Journal of Network and Computer Applications 204 (2022) 103405

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

Machine learning (ML)-centric resource management in cloud computing: A review and future directions

Tahseen Khan [a,b], Wenhong Tian [a,b,*], Guangyao Zhou [a,b], Shashikant Ilager [c], Mingming Gong [d], Rajkumar Buyya [e,a]

[a] School of Information and Software Engineering, University of Electronic Science and Technology of China, China
[b] Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China
[c] Vienna University of Technology (TU Wien), Austria
[d] School of Mathematics and Statistics, The University of Melbourne, Australia
[e] Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia

ARTICLE INFO

Keywords:
Resource management
Cloud computing
Data centres
Machine learning

ABSTRACT

Cloud computing has rapidly emerged as a model for delivering Internet-based utility computing services. Infrastructure as a Service (IaaS) is one of the most important and rapidly growing models in cloud computing. Scalability, quality of service, optimum utility, decreased overheads, higher throughput, reduced latency, specialised environment, cost-effectiveness, and a streamlined interface are some of the essential elements of cloud computing for IaaS. Traditionally, resource management has been done through static policies, which impose certain limitations in various dynamic scenarios, prompting cloud service providers to adopt data-driven, machine-learning-based approaches. Machine learning is being used to handle various resource management tasks, including workload estimation, task scheduling, VM consolidation, resource optimisation, and energy optimisation, among others. This paper provides a detailed review of machine learning-based resource management solutions. We begin by introducing background concepts of cloud computing like service models, deployment models, and machine learning use in cloud computing. Then we look at resource management challenges in cloud computing, categorise them based on various aspects of resource management types such as workload prediction, VM consolidation, resource provisioning, VM placement and thermal management, review current techniques for addressing these challenges, and evaluate their key benefits and drawbacks. Finally, we propose prospective future research directions based on observed resource management challenges and shortcomings in current approaches for solving these challenges.

1. Introduction

Cloud computing has created an environment in which consumers use software and IT infrastructure, paving the way toward the emergence of computing as a fifth utility (Buyya et al., 2018). Resource management in data centres remains a nontrivial issue in cloud computing, and it is directly dependent on the application workload. Applications were connected to specific physical servers in conventional cloud computing environments such as data centres, so these servers were often overprovisioned to handle issues related to maximum workload (Xu et al., 2017). As a result of the wasted resources and floor space, the data centre is expensive to operate in terms of resource management. On the other hand, virtualisation technology has proven that it can make data centres easier to handle. This technology offers a variety of benefits, including server consolidation and higher server utilisation. Large IT giants like Google, Microsoft, and Amazon have massive data centres with complicated resource management. Servers, virtual machines (VMs), and various management roles are all part of the resource management of these huge data centres (Bianchini et al., 2020). A server or a host is allocated multiple VMs with varying workload types in these data centres. This variable and unpredictable workload may result in a server being over-utilised and underutilised, resulting in an imbalance in resource utilisation assigned to VMs on a specific hosting server. This could lead to issues including inconsistent quality of service (QoS), unbalanced energy use, and service level agreements (SLA) violations (Singh and Kumar, 2019). According to a survey on unbalanced workload, the average CPU and memory utilisation was

* Corresponding author at: School of Information and Software Engineering, University of Electronic Science and Technology of China, China.
E-mail addresses: tahseen.khan240@gmail.com (T. Khan), tian_wenhong@uestc.edu.cn (W. Tian), guangyao_zhou@std.uestc.edu.cn (G. Zhou), shashikant.ilager@tuwien.ac.at (S. Ilager), mingming.gong@unimelb.edu.au (M. Gong), rbuyya@unimelb.edu.au (R. Buyya).
URL: http://www.buyya.com (R. Buyya).

https://doi.org/10.1016/j.jnca.2022.103405
Received 17 August 2021; Received in revised form 8 March 2022; Accepted 21 April 2022
Available online 6 May 2022

17.76% and 77.93%, respectively. A similar study in the Google data centre found that a Google cluster's CPU and memory utilisation could not exceed 60% and 50%, respectively (Kumar et al., 2020b). Due to the imbalanced workload, a data centre's productivity suffers, resulting in increased energy consumption. It is proportional to the data centre's operational costs and financial loss. This excessive energy consumption directly impacts carbon footprints, which should be reduced because an ideal machine absorbs more than half of the maximum energy consumption (Barroso et al., 2013). According to an EIA (Energy Information Administration) survey, data centres consumed around 35 TWh (Tera Watt-hour) of energy in 2015, and this figure is expected to rise to 95 TWh by 2040 (Kumar et al., 2021).

The resource use can be balanced by reducing the number of active servers; thus, the optimal mapping between VMs and servers must be discovered (Li et al., 2013). This is a challenging and NP-complete problem class. As a result, an intelligent resource management strategy is needed to meet QoS requirements and increase the data centre efficiency (Kumar and Singh, 2020). The intelligent mechanisms will generate future insights, which can aid applications in mapping to machines with better resource utilisation (Kumar et al., 2020a). However, the nonlinear and variable behaviour of VM workloads poses a significant challenge for future predictions. This insight can be obtained using two different approaches: historical workload-based prediction methods, which generate insight by learning trends from historical workload data, and homeostatic based prediction methods, which provide an upcoming future workload insight by subtracting the previous workload from the current workload (Kumar and Singh, 2018). Furthermore, the previous workload's mean may be static or dynamic. Both methods have advantages and disadvantages, but historical load-based forecasts are considered simpler and are well-known in this field.

The allocation of physical resources based on an estimate to increase resource utilisation and energy efficiency is known as resource provisioning. This estimation based on future resource behaviour prediction can help with more effective resource provisioning (Khan et al., 2021). Thus, intelligent resource management will play a critical role in optimising the data centre's SLA, energy usage, and operating costs by conducting effective and intelligent resource provisioning. Resource management in data centres encompasses a variety of activities, including resource provisioning, reporting, workload scheduling, and a variety of other functions, like thermal management (Ilager et al., 2020). Many of these activities revolve around resource provisioning. Resource provisioning aims to assign cloud resources to VMs based on end-user requests while maintaining a minimum of SLA violations, such as availability, reliability, response time limit, and cost limit (Shahidinejad et al., 2020). It should assign resources following end-user demands and prevent over or under-provisioning, such as allocating more or fewer resources to VMs. This resource allocation technique can be carried out in two ways: proactive and reactive. In proactive approaches, resource provisioning is focused on prior workload prediction, estimated by learning trends from historical workload, while reactive approaches are carried out after resource demand arrives. As a result, it is inferred that historical-based prediction methods' expertise can be effectively incorporated in proactive approaches to provide intelligent dynamic resource scaling, which contributes to intelligent dynamic resource management. In addition, other functions, such as VM consolidation and task scheduling, can be performed based on forecasts to optimise resource utilisation, energy consumption and increase QoS.

Machine learning (ML) techniques are widely used in a variety of fields, including computer vision, pattern recognition, and bioinformatics (Injadat et al., 2021). Large-scale computing systems have benefited from the advancement of machine learning algorithms (Mao et al., 2019). Google recently released a report detailing its efforts to optimise electricity, reduce costs, and improve efficiency (Jeff, 2018). ML has drawn attention to dynamic resource scaling by providing data-driven methods for future insights, regarded as a promising approach for predicting workload quickly and accurately. The use of ML on cloud computing platforms can be classified in five classes (Pop, 2016).

- Machine Learning environments from the cloud: Providers in this category provide computer clusters pre-installed with statistics software, such as R system, Octave, or Mapple, utilising public cloud providers like Amazon EC2, Rackspace, and others. Customers are relieved of the stress of installing and administering their clusters using these solutions, which provide scalable, high-performance resources in the cloud.
- Plugins for Machine Learning tools: In this class, participants may construct a Hadoop cluster in the cloud and conduct time-consuming operations over massive datasets using plugins for statistics programmes (e.g. R system, Python). The majority of the focus was directed toward R, which has several extensions, instead of Python, which has had less work to enable distributed processing until lately.
- Distributed Machine Learning libraries: This category contains complicated libraries that operate in various distributed configurations (Hadoop, Dryad, MPI). They enable users to utilise pre-built algorithms or create their own, then executed in parallel over a cluster of computers.
- Complex Machine Learning systems: Several business intelligence and data analytics solutions are presented in this class, all of which have a set of common features: All of them are deployable on-premises or in the cloud, offering a comprehensive set of graphical tools for analysing, exploring, visualising massive volumes of data, and using Apache Hadoop as a processing engine and/or storage environment. There are differences in how data is integrated, processed, and supported data sources and system complexity.
- Software as a Service provider for Machine Learning: This class focuses on providers of machine learning platform-as-a-service (or software-as-a-service). They primarily provide services through RESTful APIs, with the option of installing the solution on-premise (Myrrix) in some (rare) circumstances, as opposed to the solutions in the previous section, which are primarily deployable systems on private data centres. Predictive modelling (BigML, Google Prediction API, Eigendog) is the most popular class of ML issues among these systems.

The deployment of machine learning algorithms on clouds also offers various opportunities to use ML for more efficient resource management. As a result, this article focuses on the review based on challenges discovered in state-of-the-art research in resource management by using ML algorithms, including various resource management tasks such as provisioning, VM consolidation and other management approaches. Then we will discuss the advantages and limitations of various state-of-the-art research studies that use machine learning algorithms in resource management. We will also discuss the experimental settings, used data sets, and performance improvements. Finally, we propose future research directions based on identified challenges and limitations in current research. Fig. 1shows the cloud computing components while using machine learning. A resource management system (RMS) works with both users and ML prediction components to efficiently manage the cloud infrastructure's underlying resources. Data collection, ML models, training and validation of ML models, and eventually deployment of models for run time use are all components of the ML prediction module.

1.1. Aim and motivation of research

Resource management is a difficult task in cloud operations because multi-tenant end-users demand nonlinear workloads, leading to many over- and under-utilised servers. It directly affects whether electricity is over-or under-utilised, resulting in a high operating cost. As a result, intelligent resource management can benefit from a prior estimate of workload based on historical data. Static policies are often used in cloud computing systems to manage resources. They have two

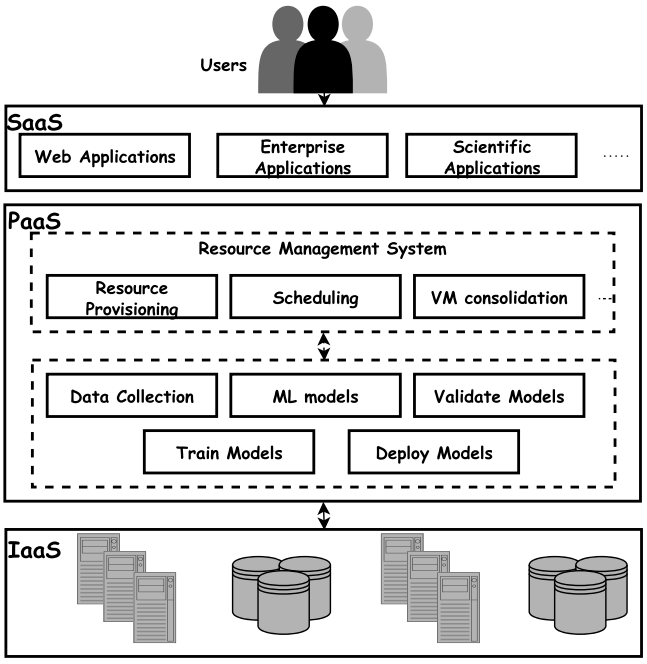

Fig. 1. A high-level view and components of resource management in cloud computing using machine learning.

flows: they are based on a static threshold value adjusted in offline mode. They appear to require reactive behaviour, resulting in excessive overheads and delayed customer responses. These strategies fail in a dynamic context, for example, when the load reaches the static threshold and rapidly drops, indicating that VM migration is unnecessary in the case of VM consolidation. Furthermore, they are unable to capture the dynamics of technology and workload in complex dynamic environments (such as Cloud and Edge) and therefore fail to move through (Ilager et al., 2020). Machine learning has supplanted static heuristics with dynamic heuristics that adapt to the actual production workload to address these disadvantages (Yadwadkar, 2018; Mao et al., 2016). Predictive management is made possible by machine learning techniques, which provide future insight based on historical data. As a result, A data-driven Machine Learning (ML) model in an ML-centric RMS can forecast future workload demand and control the auto-scaling of resources accordingly. Such strategies are highly beneficial for both consumers and service providers who want to improve their QoS and keep their competitive edge in the market. For cloud resource management, modern ML methods, such as a random forest (Cao et al., 2018) and neural networks (Chen et al., 2018), has been shown to make more reliable predictions than traditional time-series analysis methods, such as the Autoregressive Integrated Moving Average (ARIMA) model. Several ML algorithms have been developed to predict prior workload for intelligent resource management. Furthermore, several IT behemoths have begun to investigate machine learning-based resource management in production (Cortez et al., 2017; Gao, 2014). Google optimises fan speeds and other energy knobs using a neural network (Gao, 2014). Microsoft Azure makes use of a framework resource central to provide online forecasts of different workloads using various ML Gradient Boosting Trees (Bianchini et al., 2020). Despite these previous attempts and opportunities, the best way to incorporate machine learning into cloud resource management is currently uncertain. As a result, it has become critical to present research that addresses current challenges and suggests potential future research directions while also highlighting the benefits and limitations of current research. An abbreviation table has been shown in Table 1.

