JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Survey on Artificial Intelligence and Machine Learning for Resources Allocation in Cloud Data Centers

[1]ZEENATH SULTANA, [2]Dr. RAAFIYA GULMEHER
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
Faculty of Engineering and Technology,
[1]Khaja Bandanawaz University, Kalaburagi, India

Abstract: In modern data centers, efficient resource scheduling and management of available resources is a major concern. Cloud computing has quickly 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 areas of cloud computing. In this service model, cloud providers provide resources to users/machines such as virtual machines, raw (block) storage, firewalls, load balancers, and network devices. One of the most critical aspects of cloud computing for IaaS is resource management. Scalability, quality of service, optimal utility, reduced overheads, increased throughput, reduced latency, specialized environment, cost effectiveness, and a streamlined interface are some of the benefits of resource management for IaaS in cloud computing. Traditionally, resource management has been accomplished through static policies, which impose certain constraints in a variety of dynamic scenarios, prompting cloud service providers to adopt data-driven, machine-learning-based approaches. Machine learning is employed to handle a wide range of resource management tasks, such as workload estimation, task scheduling, Virtual Machine (VM) consolidation, resource optimisation, and energy optimisation, to name a few. This paper provides a detailed comprehensive survey of background works that have used machine learning (ML) techniques to solve real-world problems in the cloud computing area, and they have used ML algorithms to optimize various objectives related to the cloud computing environment. We also discuss the role of Artificial Intelligence and ML in cloud computing and identify challenges in cloud data centres. Finally, we will discuss open issues and potential future research directions.

***IndexTerms* – Artificial Intelligence, Cloud data center, cloud computing, resources allocations, Machine learning, Virtual Machines**

I. INTRODUCTION

Currently, industry and academia are transferring their applications to the cloud. Cloud computing provides a platform for application developers to run their applications in the cloud without having to worry about server setups and configurations. Cloud providers, on the other hand, are constantly looking for ways to provide better services to developers while also considering efficiency. Cloud computing is typically supported by massive data centres containing thousands of powerful computers. These data centres can exist in a dedicated facility that houses all hardware and software belonging to cloud servers. On the other hand, data centres can be shared, with the owner leasing out cloud services as part of the facility to service providers.

The data center will be designed to house as many customers as possible in this scenario, and as a result, it will house numerous different applications. In order to solve a number of operational and maintenance issues in the data center, virtualization is implemented, providing users with virtual platforms instead of physical ones. The process of allocating resources to the virtualization unit, such as virtual machines or containers in the cloud, such as CPU, memory, storage, and network bandwidth, is known as resource management. A definitive plan for effective resource management does not exist. Depending on their primary goals, different cloud providers may have different approaches to resource management. Most of the time, the main goals are to cut down on the amount of time it takes to finish a job, cut down on the amount of time it takes to make a start, make more use of resources, cut costs, and The aforementioned goals are extremely significant. A 100 millisecond increase in response time, for instance, costs Amazon 1% of its revenue, according to a report. Google discovered that Google's traffic will decrease by 20% with a delay of 0.5 seconds in search generation [1]. The significance of maintaining Service Level Objectives (SLOs) is demonstrated in this report. According to [2], power accounts for approximately 70% of all data center operating costs, highlighting the significance of reducing energy consumption.

A survey on unbalanced workload found that the average utilization of CPU and memory was 17.76% and 77.93%, respectively. A similar study conducted in the Google data center found that a Google cluster's utilization of CPU and memory could not exceed 60% and 50%, respectively [5]. A data center's productivity declines as a result of the uneven workload, which increases energy consumption. It is proportional to the financial loss and operational costs of the data center. This inordinate energy utilization straightforwardly affects carbon impressions, which ought to be diminished on the grounds that an ideal machine retains the greater part of the most extreme energy utilization [6]. Data centers used approximately 35 Twh (Tera Watt hours) of energy in 2015, according to an EIA survey, and this number is expected to rise to 95 Twh by 2040.

In cloud computing, scheduling and distributing resources take into account available infrastructure, service level agreements, costs, and energy consumption. For instance, a cloud service provider ensures excellent Quality of Service (QoS) and user satisfaction while managing resources in accordance with the on-demand pricing model [2]. In a similar vein, resources must be allocated in such a way that each application receives the resources it requires without exceeding the cloud environment's capacity. Similar to resource allocation, which enables service providers to allocate resources for each module, resource allocation is responsible for dealing with the issue of applications starving [3].

Cloud computing provides consumers with high-quality services at a low cost [4]. Whereas, in terms of storage capacity, data centres provide a large number of resources and distributed computing models ready to assist with request resource allocation, which leads to non-ideal resource assignment. Energy utilisation is another issue that large data centres face. It has been observed that vitality consumes more than 20% of large data centres. Reduced energy consumption can save resources suppliers a significant amount of energy and money [5]. Utilizing the hardware resources in an elastic manner and shutting down servers that are not in use is the simplest and most effective approach. However, careful planning is required to ensure that data centers do not run out of resources as requests come in. Using a variety of approaches, including heuristics and algorithms, researchers have attempted to propose solutions in this field. These approaches have the drawbacks of not being workload-specific, not being able to adapt to shifts in the workload because they lack dynamicity, and requiring prior knowledge of the workload to adjust parameters. However, some researchers attempted to resolve this issue using machine learning. Since it is responsibility explicit, it can deal with dynamicity in responsibility conduct, and it needn't bother with any responsibility specialization [4-7].

