS: Supervision

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability No data was used for the research described in the article.

Declarations

Competing interests The authors declare no competing interests.

References

1. Seddiki, D., Galan, S.G., Exposito, E.M., Ibanez, M.V., Marciniak, T., De Prado, R.J.P.: Sustainability-based framework for virtual machines migration among cloud data centers. In: 2021 15th Int. Conf. Signal Process. Commun. Syst. ICSPCS 2021 - Proc. (2021)
2. Eloghdadi, H.M., Ramadan, R.A.: Resource scheduling for offline cloud computing using deep reinforcement learning. IJCSNS Int. J. Comput. Sci. Netw. Secur. **19**, 54 (2019)
3. Aladwani, T.: Types of Task Scheduling Algorithms in Cloud Computing Environment. IntechOpen, London (2020)
4. Wickremasinghe, B., Calheiros, R.N., Buyya, R.: CloudAnalyst: a cloudsim-based visual modeller for analysing cloud computing environments and applications. Proc. Int. Conf. Adv. Inf. Netw. Appl. AINA. (2010). https://doi.org/10.1109/AINA.2010.32
5. Rodriguez, M.A., Buyya, R.: A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS cloud computing environments. Concurr. Comput. **29**, 1–23 (2017). https://doi.org/10.1002/cpe.4041
6. Li, W., Wu, J., Cao, J., Chen, N., Zhang, Q., Buyya, R.: Blockchain-based trust management in cloud computing systems: a taxonomy, review and future directions. J. Cloud Comput. (2021). https://doi.org/10.1186/s13677-021-00247-5
7. Ali, M.R., Ahmad, F., Chaudary, M.H., Khan, Z.A., Alqahtani, M.A., Alqurni, J.S., Ullah, Z., Khan, W.U.: Petri Net based modeling and analysis for improved resource utilization in cloud computing. PeerJ Comput. Sci. **7**, 1–22 (2021). https://doi.org/10.7717/PEERJ-CS.351
8. Balashunmugaraja, B., Ganeshbabu, T.R.: Privacy preservation of cloud data in business application enabled by multi-objective red deer-bird swarm algorithm. Knowl.-Based Syst. **236**, 107748 (2022). https://doi.org/10.1016/j.knosys.2021.107748
9. Alnumay, W., Ghosh, U.: A trust-based predictive model for mobile ad hoc network in internet of things. Sensors **19**, 1–14 (2019). https://doi.org/10.3390/s19061467
10. Yadav, C., Patro, B.D.K., Yadav, V.: Authentication, access control, VM allocation and energy efficiency towards securing computing environments in cloud computing. Ann. Roman. Soc. Cell Biol. **25**, 17939–17954 (2021)
11. Bal, P.K., Mohapatra, S.K., Das, T.K., Srinivasan, K., Hu, Y.C.: A joint resource allocation, security with efficient task scheduling in cloud computing using hybrid machine learning techniques. Sensors **22**, 1242 (2022). https://doi.org/10.3390/S22031242
12. Syarif, I., Zaluska, E., Prugel-Bennett, A., Wills, G.: Application of Bagging, Boosting and Stacking to Intrusion Detection. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-31537-4_46
13. Tabrizchi, H., Kuchaki Rafsanjani, M.: A survey on security challenges in cloud computing: issues, threats, and solutions. J. Supercomput. **76**, 9493–9532 (2020). https://doi.org/10.1007/s11227-020-03213-1
14. Kumar, J., Singh, A.K.: Performance assessment of time series forecasting models for cloud datacenter networks' workload prediction. Wirel. Pers. Commun. **116**, 1949–1969 (2021)
15. Sayadnavard, M.H., Toroghi Haghighat, A., Rahmani, A.M.: A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers. Eng. Sci. Technol. **2**, 6 (2021). https://doi.org/10.1016/j.jestch.2021.04.014
16. Elrotub, M., Gherbi, A.: Virtual machine classification-based approach to enhanced workload balancing for cloud computing applications. Procedia Comput. Sci. **130**, 683–688 (2018). https://doi.org/10.1016/j.procs.2018.04.120
17. Ramesh, R.K., Wang, H., Shen, H., Fan, Z.: Machine learning for load balancing in cloud datacenters. In: Proc. - 21st IEEE/ACM Int. Symp. Clust. Cloud Internet Comput. CCGrid 2021. pp. 186–195 (2021). https://doi.org/10.1109/CCGrid51090.2021.00028
18. Kumar, J., Singh, A.K., Buyya, R.: Self Directed Learning Based Workload Forecasting Model for Cloud Resource Management. Elsevier, Amsterdam (2021)