Table 1
List of abbreviations used in this paper.

Abbreviations	Full description
ML	Machine Learning
RMS	Resource Management System
IaaS	Infrastructure as a Service
QoS	Quality of Services
VMs	Virtual Machines
IT	Information Technology
SLA	Service Level Agreements
EIA	Energy Information Administration
TWh	Tera Watt-hour
NIST	National Institute of Standards and Technology
AWS	Amazon Web Services
SaaS	Software as a Service
PaaS	Platform as a Service
AI	Artificial Intelligence
SSL	Semi-Supervised Learning
RL	Reinforcement Learning
MAE	Mean Absolute Error
LLC	Last-Level-Cache
RC	Resource Central
DL	Deep Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network

1.2. Research questions

- How to reduce the time complexity of ML algorithms in ML-based resource management in data centres?
- How can the accuracy of workload prediction using ML algorithms be improved?
- How can training time be reduced while developing an ML model?
- How can VMs collaborate in similar groups to estimate the state of energy consumption?
- How to reduce energy consumption significantly?

1.3. Our contributions

The following are the main contributions of our work:

- We present a review of ML-based resource management approaches in cloud computing based on identified challenges in the state-of-the-art research.
- We identify the advantages and drawbacks of these methods, as well as their experimental configuration, data sets used, and performance improvements.
- We propose potential future research directions based on identified challenges and limitations in the state-of-the-art research to strengthen resource management

1.4. Related surveys

A few studies have investigated machine learning-based resource management in cloud computing. Sun et al. (2016) provided a detailed survey of the most important data centre resource management research activities to improve resource usage. After that, the article summarises two major components of the resource management platform and addresses the benefits of predicting workload accurately in resource management. Manvi and Shyam (2014) focused on resource provisioning, resource allocation, mapping, and resource adaptation, among other essential resource management techniques. Zhang et al. (2016) surveyed the state of the algorithms, organised them into categories, and addressed closely related topics such as virtual machine migration, forecast methods, stability, and their availability. Braiki and Youssef (2019) considerable improvements to previous work based on approach optimisation, techniques, and objective models. Jennings and

Table 2
A comparison with relevant existing surveys.

Study	Year	Domain	Key contributions	Challenge	Shortcomings of existing solutions	ML-centric	ML-based Future directions
Manvi and Shyam (2014)	2014	Resource Management	Examines some of the essential resource management approaches, including resource provisioning, resource allocation, resource mapping, and resource adaptability	✓	×	×	×
Jennings and Stadler (2015)	2015	Resource Management	Evaluates the recent literature, encompassing over 250 papers and emphasising major findings	✓	×	×	×
Sun et al. (2016)	2016	Resource Management	Provides a detailed assessment of the most significant research activity on data centre resource management that attempts to maximise resource use	×	×	✓	×
Usmani and Singh (2016)	2016	VM placement	Provides an in-depth examination of cutting-edge VM placement and consolidation approaches	×	×	×	×
Zhang et al. (2016)	2016	Resource Provisioning	Surveys more than 150 articles	×	✓	×	×
Braiki and Youssef (2019)	2019	Resource Management	Presents substantial solutions to previous work proposed for cloud infrastructure	×	×	×	×
Helali and Omri (2021)	2021	Consolidation	Investigates the issue of consolidating data centres inside distributed cloud platforms	×	×	×	×
Nayak et al. (2021)	2021	Resource Management	Represents a short review on renewable energy-based resource management	×	×	×	×
Dewangan et al. (2021)	2021	Resource Management	Provides a thorough examination of several resource providing systems using concerted parameters	✓	✓	×	×
Mijuskovic et al. (2021)	2021	Resource Management	Addresses challenges in resource management and classifies current contributions	✓	×	×	×
Our study	2021	Resource Management	Provides a detailed review of machine learning-based resource management solutions	✓	✓	✓	✓

Stadler (2015) lays forth a conceptual framework for cloud resource management and uses it to organise the state-of-the-art review. Usmani and Singh (2016) presented a detailed assessment of the most up-to-date VM placement and consolidation techniques utilised in the green cloud, focusing on increasing energy efficiency. Helali and Omri (2021) presented a broad overview of IT consolidation at various levels of cloud services and a virtualised data centre and consolidation overview.

A summary of related works is given in Table 2 from which we observe that related works do not go into great detail about machine learning-based resource management, nor do they go into great detail about the challenges and issues that exist in the existing state-of-the-art and future research directions. As a result, it is now important to present a thorough survey that addresses various machine learning algorithms used in the resource management scenario for a data centre and their shortcomings, challenges, and potential directions, as per our vision. Hence, this article can help researchers evaluate the current machine learning scenarios in cloud resource management and their shortcomings before moving forward with their new ideas in this direction.

1.5. Article structure

The rest of the paper is organised as follows: The background details and definitions for cloud computing components and machine learning are given in Section 2. Section 3 discusses the challenges of machine learning-based resource management in cloud computing systems and the benefits and drawbacks of current research. Section 4 proposes future research directions based on the challenges and limitations pointed out in state-of-the-art research, and Section 5 concludes the paper.

2. Background and terminologies

2.1. Cloud computing

Cloud computing provides resources over the Internet, such as memory, CPU, bandwidth, disc, and applications/services. The National Institute of Standards and Technology (NIST) (Mell, 2011) states that "Cloud computing is a model for providing on-demand network access to a common pool of configurable computing resources (e.g., networks, servers, storage, software, and services) that can be quickly provisioned and released with minimal management effort or service provider involvement. There are five core features, three service models, and four deployment options in this cloud model". Based on the literature, two more characteristics have been included.

This computing model uses a client–server architecture to centralise application deployment and computation offloading. Cloud computing is cost-effective in application delivery and maintenance on both the client and server sides and flexible in resource provisioning and detaching services from related technologies. Cloud computing and

its supporting technology have been investigated for years. Many advanced computing systems have been released to the market, including Alibaba Cloud, Microsoft Azure, Adobe Creative Cloud, ServerSpace, Amazon Web Services (AWS), and Oracle Cloud.

2.2. Core features of cloud computing

- On-demand self-service: A client can query one or more services as needed and pay using a "pay-and-go" system without interacting with living beings via an online control centre.
- Broad network access: Resources and services in different cloud provider areas can be accessed from several locations and provisioned by incompatible thin and thick clients using standard mechanisms. This trait is often referred to as "easy-to-access standardised mechanisms" and "global reach capability" (Hamdaqa and Tahvildari, 2012; Yakimenko et al., 2009).
- Resource pooling: It offers a set of resources that act as if they were one blended resource (Wischik et al., 2008). In other words, the client is not aware of the location of the provided services and is not expected to be. This strategy enables vendors to dynamically include various real or virtual services in the cloud.
- Rapid elasticity: Elasticity is just another word for scalability; it refers to the ability to scale resources up or down as required. Clients can demand as many services and resources as they want at any time. Because of this consistency, Amazon, a well-known cloud service provider, named one of its most popular and commonly used services the Elastic Compute Cloud (Amazon, 2010).
- Measured service: Various facets of the cloud should be automatically controlled, monitored, optimised, and documented at several abstract levels for both vendors and customers (Krishnaveni et al., 2021).
- Multi-Tenacity: The Cloud Security Alliance proposes this idea as the fifth cloud characteristic. Multi-tenacity implies that models for policy-driven compliance, segmentation, separation, governance, service levels and chargeback/billing for various customer categories are needed (Espadas et al., 2013).
- Auditability and certifiability: Services must plan logs and trails to assess the degree to which laws and policies are followed (Hamdaqa and Tahvildari, 2012).

2.3. Cloud computing service models

- Software as a Service (SaaS) (Piraghaj et al., 2017): Using this service model, a client can access the service provider Cloud-hosted applications. Web portals are used to access applications. Since providers have access to the applications, this model has made production and testing easier for them.
- Platform as a Service (PaaS) (Jula et al., 2014): In this service model, the service provider provides basic requirements including network, servers, and operating system to enable the client to build acquired applications and manage their configuration settings.
- Infrastructure as a Service (IaaS) (Whaiduzzaman et al., 2014): The user has created all necessary applications and only requires a simple infrastructure. Vendors may include processors, networks, and storage as facilities with customer provisions in such cases.

2.4. Deployment models for cloud computing

- Public cloud (Toosi et al., 2014): This is the most popular cloud computing model, in which the cloud owner, in the majority of cases, provides public services over the Internet based on predetermined rules, regulations, and a business model. With a significant number of commonly used resource base, providers can provide consumers with a range of choices for choosing appropriate resources while maintaining QoS.
- Private cloud (Jadeja and Modi, 2012): A private cloud is created and configured to provide a company or institute with most of the advantages of a public cloud. Setting up such a system would result in fewer security problems due to corporate firewalls. The high costs of establishing a private cloud are fatal because of the business that manages it is accountable for all facets of the scheme.
- Community cloud (Dillon et al., 2010): A variety of organisations form a group and share cloud computing with their community members' customers based on common criteria, concerns, and policies. The required cloud computing infrastructure can be provided by a third-party service provider or a group of community members. The most important benefits of a community cloud are cost savings and cost-sharing among community members, and high protection.
- Hybrid cloud (Tuli et al., 2020): Combining two or more independent public, private, or community clouds resulted in the creation of a new cloud model known as hybrid cloud, in which constituent services and infrastructure maintain their special features while also requiring standardised or agreed-upon functionalities to enable them to communicate in terms of application and data interoperability and portability.

2.5. Machine learning

Machine learning covers the subject of how to design machines that improve the performance of themselves automatically via experience. It is one of today's most rapidly expanding technological topics, located at the crossroads of computer science and statistics, as well as at the heart of artificial intelligence and data science (Jordan and Mitchell, 2015). One of its basic assumptions is that it is possible to construct algorithms that can predict potential, previously unseen values using training data and statistical techniques. Machine learning has come a long way in the last two decades, from a research project to a widely used commercial technology. In particular, recent advances in machine learning stem from deep learning (DL), which is part of a broader family of machine learning approaches based on artificial neural networks. Compared to shallow learning, deep learning approaches are able to automatically learn high-level abstract representations from raw data, which reduces the efforts in feature engineering and improves the prediction performance significantly. Machine learning/deep learning has emerged as the preferred tool for designing functional apps for computer vision (Janai et al., 2020), speech recognition (Deng and Li, 2013), natural language processing (Olsson, 2009), robot control (Chin et al., 2020), self-driving cars (Stilgoe, 2018), effective web search (Bhatia and Kumar, 2008), purchase recommendations (Hastie et al., 2009) and other applications in the field of artificial intelligence (AI). Many AI system developers now understand that, for many applications, training a system by showing it examples of desired input–output actions is much simpler than programming it manually by predicting the desired answer for all possible inputs. This success is primarily owing to sophisticated model architectures, efficient optimisation techniques, accessibility of massive data, and increased efficiency in the processing power of servers and GPUs (Goodfellow et al., 2016). According to the supervision information provided in the learning process, machine learning can be roughly categorised as supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms take labelled data (feature–label pairs) as input and outputs a model that could predict the labels of future features. Typical supervised learning approaches include regression, classification, and ordinal regression, categorised by the type of labels. Unsupervised learning aims to learn the data distribution of unlabelled data via discrete mixture models (clustering (Hartigan and Wong, 1979; Guha et al., 2000)) or continuous latent factor models (dimension reduction (Ding et al., 2002; Kingma and Welling, 2013)). Semi-supervised learning aims to learn models from

both labelled and unlabelled data. Reinforcement learning is concerned with how intelligent agents ought to make decisions in an environment to maximise the notion of cumulative reward. A detailed explanation of these learning problems is as follows.

- Supervised Learning (Sen et al., 2020): Every data sample in supervised learning is made up of several input features and a label. The learning process is designed to get close to a mapping function that links the features to the label. Following that, the mapping function can be used to make predictions of the label for the data given new input features. This is the most widely used machine learning scheme, and it has been used for a lot of things. The classification task, which involves classifying an object based on its characteristics. This is a regression task if the supervised learning task is to forecast a continuous variable.
- Unsupervised Learning (Celebi and Aydin, 2016): Unsupervised learning in comparison to supervised learning is when we only have input features but no labels to go with them. As a result, the purpose of unsupervised learning is to learn the data distribution and demonstrate how the data points vary from one another. The clustering problem, which is to discover data groupings, such as grouping VMs based on their resource use patterns (Khan et al., 2022), is a good example of unsupervised learning.
- Semi-supervised learning (Van Engelen and Hoos, 2020): A branch of machine learning attempts to integrate these two activities. SSL algorithms usually try to increase efficiency in one of these two tasks by incorporating knowledge from the other. For example, when dealing with a classification problem, additional data points with unknown labels may help in the classification process. On the other hand, knowing that some data points belong to the same class will help with the learning process for clustering methods.
- Reinforcement Learning (Kober et al., 2013): RL varies from supervised and unsupervised learning in several ways. When using reinforcement learning to train an agent, it is unnecessary to use labelled input/output pairs or explicit correction on sub-optimal options. Instead, the agent attempts to find an equilibrium between exploration and exploitation by interacting with the environment. The translator rewards the agent for successful decisions or behaviour. Otherwise, it would be sanctioned.