In cloud computing, virtualization is regarded as an essential management tool that enables on-demand resource allocation and provisioning (Adnan et al., 2012)[8]. Virtualization seeks to shield users from physical and low-level system components. As a result, multiple applications can be hosted and run simultaneously, allowing for efficient resource configuration and use. Finally, these paradigms enable quick recovery and system fault tolerance. As a result, virtualization is a potent cloud computing backend technology that enables the high-level scalability, adaptability, and availability of cloud computing features. With the guide of these vital empowering advances, the cloud framework assets (for example host, stockpiling, and organization limit) in cloud server farms are virtualized and conveyed to cloud clients as Virtual Machines (VMs). As a result, the infrastructure resources (hosts) that should be assigned to incoming virtual machines are linked to the resource allocation for this environment. In most data centers, server consolidation is used to get the most use out of the physical resources. A good mapping of virtual machines to the hardware resources that are available is discovered by this resource management plan[7-15]. The available number of CPU cores, the required storage hard desk space, the physical memory, and the allotted network bandwidth are all met by this scheme's multidimensional mapping. In particular, it is a method of allocating resources to maximize the utilization of cloud physical resources. To put it another way, virtualization technology is a good way to manage dynamic resources on a cloud computing platform [16, 20]. By encapsulating the service in virtual machines and mapping it to each physical server, the cloud computing heterogeneity and platform irrelevance issue can be addressed more effectively. Despite the widespread acceptance of cloud technology as a means of providing multiple services, it faces numerous challenging issues regarding resource management, power consumption, security, service quality, and big data. Data centers' frequent flaws include their relatively high operating costs and power consumption, both of which have significant environmental effects. With the growing demand for energy and the need for a lot of computational power, addressing this shortcoming has become one of the top priorities. As a result, effective solutions to achieving a balance between the performance and quality of service provided at data centers and reducing power consumption are in high demand [21-26].

The challenges associated with resource management include managing uncertainties related to the workload and the system, effectively allocating diverse resources, and scheduling jobs mapped for a particular resource.

Before achieving this, several challenges arise, including: Dynamism, energy efficiency, scalability, and a service level agreement. Additionally, the requirement for multiple data centers brings up a number of issues, including scheduling, load balancing, multiple levels of abstraction, and sustainability issues. The goal of this study is to find the best way for cloud infrastructure providers to manage their resources in an energy- and SLA-compliant manner. This research contributes to the energy-efficient resource allocation of VMs to physical machines because virtualization, in which each physical resource is served by multiple Virtual Machines (VMs), is the core technology of cloud computing. CPU utilization is the SLA metric that is taken into account in this study. Since the pattern of demand for resources in cloud computing is always changing, this study uses resource prediction to allocate resources effectively. Machine learning and artificial intelligence (AI) are two technologies that have demonstrated their ability to alter a variety of fields. Organizations will be able to accelerate digital transformation and improve performance and efficiency by utilizing AI and ML in the cloud [27-29]. Machine learning (ML) methods are frequently used in fields such as computer vision, pattern recognition, and bioinformatics. The advancement of machine learning algorithms has benefited large-scale computing systems [30]. Google recently published a report outlining their efforts to optimise electricity, lower costs, and improve efficiency. By providing data-driven techniques for future insights, ML has drawn attention to dynamic resource scaling, which is regarded as a promising approach for predicting workload quickly and accurately.

As a result, this paper will focus on a review of challenges discovered in cutting-edge research in resource management using ML algorithms, including provisioning, VM consolidation, thermal prediction, and other management approaches. Then we'll discuss the benefits and drawbacks of various cutting-edge research studies in resource management that employ machine learning algorithms. We also discuss the role of AI and ML in cloud data centres, as well as open issues. Finally, we look into some of the potential future directions that researchers can pursue in order to improve state-of-the-art techniques.

II. Literature Survey

A few surveys have been conducted in this area. For instance, in this recent survey [31], they gave an in-depth study of resource management, but they are more focused on research articles in the context of cloud performance management and do not provide readers with a clear picture of the correlation between ML techniques and cloud computing specific objectives. Another survey [32] looked at resource management with an emphasis on energy consumption optimisation, which is not the focus of this paper. However, unlike those, the goal of this paper is on providing a literature review classified by the ML techniques used and comparing them. The recent related works carried out by various researchers are presented as follows:

Dhaya R. et al (2022) have investigated the Energy-Efficient Resource Allocation and Migration in Private Cloud Data Centre, Investigation of the Virtual Machine allocation method [33]. They focused on system structure investigations in framework of energy-efficient distribution of resources in private cloud data center architecture. On the other hand, they want to equip private cloud providers with the current design and performance analysis for energy-efficient resource allocation.

Madhusudhan H S et al (2021), has proposed the Hybrid VMP technique using Random Forest (RF) and Genetic Algorithm (GA) for Resource Allocation in Cloud Infrastructure [34]. The aim is to reduce the energy consumption of the data center and Maximizing resource utilization of the physical machines. Also this method reduces the time to find the optimal solution.

Priya Baldoss et al [2021] **have proposed the method of** Optimal-Resource Allocation and Quality of Service (QoS) Prediction in Cloud computing [35]. Also they investigated the prediction of the runtime of VMs dependent on metadata, accessible at start-up for finding the solutions to balance the workload. Finally carried out the comparsion of proposed method with conventional method in cloud computing.