19. Gill, S.S., Garraghan, P., Stankovski, V., Casale, G., Thulasiram, R.K., Ghosh, S.K., Ramamohanarao, K., Buyya, R.: Holistic resource management for sustainable and reliable cloud computing: an innovative solution to global challenge. J. Syst. Softw. **155**, 104–129 (2019). https://doi.org/10.1016/J.JSS.2019.05.025
20. Zharikov, E., Telenyk, S., Bidyuk, P.: Adaptive workload forecasting in cloud data centers. J. Grid Comput. **18**, 149–168 (2020). https://doi.org/10.1007/s10723-019-09501-2
21. Thakur, N., Singh, A., Sangal, A.L.: Cloud services selection: a systematic review and future research directions. Comput. Sci. Rev. **46**, 100514 (2022). https://doi.org/10.1016/j.cosrev.2022.100514
22. Mashuqur Rahman Mazumder, A.K.M., Aslam Uddin, K.M., Arbe, N., Jahan, L., Whaiduzzaman, M.: Dynamic task scheduling algorithms in cloud computing. In: Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019. pp. 1280–1286. Institute of Electrical and Electronics Engineers Inc. (2019)
23. Kaur, G., Bala, A.: OPSA: an optimized prediction based scheduling approach for scientific applications in cloud environment. Clust. Comput. **24**, 1955–1974 (2021)
24. Al-Maytami, B.A., Fan, P., Hussain, A., Baker, T., Liatsist, P.: A task scheduling algorithm with improved makespan based on prediction of tasks computation time algorithm for cloud computing. IEEE Access. **7**, 160916–160926 (2019). https://doi.org/10.1109/ACCESS.2019.2948704
25. Gao, J., Wang, H., Shen, H.: Machine learning based workload prediction in cloud computing. In: Proc. - Int. Conf. Comput. Commun. Networks, ICCCN. 2020-Aug 1–9 (2020). https://doi.org/10.1109/ICCCN49398.2020.9209730
26. Baig, S.U.R., Iqbal, W., Berral, J.L., Erradi, A., Carrera, D.: Adaptive prediction models for data center resources utilization estimation. IEEE Trans. Netw. Serv. Manag. **16**, 1681–1693 (2019). https://doi.org/10.1109/TNSM.2019.2932840
27. Baig, S.R., Iqbal, W., Berral, J.L., Carrera, D.: Adaptive sliding windows for improved estimation of data center resource utilization. Future Gener. Comput. Syst. **104**, 212–224 (2020). https://doi.org/10.1016/j.future.2019.10.026
28. Peng, Z., Lin, J., Cui, D., Li, Q., He, J.: A multi-objective tradeoff framework for cloud resource scheduling based on the Deep Q-network algorithm. Clust. Comput. **23**, 2753–2767 (2020). https://doi.org/10.1007/S10586-019-03042-9/FIGURES/9
29. Matrouk, K., Alatoun, K.: Scheduling algorithms in fog computing: a survey. Int. J. Netw. Distrib. Comput. **9**, 59–74 (2021). https://doi.org/10.2991/IJNDC.K.210111.001
30. Zhang, N., Yang, X., Zhang, M., Sun, Y., Long, K.: A genetic algorithm-based task scheduling for cloud resource crowdfunding model. Int. J. Commun. Syst. (2018). https://doi.org/10.1002/dac.3394
31. Masdari, M., Salehi, F., Jalali, M., Bidaki, M.: A survey of PSO-based scheduling algorithms in cloud computing. J. Netw. Syst. Manag. **25**, 122–158 (2017). https://doi.org/10.1007/s10922-016-9385-9
32. Farid, M., Latip, R., Hussin, M., Hamid, N.A.W.A.: A survey on QoS requirements based on particle swarm optimization scheduling techniques for workflow scheduling in cloud computing. Symmetry (Basel). (2020). https://doi.org/10.3390/SYM12040551
33. Abdullah, L., Li, H., Al-Jamali, S., Al-Badwi, A., Ruan, C.: Predicting multi-attribute host resource utilization using support vector regression technique. IEEE Access. **8**, 66048–66067 (2020). https://doi.org/10.1109/ACCESS.2020.2984056
34. Kaur, K., Garg, S., Aujla, G.S., Kumar, N., Zomaya, A.Y.: A multi-objective optimization scheme for job scheduling in sustainable cloud data centers. IEEE Trans. Cloud Comput. **10**, 172–186 (2022). https://doi.org/10.1109/TCC.2019.2950002