3. ML-centric resource management: State-of-the art and challenges

In this section, we discuss challenges identified in ML-based resource management in state-of-the-art research. In addition, we explore current approaches to addressing these challenges and their advantages and limitations. We categorise these challenges based on their types, such as resource provisioning, VM consolidation, thermal management, and workload prediction.

3.1. Workload prediction

3.1.1. ML in energy consumption prediction

Most cloud service providers' tools calculate and estimate the energy usage of a host or a group of hosts in offline mode, but performing this role in real-time running applications is a challenge. Furthermore, because of the nonlinear workload in various hosts, a single ML algorithm cannot be considered to perform this task well. According to Reiss et al. (2012), a Google cluster or node does not use more than 60% and 50% of its CPU and memory, respectively. As a result, ensemble learning can be a key component of providing accurate predictions in a cloud architecture.

Subirats and Guitart (2015) introduced an ensemble learning method for forecasting future energy efficiency in virtual machine resources, such as CPU utilisation, infrastructure, and service levels in a cloud computing environment. Ensemble learning, which uses four different prediction approaches such as moving average, exponential smoothing, linear regression, and double exponential smoothing, is the key benefit of their work. They predict the next use of VM resources, such as CPU consumption, in each time iteration and calculate the mean absolute error (MAE) of all iterations to pick the best performing model predictions for measuring and forecasting energy efficiency and ecological efficiency in an IaaS setting in real-time. However they do not consider metrics like Last-level-cache (LLC) and disc throughput for prediction, which have an effect on a host's energy consumption at the VM level as mentioned in Sayadnavard et al. (2021). Furthermore, the accuracy of the chosen model is workload-specific, i.e., interactive and batch workloads, rather than being generalised for all data.

3.1.2. Performance and online profiling of workload

The main components of large commercial providers' workloads are not well addressed in cloud resource management research. For example, they do not look into VMs' lifetime virtual resource consumption. The majority of research focuses on offline workload profiling, which is infeasible because the input workload may not be available until the VMs are not running in production. On the other hand, online profiling is challenging because it is difficult to determine when a random VM has exhibited representative behaviour as mentioned by Bianchini et al. (2020). Resource management can be more effective if the different workload characteristics are accurately predicted with minimal time complexity. As a result, prediction algorithms face another challenge in terms of accuracy and time complexity.

On Microsoft Azure compute fabric, Bianchini et al. (2020) presented a machine learning-based prediction system. Through a rest API, this system can learn behaviour from historical data and provide predictions online to various resource managers, such as Server health managers, migration managers, Container schedulers, and energy capping managers. They also released detailed Microsoft Azure real-world workload traces from this system, which show that several VMs consistently have peak CPU utilisation in various ranges. In the event of oversubscribed servers, they changed Azure's VM scheduler to use RC benefit predictions.

This forecast-based schedule helps to avoid overuse and exhaustion of physical resources. However, (1) they did not consider memory utilisation in released traces or the predictive system RC, even though memory utilisation plays a significant role in physical resource exhaustion. (2) They analysed CPU utilisation time series to determine whether a VM is interactive or delay-insensitive, categorised the workload into these two categories, and used Extreme Gradient Boosting Tree (EGBT) to perform supervised classification of these VM workloads. However, they did not consider the case of a distributed data centre, where data is dispersed and may have partial labels for these two classes; in this case, there will be insufficient labels to train this algorithm.

3.1.3. Prediction accuracy in auto-scaling of web applications

Auto-scaling determines when and how resources are allocated for cloud-based applications (Persico et al., 2017). Auto-scaling is done in two ways: reactive and proactive. The reactive approach allocates resources when system events such as CPU utilisation, number of requests, and queue length exceed a fixed threshold. The proactive approach is in charge of anticipating the number of resources required ahead of time to avoid unneeded events. Furthermore, proactive strategies include predictions based on traditional statistical time-series analysis models, which do not fit all cases in terms of prediction accuracy, making it a challenging task. Furthermore, traditional statistical methods have the following drawbacks: (1) It is based on rule-based programming, which is formalised as a relationship between variables; (2) It is often based on a dataset consisting of a few attributes, as the methods are not scalable to high-dimensional data; (3) It relies on assumptions like linearity, normality, no multicollinearity,

homoscedasticity, and so on; (4) The majority of the ideas in traditional statistics are generated from the sample, population, and hypothesis; (5) It is a math-intensive subject that relies on the coefficient estimator and necessitates a thorough knowledge of a dataset.

Messias et al. (2016) used a genetic algorithm to combine the advantages of individual ML models to obtain the best performing prediction results for web application auto-scaling. Each time-series prediction model used in the system is fitted with a suitable weight using a genetic algorithm. The primary benefits of their work are that (1) Auto-scaling can adapt to any new workload as its characteristics change over time. (2) This approach is unaffected by the type of prediction models used. (3) It is simple to adapt to various more advanced prediction models. However, this approach has a high time complexity, affecting the response time of any web application hosted in cloud infrastructure, violating SLAs.

3.1.4. Time-series prediction data

The workload in modern data centres follows a time-series pattern. As a result, models for time series prediction should be trained on historical data, as it is presumed that future trends would be identical to those seen previously. However, data centres experience very nonlinear workload variations, which is why new trends often emerge, making it difficult for the model to learn precisely. Due to the lack of a single model suitable for all types of time series prediction data, an ensemble approach is being used to address this issue (Wolski, 1998). Furthermore, most ensemble models for time series prediction are based on a collection of fixed predictors, either homogeneous or heterogeneous, which makes it difficult for the models to learn pattern change in time series prediction.

Cao et al. (2014) suggested a new ensemble method that can dynamically update the predictors in the ensemble approach to respond to trend changes in time-series prediction quickly. The ensemble method dynamically adjusts the models, which is the key benefit of this work. It is adaptable, as new models can be quickly added and removed depending on how well it performs with a nonlinear workload. They set a threshold value of 5 and a floor limit of 0 to determine which predictor is performing well and which is not. Every predictor is given a score, which rises and falls in response to the predictor's results. This predictor is selected as a representative predictor if its score exceeds the threshold value and is discarded if it meets the floor limit. On the other hand, these fixed parameters yield satisfactory results for their chosen dataset, resulting in a non-generalised approach.

3.1.5. Training data

In modern cloud environments, virtual resources such as virtual CPUs (vCPUs) and memory (vRAMs) have a nonlinear resource demand, resulting in complex resource utilisation behaviour. As a result, optimisation of virtual resource performance is required with this high amount of daily workload. Large corporations such as Amazon, Alibaba, and others have occasionally failed due to a lack of resource management planning. As a result, predicting virtual resources (such as vCPU and vRAM) is a challenging task. Furthermore, resource forecasting presents some challenges: (1) The prediction of these resources should be dynamic to respond to changing workload patterns over time; (2) The data for training should be chosen in such a way that it has the most significant impact on the target variable so that the model can learn to predict it effectively.

(Shyam and Manvi, 2016) It proposed a model that took into account a variety of parameters in a virtualised platform to reliably predict virtual resources with the least amount of SLA violations. This method was based on a Bayesian approach that identified various variables and considered the best training data. The key benefit of their work is that it systematically detects dependencies in a variable based on the study of nonlinear workloads from various data centres such as Amazon, EC2, and Google. However, (1) they do not take into account the combination of several application types, (2) Since it relies on the dependencies of a specific problem, this approach lacks generalisation, (3) For prediction, this method ignores high-level metrics, including transaction throughput and latency of underlying resources, such as vCPU cores (see Tables 3–5).

3.2. Runtime VM management

3.2.1. Multiple resource usage in VM consolidation

VM consolidation approaches attempt to consolidate more VMs on fewer hosts to turn off the remaining hosts and save energy. Most researchers used current CPU utilisation to determine whether a host was overloaded or not in this process. This may result in unnecessary VM migration and host power mode transition, lowering the consolidation process' efficiency. The destination host for migrating VMs is the host with the highest CPU utilisation, but due to the lack of future estimation, this may result in overutilisation. As a result, future resource utilisation estimation can address this issue. Aside from CPU utilisation, other resource consumption, such as memory and disc, can cause the host to become overloaded, making the consolidation process challenging.

Haghshenas and Mohammadi (2020) proposed an intelligent VM consolidation technique to reduce energy consumption. Based on historical data, this technique predicted resource utilisation in the past and used that prediction to choose a host with higher utilisation in advance for VM migration. A dynamic consolidation procedure was used to address this issue. To predict the future usage of all VMs, a machine learning method called Linear Regression (LR) was used. This task was carried out using real workload traces from PlanetLab VMs (Chun et al., 2003). They used the CloudSim toolkit (Calheiros et al., 2011) to model a data centre and implement their VM migration strategy to save energy. Their work had the main benefit of taking into account time overheads while lowering energy consumption on a larger simulated benchmark with 7600 hosts. However, if this approach is used in real-world workload production, the time overhead is a significant factor that is also affected by the ML algorithm's data training time. However, they considered the LR method, which relies on various features to predict the target variable, may make it time-consuming and potentially affect the data centre's response time.

3.2.2. Multi-dimensional resource requirement

Flexible resource provisioning frameworks are needed in cloud data centres to manage host load based on various requirements. As a result, data centres conduct dynamic resource provisioning, which uses prediction models to estimate the number of resources needed in advance for varying workloads over time. It aims to predict future VM request workloads by looking at previous usage trends. However, since VM requests include a variety of virtual resources such as CPU, memory, disc, and network throughput, it is extremely challenging and complex to forecast demand for each form of resource separately. In the case of choosing an ML prediction model, the multi-resource existence of a VM presents a specific challenge. Furthermore, different cloud users can make different requests for cloud resources. As a result, forecasting the demand for each form of resource is difficult and impractical.

(Ismaeel and Miri, 2015) They have proposed a model for dividing VM clusters into different categories and then developing prediction models for each cluster. The key benefit of their work is that (1) they use Extreme Learning Machines (ELMs) (Darges et al., 2022), which can find the best weight for the predictor in a single step. (2) They avoid issues like stopping conditions, learning rate selection, learning epoch scale, and local minimums of gradient-based learning methods like NN and ANFIS using ELMs. (3) As it deals with nonlinear processes, this work can handle the linear behaviour of the LR method. (4) It predicts VM requests in each cluster using a single network. (5) Every cluster can have its prediction network. However, in K-means clustering, they set the number of clusters to 3, resulting in a model with a fixed number of VM clusters.

Table 3
State-of-art research: A summary of experimental configurations, data sets and their targeted performance improvement.

