Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 12 Issue 01, 2023

Deep Learning and Optimization Approaches in A Proactive Auto-scaling and Energy-Efficient VM Allocation Framework Using an Online Multi-Resource Capsule Shuffle Attention Network for Cloud Data Centres: A Review

Rashmita Khadimzada

Associate Professor, Department of Computer Science and Engineering, Siam Delta Engineering Institute, Thailand

Email: rashmita.khadimzada@sdei-th.edu

Peer Review Information

Submission: 05 Jan 2023

Revision: 26 Jan 2023

Acceptance: 11 Feb 2023

Keywords

Cloud Computing, Auto-Scaling, Virtual Machine Allocation, Deep Learning, Capsule Neural Networks, Energy-Efficient Data Centres.

Abstract

Cloud computing has become a fundamental component of modern digital infrastructure, offering scalable and flexible access to computational resources through distributed cloud data centres. As the demand for digital services continues to grow, efficient management of large-scale workloads while ensuring service performance, energy efficiency, and sustainability has become a critical challenge. One of the primary issues in cloud environments is the dynamic allocation of virtual machines (VMs) to physical servers, where inefficient allocation can lead to underutilization of resources, increased energy consumption, and violations of service level agreements (SLAs). To address these challenges, proactive auto-scaling and energy-efficient VM allocation frameworks have gained significant attention. Unlike traditional reactive methods, proactive approaches leverage predictive models to anticipate workload variations and allocate resources in advance, thereby improving system responsiveness. Recent advancements integrate deep learning techniques with optimization algorithms to enhance prediction accuracy and allocation efficiency. Models such as convolutional neural networks, recurrent neural networks, and capsule networks effectively capture complex workload patterns, while attention mechanisms further refine predictions by focusing on critical resource features, enabling more intelligent and efficient cloud resource management.

Introduction

Cloud computing has revolutionized the way computing resources are delivered and consumed in modern information systems. By providing on-demand access to scalable computing infrastructure, cloud computing enables organizations to deploy applications and services without investing in expensive hardware resources. Cloud data centres host large clusters of servers that provide computing power, storage capacity, and networking capabilities for various cloud services. With the increasing demand for cloud-based applications, efficient management of cloud resources has become a critical challenge for cloud service providers.

One of the primary issues faced by cloud data centres is the dynamic nature of workloads. Applications hosted in cloud environments experience frequent fluctuations in resource demands due to varying user activities, application workloads, and system conditions. As

a result, cloud systems must continuously adjust the allocation of computing resources to maintain service performance and reliability. Virtualization technology plays a crucial role in enabling flexible resource management by allowing multiple virtual machines to run on a single physical server.

Virtual machine allocation is a key process in cloud resource management. The objective of VM allocation is to distribute workloads across physical servers in a way that maximizes resource utilization while minimizing energy consumption and operational costs. Inefficient VM placement strategies may lead to server overloading or underutilization, resulting in degraded system performance and increased energy consumption. Therefore, designing efficient VM allocation algorithms is essential for maintaining the performance and sustainability of cloud data centres.

Energy consumption is another major concern for large-scale cloud infrastructures. Data centres require substantial amounts of electricity to power servers, cooling systems, and networking equipment. Studies have shown that energy costs represent a significant portion of the operational expenses of cloud service providers. Consequently, reducing energy consumption while maintaining system performance has become an important research objective in cloud computing.

Auto-scaling mechanisms have been widely adopted to address workload fluctuations in cloud systems. Auto-scaling allows cloud infrastructures to dynamically increase or decrease the number of active virtual machines based on workload demands. Traditional auto-scaling approaches are often reactive, meaning that resources are allocated only after workload changes occur. However, reactive scaling may lead to delays in resource provisioning and temporary service degradation.

Literature Review

Saxena and Singh (2021) proposed a proactive auto-scaling framework that utilizes a multi-resource neural network model to predict cloud workload demands. The system analyses historical resource utilization data to forecast CPU, memory, and network usage simultaneously. Based on these predictions, the framework dynamically allocates virtual machines before workload spikes occur. Experimental results showed that the proactive approach significantly reduces SLA violations and improves resource utilization compared with reactive scaling strategies.

Zhao et al. (2021) introduced a reinforcement learning-based VM allocation framework for cloud data centres. The proposed model treats VM placement as a sequential decision-making problem and learns optimal resource allocation policies through continuous interaction with the cloud environment. Simulation results demonstrated improved energy efficiency and load balancing compared with heuristic-based allocation methods.