35. Alboaneen, D., Tianfield, H., Zhang, Y., Pranggono, B.: A metaheuristic method for joint task scheduling and virtual machine placement in cloud data centers. Future Gener. Comput. Syst. **115**, 201–212 (2021). https://doi.org/10.1016/j.future.2020.08.036
36. Nawrocki, P., Osypanka, P., Nawrocki, P., Osypanka, P.: Cloud resource demand prediction using machine learning in the context of QoS parameters. J. Grid Comput. **19**, 1–20 (2021). https://doi.org/10.1007/S10723-021-09561-3
37. Mohamed, A., Hamdan, M., Khan, S., Abdelaziz, A., Babiker, S.F., Imran, M., Marsono, M.N.: Software-defined networks for resource allocation in cloud computing: a survey. Comput. Netw. **195**, 108151 (2021). https://doi.org/10.1016/j.comnet.2021.108151
38. Ouhame, S., Hadi, Y., Ullah, A.: An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model. Neural Comput. Appl. **33**, 10043–10055 (2021). https://doi.org/10.1007/s00521-021-05770-9
39. Cao, R., Yu, Z., Marbach, T., Li, J., Wang, G., Liu, X.: Load prediction for data centers based on database service. Proc. Int. Comput. Softw. Appl. Conf. **1**, 728–737 (2018). https://doi.org/10.1109/COMPSAC.2018.00109
40. Su, M., Wang, G., Choo, K.K.R.: Prediction-based resource deployment and task scheduling in edge-cloud collaborative computing. Wirel. Commun. Mob. Comput. **222**, 1–17 (2022). https://doi.org/10.1155/2022/2568503
41. Mahmoud, Q.H.: Analysis of job failure and prediction model for cloud computing using machine learning. Sensors. **22**, 1–25 (2022)
42. Morariu, C., Borangiu, T.: Time series forecasting for dynamic scheduling of manufacturing processes. (2018)
43. Dezhabad, N., Ganti, S., Shoja, G.: Cloud workload characterization and profiling for resource allocation. In: Proceeding 2019 IEEE 8th Int. Conf. Cloud Networking, CloudNet 2019. (2019). https://doi.org/10.1109/CLOUDNET47604.2019.9064138
44. Kaur, G., Bala, A.: Prediction based task scheduling approach for floodplain application in cloud environment. Computing **103**, 895–916 (2021)
45. Sharkh, M.A., Xu, Y., Leyder, E.: CloudMach: Cloud Computing Application Performance Improvement through Machine Learning. Can. Conf. Electr. Comput. Eng. 2020-Augus, 1–6 (2020). https://doi.org/10.1109/CCECE47787.2020.9255686
46. Yeung, G., Borowiec, D., Yang, R., Friday, A., Harper, R., Garraghan, P.: Horus: Interference-aware and prediction-based scheduling in deep learning systems. IEEE Trans. Parallel Distrib. Syst. **33**, 88–100 (2022). https://doi.org/10.1109/TPDS.2021.3079202
47. Jassas, M.S., Mahmoud, Q.H.: Analysis of job failure and prediction model for cloud computing using machine learning. Sensors. **22**, 2035 (2022)
48. Wang, C., Li, J., He, Y., Xiao, K., Zhang, H.: Destination prediction-based scheduling algorithms for message delivery in IoVs. IEEE Access. **8**, 14965–14976 (2020). https://doi.org/10.1109/aCCESS.2020.2966494
49. Yu, J., Gao, M., Li, Y., Zhang, Z., Ip, W.H., Yung, K.L.: Workflow performance prediction based on graph structure aware deep attention neural network. J. Ind. Inf. Integr. **27**, 100337 (2022). https://doi.org/10.1016/j.jii.2022.100337
50. Yan, J., Rui, L.L., Yang, Y., Chen, S., Chen, X.: Resource Scheduling Algorithms for Burst Network Flow in Edge Computing. Lect. Notes Electr. Eng. 808 LNEE, pp. 1569–1578 (2022). https://doi.org/10.1007/978-981-16-6554-7_173
51. Chen, L., Zhang, W., Ye, H.: Accurate workload prediction for edge data centers: Savitzky-Golay filter, CNN and BiLSTM