Study	Experiments configuration	Dataset	Performance improvement
Garg et al. (2014)	Simulation using CloudSim with 1500 physical nodes	Grid Workload Archive (GWA) (Iosup et al., 2008) and PlanetLab (Chun et al., 2003)	Reduces the number of servers utilised by 60% compared to other strategies
Yang et al. (2014)	Simulation using real VM workload	NASA NPB, IOzone and Cachebench	Predicts VM power usage with an average error of 5% and 4.7% compared to actual power measurement models
Calheiros et al. (2014)	Simulation using CloudSim with 1000 hosts	Wikimedia Foundation	Achieves efficiency in resource utilisation up to 91% guaranteeing QoS
Cao et al. (2014)	Colected CPU load from 12 different hosts	Private cloud environment	Improvement in prediction by 4.81%, 5.92% and 7.37% for BEST MRE, 50% MREs and 80% MREs
Ismaeel and Miri (2015)	Experimentation using real workload VM traces	Google Cluster data (Reiss et al., 2011)	Produces lower RMSE value than other approaches
Subirats and Guitart (2015)	Experimentation for predictions for a different types of workloads	Workloads generated using SPECweb2005	It improves the precision of the forecasts of the energy efficiency while running different workload types benchmark
Verma et al. (2016)	Simulation using two data centres and three hosts per data centre	8 VMs in modelled data centres in CloudSim	Significant allocations of VMs to the host with full capacity
Messias et al. (2016)	Experimentation using real web logs	FIFA world cup 98 Web servers (Arlitt and Jin, 2000), NASA Web servers and ClarkNet Web server (Arlitt and Williamson, 1997)	Significant prediction results
Shyam and Manvi (2016)	Simulation using SamIam Bayesian network	Amazon EC2 and Google CE data centres	Workload predicted with accuracies greater than 80%
Nguyen et al. (2017)	Google Cluster Data PlanetLab	Simulation using CloudSim with 800 hosts	Significantly reduces energy consumption and VM migrations
Shaw et al. (2019)	Simulation using CloudSim with 800 hosts	PlanetLab (Chun et al., 2003)	Reduces energy up to 18% and service violation up to 34% compared to its baseline
Bianchini et al. (2020)	Online experimentation using real VM traces	Microsoft Azure Trace (Cortez et al., 2017)	Significant prediction accuracies for different workload
Haghshenas and Mohammadi (2020)	Simulation using CloudSim with 7600 hosts	PlanetLab (Chun et al., 2003)	It reduces the energy consumption up to 38% compared to other work. It takes 5% less time overhead to execute for a modelled data centre
Ilager et al. (2021)	Simulation using CloudSim with 75 hosts	Private cloud data from University of Melbourne	Reduces peak temperature by 6.5 °C and consumes 34.5% less energy compared to its baseline

3.2.3. Energy metering at software-level

Modern servers have multiple energy metres to monitor energy usage. Still, they are unable to monitor the energy of a single virtual machine, which is difficult to do since measuring power at the software level is difficult. And, according to the energy budget in data centres, energy consumption has become a difficult factor to consider for a successful VM consolidation phase. The previous study only looked at server resource utilisation for VM consolidation, which contradicted the energy capping mechanism by increasing across the levels of specific servers during the process, which violated energy constraints. The term "energy capping" refers to a process introduced at the hardware level. As a result, lowering the CPU frequency reduces the energy consumption of the combined server, which violates the energy constraints. Hence, lowering the server's CPU frequency due to the load of one VM affects all other operating VMs at the same time. Therefore, the efficiency in workloads running in VMs degrades, breaching SLAs and the isolation property of virtualisation. VM consolidation and energy capping are the two most common methods in data centres, but neither allows for accurate energy usage monitoring for individual VMs.

Yang et al. (2014) proposed the iMeter energy consumption prediction model, which is based on the Support Vector Regressor (Smola and Schölkopf, 2004) machine learning method (SVR). They used principal component analysis (PCA) to identify the most associated components that influenced VM energy consumption and projected individual VM and multiple consolidated VM energy consumption for various workloads. However, predicting the energy consumption of a single VM is difficult due to the various types of cloud resources residing in the VM, such as CPU, memory, and IO, and the fact that different cloud end users can demand other volumes of the same resources at the same time. Furthermore, the resource manager must make individual decisions for VMs, which slows down end-user response time and violates QoS.

3.2.4. Usage level management

The current resource utilisation prediction causes unreliable overloaded host detection, especially when a current resource utilisation exceeds a threshold value. The challenge arises in deciding whether VMs allocated to this host should be migrated because the load decreases rapidly after a very short period, leading to a false hot detection point, i.e., false overloaded host detection. However, when the duration of load degradation is large enough, VMs need to be migrated to avoid over utilisation. Such kind of VM consolidation mechanism poses a unique challenge to the resource management system to prevent unnecessary VM migration overhead.

Nguyen et al. (2017) proposed a VM consolidation strategy based on multiple usage prediction and multi-step prediction to limit unnecessary VM migrations to avoid overheads and wasted energy consumption in data centres. Thus, this mechanism was computed to estimate the

Table 4
State-of-art research: Objectives, Advantages and Limitations.

Study	Objectives	Advantages	Limitations
Bianchini et al. (2020)	Online profiling of workload	Predictions are provided online	Memory use is not taken into account, nor is the case of distributed data centres
Haghshenas and Mohammadi (2020)	VM consolidation	Time overhead is considered	Prediction relies on multiple features
Shaw et al. (2019)	VM placement	Dynamic VM placement based on CPU utilisation and network bandwidth	Disc throughput is not considered
Ilager et al. (2021)	Thermal management	Peak temperature is reduced significantly	Algorithm overhead
Nguyen et al. (2017)	VM consolidation based on multiple resource usage	Combination of current and future resource utilisation is considered	Overloaded host in the current period is not taken into account
Yang et al. (2014)	Energy consumption prediction	Energy metering at software-level, i.e., VM-level	Decision could be taken for an individual VM only based on predicted energy consumption in RMS
Garg et al. (2014)	Resource management strategy based on SLAs	Historical CPU utilisation data with SLA penalties is used	Deviation of prediction from actual value, Highly nonlinear workload is not considered
Calheiros et al. (2014)	QoS aware workload prediction	Predicted requests are considered to provision VM dynamically	Future estimation is provided for a static time-interval
Verma et al. (2016)	Resource demand prediction and provision strategy	Classification of service tenants based on a binary problem	Information of how binaries are assigned to service tenants are not available; assigning binaries could be time-consuming, Supervised classification could have some limitations in case of partially labelled data
Subirats and Guitart (2015)	Energy consumption prediction based on ensemble learning	Ensemble learning is considered	Last-level-cache (LLC), disc throughput are not considered, Accuracy is workload specific
Messias et al. (2016)	Auto-Scaling of web applications	Auto-scaling can adapt to any new workload, Independent of type of prediction models, and It can adapt more advanced prediction models	High time complexity
Cao et al. (2014)	Time-series prediction	Ensemble approach can dynamically adjust the models	Non-generalised approach
Shyam and Manvi (2016)	Prediction of virtual resources	Detection of dependencies comprehensively in a variable based on the analysis of nonlinear workloads	Combination of several application types is not considered, Non-generalised approach, Transaction throughput and latency are not taken into account
Ismaeel and Miri (2015)	VM categorisation	Use of Extreme learning machines (ELMs) can deal with nonlinear processes, Use a single network for prediction. Every cluster can have its network for prediction	Static number of VM clusters.

long-term utilisation of several resources such as CPU memory based on the historical data for a particular PM. In VM consolidation, the main task is to detect overloaded and underloaded hosts. Thus, they considered current and predicted resource utilisation to identify the overloaded and underloaded hosts. An efficient multiple usage prediction algorithms was presented to compute the long-term utilisation of different resource types based on local historical data. Furthermore, a VM consolidation based on multiple usage prediction was proposed to reduce energy consumption by limiting the unnecessary VM migrations from overloaded hosts. Hence, the current and predicted resource utilisation combination plays an important role in reliable overloaded and underloaded host detection. According to this, a host is considered overloaded if it follows two constraints: (1) if the host is overloaded in both current and predicted resource utilisation, and (2) if the host is in normal condition and will be overloaded in a future period. And VM consolidation was performed based on the detected overloaded hosts by following these two constraints. However, they did not consider the case. If a host is overloaded in a current period but will not be overloaded in the future period, then what about the overloaded host in the current period. This point should be considered in the VM consolidation scheme.

3.3. VM placement

3.3.1. Cloud network traffic

The current research in VM allocation involves many solutions to allocate a single VM to a host and allocate various VM resources by ensuring that every host has sufficient capacity to run the workload. This approach leads to inefficient resource utilisation as the application workload varies from time to time with a mix of high and low resource utilisation. The challenges arise when different applications exhibit different resource demands and are allocated to suitable VMs in data centres that cause varying resource demand patterns. Moreover, many VM placement solutions consider only current resource utilisation like CPU demands. However, varying workload continuously poses a challenge to such solutions. Future resources like CPU demand can be more effective for VM placement strategies. In addition to CPU resource demand, cloud network bandwidth is also another challenging factor inefficient resource management in data centres (Genez et al., 2015; Duggan et al., 2017). As Networking (2016) reported that there will be 51,774 GB/s amount of internet traffic would be produced because of computing as a service via cloud computing and this would affect cloud networks as well. And this key factor affects the VM migration time in case of dynamic VM placement and violates SLAs (Verma et al., 2008).

Table 5
Machine learning-centric resource management challenges and future research directions.

X	Challenges (Section 3.X)	Future Research Directions (Section 4.X)
1	Online profiling of nonlinear workload, Prediction accuracy, Time complexity	More precise estimate of prior workload using advanced ML models, Prediction of memory utilisation in physical resource exhaustion along with CPU utilisation, Semi-supervised classification in categorising VMs
2	Excessive VM migrations, Host overutilisation, Memory and disc utilisation in VM consolidation	Overloaded host detection based on the combination of CPU, memory, and bandwidth utilisation, Workload prediction using DL methods like LSTM, GRU, etc.
3	Non-linear resource utilisation, Various resource demands patterns, Cloud network bandwidth	Consideration of disc throughput along with CPU and bandwidth utilisation in VM placement heuristics
4	To cool down the host, Cost of the cooling system, Thermal management	Prior CPU estimation-based resource provisioning, Use of GRU for inlet temperature prediction only
5	Rapid degradation of load, Unnecessary VM migration overhead	Development of ML algorithm with dynamic resource utilisation threshold
6	Prediction of energy consumption at VM level, Performance degradation due to lower the CPU frequency of server	Prediction of energy consumption state at VM level using clustering analysis
7	Resource wastage, Resource prediction in the presence of computationally intensive applications	To involve current and future requirements of resources like CPU, memory and bandwidth and SLAs such as compute-intensive non-interactive jobs and transactional applications in VM dynamic consolidation, To consider a combination of provisioned and utilised resources like CPU and memory in dynamic resource provisioning
8	High utilisation, The exact number of resources in the presence of varying load	To deal with reactive approaches in resource provisioning, Adhoc decisions in dynamic resource provisioning, To predict the peak utilisation of resources using different ML models, Ensemble learning, Estimating future web requests with a dynamic time interval
9	To obtain historical data, Amount of resources, Varying resource requirements	To classify service tenants using clustering or semi-supervised clustering
10	Prediction of energy consumption in real-time production	To consider memory, disc, and network components system in energy consumption prediction, To inspect nonlinear relationships such as polynomial or exponential between virtual resource and energy consumption, Combine information provided by an individual model, To keep track of the value of the parameter of each model from the record, To feed the ML model with average workload performance for training
11	Prediction accuracy in proactive approaches, Limitations of statistical learning over machine learning	To use ML methods to forecast workload instead of statistical methods, To use feature selection methods such as wrappers, filters, the embedded method in ML models
12	Arrival of new patterns in workload, No single ML model for all-time series, Fixed prediction models in ensemble learning approaches	Generalised ensemble framework, Novel models incorporating both global and local parameters, Ensemble learning, Prediction using advanced neural networks like Temporal Convolution Networks (TCN)
13	Dynamic resource prediction, optimal data training in ML model	Optimisation of hyperparameters of ML models using heuristics like Grid Search, Random Search, Bayesian Optimisation, Gradient-based Optimisation, and Evolutionary Optimisation
14	Multiple virtual resources, Demand prediction of each type of virtual resource	To categorise the VMs using advanced clustering approaches like cluster ensemble involving clustering accuracy, time complexity and resource usage (CPU and memory utilisation) as model evaluation criteria

(Shaw et al., 2019) They proposed a network-aware predictive VM placement heuristic to reduce energy consumption and SLA violations by considering CPU demand and network bandwidth. The main advantage of their work was to design a dynamic VM placement strategy based on the prediction of both CPU utilisation and network bandwidth because estimating network bandwidth in case of large VM migration contributes to making decisions with improved scheduling and makes VM placement efficient reliable. Thus, VM placement strategies should consider future insights of resources to balance limited resource availability and energy-efficient management. However, they did not consider another aspect, disc throughput, that may also affect VM migration time (Brewer et al., 2016).

3.4. Thermal management

3.4.1. Host temperature

In modern cloud data centres, minimising host temperature is a challenging issue. This is caused by the released heat in the process of energy consumption by the host. The cooling systems are deployed to rid this dissipated heat to keep the host's temperature below the threshold. This increased temperature directly affects the cost of the cooling system and has become a challenging issue to resolve in resource management systems. It also creates hot spots in the system and is responsible for several system failures. Thus, thermal management is necessary and challenging due to this dynamic behaviour of the host's temperature.

Ilager et al. (2021) proposed a thermal aware predictive scheduling approach to reduce the peak temperature of a host and energy consumption. Since most data centres have monitoring sensors to record several parameters such as resource usage, energy consumption, thermal reading, and fan speed readings, this kind of data was collected from the University of Melbourne's private cloud data centre. They predicted host temperature using several machine learning algorithms. They proposed a thermal aware scheduling algorithm to minimise the peak temperature of hosts while migrating VMs to the fewest hosts to reduce energy consumption. In this approach, the prediction model

is invoked to predict the host temperature, and further scheduling is guided. The main advantage of their work is that they reduce the peak temperature up to 6.5° and 34% energy consumption in comparison to existing algorithms, and it was reported (Gao, 2014) that reducing even one degree in temperature can save up to millions of dollars in a large-scale data centre. They consider the host's ambient temperature for prediction instead of CPU temperature that combines inlet temperature and CPU temperature; however, it may increase the algorithm overhead.