Shanky Goyal, et al [2021] have developed the Optimized Framework using Whale Optimization Algorithm(WOA) for Energy-Resource Allocation in a Cloud computing[36]. They address the power consumption issue in cloud computing infrastructure and investigated the particle swarm optimization (PSO), cat swarm optimization (CSO), BAT, cuckoo search algorithm (CSA) optimization algorithm and the whale optimization algorithm (WOA) for balancing the load, energy efficiency, and better resource scheduling to make an efficient cloud environment. As a result the WOA is more cost effective than other algorithms because it consumes less energy and takes less time to respond and execute.

E. I. Elsedimy et al [2021], have examined Toward Improving the Energy Effectiveness and Limiting the SLA Infringement in Cloud Server farms [37]. To investigate and proposed the solutions to overcome the VMP problem such as VMPMOPSO using Multi objective algorithm based on Particle Swarm Optimization (PSO) method. Results of this work demonstrated that this solution is more efficient as compare FDD, MGGA, and VMPACS algorithms.

D. Viknesh Kumar[2020] has proposed the Multi-Cloud Framework based on Machine Learning for Resource Allocation in cloud[38]. This solution provides performance improvement of existing and future cloud storage with QoS Resource Distribution. The ML-RA method offers all data transfer from IOT to Cloud. As a results, higher than that of present age values were derived from the ML-RA method.

Arwa Mohamed [2020] has proposed the supervisor Controller based Software Defined Cloud Data Center (SC-boSD-CDC) framework for Dynamic Resource prediction and Allocation in Cloud data center[39]. Also a Genetic Algorithm (GA) has been applied to solve the multi-objective problem of Cloud Data Center (CDC). Further a Virtual Machines (VMs) placement algorithm has been utilised to cloud computing resources allocation and to select suitable bandwidth links among switches. As a result, it is seen that the increase in CPU and memory utilization and overall power consumption reduction.

Álvaro López García, et al[2020] has proposed DEEP-Hybrid-DataCloud for ML Workloads and Applications[40]. This method allows transparent access to existing e-Infrastructures and effectively exploiting distributed resources for the most compute-intensive tasks. Also standardized API have been developed for machine learning models and offers a set of cloud oriented services with the use of server less architecture and DevOps approach to allow easy share, publish and deploy of the ML models.

Kanav S. et al[2020] has introduced the solution for cloud storage mechanism using Machine Learning and this approach were efficient and secure[41]. The encryption mechanism named as Rivest–Shamir–Adleman (RSA) with Triple Data Encryption Standard (DES) approach were employed to solve the security issues and data storage issue is solved using Modified Best Fit Decreasing (MBFD) with Whale Optimization algorithm (WOA)&Artificial Neural Network (ANN) approach. As a result, it is seen that proposed system outperforms in terms of energy consumption, delay, and Service Level Agreement (SLA) violation.

Z Chen et al [2019] has been proposed advantage actor-critic based reinforcement learning (RL) framework for Learning-Based Resource Allocation in the Cloud Data Center[42]. The simulation of proposed system have been done using Google cluster-usage traces and result have shown that the proposed method efficient in cloud resource allocation. Also it is seen that the proposed method outperforms over classic resource allocation algorithms and achieves faster convergence speed as compare to traditional policy gradient method.

Sambit K.M et al [2018] has investigated and proposed the mapping algorithm for Energy efficient virtual machines-placement in the cloud data center [43]. They provided the solution to solve optimal mapping of tasks to VMs and VMs to PMs(physical machines) issues. The proposed algorithm is executed on CloudSim simulator and results indicate that the energy consumption is reduced and makespan and task rejection rate is minimized. This demonstrates that the proposed algorithm outperforms over other standard algorithms.

Yang, R,et al . (2018) has proposed the ML based Intelligent Resource Scheduling at Scale to solve scheduling problems [44]. Also they proposed the ML to autonomously exploit and understand workloads and also environments. The results indicate that the proposed method is efficient and provides architecture optimization and improvement in efficiency. Also the prediction accuracy of node obtained is 92.86%.

R. Dharani et al [2016] has investigated configuration of different layers such as IaaS and SaaS in cloud and proposed **an** efficient cloud resource allocation optimization algorithm for Resource Allocation and Scheduling[45]. The iterative algorithm is employed to provide mechanism of optimization of resource allocation. As a result, our method computes scheduling plans that produce make span as good as the best known method while dramatically reducing monetary costs.

Mehiar Dabbagh et al [2015] has investigated Energy Efficient Resource Allocation and proposed the Provisioning Framework using ML for the Cloud Data Centers[46]. This method predicts no. of requests from VM and associated amount of CPU and memory resources requests. The Proposed system offers reduce energy consumption and provides accurate estimations of no. of physical machines (PMs) of cloud data centers. The result confirms that proposed framework makes the substantial energy savings.

Parvathy S. Pillai [2014] has proposed the Uncertainty Principle of Game Theory for Resource Allocation mechanism for machines on cloud[47]. They carried out the comparison of results obtained with proposed mechanism with other existing resource allocation methods. The result shows that the proposed mechanism provides better resource utilization and higher request satisfaction.

As a result of reading this survey, the reader will have a general understanding of why researchers chose specific ML techniques for specific goals, what the drawbacks of their approaches are, and also how they addressed gaps in previous works. Hopefully, this will provide a clear picture of current research in the use of AI and ML techniques for resource management in cloud data centres.

III. CLOUD DATA CENTER

The term "cloud" can refer to a collection of services, such as a global or local network of servers with specialised functions. The cloud is not a physical entity, but rather a collection or network of remote servers linked together to perform a single task. A data centre is a facility or location that houses networked computers and associated components (such as telecommunications and storage) that assist businesses and organisations in handling large amounts of data. These data centres enable data to be organised, processed, stored, and transmitted across business applications.