with attention mechanism. Appl. Intell. **52**, 13027–13042 (2022). https://doi.org/10.1007/S10489-021-03110-X/FIGURES/12

52. Praveen, S.P., Thirupathi, K., Janakiramaiah, B.: Effective allocation of resources and task scheduling in cloud environment using social group optimization. Arab. J. Sci. Eng. **43**, 4265–4272 (2018). https://doi.org/10.1007/s13369-017-2926-z
53. Mortazavi-Dehkordi, M., Zamanifar, K.: Efficient deadline-aware scheduling for the analysis of big data streams in public cloud. Clust. Comput. **23**, 241–263 (2020). https://doi.org/10.1007/S10586-019-02908-2/FIGURES/14
54. Cheng, F., Huang, Y., Tanpure, B., Sawalani, P., Cheng, L., Liu, C.: Cost-aware job scheduling for cloud instances using deep reinforcement learning. Clust. Comput. **25**, 619–631 (2022). https://doi.org/10.1007/S10586-021-03436-8/FIGURES/5
55. Theodoropoulos, T., Makris, A., Boudi, A., Taleb, T., Herzog, U., Rosa, L., Cordeiro, L., Tserpes, K., Spatafora, E., Romussi, A., Zschau, E., Kamarianakis, M., Protopsaltis, A., Papagiannakis, G., Dazzi, P.: Cloud-based XR services: a survey on relevant challenges and enabling technologies. J. Netw. Netw. Appl. **2**, 1–22 (2022). https://doi.org/10.33969/J-NANA.2022.020101
56. Hamad, R.M.H., Al Fayoumi, M.: Modernization of a classical data center (CDC) vs. adoption in cloud computing calculate total cost of ownership for both cloud and CDC - Jordanian Case Study. ACIT 2018 - 19th Int. Arab Conf. Inf. Technol. pp. 1–8 (2019). https://doi.org/10.1109/ACIT.2018.8672686
57. Vazquez, C., Krishnan, R., John, E.: Time series forecasting of cloud data center workloads for dynamic resource provisioning. J. Wirel. Mob. Netw. **6**, 87–110 (2015). https://doi.org/10.22667/JOWUA.2015.09.31.087
58. Sakpal, M.: 7 Workloads That Should Be Moved to Cloud Right Now, https://www.gartner.com/smarterwithgartner/7-workloads-that-should-be-moved-to-cloud-right-now
59. Jyoti, A., Shrimali, M., Tiwari, S., Singh, H.P.: Cloud computing using load balancing and service broker policy for IT service: a taxonomy and survey. J. Ambient Intell. Humaniz. Comput. **11**, 4785–4814 (2020). https://doi.org/10.1007/s12652-020-01747-z
60. Mehmood, T., Latif, S., Malik, S.: Prediction of cloud computing resource utilization. In: 2018 15th Int. Conf. Smart Cities Improv. Qual. Life Using ICT IoT, HONET-ICT 2018. 38–42 (2018). https://doi.org/10.1109/HONET.2018.8551339
61. Shen, S., Van Beek, V., Iosup, A.: Statistical characterization of business-critical workloads hosted in cloud datacenters. In: Proc. - 2015 IEEE/ACM 15th Int. Symp. Clust. Cloud, Grid Comput. CCGrid 2015. pp. 465–474 (2015). https://doi.org/10.1109/CCGrid.2015.60
62. Zia Ullah, Q., Hassan, S., Khan, G.M.: Adaptive resource utilization prediction system for infrastructure as a service cloud. Comput. Intell. Neurosci. **2017**, 1–12 (2017). https://doi.org/10.1155/2017/4873459
63. Bashir, S., Mustafa, S., Ahmad, R.W., et al.: Multi-factor nature inspired SLA-aware energy efficient resource management for cloud environments. Clust. Comput **26**, 1643–1658 (2023). https://doi.org/10.1007/s10586-022-03690-4
64. Shukur, H., Zeebaree, S., Zebari, R., Zeebaree, D., Ahmed, O., Salih, A.: Cloud computing virtualization of resources allocation for distributed systems. J. Appl. Sci. Technol. Trends. **1**, 98–105 (2020). https://doi.org/10.38094/jastt1331
65. Kaur, G., Bala, A.: A survey of prediction-based resource scheduling techniques for physics-based scientific applications. Mod. Phys. Lett. B (2018). https://doi.org/10.1142/S0217984918502950