3.5. Resource provisioning

3.5.1. SLA-based VM management

Over-provisioning has long been used in data centres to prevent the worst-case scenario of peak load utilisation while still meeting SLA obligations ?. However, during regular hours, the hosts use very little energy, resulting in resource waste. Reiss et al. (2012) studied actual workload traces of VMs' resource utilisation from the Google data center and found that the average CPU and memory utilisation were less than 60% and 50%, respectively. Overprovisioning of services, as a result, results in additional maintenance costs in host cooling and administrative activities (Sun et al., 2016). Research has aimed to solve this difficult problem by using dynamic resource provisioning of resources in virtualisation technology. Still, it primarily focuses on a particular form of SLA or application, such as transactional workload. However, computationally intensive applications are increasingly becoming a part of enterprise data centres, which run multiple types of applications on multiple VMs without considering SLA criteria, such as the deadline that results in an under-utilised host. In the case of resource estimation, this factor presents a unique challenge.

Garg et al. (2014) suggested a novel resource management approach that took into account various types of SLA specifications for various applications operating on various VMs. This approach addresses two types of applications: non-interactive compute-intensive jobs and transactional applications. Both types of applications had a wide range of SLA criteria and specifications. The key benefit of their work was that they used historical CPU utilisation data combined with SLA penalties to forecast potential insight, allowing them to make complex placement decisions in response to shifts in transactional workload and scheduled jobs, taking into account CPU cycles in case of under-utilisation during usual or off-peak periods. The sample of VM CPU usage was used to train an artificial neural network (ANN) to predict VM CPU usage for the next two hours, with the result plotted against actual usage. The X-axis was distributed at a regular interval of 5 min. We saw some shortcomings in their work at this point: (1) When there is a wide variance in preparation, the ANN forecast deviates from the actual value in some situations, (2) In a few instances, it also predicts low CPU utilisation from the actual value, (3) They did not take into account highly nonlinear data. The testing data had no nonlinear variation, and non-linearity in workload is a major issue nowadays, as data centres have very high non-linearity in workload, which leads to a variety of issues such as high energy consumption, inconsistent QoS and SLA violations (Kumar et al., 2020b).

3.5.2. QoS-aware resource provisioning

The pattern of evaluating applications deployed on running VMs in modern data centres varies from time to time, i.e., many users attempt to access the application simultaneously. As a result, in the cloud, static resource allocation to SaaS applications is inefficient because it results in nonlinear resource use during low demand and high utilisation periods. When demand is low, available resources are wasted, resulting in excessive overheads and costs for the cloud service provider; when demand is high, available resources can be inadequate, resulting in weak QoS. This problem can be solved with dynamic resource provisioning. Still, in this case, the difficulty is determining the correct number of resources to deploy in a given period to satisfy QoS requirements when varying workload is available. This challenge is being addressed in two ways: reactively and proactively. The latter has been significantly modified because it is dependent on future load variations before their occurrence, i.e., estimating the QoS parameters in advance.

Calheiros et al. (2014) proposed an ARIMA-based workload prediction model. The main benefit of their work was that the expected requests were used to dynamically provision VMs in an elastic cloud environment while taking into account QoS parameters such as response time and rejection rate. The accuracy of forecast user requests was also assessed to see how it affected resource use and QoS parameters. However, we would like to draw your attention to the following limitation in this work. They gathered historical web request data from the Wikimedia Foundation (Amekraz and Hadi, 2018) and fed it into a component of their proposed model called *Workload Analyser*. The ARIMA model was used in this component to provide a future estimation for a specific time interval that can be adjusted for a specific application. The time interval should be long enough to allow for the placement of a new VM for optimal system utilisation. This static time interval may cause issues if a VM deployment time is less than this static time interval, as the extra remaining time may affect QoS parameters such as response time.

3.5.3. Varying patterns of service tenant in resource allocation

Resource demand prediction in a multi-tenant service cloud environment requires historical data to learn the past profiles of service tenants, which is challenging due to the need to update the prediction model regularly because the profiles or trends of service tenants change. Another challenge is maintaining the number of resources required by a service tenant to conduct its operations, which is dependent on many factors, including (1) the operation type, (2) the specific period when the operation is conducted, and (3) the load faced by the service tenant at a specific time. As a result, it presents a challenge because a service tenant's resource requirements can shift. This is a critical topic to address when dealing with resource provisioning using proactive methods for a single service tenant and multiple service tenants.

In multi-tenant service clouds, (Verma et al., 2016) a dynamic resource demand prediction and provisioning approach was proposed to assign resources in advance. They divided the service tenants into groups based on whether or not their resource use would rise in the future. As a result, the proposed system forecasts resource demand with priority for only those service tenants whose resource demand was expected to increase, reducing the time required for prediction, which may affect the total time of all operations, thereby affecting QoS. Furthermore, the proposed mechanism used the Best-fit decreasing heuristic method to determine the efficiency of maximum PMs utilisation by combining the service tenants with the matched VMs and allocating them to physical machines (PMs). The most significant aspect of this research is that it classifies service tenants based on a binary issue of whether resource demand will increase or not and then predicts resource demand for tenants whose resource demand will increase, resulting in a decrease in computational time and cost of prediction. However, (1) we cannot determine on what basis they mark binaries (0, 1) with the service tenants' characteristics, even though labelling data is needed to classify it using supervised learning techniques. (2) If we presume that the service tenants' features were labelled with binaries based on some condition, then labelling the data in a large-scale multi-tenant cloud would be time-consuming and increase the prediction cost. (3) Some data may be accessible without labels in a large-scale distributed multi-tenant cloud, in which case supervised classification would not work.

In summary, ML techniques, although presents significant challenges in adopting to resource management of Cloud, with the advancements of fundamental ML algorithms and tools, we envision Ml-centric resource management will become pervasive. In the next section, we discuss potential research directions in regard to utilising ML techniques in various cloud resource management tasks.

4. Future research directions

4.1. Workload prediction

4.1.1. ML in energy consumption prediction

Apart from the CPU, a system power model includes memory, disc, and network components so that these components could be considered as well. The current study looks at the linear relationship between these metrics and energy consumption; however, nonlinear relationships, such as polynomial or exponential, could be explored in the future. In addition, the best individual model is chosen in an ensemble learning approach, which may or may not be the best solution. Another option is to combine the information provided by each model and analyse the results. This can be accomplished by estimating the average using weights based on each predictor's mean average error. Furthermore, each workload type requires its own set of configuration parameters. The future research direction is to keep track of the value of the parameter of each model from the record that has increased the maximum utilisation of resources and to use them in real-time scenarios to adapt the models to the workload type of each VM. In addition, the forecast accuracy is also affected by a sudden change in the use of resources. Therefore, a further future research direction is to feed the ML model with average workload performance, such as CPU utilisation.

4.1.2. Performance and online profiling of workload

The efficiency of the intelligent resource management system is determined by many factors, including the accuracy and time complexity of the prediction model. Huge corporations such as Google, Microsoft, Amazon, and others are in charge of extremely complex data centres with a wide range of workloads. As a result, in the presence of such a highly variable or nonlinear workload for VMs, a more accurate estimation of prior workload is a future research direction by employing more sophisticated ML and DL modes. Furthermore, the time complexity of an algorithm is a measurement of its performance in terms of the time it takes to run the input code. As a result, the algorithm should be designed to be as simple as possible in terms of time complexity. Furthermore, online profiling is necessary to prevent VM blackouts until they are running in development, as well as various resource utilisation such as CPU and memory, which are major contributors to physical resource exhaustion and should be considered for prediction. Cortez et al. (2017), Bianchini et al. (2020) conducted online workload profiling and provided an analysis to determine if a virtual machine is interactive or delay-insensitive. To categorise VMs into these two groups, they used supervised classification. In this situation, semi-supervised learning (Zhu and Goldberg, 2009) may play a vital role and maybe a potential research direction to train the data with these partial labels and perform classification with promising accuracy in large-scale distributed data centres.

4.1.3. Prediction accuracy in auto-scaling of web applications

Machine learning models may be used to predict workload in the future, which has many advantages: (1) Machine Learning learns from data without the need for explicit programming. (2) Machine Learning can learn from billions of observations and features, (3) Machine Learning relies less on assumptions and, in most cases, disregards them. (4) Machine Learning emphasises predictions, supervised learning, unsupervised learning, and semi-supervised learning (5) Machine Learning uses iterations to identify patterns in a dataset, requiring far less human effort. The training of multiple features is needed to predict the target variable, which increases the time complexity of machine learning methods like regression. As a result of the existence of redundant features, ML methods suffer from latency and computational complexity problems when processing multiple features. In such datasets, the number of functions, feature dependency, number of records, feature types, and nested feature categories substantially increase ML methods' processing time. As a result, future research should concentrate on using suitable feature selection methods, such as wrappers, filters, embedded methods, and enhanced versions (Majeed, 2019), to effectively overcome the computation speed versus accuracy trade-off when processing large and complex datasets.

4.1.4. Time-series prediction data

Developing a generalised ensemble framework for any type of dataset in cloud time-series workload data is a future research direction. Deep learning (DL), in general, is a rapidly expanding and broad research field that involves novel architectures. However, researchers are never sure when to adapt which methods to which situations. Hewamalage et al. (2021) used global NN models, which are prone to outlier errors in some time series. As a result, novel models incorporating both global and local parameters for individual time series must be developed in the form of hierarchical models. These models can be combined with ensembling, which involves training multiple models with the same dataset in different ways. Furthermore, CNNs have long been used for image processing, but they are now being used to forecast time series data. According to Lai et al. (2018), Shih et al. (2019), traditional RNN models are ineffective at modelling seasonality in time series forecasting. As a result, they combine CNN filters for local dependencies and a custom attention score function for long-term dependencies. To capture seasonality patterns, Lai et al. (2018) has also tried recurrent skip connections. Oord et al. (2016) developed Dilated Causal Convolutions to effectively capture long-range dependencies along the temporal dimension. They have recently been used in conjunction with CNNs to solve problems involving time series forecasting. Temporal Convolution Networks (TCN), which combine dilated convolutions and residual skip connections, have also been introduced as more advanced CNNs (Borovykh et al., 2017). According to Bai et al. (2018) TCNs are promising NN architectures for sequence modelling tasks, in addition to being efficient in training. As a result, using CNNs instead of RNNs could provide a competitive advantage for forecasting practitioners. As a consequence, these potentially advanced neural networks could be used in the future to forecast workload time series in cloud infrastructure.

4.1.5. Data training

Optimising machine learning hyperparameters aims to find the hyperparameters for a particular machine learning algorithm that achieves the best performances on validation data. The hyperparameters are set by the experts before the training, contrary to the model parameters. The number of trees in a random forest, for example, is a hyperparameter, whereas the weights in a neural network are model parameters learned during training. Size and decay are support vector machine hyperparameters (SVM) and k in k-nearest neighbours (KNN), respectively. Furthermore, hyperparameter optimisation returns an optimal model that reduces a predefined loss function and, as a result, improves the accuracy on given independent data by finding a combination of hyperparameters. Hyperparameters can thus have a direct effect on machine learning algorithm training. Therefore, it is critical to understand how to optimise them to achieve maximum performance. This points to a future research direction of optimising the hyperparameters of ML algorithms for achieving optimal dataset training. This can be accomplished by employing common heuristics such as Grid Search, Random Search, Bayesian Optimisation, Gradient-based Optimisation, and Evolutionary Optimisation (Feurer and Hutter, 2019).

4.2. Runtime VM management

4.2.1. Multiple resource usage in VM consolidation

A host is considered overloaded during the VM consolidation phase if CPU utilisation reaches a throughput threshold, such as 80% (Nguyen et al., 2017). However, other resource utilisation, such as memory and bandwidth use (Abdelsamea et al., 2017), leads to host overloading.

As a result, detecting overloaded hosts using a combination of CPU, memory, and bandwidth use is a potential research direction in the VM consolidation phase. For an efficient VM consolidation operation, the estimation of current and future CPU, memory, and bandwidth use should be addressed. The current study (Abdelsamea et al., 2017; Haghshenas and Mohammadi, 2020) involves a variety of machine learning algorithms, such as linear regression and multiple regression, in which the model's training is based on multiple features to simulate a target variable, such as CPU utilisation. The training time of multiple features will affect the VM migration time in the VM consolidation process, which affects QoS and SLAs in large-scale distributed data centres where millions of VMs are running in production. As a result, dealing with the training time of ML models is a potential future research direction. Different deep learning (DL) approaches, such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014), can deal with training time by avoiding the overheads of multiple features by using a single feature, such as a vector of CPU utilisation, as an input for training to predict its next state in the future.