A data centre, also known as a data center, is a facility comprised of networked computers, storage systems, and computing infrastructure that businesses and other organisations use to organise, process, store, and broadcast large amounts of data. A data centre is a focal point and critical asset for everyday operations because a business typically relies heavily on applications, services, and data within it. To secure and protect in-house, onsite resources, enterprise data centres are increasingly incorporating cloud computing resources and facilities. As more businesses turn to cloud computing, the lines between cloud providers' and enterprise data centres become blurred.

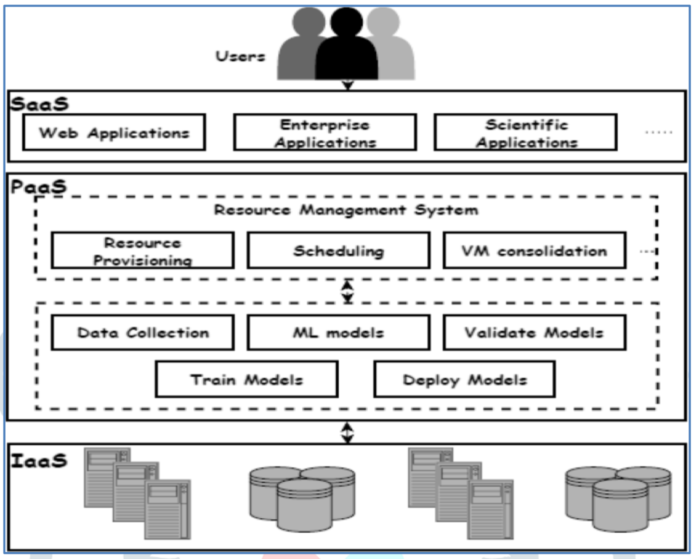

Figure 1: ML-based Cloud Computing Components [48]

Cloud servers are typically housed in data centres or server farms, and there is a vast network of data centres all over the world that are linked by the internet. Cloud providers with big names like Google, Microsoft, and AWS have their own networks of data centres that form their own cloud platforms. You can rent both space and infrastructure in a cloud data centre. Cloud providers will run large data centres that are fully secure and compliant. You can gain access to this infrastructure by using various services that provide greater flexibility in usage and payment. Figure 1 presents about the cloud computing components based on ML methods.

With the increasing importance of data in today's businesses, data management is critical for managing and governing large data sets for business growth. Companies are using advanced analytics and automation tools to process massive amounts of data. They are also leveraging well-equipped data centres for better data management. Data centres provide seamless data backup and recovery services while also supporting cloud storage applications and transactions. Companies are turning to emerging technologies like artificial intelligence and machine learning to advance their data centre infrastructure due to their distinct capabilities to business data storage. Machine learning, a sophisticated subset of artificial intelligence, can examine and discover patterns in massive amounts of data. It has the potential to optimise every aspect of datacenter operations, including planning and design, uptime maintenance, IT workload management, and cost control. AI and machine learning are expected to vastly improve data centre efficiency. According to IDC, embedded AI functionality will enable 50% of IT assets in data centres to run autonomously [49]. AI cloud computing is a concept used by several businesses that combines Artificial Intelligence with cloud computing.AI hardware and software (including open source) could be merged to enable enterprises to access and leverage AI by providing AI software-as-a-service on hybrid cloud infrastructure. An AI-powered cloud environment learns from data to predict and resolve issues before users even observe them[50].

IV. ROLE OF AI & ML IN CLOUD ENVIRONMENT

AI and ML are two technologies that have shown to be transformative in a variety of fields. Organizations will be able to improve performance and efficiency while driving digital transformation by utilising AI and ML in the cloud. The majority of businesses have already begun to use the public cloud. Moreover, by incorporating AI into the cloud, organisations hope to increase productivity and provide better IT solutions to their customers. John McCarthy coined the phrase for the first time in 1956. With each passing year, this terminology grew in popularity. Presently, it has evolved into a mainstream technology that is widely used to solve complex problems in a variety of industries. The global AI market is expected to be nearly $60 billion by 2025, with the global ML market expanding at a 42.08% CAGR between 2018 and 2024. Artificial Intelligence and Machine Learning are the presenter's most emerging technologies for developing intelligent systems. Despite the fact that these two technologies are related, there is a minor difference between them. AI is a wider concept that produces intelligent machines that can stimulate human thinking capability and behaviour, whereas ML is a subset of AI that allows machines to learn from data without requiring any coding. In a nutshell, ML is a technology that aids in the development of AI. However, AI does not have to be developed using ML, despite the fact that it makes Artificial Intelligence much more convenient [50]. The figure 2 depicts classification of ML methods for resource allocation in cloud data centers. The three most common types of machine learning models used in resource management research are supervised, unsupervised, and reinforcement learning (RL).