66. Vashistha, A., Sharma, C.M., Mahapatra, R.P., Chariar, V.M., Sharma, N.: Sustainable technical debt-aware computing model for virtual machine migration (TD4VM) in IaaS cloud. Wirel. Commun. Mob. Comput. **2022**, 1–12 (2022). https://doi.org/10.1155/2022/6709797
67. Ebadifard, F., Babamir, S.M.: Autonomic task scheduling algorithm for dynamic workloads through a load balancing technique for the cloud-computing environment. Clust. Comput. **24**, 1075–1101 (2021). https://doi.org/10.1007/S10586-020-03177-0/TABLES/12
68. Stavrinides, G.L., Karatza, H.D.: The impact of workload variability on the energy efficiency of large-scale heterogeneous distributed systems. Simul. Model. Pract. Theory. **89**, 135–143 (2018). https://doi.org/10.1016/j.simpat.2018.09.013
69. Shen, S., Van Beek, V., Iosup, A., Shen, {s, Nl, A.I.: Delft University of technology parallel and distributed systems report series statistical characterization of business-critical workloads hosted in cloud datacenters. (2014)
70. Calheiros, R.N., Ranjany, R., Buyya, R.: Virtual machine provisioning based on analytical performance and QoS in cloud computing environments. Proc. Int. Conf. Parallel Process. pp. 295–304 (2011). https://doi.org/10.1109/ICPP.2011.17
71. Salot, P.: A survey of various scheduling algorithm in cloud computing environment. Int. J. Res. Eng. Technol. **2**, 131–135 (2013). https://doi.org/10.15623/ijret.2013.0202008
72. Tibermacine, O., Tibermacine, C., Kerdoudi, M.L.: Reputation evaluation with malicious feedback prevention using a HITS-based model. In: Proc. - 2019 IEEE Int. Conf. Web Serv. ICWS 2019 - Part 2019 IEEE World Congr. Serv. 180–187 (2019). https://doi.org/10.1109/ICWS.2019.00039
73. Lundin, R.: The advantages of keeping mission critical workloads on-premises vs going to the cloud. (2018)
74. Liu, C., Liu, C., Shang, Y., Chen, S., Cheng, B., Chen, J.: An adaptive prediction approach based on workload pattern discrimination in the cloud. J. Netw. Comput. Appl. **80**, 35–44 (2017). https://doi.org/10.1016/j.jnca.2016.12.017
75. Li, J., Feng, L., Fang, S.: An greedy-based job scheduling algorithm in cloud computing. J. Softw. **9**, 921–925 (2014). https://doi.org/10.4304/jsw.9.4.921-925
76. Xavier, S., Lovesum, S.P.J.: A survey of various workflow scheduling algorithms in cloud environment. Int. J. Sci. Res. Publ. **3**, 2–4 (2013)
77. Kaur, G., Bala, A.: An efficient resource prediction–based scheduling technique for scientific applications in cloud environment. Concurr. Eng. Res. Appl. **27**, 112–125 (2019). https://doi.org/10.1177/1063293X19832946
78. Gautam, J. V., Prajapati, H.B., Dabhi, V.K., Chaudhary, S.: A survey on job scheduling algorithms in Big data processing. Proc. 2015 IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2015. (2015). https://doi.org/10.1109/ICECCT.2015.7226035
79. Bi, J., Li, S., Yuan, H., Zhou, M.C.: Integrated deep learning method for workload and resource prediction in cloud systems. Neurocomputing **424**, 35–48 (2021). https://doi.org/10.1016/J.NEUCOM.2020.11.011
80. Saxena, D., Singh, A.K.: Auto-adaptive Learning-Based Workload Forecasting in Dynamic Cloud Environment. Taylor & Francis, London (2020)
81. Banerjee, S., Roy, S., Khatua, S.: Efficient resource utilization using multi-step-ahead workload prediction technique in cloud. J. Supercomput. **77**, 10636–10663 (2021). https://doi.org/10.1007/s11227-021-03701-y
82. Saxena, S., Sivalingam, K.M.: Slice admission control using overbooking for enhancing provider revenue in 5G Networks. In: Proc. IEEE/IFIP Netw. Oper. Manag. Symp. 2022 Netw. Serv. Manag. Era Cloudification, Softwarization Artif. Intell. NOMS 2022. (2022). https://doi.org/10.1109/NOMS54207.2022.9789905