4.2.2. Multi-dimensional resource requirement

As stated in Section 3.2.2, there is a future research direction to categorise the VMs and develop a prediction model for each cluster to address the multi-resource demand challenges. However, using a clustering algorithm such as K-means can limit the number of clusters available, causing a VM to be placed in the incorrect cluster. A clustering ensemble can be a better approach than clustering because it aims to combine multiple clustering algorithms to produce a final consensus solution that is more robust and accurate than a single clustering algorithm (Alqurashi and Wang, 2019). In this literature (Boongoen and Iam-On, 2018) mentions several clustering ensemble methods. Furthermore, in a recent work (Kadhim et al., 2019), two additional evaluation criteria, such as time complexity and resource usage (CPU and memory usage), were considered to evaluate the novel clustering ensemble, in addition to clustering accuracy. Thus, advanced clustering methods such as clustering ensembles can be used in the future to achieve the best clusters with the highest precision, least time complexity, and least resource consumption.

4.2.3. Energy metering at software-level

Many power management decisions, such as power capping, will benefit from the visibility of energy usage at the host and VM levels. At the host level, energy consumption is simple to predict or calculate since modern data centres have several built-in sensors that track it. Still, it is difficult to measure at the VM level because to measure the energy consumption induced by memory, and we must collect LLC (last-level-cache) events raised by each VM on each core, which is difficult to do (Kansal et al., 2010; Zhao-Hui and Qin-Ming, 2012). Rather than calculating or predicting energy consumption at the VM level, clustering analysis may be used to determine the status of VMs in terms of energy consumption, such as low, moderate, or critical. Thus, dividing VMs by conducting clustering analysis based on highly co-related features with energy consumption at the VM-level is a potential research direction, and there would be no need to obtain host-level features. ML techniques such as ChiSquare Score, Fisher Score, Gini Index, and Correlation-based Feature Selection (CFS) can be used to find the correlation with energy consumption (Vora and Yang, 2017). Then, using a clustering algorithm or a clustering ensemble (Kadhim et al., 2019), a clustering analysis can be performed to determine which VMs are in low and critical energy-consuming states. By doing so, a group of VMs can be managed together in a data centre's resource management system, potentially reducing response time and improving QoS.

4.2.4. Usage level management

The overloaded host detection's static threshold can result in unreliable VM migration. If the utilisation of a VM's resources degrades in a short period, there is no need to migrate the VM. In this case, the algorithm should have a dynamic resource utilisation threshold that automatically prevents VM migration when it reaches the fixed threshold, taking into account near-future data. This is the future research direction for efficient VM migration in VM consolidation. Furthermore, VMs should be migrated if the near future information has a long period of load degradation.

4.3. VM placement

4.3.1. Cloud network traffic

The problem of varying patterns of various types of workloads when considering current resource utilisation in VM allocation on a host is a challenge. As a result, predicting potential resource demand, such as CPU and network bandwidth, has proven to be an alternative approach (Shaw et al., 2019). However, disc throughput is a significant factor to consider in addition to these resources. In VM placement heuristics, taking disc throughput into account is a new research direction. It calculates the amount of data that can be stored, read, and written per second. Brewer et al. (2016) published a report stating that disc tail latency, especially reads, is a key factor when delivering online services where a user is waiting for a response. As a result, disc throughput can play a role in VM migration time, affecting tail latency time and violating SLAs. Therefore, according to our vision, a prior maximum estimate of disc throughput will play a critical role in avoiding delay.

4.4. Thermal management

4.4.1. Host temperature

Ilager et al. (2021) proposed a scheduling algorithm to minimise the host temperature driven by the host temperature prediction computed using several ML algorithms. Consequently, estimating host temperature ahead of time can help with thermal management decisions like VM migration to reduce host temperature, i.e., CPU temperature. Ilager et al. (2021), on the other hand, it took into account the ambient temperature for prediction, which is a combination of CPU and inlet temperature. This could increase algorithm overhead. Furthermore, they discovered that the host's CPU temperature is primarily affected by CPU load and power consumption. As a result, it is being waited for the CPU to become overloaded, causing the temperature to rise, resulting in additional cooling costs for the host. As a potential research topic, Prior CPU estimation-based resource provisioning can prevent the CPU from overloading and save energy. Then we will only have to deal with the inlet temperature, which may reduce the thermal management algorithm's overhead. Furthermore, several ML algorithms necessitate a significant amount of training time due to the training of multiple features, which can slow down VM migration. It will cause VM migration to be delayed, which will slow down host temperature degradation and add to the cost. Thus, using an ML or DL method like GRU, where the inlet temperature can be used as an input to train a model that can predict its future state using single feature training, could be an alternative. Doing so can avoid an overhead algorithm, a delay in VM migration, a delay in minimising the host temperature.

4.5. Resource provisioning

4.5.1. SLA-based VM management

Future research directions for avoiding nonlinear resource utilisation in modern data centres include dynamic resource provisioning and dynamic VM consolidation, which take into account various types of VM resources such as CPU, memory, and bandwidth, current and future resource needs, and SLAs such as compute-intensive non-interactive

jobs and transactional applications. Both of these methods rely heavily on accurate resource prediction. Garg et al. (2014), for example, they provided long-term CPU utilisation forecasts that differed significantly from actual test phase data due to a substantial shift in CPU utilisation during the training phase, which is critical for dealing with nonlinear utilisation in modern data centres. Future research will focus on optimising hyperparameters used in Artificial Neural Network (ANN) learning, such as mini-batch size, epochs, and several neurons. The model is said to work better if trained on the data in an optimised manner. The observation of the validation and loss graphs estimated with these optimised hyperparameters may indicate that the model has learned a lot when both plots begin moving closely and consistently, and learning should be stopped at these optimised parameters.

4.5.2. QoS-aware resource provisioning

This study uses constructive dynamic resource provisioning based on workload estimation using historical data to improve QoS parameters like response time and rejection rate. Future research could deal with it reactively, with resource provisioning occurring after resource demand, such as the number of requests, has arrived. Furthermore, according to the current study (Calheiros et al., 2014), the error in request prediction can be mitigated by Adhoc decisions in dynamic resource provisioning, which can help to boost poor QoS efficiency. Furthermore, there is a potential research direction to forecast peak CPU use using more sophisticated ML models such as XGBoost (Chen and Guestrin, 2016), LSTM (Hochreiter and Schmidhuber, 1997), and GRU (Cho et al., 2014) in a correct manner that cannot be equipped with the ARIMA model. Furthermore, no single machine learning algorithm can suit any non-linear workload with time-series data, necessitating an ensemble learning approach in which various ML and DL methods can be used in the future. After that, the best-performing model can be selected for potential use. Calheiros et al. (2014), as discussed in Section 3.5.2, estimates web requests based on a static time interval that can affect response time. Therefore, pAs a result, it can be addressed by estimating future web requests with a dynamic time interval that adjusts automatically based on the VM deployment time. In such a way that the time interval of estimation can be equivalent to the VM deployment time, and the remaining time can be avoided if the VM deployment time is much shorter than this static time interval that affects the QoS parameter as the response time. Prior estimation of VM deployment time based on historical data should be computed and used in the above-mentioned case to satisfy the condition of equivalence with the estimated time of the request prediction.

In summary, ML techniques provide numerous opportunities to apply them for various resource management tasks as described in this section. However, ensuring availability of quality data, careful selection of suitable ML models, and performance guarantee is necessary to successfully deploy them on real Cloud environments.

4.5.3. Varying patterns of service tenant in resource allocation

Clustering analysis, which does not require any data labelling, could classify service tenants as a future research direction. Based on historical resource demands, similar patterns of service tenants can be automatically obtained. By observing the similarity between data using clustering, service tenants with high and low resource demand can be distinguished, and predictions for those with high resource demand can be provided using ML and DL regression techniques. In the case of a distributed data centre where data is dispersed and partial labels are available, a concept is known as semi-supervised clustering (Śmieja et al., 2020) can be used, in which unsupervised data is given a little supervision using partial labels and techniques such as instance-level constraints (Wagstaff and Cardie, 2000) and relative distance constraints (Cho et al., 2014).

4.6. Answering of research questions

In this subsection, we answer the research questions raised in Section 1.2 based on the detailed study done in this article.

- The temporal complexity of an ML algorithm is the number of operations it must perform to achieve its goal in relation to the size of the input. To put it another way, it takes time to finish the task. The desire to minimise the complexity of a model can occur for a variety of reasons, the most common of which is to reduce computational needs. However, complexity cannot be reduced arbitrarily because that is the only model that produced good results after several iterations of training and testing. It is important to optimise the ML models to reduce the number of training rounds and achieve the required accuracy by carefully selecting the training parameters. This matter is currently being researched and Koning et al. (2019) proposes a solution to this difficulty for CNNs used for exoplanet identification.
- To improve the accuracy of prediction models, several techniques are employed, such as using LSTM encoders with an attention mechanism can improve workload prediction accuracy. An automatic decision mechanism for input weights and hidden biases is used in an extreme learning machine that requires many hidden neurons to obtain decent results. Using a neural network and a self-adaptive differential evolution method is another option. Another alternative is to use a clustering-based workload prediction method, which divides all tasks into groups and then trains a prediction model for each one.
- The training time of an ML algorithm can be reduced using several techniques including reducing the input size to the necessary dimensions, ensuring that critical features are not lost. Prepossessing the data to make it zero mean and normalise it by dividing it by the standard deviation or the difference between the maximum and minimum values. In addition, maintaining a network depth and width that is neither too large nor too little. Alternatively, always utilise the theoretically proven standard architecture and initialise weights using tried-and-true methods like Xavier Initialisation. An appropriate learning rate should be determined by trying several and selecting the one that reduces error the most in relation to the number of epochs. Also, while executing gradient descent, employ the learning rate decay approach to ensure you do not skip a solution. Always double-check the epochs. There is no point in taking more epochs if you cannot enhance your mistake or accuracy beyond a certain point. The batch size should be determined by the amount of RAM available and the number of CPUs/GPUs. If the batch cannot be fully loaded in memory, operations will be slowed owing to paging between memory and the filesystem. Use batch normalisation to treat and process data through the pipeline (feedforward). This data transformation aids in the faster learning of weights, resulting in faster optimisation.
- To reduce the operational complexity in VM management, VMs can be categorised in similar groups by using clustering algorithms (Jain et al., 1999) and cluster ensembles (Ghosh and Acharya, 2011) and resource management actions can be enforced on the group of VMs instead of building individual models for each VM.
- The energy efficient Clouds can be achieved by variety of resource management measures. By limiting the number of active servers, the energy consumption of servers can be reduced. This is commonly accomplished by scheduling optimisation, which is a common strategy for green clouds and is considered (Xian et al., 2007) more efficient in terms of cost, utilised resources, and scalability than hardware optimisation. To reduce the amount of power utilised, it is necessary to find a proper mapping between demands for VMs and actual servers.

5. Summary and conclusions

Cloud computing systems are immensely complex, huge in scale, and diverse, allowing for the development of highly networked resource-intensive corporate, scientific, and personal applications. In such a complex infrastructure, holistic resource management has become a difficult undertaking. In today's cloud computing context, state-of-the-art rule-based or heuristic resource management systems are insufficient. RMS rules must deal with vast size, heterogeneity, and shifting workload demands. Therefore, we need data-driven AI techniques that draw critical insights from data, learn from surroundings, and make resource management decisions based on that learning. In this paper, we discuss the challenges in resource management in a cloud computing environment, the various ML approaches that have been used to solve these challenges in recent years, and their benefits and drawbacks. In recent years, there has been a significant increase in the number of studies looking at how to use machine learning techniques to conduct workload prediction, energy consumption prediction, and other tasks. Different ML methods are used to deal with various types of problems. Finally, based on the challenges and drawbacks identified in the state-of-the-artwork, new potential future research directions are proposed to strengthen the current ML methods for resource management in cloud-based systems. The overall knowledge provided in this paper aids clouds researchers in comprehending cloud resource management and the significance of machine learning techniques.

Our investigation shows that machine learning models can be used in cloud computing systems to achieve various optimisation goals and deal with complex tasks. ML approaches also open up a new avenue for intelligent resource and application management. This article illustrates the progress of machine learning approaches in current research and helps readers understand the research gap in this field. To improve system efficiency, one promising way is to use advanced machine learning techniques such as reinforcement learning and deep learning to perform intelligent resource management.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank anonymous reviewers for their suggestions on improving our paper. This research is partially supported by Natural Science Foundation of China with ID 61672136 and Melbourne-Chindia Cloud Computing (MC3) Research Network.

References

Abdelsamea, Amany, El-Moursy, Ali A, Hemayed, Elsayed E, Eldeeb, Hesham, 2017. Virtual machine consolidation enhancement using hybrid regression algorithms. Egypt. Inform. J. 18 (3), 161–170.

Alqurashi, Tahani, Wang, Wenjia, 2019. Clustering ensemble method. Int. J. Mach. Learn. Cybern. 10 (6), 1227–1246.

Amazon, E.C., 2010. Amazon Elastic Compute Cloud (Amazon EC2). Vol. 5. pp. 18–23, Amazon Elastic Compute Cloud (Amazon EC2).

Amekraz, Zohra, Hadi, Moulay Youssef, 2018. Higher order statistics based method for workload prediction in the cloud using ARMA model. In: 2018 International Conference on Intelligent Systems and Computer Vision. ISCV, IEEE, pp. 1–5.