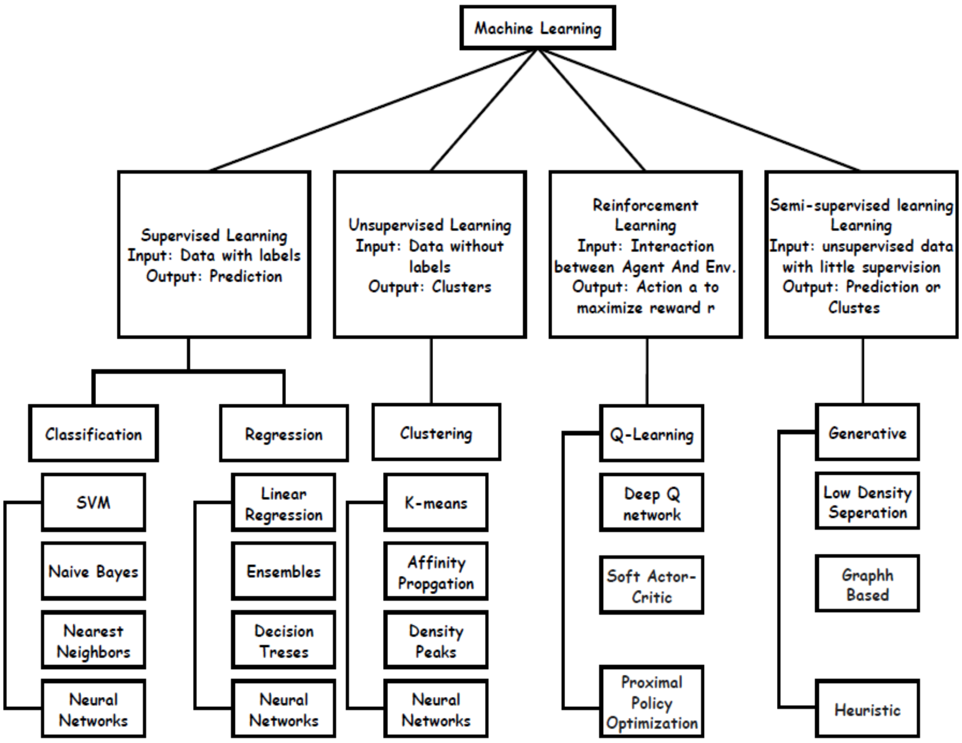

Figure 2: Classification of Machine Learning Techniques [48]

AI and machine learning have made significant contributions to the growth of the cloud computing industry. AL and ML aid in the automation of routine activities within IT infrastructure, increasing productivity and improving organisational performance and efficiency. AI capabilities enabled by cloud computing enable businesses to become more efficient, strategic, and insight-driven, all while enjoying greater flexibility, agility, and cost reduction[50].

Let us look at some of the aspects AI and machine learning have altered the cloud landscape:

***Lower costs*:** Adoption of the cloud allows businesses to only pay for what they use rather than setting up and managing large data centres, as is the case with traditional infrastructure costs. This upfront cost can be used to fund the more strategic development of AI tools and accelerators, which will generate more significant revenues while saving the enterprise basic costs. Furthermore, using auto-scaling groups, AI applications can analyze data without human intervention and gather understanding from it.

Enhanced data management: Data management is a time-consuming task. It is one of the most significant challenges that any enterprise faces in this data-driven world. Cloud-based AL and ML applications aid in data management by identifying, updating, cataloguing, and providing real-time data insights to customers. You can also detect fraudulent activities and notice unusual patterns in the system by using Artificial Intelligence tools. Banking and financial institutions use this technology to remain relevant and secure in high-risk environments.

Intelligent automation: AI-driven cloud computing allows businesses to automate complex and repetitive tasks to boost productivity and analyze data without human intervention, in furthermore to being more efficient, strategic, and insight-driven. AI and machine learning can also be used to monitor and manage workflows in information technology departments. As AI handles complex tasks, IT teams can concentrate on strategic operations.

Availability of Advanced Infrastructure: AI applications typically perform exceptionally well when run on servers with multiple GPUs (Graphical Processing Units). Such systems are prohibitively expensive for many organisations. AI as a service is thus a more cost-effective option for these organisations.

***Team integration:* By m**erging artificial intelligence with cloud computing can also help with integration between DevOps teams, because the cloud allows for information sharing across sectors and allows teams to work more efficiently.

AI and machine learning aid in the automation of tasks within IT infrastructure. As a result, every organisation is thinking about incorporating these technologies into their system. Such technologies must be widely adopted for a business to evolve and remain competitive in the market.

By merging AI and the cloud, a business can provide greater efficiency, productivity, and security, both in terms of the information handled and the accuracy of AI-structured processes and procedures.

V. Open Issues, Challenges and Future Research Directions

In this section, we present about the issues and challenges identified in Machine learning based resource allocation in the state-of-art research and future research directions.

Issues and Challenges:

Cloud data centres' primary issues include energy efficiency, robustness, and scalability (DCs). Researchers and industry are working hard to find viable solutions to the challenges that DCs face. Cloud computing services enable the allocation of VM to various tasks based on the demand of cloud consumers. According to previous research, there are a few aspects of resource allocation in cloud computing that need to be addressed.

Strategic: Strategic resource allocation techniques are used to increase or decrease resource allocation based on the ever-changing demands of cloud consumers. Strategic-based resource allocation has already addressed the aspects of meeting the fluctuating demands of users through prediction as well as utilizing artificial intelligence to allocate resources.

One important area of future research will be to discover the details for detecting resource and workload for improved mappings for job execution and scheduling. To that end, workloads should be executed efficiently in order to be flexible, scalable, and optimal, avoiding resource under and overutilization.

Furthermore, the use of artificial intelligence algorithms in resource allocation reduces error chances and failure rate to nearly zero, resulting in better precision and accuracy for resource allocation in cloud computing. Even so, artificial-based resource allocation must also consider cost and improve the method to make it suitable for larger systems.