83. Sharma, O., Saini, H.: VM consolidation for cloud data center using median based threshold approach. Procedia Comput. Sci. **89**, 27–33 (2016). https://doi.org/10.1016/j.procs.2016.06.005
84. Zhang, Q., Yang, L.T., Yan, Z., Chen, Z., Li, P.: An efficient deep learning model to predict cloud workload for industry informatics. IEEE Trans. Ind. Inf. **14**, 3170–3178 (2018). https://doi.org/10.1109/TII.2018.2808910
85. Zhu, Y., Zhang, W., Chen, Y., Gao, H.: A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment. Eurasip J. Wirel. Commun. Netw. **2019**, 1–18 (2019). https://doi.org/10.1186/S13638-019-1605-Z/FIGURES/12
86. Xu, M., Song, C., Wu, H., Gill, S.S., Ye, K., Xu, C.: esDNN: deep neural network based multivariate workload prediction in cloud computing environments. ACM Trans. Internet Technol. **22**, 1–24 (2022). https://doi.org/10.1145/3524114
87. Patel, E., Kushwaha, S., Patel, E., Kushwaha, D.S.: A hybrid CNN-LSTM model for predicting server load in cloud computing. J. Supercomput. **78**, 10327–10356 (2022). https://doi.org/10.1007/s11227-021-04234-0
88. Kumar, J., Singh, A.K.: Workload prediction in cloud using artificial neural network and adaptive differential evolution. Future Gener. Comput. Syst. **81**, 41–52 (2018). https://doi.org/10.1016/J.FUTURE.2017.10.047
89. Amekraz, Z., Hadi, M.Y.: CANFIS: a chaos adaptive neural fuzzy inference system for workload prediction in the cloud. IEEE Access. **10**, 49808–49828 (2022). https://doi.org/10.1109/ACCESS.2022.3174061
90. Kumar, J., Saxena, D., Singh, A.K., Mohan, A.: BiPhase adaptive learning-based neural network model for cloud datacenter workload forecasting. Sensors **24**, 14593 (2020)
91. Sarker, I.H.: Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput. Sci. **2**, 1–20 (2021). https://doi.org/10.1007/S42979-021-00815-1/FIGURES/13
92. Murad, S.S., Abdal, R.B., Alsandi, N., Faraj, R., SalamMurad, S., Badeel, R., Salih, N., Alsandi, A., Alshaaya, R.F., Ahmed, R.A., Muhammed, A., Derahman, M.: Optimized MIN-MIN task scheduling algorithm for scientific workflows in a cloud environment. J. Theor. Appl. Inf. Technol. **31**, 480 (2022)
93. Banik, S., Sharma, N., Mangla, M., Mohanty, S.N., Shitharth, S.: LSTM based decision support system for swing trading in stock market. Knowl.-Based Syst. **239**, 107994 (2022). https://doi.org/10.1016/j.knosys.2021.107994
94. Prakash, K.B., Imambi, S.S., Ismail, M., Kumar, T.P., Pawan, Y.V.R.N.: Analysis, prediction and evaluation of COVID-19 datasets. Int. J. Emerg. Trends Eng. Res. **8**, 2199–2204 (2020)
95. Yu, K., Yang, Z., Wu, C., Huang, Y., Xie, X.: In-hospital resource utilization prediction from electronic medical records with deep learning. Knowledge-Based Syst. **223**, 107052 (2021). https://doi.org/10.1016/j.knosys.2021.107052
96. Banerjee, A., Pasea, L., Harris, S., Gonzalez-Izquierdo, A., Torralbo, A., Shallcross, L., Noursadeghi, M., Pillay, D., Sebire, N., Holmes, C., Pagel, C., Wong, W.K., Langenberg, C., Williams, B., Denaxas, S., Hemingway, H.: Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. Lancet **395**, 1715–1725 (2020). https://doi.org/10.1016/S0140-6736(20)30854-0
97. Heidari, M., Garnaik, P.P., Dutta, A.: The valorization of plastic via thermal means: Industrial scale combustion methods. Plast. to Energy Fuel, Chem. Sustain. Implic. 295–312 (2018). https://doi.org/10.1016/B978-0-12-813140-4.00011-X
98. Gao, M., Li, Y., Yu, J.: Workload Prediction of Cloud Workflow Based on Graph Neural Network. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 12999 LNCS, pp. 169–189 (2021). https://doi.org/10.1007/978-3-030-87571-8_15/COVER
99. Valarmathi, K., Kanaga Suba Raja, S.: Resource utilization prediction technique in cloud using knowledge based ensemble random forest with LSTM model. Concurr. Eng. Res. Appl. **29**, 396–404 (2021)
100. El Motaki, S., Yahyaouy, A., Gualous, H., Sabor, J.: A new weighted fuzzy C-means clustering for workload monitoring in cloud datacenter platforms. Clust. Comput. **24**, 3367 (2021)
101. Karim, M.E., Maswood, M.M.S., Das, S., Alharbi, A.G.: BHyPreC: a novel Bi-LSTM based hybrid recurrent neural network model to predict the CPU workload of cloud virtual machine. IEEE Access. **9**, 131476–131495 (2021). https://doi.org/10.1109/ACCESS.2021.3113714
102. Tseng, F.H., Wang, X., Chou, L.D., Chao, H.C., Leung, V.C.M.: Dynamic resource prediction and allocation for cloud data center using the multiobjective genetic algorithm. IEEE Syst. J. **12**, 1688–1699 (2018). https://doi.org/10.1109/JSYST.2017.2722476
103. Jiang, H., Haihong, E., Song, M.: Multi-prediction based scheduling for hybrid workloads in the cloud data center. Clust. Comput. **21**, 1607–1622 (2018). https://doi.org/10.1007/s10586-018-2265-1
104. Chen, W., Ye, K., Wang, Y., Xu, G., Xu, C.Z.: How Does the Workload Look Like in Production Cloud? Analysis and Clustering of Workloads on Alibaba Cluster Trace. Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS. 2018-December, pp. 102–109 (2019). https://doi.org/10.1109/PADSW.2018.8644579
105. Chen, Z., Hu, J., Min, G., Zomaya, A.Y., El-Ghazawi, T.: Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning. IEEE Trans. Parallel Distrib. Syst. **31**, 923–934 (2020). https://doi.org/10.1109/TPDS.2019.2953745
106. Shyam, G.K., Manvi, S.S.: Virtual resource prediction in cloud environment: a Bayesian approach. J. Netw. Comput. Appl. **65**, 144–154 (2016). https://doi.org/10.1016/J.JNCA.2016.03.002
107. Qaddoum, K.S., Elemam, N.N., Abualhaj, M.A., Qaddoum, K.: Elastic neural network method for load prediction in cloud computing grid. Int. J. Electr. Comput. Eng. **9**, 1201–1208 (2019). https://doi.org/10.11591/ijece.v9i2.pp1201-1208
108. Alsharif, M.H., Younes, M.K., Kim, J.: Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry. **11**, 240 (2019). https://doi.org/10.3390/SYM11020240
109. Idrees, S.M., Alam, M.A., Agarwal, P.: A prediction approach for stock market volatility based on time series data. IEEE Access. **7**, 17287–17298 (2019). https://doi.org/10.1109/ACCESS.2019.2895252
110. Liang, H.: An intelligent prediction for sports industry scale based on time series algorithm and deep learning. Comput. Intell. Neurosci. **2022**, 9649825 (2022). https://doi.org/10.1155/2022/9649825
111. Barrera-Animas, A.Y., Oyedele, L.O., Bilal, M., Akinosho, T.D., Delgado, J.M.D., Akanbi, L.A.: Rainfall prediction: a comparative analysis of modern machine learning algorithms for time-series forecasting. Mach. Learn. Appl. **7**, 100204 (2022). https://doi.org/10.1016/J.MLWA.2021.100204
112. Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V.: Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. Chaos Solitons Fract. **140**, 110227 (2020). https://doi.org/10.1016/j.chaos.2020.110227
113. Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep learning. Electron. Mark. **31**, 685–695 (2021). https://doi.org/10.1007/S12525-021-00475-2/TABLES/2
114. Chen, H., Fu, X., Tang, Z., Zhu, X.: Resource monitoring and prediction in cloud computing environments. In: Proc. - 3rd Int.