Arlitt, Martin, Jin, Tai, 2000. A workload characterization study of the 1998 world cup web site. IEEE Netw. 14 (3), 30–37.

Arlitt, Martin F., Williamson, Carey L., 1997. Internet web servers: Workload characterization and performance implications. IEEE/ACM Trans. Netw. 5 (5), 631–645.

Bai, Shaojie, Kolter, J. Zico, Koltun, Vladlen, 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Barroso, Luiz André, Clidaras, Jimmy, Hölzle, Urs, 2013. The datacenter as a computer: An introduction to the design of warehouse-scale machines. Synth. Lect. Comput. Archit. 8 (3), 1–154.

Bhatia, Mahinder Pal Singh, Kumar, Akshi, 2008. Information retrieval and machine learning: supporting technologies for web mining research and practice. Webology 5 (2), 5.

Bianchini, Ricardo, Fontoura, Marcus, Cortez, Eli, Bonde, Anand, Muzio, Alexandre, Constantin, Ana-Maria, Moscibroda, Thomas, Magalhaes, Gabriel, Bablani, Girish, Russinovich, Mark, 2020. Toward ML-centric cloud platforms. Commun. ACM 63 (2), 50–59.

Boongoen, Tossapon, Iam-On, Natthakan, 2018. Cluster ensembles: A survey of approaches with recent extensions and applications. Comp. Sci. Rev. 28, 1–25.

Borovykh, Anastasia, Bohte, Sander, Oosterlee, Cornelis W., 2017. Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691.

Braiki, Khaoula, Youssef, Habib, 2019. Resource management in cloud data centers: a survey. In: 2019 15th International Wireless Communications & Mobile Computing Conference. IWCMC, IEEE, pp. 1007–1012.

Brewer, Eric, Ying, Lawrence, Greenfield, Lawrence, Cypher, Robert, T'so, Theodore, 2016. Disks for data centers.

Buyya, Rajkumar, Srirama, Satish Narayana, Casale, Giuliano, Calheiros, Rodrigo, Simmhan, Yogesh, Varghese, Blesson, Gelenbe, Erol, Javadi, Bahman, Vaquero, Luis Miguel, Netto, Marco A.S., Toosi, Adel Nadjaran, Rodriguez, Maria Alejandra, Llorente, Ignacio M., Vimercati, Sabrina De Capitani Di, Samarati, Pierangela, Milojicic, Dejan, Varela, Carlos, Bahsoon, Rami, Assuncao, Marcos Dias De, Rana, Omer, Zhou, Wanlei, Jin, Hai, Gentzsch, Wolfgang, Zomaya, Albert Y., Shen, Haiying, 2018. A manifesto for future generation cloud computing: Research directions for the next decade. ACM Comput. Surv. (ISSN: 0360-0300) 51 (5), http://dx.doi.org/10.1145/3241737.

Calheiros, Rodrigo N, Masoumi, Enayat, Ranjan, Rajiv, Buyya, Rajkumar, 2014. Workload prediction using ARIMA model and its impact on cloud applications' QoS. IEEE Trans. Cloud Comput. 3 (4), 449–458.

Calheiros, Rodrigo N, Ranjan, Rajiv, Beloglazov, Anton, De Rose, César AF, Buyya, Rajkumar, 2011. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. - Pract. Exp. 41 (1), 23–50.

Cao, Jian, Fu, Jiwen, Li, Minglu, Chen, Jinjun, 2014. CPU load prediction for cloud environment based on a dynamic ensemble model. Softw. - Pract. Exp. 44 (7), 793–804.

Cao, Rui, Yu, Zhaoyang, Marbach, Trent, Li, Jing, Wang, Gang, Liu, Xiaoguang, 2018. Load prediction for data centers based on database service. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference. Vol. 1. COMPSAC, IEEE, pp. 728–737.

Celebi, M. Emre, Aydin, Kemal, 2016. Unsupervised Learning Algorithms. Springer.

Chen, Tianqi, Guestrin, Carlos, 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Chen, Shengwei, Shen, Yanyan, Zhu, Yanmin, 2018. Modeling conceptual characteristics of virtual machines for CPU utilization prediction. In: International Conference on Conceptual Modeling. Springer, pp. 319–333.

Chin, Keene, Hellebrekers, Tess, Majidi, Carmel, 2020. Machine learning for soft robotic sensing and control. Adv. Intell. Syst. 2 (6), 1900171.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, Bengio, Yoshua, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chun, Brent, Culler, David, Roscoe, Timothy, Bavier, Andy, Peterson, Larry, Wawrzoniak, Mike, Bowman, Mic, 2003. Planetlab: an overlay testbed for broad-coverage services. ACM SIGCOMM Comput. Commun. Rev. 33 (3), 3–12.

Cortez, Eli, Bonde, Anand, Muzio, Alexandre, Russinovich, Mark, Fontoura, Marcus, Bianchini, Ricardo, 2017. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In: Proceedings of the 26th Symposium on Operating Systems Principles. pp. 153–167.

Darges, John, Alexanderian, Alen, Gremaud, Pierre, 2022. Extreme learning machines for variance-based global sensitivity analysis. arXiv preprint arXiv:2201.05586.

Deng, Li, Li, Xiao, 2013. Machine learning paradigms for speech recognition: An overview. IEEE Trans. Audio Speech Lang. Process. 21 (5), 1060–1089.

Dewangan, Bhupesh Kumar, Agarwal, Amit, Choudhury, Tanupriya, Pasricha, Ashutosh, Chandra Satapathy, Suresh, 2021. Extensive review of cloud resource management techniques in industry 4.0: Issue and challenges. Softw. - Pract. Exp. 51 (12), 2373–2392.

Dillon, Tharam, Wu, Chen, Chang, Elizabeth, 2010. Cloud computing: issues and challenges. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications. Ieee, pp. 27–33.

Ding, Chris, He, Xiaofeng, Zha, Hongyuan, Simon, Horst D., 2002. Adaptive dimension reduction for clustering high dimensional data. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, pp. 147–154.

Duggan, Martin, Duggan, Jim, Howley, Enda, Barrett, Enda, 2017. A network aware approach for the scheduling of virtual machine migration during peak loads. Cluster Comput. 20 (3), 2083–2094.

Espadas, Javier, Molina, Arturo, Jiménez, Guillermo, Molina, Martín, Ramírez, Raúl, Concha, David, 2013. A tenant-based resource allocation model for scaling software-as-a-service applications over cloud computing infrastructures. Future Gener. Comput. Syst. 29 (1), 273–286.

Feurer, Matthias, Hutter, Frank, 2019. Hyperparameter optimization. In: Automated Machine Learning. Springer, Cham, pp. 3–33.

Gao, J., 2014. Machine learning applications for data center optimization (Google White paper).

Garg, Saurabh Kumar, Toosi, Adel Nadjaran, Gopalaiyengar, Srinivasa K, Buyya, Rajkumar, 2014. SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter. J. Netw. Comput. Appl. 45, 108–120.

Genez, Thiago AL, Bittencourt, Luiz F, da Fonseca, Nelson LS, Madeira, Edmundo RM, 2015. Estimation of the available bandwidth in inter-cloud links for task scheduling in hybrid clouds. IEEE Trans. Cloud Comput. 7 (1), 62–74.

Ghosh, Joydeep, Acharya, Ayan, 2011. Cluster ensembles. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (4), 305–315.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, Bengio, Yoshua, 2016. Deep Learning. Vol. 1. No. 2. MIT press Cambridge.

Guha, Sudipto, Rastogi, Rajeev, Shim, Kyuseok, 2000. ROCK: A robust clustering algorithm for categorical attributes. Inf. Syst. 25 (5), 345–366.

Haghshenas, Kawsar, Mohammadi, Siamak, 2020. Prediction-based underutilized and destination host selection approaches for energy-efficient dynamic VM consolidation in data centers. J. Supercomput. 1–18.

Hamdaqa, Mohammad, Tahvildari, Ladan, 2012. Cloud computing uncovered: a research landscape. In: Advances in Computers. Vol. 86. Elsevier, pp. 41–85.

Hartigan, John A., Wong, Manchek A., 1979. AK-means clustering algorithm. J. R. Stat. Soc. Ser. C. Appl. Stat. 28 (1), 100–108.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Helali, Leila, Omri, Mohamed Nazih, 2021. A survey of data center consolidation in cloud computing systems. Comp. Sci. Rev. 39, 100366.

Hewamalage, Hansika, Bergmeir, Christoph, Bandara, Kasun, 2021. Recurrent neural networks for time series forecasting: Current status and future directions. Int. J. Forecast. 37 (1), 388–427.

Hochreiter, Sepp, Schmidhuber, Jürgen, 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Ilager, Shashikant, Muralidhar, Rajeev, Buyya, Rajkumar, 2020. Artificial intelligence (AI)-centric management of resources in modern distributed computing systems. In: IEEE Cloud Summit.

Ilager, S., Ramamohanarao, K., Buyya, R., 2021. Thermal prediction for efficient energy management of clouds using machine learning. IEEE Trans. Parallel Distrib. Syst. 32 (5), 1044–1056. http://dx.doi.org/10.1109/TPDS.2020.3040800.

Injadat, MohammadNoor, Moubayed, Abdallah, Nassif, Ali Bou, Shami, Abdallah, 2021. Machine learning towards intelligent systems: applications, challenges, and opportunities. Artif. Intell. Rev. 1–50.

Iosup, Alexandru, Li, Hui, Jan, Mathieu, Anoep, Shanny, Dumitrescu, Catalin, Wolters, Lex, Epema, Dick HJ, 2008. The grid workloads archive. Future Gener. Comput. Syst. 24 (7), 672–686.

Ismaeel, Salam, Miri, Ali, 2015. Using ELM techniques to predict data centre VM requests. In: 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing. IEEE, pp. 80–86.

Jadeja, Yashpalsinh, Modi, Kirit, 2012. Cloud computing-concepts, architecture and challenges. In: 2012 International Conference on Computing, Electronics and Electrical Technologies. ICCEET, IEEE, pp. 877–880.

Jain, Anil K., Murty, M. Narasimha, Flynn, Patrick J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.

Janai, Joel, Güney, Fatma, Behl, Aseem, Geiger, Andreas, et al., 2020. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Found. Trends® Comput. Graph. Vis. 12 (1–3), 1–308.

Jeff, D., 2018. ML for system, system for ML, keynote talk in workshop on ML for systems, NIPS.

Jennings, Brendan, Stadler, Rolf, 2015. Resource management in clouds: Survey and research challenges. J. Netw. Syst. Manage. 23 (3), 567–619.

Jordan, Michael I., Mitchell, Tom M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349 (6245), 255–260.

Jula, Amin, Sundararajan, Elankovan, Othman, Zalinda, 2014. Cloud computing service composition: A systematic literature review. Expert Syst. Appl. 41 (8), 3809–3824.

Kadhim, Mustafa R., Tian, Wenhong, Khan, Tahseen, 2019. Rapid clustering with semi-supervised ensemble density centers. In: 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing. IEEE, pp. 230–235.

Kansal, Aman, Zhao, Feng, Liu, Jie, Kothari, Nupur, Bhattacharya, Arka A, 2010. Virtual machine power metering and provisioning. In: Proceedings of the 1st ACM Symposium on Cloud Computing. pp. 39–50.

Khan, Tahseen, Tian, Wenhong, Ilager, Shashikant, Buyya, Rajkumar, 2021. Workload forecasting and energy state estimation in cloud data centers: ML-centric approach. Future Gener. Comput. Syst..

Khan, Tahseen, Tian, Wenhong, Ilager, Shashikant, Buyya, Rajkumar, 2022. Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. Future Gener. Comput. Syst. 128, 320–332.

Kingma, Diederik P., Welling, Max, 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kober, Jens, Bagnell, J. Andrew, Peters, Jan, 2013. Reinforcement learning in robotics: A survey. Int. J. Robot. Res. 32 (11), 1238–1274.

Koning, Sebastiaan, Greeven, Caspar, Postma, Eric, 2019. Reducing artificial neural network complexity: A case study on exoplanet detection. arXiv preprint arXiv:1902.10385.

Krishnaveni, S., Sivamohan, S., Sridhar, S.S., Prabakaran, S., 2021. Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. Cluster Comput. 1–19.

Kumar, Jitendra, Singh, Ashutosh Kumar, 2018. Workload prediction in cloud using artificial neural network and adaptive differential evolution. Future Gener. Comput. Syst. 81, 41–52.

Kumar, Jitendra, Singh, Ashutosh Kumar, 2020. Cloud datacenter workload estimation using error preventive time series forecasting models. Cluster Comput. 23 (2), 1363–1379.

Kumar, Jitendra, Singh, Ashutosh Kumar, Buyya, Rajkumar, 2020a. Ensemble learning based predictive framework for virtual machine resource request prediction. Neurocomputing.