Target Resource: Target resource-based resource allocation identifies the specific resource for which the allocation technique is intended. The aspects of the VM position on a physical machine and network aware resource allocation were previously covered.

The network-aware resource allocation must priorities minimizing communication between VMs from different sub-data centres (or servers). The techniques should primarily aim to meet the needs of the consumers while also lowering communication costs by determining the shortest path between the VMs.

Optimization: The demands of customers are increasing almost daily, necessitating an effective strategy for allocating resources to meet these requirements. By guaranteeing QoS to cloud users, optimization-based resource allocation solves this problem. In optimization-based resource allocation, the researchers have already proposed methods for better resource utilization and the assurance of quality of service.

By taking into account the SLA negotiation procedure in cloud computing environments, optimization-based resource allocation techniques ought to concentrate on increasing total profit efficiency and customer satisfaction. In addition, system failures must be taken into account when considering the penalty limit.

*Scheduling:*The scheduling of resources ensures the effective and efficient utilization of resources, as well as the early identification of resource capacity. The work on allocating resources to tasks based on priority, cost, and CPU time (RR) has already been completed in scheduling-based resource allocation.

Different scheduling criteria must be reassessed in order to implement the various resource scheduling algorithms. In addition, based on previous research, we felt the need to test the resource scheduling algorithms in a real-world setting. We discovered that dynamic resource scheduling is an issue.

Power: A power-based resource allocation approach seeks to reduce energy consumption, heat generation, and resource waste. Techniques for consuming less energy and producing more heat have already been proposed in power-based resource allocation. A comprehensive study of power-based resource allocation techniques is required, particularly with regard to green data centre optimization. Low energy aware technologies use low power energy-efficient hardware equipment to reduce energy utilisation and peak power consumption.

The majority of researchers were using supervised learning. The majority of the works that used supervised learning used recent workloads to predict current/future workload. Their primary goals are to reduce the number of servers or virtual machines to save energy and money. A job should be scheduled on a VM/server with sufficient resources. Adding the new job to it should not interfere with other jobs on that VM/server. The ultimate VM/server is chosen in a greedy manner from among all the VMs/servers that satisfy the above conditions, which does not always guarantee a return to the most optimum choice.

Following supervised learning, reinforcement learning is the most commonly investigated method [63]. In most RL works, the state-space consists of job DAGs, and the action space consists of scheduling jobs and specifying the level of job parallelism. Their primary goal is to reduce job completion time. The problem with these solutions is that their RL methods are not online, so whenever there is a change in the workload, re-training the model and responding to that change will be delayed. In those cases, they should use a method other than RL.

Unsupervised learning is the least suitable for resource management because it divides workloads into clusters, and many workloads end up in the same cluster. As a result, the system allocates the same resources to all cluster members.

The only time unsupervised learning is beneficial is when the workload changes.

Researchers studying online learning with RL are a promising direction for future work. They can make use of meta-learning. It will be interesting to consider more than two conflicting objectives by using multiple agents, and researchers should also investigate pre-emptible jobs by using a specific agent to handle job preemption decisions. It would be preferable if RL models did not rely on the job resource usage profile because it is not always accurate; instead, supervised learning should be used to estimate each job resource profile.

Future Research Directions:

We propose potential future research directions based on identified challenges and limitations to improve resource management in cloud data centres. The following are some of the key topics identified for future research.

1. Performance and Online Profiling of Workload
2. Multiple Resource Usage in VM consolidation
3. Cloud Network Traffic
4. Host Temperature
5. False Host Overloaded Detection
6. Energy metering at Software-Level
7. SLA-based VM Management
8. QoS-Aware Resource Provisioning
9. Varying Patterns of a Service Tenant in Resource Allocation
10. Single ML model in energy consumption prediction
11. Prediction Accuracy in Auto-Scaling of web applications
12. Time-Series Prediction Data
13. Data Training
14. VM Multi Resource

We have discussed the challenges of machine-learning-based resource management in a cloud computing environment, as well as the various methods that have been used in recent years to address these issues, in addition to their benefits and drawbacks. There has been a significant increase in the number of research studies at how to employ machine learning techniques to predict workload, energy consumption, and other tasks in recent years. These techniques employ various ML methods to address a wide range of problems. Finally, new potential future research directions are proposed to strengthen the current ML methods for resource management in cloud-based systems based on the challenges and drawbacks identified in the state-of-the-art work. This paper's overall knowledge assists cloud researchers in comprehending cloud resource management and the significance of ML techniques.

VI. CONCLUSION

In this paper, we gave an in-depth survey of background works that have used ML techniques to solve real-world problems in the cloud computing area, and they have used ML algorithms to optimize various cloud computing objectives. We also discussed the role of AI and ML for resource allocation in cloud data centre. Finally we as well described potential future directions for this research area, and we hope that using this literature review will assist researchers in making more improvement in this field.

Our findings demonstrated that ML models can be used in cloud computing systems to achieve various optimisation goals and handle complex tasks. The application of ML approaches also opens up a new path for intelligent resource and application management. This article demonstrates the advancement of machine learning approaches in current research and assists readers in understanding the research gap in this field. One promising approach for increasing system efficiency is to use advanced ML techniques such as reinforcement learning and deep learning (DL) to perform intelligent resource management.