Conf. Appl. Comput. Inf. Technol. 2nd Int. Conf. Comput. Sci. Intell. ACIT-CSI 2015. 288–292 (2015). https://doi.org/10.1109/ACIT-CSI.2015.58

115. Cetinski, K., Juric, M.B.: AME-WPC: advanced model for efficient workload prediction in the cloud. J. Netw. Comput. Appl. **55**, 191–201 (2015). https://doi.org/10.1016/j.jnca.2015.06.001

116. Hu, Y., Deng, B., Peng, F., Wang, D.: Workload prediction for cloud computing elasticity mechanism. In: Proc. 2016 IEEE Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2016. 244–249 (2016). https://doi.org/10.1109/ICCCBDA.2016.7529565

117. Gupta, S., Dinesh, D.A.: Resource usage prediction of cloud workloads using deep bidirectional long short term memory networks. 11th IEEE Int. Conf. Adv. Networks Telecommun. Syst. ANTS 2017. 1–6 (2018). https://doi.org/10.1109/ANTS.2017.8384098

118. Sniezynski, B., Nawrocki, P., Wilk, M., Jarzab, M., Zielinski, K.: VM reservation plan adaptation using machine learning in cloud computing. J. Grid Comput. **17**, 797–812 (2019). https://doi.org/10.1007/S10723-019-09487-X/METRICS

119. Khan, T., Tian, W., Ilager, S., Buyya, R.: Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. Future Gener. Comput. Syst. **128**, 320–332 (2022). https://doi.org/10.1016/j.future.2021.10.019

120. Gadhavi, L.J., Bhavsar, M.D.: Adaptive cloud resource management through workload prediction. Energy Syst. **13**, 601–623 (2022). https://doi.org/10.1007/S12667-019-00368-6/FIGURES/9

121. Vijindra, S.S.: Survey on scheduling issues in cloud computing. Procedia Eng. **38**, 2881–2888 (2012). https://doi.org/10.1016/j.proeng.2012.06.337

122. Mohiddin, M.K., Kohli, R., Dutt, V.B., Dixit, P., Michal, G.: Energy-efficient enhancement for the prediction-based scheduling algorithm for the improvement of network lifetime in WSNs. Wirel. Commun. Mob. Comput. (2021). https://doi.org/10.1155/2021/9601078

123. Zeng, F., Zhang, R., Cheng, X., Yang, L.: Channel prediction based scheduling for data dissemination in VANETs. IEEE Commun. Lett. **21**, 1409–1412 (2017). https://doi.org/10.1109/LCOMM.2017.2676766

124. Pratap, R., Zaidi, T.: Comparative Study of Task Scheduling Algorithms through Cloudsim. In: 2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018. pp. 397–400 (2018)

125. Zhou, K., Zhou, K., Yang, S.: Reinforcement learning-based scheduling strategy for energy storage in microgrid. J. Energy Storage. **51**, 104379 (2022). https://doi.org/10.1016/j.est.2022.104379

126. Chen, Z., Zhu, Y., Di, Y., Feng, S.: A dynamic resource scheduling method based on fuzzy control theory in cloud environment. J. Control Sci. Eng. **2015**, 1–10 (2015). https://doi.org/10.1155/2015/383209