Kumar, Jitendra, Singh, Ashutosh Kumar, Buyya, Rajkumar, 2020b. Self directed learning based workload forecasting model for cloud resource management. Inform. Sci. 543, 345–366.

Kumar, Jitendra, Singh, Ashutosh Kumar, Buyya, Rajkumar, 2021. Self directed learning based workload forecasting model for cloud resource management. Inform. Sci. 543, 345–366.

Lai, Guokun, Chang, Wei-Cheng, Yang, Yiming, Liu, Hanxiao, 2018. Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 95–104.

Li, Xin, Qian, Zhuzhong, Lu, Sanglu, Wu, Jie, 2013. Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center. Math. Comput. Modelling 58 (5–6), 1222–1235.

Majeed, Abdul, 2019. Improving time complexity and accuracy of the machine learning algorithms through selection of highly weighted top k features from complex datasets. Ann. Data Sci. 6 (4), 599–621.

Manvi, Sunilkumar S., Shyam, Gopal Krishna, 2014. Resource management for infrastructure as a service (IaaS) in cloud computing: A survey. J. Netw. Comput. Appl. 41, 424–440.

Mao, Hongzi, Alizadeh, Mohammad, Menache, Ishai, Kandula, Srikanth, 2016. Resource management with deep reinforcement learning. In: Proceedings of the 15th ACM Workshop on Hot Topics in Networks. pp. 50–56.

Mao, Hongzi, Schwarzkopf, Malte, Venkatakrishnan, Shaileshh Bojja, Meng, Zili, Alizadeh, Mohammad, 2019. Learning scheduling algorithms for data processing clusters. In: Proceedings of the ACM Special Interest Group on Data Communication. pp. 270–288.

Mell, Peter, 2011. The NIST definition of cloud computing. In N. I. O. S. A. Technology (Ed.): U.S. Department of Commerce.

Messias, Valter Rogério, Estrella, Julio Cezar, Ehlers, Ricardo, Santana, Marcos José, Santana, Regina Carlucci, Reiff-Marganiec, Stephan, 2016. Combining time series prediction models using genetic algorithm to autoscaling web applications hosted in the cloud infrastructure. Neural Comput. Appl. 27 (8), 2383–2406.

Mijuskovic, Adriana, Chiumento, Alessandro, Bemthuis, Rob, Aldea, Adina, Havinga, Paul, 2021. Resource management techniques for cloud/fog and edge computing: An evaluation framework and classification. Sensors 21 (5), 1832.

Nayak, Sanjib Kumar, Panda, Sanjaya Kumar, Das, Satyabrata, 2021. Renewable energy-based resource management in cloud computing: a review. Adv. Distrib. Comput. Mach. Learn. 45–56.

Networking, Cisco Visual, 2016. Cisco global cloud index: Forecast and methodology, 2016–2021. White Paper. Cisco Public, San Jose.

Nguyen, Trung Hieu, Di Francesco, Mario, Yla-Jaaski, Antti, 2017. Virtual machine consolidation with multiple usage prediction for energy-efficient cloud data centers. IEEE Trans. Serv. Comput..

Olsson, Fredrik, 2009. A Literature Survey of Active Machine Learning in the Context of Natural Language Processing. Swedish Institute of Computer Science.

Oord, Aaron van den, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, Kavukcuoglu, Koray, 2016. Wavenet: A generative model for raw audio. In: The 9th ISCA Speech Synthesis Workshop. ISCA, p. 125.

Persico, Valerio, Grimaldi, Domenico, Pescape, Antonio, Salvi, Alessandro, Santini, Stefania, 2017. A fuzzy approach based on heterogeneous metrics for scaling out public clouds. IEEE Trans. Parallel Distrib. Syst. 28 (8), 2117–2130.

Piraghaj, Sareh Fotuhi, Dastjerdi, Amir Vahid, Calheiros, Rodrigo N, Buyya, Rajkumar, 2017. A survey and taxonomy of energy efficient resource management techniques in platform as a service cloud. In: Handbook of Research on End-to-End Cloud Computing Architecture Design. IGI Global, pp. 410–454.

Pop, Daniel, 2016. Machine learning and cloud computing: Survey of distributed and saas solutions. arXiv preprint arXiv:1603.08767.

Reiss, Charles, Tumanov, Alexey, Ganger, Gregory R, Katz, Randy H, Kozuch, Michael A, 2012. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In: Proceedings of the Third ACM Symposium on Cloud Computing. pp. 1–13.

Reiss, Charles, Wilkes, John, Hellerstein, Joseph L., 2011. Google Cluster-Usage Traces: Format+ Schema. White Paper, Google Inc. pp. 1–14.
Sayadnavard, Monireh H, Haghighat, Abolfazl Toroghi, Rahmani, Amir Masoud, 2021. A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers. Eng. Sci. Technol. Int. J..
Sen, Pratap Chandra, Hajra, Mahimarnab, Ghosh, Mitadru, 2020. Supervised classification algorithms in machine learning: A survey and review. In: Emerging Technology in Modelling and Graphics. Springer, pp. 99–111.
Shahidinejad, Ali, Ghobaei-Arani, Mostafa, Masdari, Mohammad, 2020. Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. Cluster Comput. 1–24.
Shaw, Rachael, Howley, Enda, Barrett, Enda, 2019. An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. Simul. Model. Pract. Theory 93, 322–342.
Shih, Shun-Yao, Sun, Fan-Keng, Lee, Hung-yi, 2019. Temporal pattern attention for multivariate time series forecasting. Mach. Learn. 108 (8), 1421–1441.
Shyam, Gopal Kirshna, Manvi, Sunilkumar S., 2016. Virtual resource prediction in cloud environment: a Bayesian approach. J. Netw. Comput. Appl. 65, 144–154.
Singh, A.K., Kumar, Jitendra, 2019. Secure and energy aware load balancing framework for cloud data centre networks. Electron. Lett. 55 (9), 540–541.
Śmieja, Marek, Struski, Łukasz, Figueiredo, Mário AT, 2020. A classification-based approach to semi-supervised clustering with pairwise constraints. Neural Netw. 127, 193–203.
Smola, Alex J., Schölkopf, Bernhard, 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.
Stilgoe, Jack, 2018. Machine learning, social learning and the governance of self-driving cars. Soc. Stud. Sci. 48 (1), 25–56.
Subirats, Josep, Guitart, Jordi, 2015. Assessing and forecasting energy efficiency on cloud computing platforms. Future Gener. Comput. Syst. 45, 70–94.
Sun, Xiang, Ansari, Nirwan, Wang, Ruopeng, 2016. Optimizing resource utilization of a data center. IEEE Commun. Surv. Tutor. 18 (4), 2822–2846.
Toosi, Adel Nadjaran, Calheiros, Rodrigo N., Buyya, Rajkumar, 2014. Interconnected cloud computing environments: Challenges, taxonomy, and survey. ACM Comput. Surv. 47 (1), 1–47.
Tuli, Shreshth, Sandhu, Rajinder, Buyya, Rajkumar, 2020. Shared data-aware dynamic resource provisioning and task scheduling for data intensive applications on hybrid clouds using aneka. Future Gener. Comput. Syst. 106, 595–606.
Usmani, Zoha, Singh, Shailendra, 2016. A survey of virtual machine placement techniques in a cloud data center. Procedia Comput. Sci. 78, 491–498.
Van Engelen, Jesper E., Hoos, Holger H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.
Verma, Akshat, Ahuja, Puneet, Neogi, Anindya, 2008. pMapper: power and migration cost aware application placement in virtualized systems. In: ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing. Springer, pp. 243–264.
Verma, Manish, Gangadharan, GR, Narendra, Nanjangud C, Vadlamani, Ravi, Inamdar, Vidyadhar, Ramachandran, Lakshmi, Calheiros, Rodrigo N, Buyya, Rajkumar, 2016. Dynamic resource demand prediction and allocation in multi-tenant service clouds. Concurr. Comput.: Pract. Exper. 28 (17), 4429–4442.
Vora, Suchi, Yang, Hui, 2017. A comprehensive study of eleven feature selection algorithms and their impact on text classification. In: 2017 Computing Conference. IEEE, pp. 440–449.
Wagstaff, Kiri, Cardie, Claire, 2000. Clustering with instance-level constraints. In: AAAI/IAAI. Vol. 1097. pp. 577–584.
Whaiduzzaman, Md, Sookhak, Mehdi, Gani, Abdullah, Buyya, Rajkumar, 2014. A survey on vehicular cloud computing. J. Netw. Comput. Appl. 40, 325–344.
Wischik, Damon, Handley, Mark, Braun, Marcelo Bagnulo, 2008. The resource pooling principle. ACM SIGCOMM Comput. Commun. Rev. 38 (5), 47–52.
Wolski, Rich, 1998. Dynamically forecasting network performance using the network weather service. Cluster Comput. 1 (1), 119–132.
Xian, Changjiu, Lu, Yung-Hsiang, Li, Zhiyuan, 2007. Energy-aware scheduling for real-time multiprocessor systems with uncertain task execution time. In: 2007 44th ACM/IEEE Design Automation Conference. IEEE, pp. 664–669.
Xu, Minxian, Tian, Wenhong, Buyya, Rajkumar, 2017. A survey on load balancing algorithms for virtual machines placement in cloud computing. Concurr. Comput.: Pract. Exper. 29 (12), e4123.
Yadwadkar, Neeraja Jayant, 2018. Machine Learning for Automatic Resource Management in the Datacenter and the Cloud (Ph.D. thesis). UC Berkeley.
Yakimenko, Oleg A, Slegers, Nathan J, Bourakov, Eugene A, Hewgley, Charles W, Bordetsky, Alex B, Jensen, Red P, Robinson, Andrew B, Malone, Josh R, Heidt, Phil E, 2009. Mobile system for precise aero delivery with global reach network capability. In: 2009 IEEE International Conference on Control and Automation. IEEE, pp. 1394–1398.
Yang, Hailong, Zhao, Qi, Luan, Zhongzhi, Qian, Depei, 2014. iMeter: An integrated VM power model based on performance profiling. Future Gener. Comput. Syst. 36, 267–286.
Zhang, Jiangtao, Huang, Hejiao, Wang, Xuan, 2016. Resource provision algorithms in cloud computing: A survey. J. Netw. Comput. Appl. 64, 23–42.
Zhao-Hui, Y., Qin-Ming, J., 2012. Power management of virtualized cloud computing platform. Chinese J. Comput. 6, 015.
Zhu, Xiaojin, Goldberg, Andrew B., 2009. Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. 3 (1), 1–130.

Tahseen Khan obtained Bachelor of Science (Honours) in Physics and Master in Computer and Applications from Department of Physics and Department of Computer Science, Aligarh Muslim University, India. He is currently pursuing PhD in School of Information and Software Engineering, University of Electronic Science and Technology of China, China. His research interests include machine learning, deep learning and their applications in different areas such as Computer Vision and Cloud computing.

Prof. Wenhong Tian has a PhD from Computer Science Department of North Carolina State University, USA. He is a professor at University of Electronic Science and Technology of China. His research interests include dynamic resource scheduling algorithms and management in Cloud Data Centres, machine learning and deep learning algorithms for computer vision and natural language processing. He published about 70 journal and conference papers, and 3 English books in related areas. He is a member of ACM, IEEE and CCF.

Guangyao Zhou is currently pursuing PhD in School of Information and Software Engineering, University of Electronic Science and Technology of China, China. His research interests include machine learning, deep learning and their applications in Cloud computing.

Prof. Rajkumar Buyya is a Redmond Barry Distinguished Professor and Director of the Cloud Computing and Distributed Systems (CLOUDS). Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft, a spin-off company of the University, commercialising its innovations in Cloud Computing. He served as a Future Fellow of the Australian Research Council during 2012–2016. He has authored over 625 publications and seven textbooks, including "Mastering Cloud Computing" published by McGraw Hill, China Machine Press, and Morgan Kaufmann for Indian, Chinese and international markets, respectively. He has also edited several books, including "Cloud Computing: Principles and Paradigms" (Wiley Press, USA, Feb 2011). He is one of the highly cited authors in computer science and software engineering worldwide (h-index = 143, g-index = 317, 109800 citations). Microsoft Academic Search Index ranked Dr. Buyya as #1 author in the world (2005–2016) for both field rating and citations evaluations in the area of Distributed and Parallel Computing. "A Scientometric Analysis of Cloud Computing Literature" by German scientists ranked Dr. Buyya as the World's Top-Cited (#1) Author and the World's Most-Productive (#1) Author in the Cloud Computing. Recently, Dr. Buyya is recognised as a "2016 Web of Science Highly Cited Researcher" by Thomson Reuters and a Fellow of IEEE for his outstanding contributions to Cloud computing and Scopus Researcher of the Year 2017 with Excellence in Innovative Research Award by Elsevier. He has been recently recognised as the "Best of the World", in Computing Systems field, by The Australian 2019 Research Review. For further information on Dr. Buyya, please visit his cyberhome: www.buyya.com.