REFERENCES

[1] Y. Einav, Amazon Found Every 100ms of Latency Cost them 1% in Sales, 2019 (accessed September 25th, 2020). [Online]. Available: https://www.gigaspaces.com/blog/amazon-foundevery- 100ms-of-latency-cost-them-1-in-sales/

[2] M. Rareshide, Power in the Data Center and its Cost Across the U.S., 2017 (accessed September 25th, 2020). [Online]. Available: https://info.siteselectiongroup.com/blog/power-inthe-data-center-and-its-costs-across-the-united-states

[3] S. Gong, B. Yin, Z. Zheng, and K.-Y. Cai, “Adaptive multivariable control for multiple resource allocation of service-based systems in cloud computing”, IEEE Access, vol. 7, pp. 13817–13831, 2019. DOI: 10.1109/ACCESS.2019.2894188.

[4] S. H. H. Madni, M. Sh. A. Latiff, Y. Coulibaly, and Sh. M. Abdulhamid, “Recent advancements in resource allocation techniques for cloud computing environment: A systematic review”, Cluster Computing, vol. 20, no. 3, pp. 2489–2533, 2017. DOI: 10.1007/s10586-016-0684-4.

[5] Q. Qi and F. Tao, “A smart manufacturing service system based on edge computing, fog computing, and cloud computing”, IEEE Access, vol. 7, pp. 86769–86777, 2019. DOI: 10.1109/ACCESS.2019.2923610.

[6] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Dang, and K. Pentikousis, “Energy-efficient cloud computing”, The Computer Journal, vol. 53, no. 7, pp. 1045–1051, 2010. DOI: 10.1093/comjnl/bxp080

[7] Ali Shahidinejad, Mostafa Ghobaei-Arani, and Mohammad Masdari. Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. Cluster Computing, pages 1{24, 2020.

[8] Adnan, M.A., Sugihara, R., Gupta, R., 2012. ‘‘Energy Efficient Geographical Load Balancing via Dynamic Deferral of Workload”. In: Proceedings of the 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 188–195.

[9] Gai, K., Qiu, M., Liu, M., Zhao, H. ‘‘Smart Resource Allocation Using Artificial Intelligence in Content-Centric Cyber-Physical Systems”. In: International Conference on Smart Computing and Communication, Springer, Cham, pp. 39–52.

[10] Moreno, I.S., Xu, J., 2011. Energy-efficiency in cloud computing environments: towards energy savings without performance degradation. Int. J. Cloud Appl. Comput. (IJCAC) 1 (1), 17–33.

[11] Xiao, Z., Song, W., Chen, Q., 2013. Dynamic resource allocation using virtual machines for cloud computing environment. IEEE Trans. Parallel Distrib. Syst. (TPDS) 24 (6), 1107–1117.

[12] Yousafzai, A., Gani, A., Noor, R.M., Sookhak, M., Talebian, H., Shiraz, M., Khan, M.K., 2017. Cloud resource allocation schemes: review, taxonomy, and opportunities. Knowl. Inf. Syst. 50 (2), 347–381.

[13] Zhang, Q., Zhu, Q., Boutaba, R., 2011. ‘‘Dynamic Resource Allocation for Spot Markets in Cloud Computing Environments”. In: 2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC), pp. 178–185.

[14] Y. Zhang, J. Yao, and H. Guan, “Intelligent cloud resource management with deep reinforcement learning,” IEEE Cloud Computing, vol. 4, no. 6, pp. 60–69, 2017.

[15] Zhou, Z., Hu, Z.G., Yu, J.Y., Abawajy, J., Chowdhury, M., 2017. Energy-efficient virtual machine consolidation algorithm in cloud data centers. J. Cent. South Univ. 24 (10), 2331–2341.

[16] Kawsar Haghshenas and Siamak Mohammadi. Prediction-based underutilized and destination host selection approaches for energy-efficient dynamic vm consolidation in data centers. The Journal of Supercom- puting, pages 1{18, 2020.

[17] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. IEEE transac-tions on cybernetics, 50(8):3668{3681, 2019.

[18] Kumar P, Kumar R (2019) Issues and challenges of load balancing techniques in cloud computing: A survey. ACM Comput Surv (CSUR) 51(6):1–35

[19] Kumar J, Singh AK, Buyya R (2021) Self-directed learning-based workload forecasting model for cloud resource management. Inf Sci 543:345–366

[20] Kumar J, Singh AK, Mohan A (2021) Resource-efficient load ‐ balancing framework for cloud data center networks. ETRI J 43(1):53–63

[21] Aibin M (2020) LSTM for Cloud Data Centers Resource Allocation in Software-Defined Optical Networks. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, New York, p 0162–0167

[22] Amazon Web Services (2016) Elastic Compute Cloud (EC2) Cloud Server & Hosting AWS. [Online] Available: https:// aws. amazon. com/ ec2 Accessed 20 Apr 2022

[23] Arunarani AR, Manjula D, Sugumaran V (2019) Task scheduling techniques in cloud computing: A literature survey. Future Generation Computer Systems 91:407–415

[24] Aslam S, Shah MA (2015) Load balancing algorithms in cloud computing: A survey of modern techniques. In: 2015 National software engineering conference (NSEC). IEEE, Rawalpindi, p 30–35

[25] Baeldung (2022) A Guide to DeepLearning4J. [Online] Available at: https:// www. baeld ung. com/ deepl earni ng4j. Accessed 20 Apr 2022

[26] Cisco Systems (2016) Cisco Global Cloud Index: Forecast and Methodology. pp 1–41

[27] Gomathi B, Karthikeyan K (2013) Task scheduling algorithm based on hybrid particle swarm optimization in cloud computing. Appl Inf Techno 55:33–38

[28] Katyal M, Mishra A (2014) A comparative study of load balancing algorithms in cloud computing environment. arXiv preprint rXiv:1403.6918

[29] Shafiq DA, Jhanjhi NZ, Abdullah A, Alzain MA (2021) A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications. IEEE Access 9:41731–41744

[30] Khan T, Tian W, Zhou G, Ilager S, Gong M, Buyya R (2022) Machine learning (ML)–Centric resource management in cloud computing: A review and future directions. J Netw Comp Appl 204. https:// doi. Org 10. 1016/j. jnca. 2022. 103405

[31] S. K. Moghaddam, R. Buyya, and K. Ramamohanarao, “Performanceaware management of cloud resources: A taxonomy and future directions,” ACM Computing Surveys (CSUR), vol. 52, no. 4, pp. 1–37, 2019.