127. Atef, S., Ismail, N., Eltawil, A.B.: A new fuzzy logic based approach for optimal household appliance scheduling based on electricity price and load consumption prediction. Adv. Build. Energy Res. **16**, 262–280 (2022). https://doi.org/10.1080/17512549.2021.1873183

128. Ismail, L., Materwala, H.: EATSVM: energy-aware task scheduling on cloud virtual machines. Procedia Comput. Sci. **135**, 248–258 (2018). https://doi.org/10.1016/j.procs.2018.08.172

129. Jiang, L., Sun, X., Mercaldo, F., Santone, A.: DECAB-LSTM: deep contextualized attentional bidirectional LSTM for cancer hallmark classification. Knowl.-Based Syst. **210**, 106486 (2020). https://doi.org/10.1016/j.knosys.2020.106486

130. Banerjee, S., Adhikari, M., Kar, S., Biswas, U.: Development and analysis of a new cloudlet allocation strategy for QoS improvement in cloud. Arab. J. Sci. Eng. **40**, 1409–1425 (2015). https://doi.org/10.1007/s13369-015-1626-9

131. Hespanha, J.P., Chinchilla, R., Costa, R.R., Erdal, M.K., Yang, G.: Forecasting COVID-19 cases based on a parameter-varying stochastic SIR model. Annu. Rev. Control. **51**, 460–476 (2021). https://doi.org/10.1016/j.arcontrol.2021.03.008

132. Suleiman, B., Sakr, S., Jeffery, R., Liu, A.: On understanding the economics and elasticity challenges of deploying business applications on public cloud infrastructure. J. Internet Serv. Appl. **3**, 173–193 (2012). https://doi.org/10.1007/s13174-011-0050-y

133. Madni, S.H.H., Abd Latiff, M.S., Abdullahi, M., Abdulhamid, S.M., Usman, M.J.: Performance comparison of heuristic algorithms for task scheduling in IaaS cloud computing environment. PLoS ONE **12**, e0176321 (2017). https://doi.org/10.1371/journal.pone.0176321

134. Vijaya Kumari, C., Aharonu, M., Sunil, T.: Energy efficient resource allocation in cloud computing. Int. J. Eng. Adv. Technol. **8**, 2071–2074 (2019). https://doi.org/10.35940/ijeat.F1394.0986S319

135. Liang, B., Wu, D., Wu, P., Su, Y.: An energy-aware resource deployment algorithm for cloud data centers based on dynamic hybrid machine learning. Knowl.-Based Syst. **222**, 1070 (2021). https://doi.org/10.1016/j.knosys.2021.107020

136. Shekhawat, V.S., Gautam, A., Thakrar, A.: Datacenter Workload Classification and Characterization: An Empirical Approach. 2018 13th Int. Conf. Ind. Inf. Syst. ICIIS 2018 - Proc. pp. 1–7 (2018). https://doi.org/10.1109/ICIINFS.2018.8721402

137. Dinda, P.A.: A prediction-based real-time scheduling advisor. Proc. - Int. Parallel Distrib. Process. Symp. IPDPS 2002. pp. 88–95 (2002). https://doi.org/10.1109/IPDPS.2002.1015480

138. Zhang, W., Li, B., Zhao, D., Gong, F., Lu, Q.: Workload prediction for cloud cluster using a recurrent neural network. In: Proc. - 2016 Int. Conf. Identification, Inf. Knowl. Internet Things, IIKI 2016. 2018-Janua, 104–109 (2018). https://doi.org/10.1109/IIKI.2016.39

139. Borkowski, M., Schulte, S., Hochreiner, C.: Predicting cloud resource utilization. In: Proc. - 9th IEEE/ACM Int. Conf. Util. Cloud Comput. UCC 2016. 37–42 (2016). https://doi.org/10.1145/2996890.2996907

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Shobhana Kashyap is a Research Scholar in Department of CSE in Dr. B. R. Ambedkar National Institute of Technology, Jalandhar Punjab, India since 2019. She has completed her master's from Thapar University, Patiala India. Her current research is concerned with Machine Learning and Cloud Computing. She has 2 year teaching experience and written papers for national and international conferences.

Avtar Singh is working as an Assistant Professor in Department of CSE in Dr. B. R. Ambedkar National Institute of Technology, Jalandhar Punjab, India. He received the B.Tech. and M.Tech. degree in Computer Science Engineering from the Electro Technical University, Saint Petersburg Russia (LETI) in 1999 and 2001 respectively. In 2001, he served IT industry in Bangalore and in 2006 joined academics in leading educational institutions. His research areas of interests include Cloud Computing, Internet of Things (IoT), Parallel and Distributed Computing and Machine Learning. He is a member of IEEE and ACM organization. He has written number of papers for national and international journals and conferences. He is currently guiding 4 research scholars at Ph.D. levels and 3 master's students. He has published extensively in these areas and has supervised 8 Master's students. He has organized four STC events.