[32] K. Braiki and H. Youssef, “Resource management in cloud data centers: a survey,” in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019, pp. 1007–1012

[33] Dhaya R. , Ujwal U. J.,2 Tripti Sharma , Mr. Prabhdeep Singh ,Kanthavel R. ,Senthamil Selvan,6 and Daniel Krah, Energy-Efficient Resource Allocation and Migration in Private Cloud Data Centre Hindawi Wireless Communications and Mobile Computing Volume 2022, Article ID 3174716, 13 pages https://doi.org/10.1155/2022/3174716

[34] Madhusudhan H S , Satish Kumar T , S.M.F D Syed Mustapha , Punit Gupta , and Rajan Prasad Tripathi , Hybrid Approach for Resource Allocation in Cloud Infrastructure Using Random Forest and Genetic Algorithm Hindawi Scientific Programming Volume 2021, Article ID 4924708, 1-10 pages.

[35] Priya Baldoss and Gnanasekaran Thangavel, Optimal Resource Allocation and Quality of Service Prediction in Cloud, Computers, Materials & Continua, CMC, 2021, vol.67, no.1

[36] Shanky Goyal, Shashi Bhushan , Yogesh Kumar , Abu ul Hassan S. Rana , Muhammad Raheel Bhutta 5 , Muhammad Fazal Ijaz and Youngdoo Son, An Optimized Framework for Energy-Resource Allocation in a Cloud Environment based on the Whale Optimization Algorithm MDPI Sensors 2021, 21, 1583.

[37] E. I. Elsedimy and Fahad Algarni, Toward Enhancing the Energy Efficiency and Minimizing the SLA Violations in Cloud Data Centers, Hindawi Applied Computational Intelligence and Soft Computing Volume 2021, Article ID 8892734, 14 pages.

[38] D. Viknesh Kumar, Multi-Cloud Framework On Machine Learning Resource Allocation, ICTACT Journal On Data Science And Machine Learning, June 2020, Volume: 01, Issue: 03

[39] Arwa Mohamed, Mosab Hamdan, Ahmed Abdelazizb , Sharief F. Babiker, Dynamic Resource Allocation In Cloud Computing Based On Software-Defined Networking Framework, Open Journal of Science and Technology. Vol. 3 No. 3, 2020

[40] Álvaro López García, et al A Cloud-Based Framework for Machine Learning Workloads and Applications, IEEE Access, VOLUME 8, 2020

[41] Kanav Sadawarti, Satish Saini, Machine Learning Based Efficient and Secure Storage Mechanism in Cloud Computing, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5, March 2020.

[42] Z Chen, J Hu and G Min, Learning-Based Resource Allocation in Cloud Data Center Using Advantage Actor-Critic, Institute of Electrical and Electronics Engineers (IEEE)' 2019

[43] Sambit Kumar Mishraa, Deepak Puthal b, Bibhudatta Sahooa, Prem Prakash Jayaramanc, Song Jund, Albert Y. Zomaya e, Rajiv Ranjand, Energy-efficient VM-placement in cloud data center, Elsevier, Sustainable Computing: Informatics and Systems 20 (2018) 48–55.

[44] Yang, R, Ouyang, X, Chen, Y et al. (2 more authors) (2018) Intelligent Resource Scheduling at Scale: a Machine Learning Perspective. In: IEEE International Symposium on Service Oriented System Engineering. 2018 IEEE SOSE, 26-29 Mar 2018, Bamberg, Germany. IEEE , pp. 132-141. ISBN 978-1-5386-5207-7.

[45] R. Dharani , Dr. M. Kalaiarasu, Efficient Resource Allocation And Scheduling In Cloud Computing Environment, International Journal Of Research In Computer Applications And Robotics, Vol.4 Issue 3, Pg.: 48-55 March 2016.

[46] Mehiar Dabbagh, Bechir Hamdaoui, Mohsen Guizani and Ammar Rayes, Energy-Efficient Resource Allocation and Provisioning Framework for Cloud Data Centers, IEEE Transactions on Network and Service Management · December 2015

[47] Parvathy S. Pillai, and Shrisha Rao, Resource Allocation in Cloud Computing Using the Uncertainty Principle of Game Theory, IEEE Systems Journal, 2014.

[48] Tahseen Khana, Wenhong Tiana, Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions, https://arxiv.org/pdf/2105.05079.pdf

[49] Vivek Kumar, January 26, 2021 https://www.analyticsinsight.net/powering-data-centers-with-ai-and-machine-learning/

[50] Blogs / By Som D Role of Artificial Intelligence & Machine Learning in Cloud Environment https://www.rapyder.com/blogs/role-artificial-intelligence-machine-learning-cloud-environment/