Zhang et al. (2022) proposed a deep learning-based multi-resource prediction model using long short-term memory networks. The model predicts future resource demands based on historical workload patterns and enables proactive resource allocation. The study showed that LSTM-based prediction models significantly improve resource utilization and reduce response time in cloud applications.

Sinha et al. (2022) developed a hybrid optimization algorithm combining ant colony optimization and genetic algorithms for VM placement. The hybrid approach explores multiple candidate solutions to identify optimal VM placement strategies that minimize energy consumption and improve load balancing across cloud servers.

Dasgupta et al. (2023) proposed a capsule neural network-based resource prediction framework for cloud data centres. The model captures hierarchical relationships between multiple resource parameters and improves prediction accuracy compared with conventional neural networks. The framework demonstrated improved performance in proactive auto-scaling and energy-efficient resource allocation tasks.

Beloglazov et al. (2020) proposed an energy-aware dynamic virtual machine consolidation framework for cloud data centres aimed at reducing power consumption while maintaining system performance. The framework monitors server utilization levels and migrates virtual machines when physical hosts become underutilized or overloaded. By consolidating workloads onto fewer servers during low demand periods, idle machines can be switched to low-power states. Experimental results showed that the proposed consolidation strategy significantly reduces energy consumption and operational costs while maintaining service availability. However, frequent VM migration operations may introduce network overhead and temporary performance degradation.

Mishra and Sahoo (2021) introduced a machine learning-based workload prediction framework for proactive cloud auto-scaling. The system analyses historical workload patterns to forecast resource demands and dynamically allocate virtual machines. The predictive model uses regression techniques to estimate future CPU and memory utilization. Based on these predictions,

the framework adjusts the number of active virtual machines to maintain service performance. Experimental evaluation demonstrated improved prediction accuracy and reduced SLA violations compared with traditional reactive auto-scaling methods. However, the authors noted that prediction accuracy depends on the quality and size of historical datasets used for training.

Chen et al. (2021) proposed a deep neural network-based auto-scaling framework for cloud applications. The system uses convolutional neural networks to extract workload features and predict resource utilization patterns. The predicted resource demand enables proactive allocation of virtual machines, improving system performance during workload spikes. Simulation results indicated that the proposed model improves resource utilization and reduces response time in cloud services. However, training deep learning models requires significant computational resources, which may limit scalability in extremely large cloud infrastructures.

Kumar and Kumar (2021) developed a hybrid optimization algorithm combining Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) for energy-efficient VM allocation. The hybrid algorithm evaluates multiple candidate VM placement configurations and selects the optimal solution that minimizes energy consumption and balances workload distribution across servers. Experimental results showed that the hybrid GA–PSO algorithm improves resource utilization and reduces energy consumption compared with conventional VM allocation methods. Nevertheless, the computational complexity of hybrid optimization algorithms remains a challenge for large-scale cloud environments.

Wang et al. (2023) proposed an attention-based deep learning framework for multi-resource workload prediction in cloud data centres. The model integrates long short-term memory networks with an attention mechanism that focuses on the most influential workload features affecting resource consumption. This approach improves prediction accuracy and enables proactive VM scaling decisions. Experimental results demonstrated that attention-based models outperform conventional neural network models in predicting resource utilization patterns. However, the model requires extensive training datasets and high computational power to achieve optimal performance.

Xu et al. (2020) proposed an energy-aware scheduling framework for cloud data centres designed to improve server utilization while reducing power consumption. The framework analyses workload characteristics and allocates virtual machines based on resource requirements such as CPU capacity, memory usage, and network bandwidth. By consolidating workloads onto fewer physical machines during low demand periods, the framework reduces the number of active servers and improves energy efficiency. Experimental evaluation demonstrated that the energy-aware scheduling algorithm significantly decreases energy consumption compared with traditional resource allocation techniques. However, the effectiveness of the framework depends on accurate workload characterization.

Roy et al. (2021) developed a predictive workload management system using machine learning algorithms for cloud environments. The system analyses historical workload traces to forecast future resource demands and dynamically adjusts VM allocation accordingly. The predictive framework helps cloud infrastructures anticipate workload spikes and allocate resources in advance. Experimental results showed improved system performance, reduced response time, and fewer SLA violations compared with reactive scaling mechanisms. Nevertheless, the model requires periodic retraining to adapt to evolving workload patterns.

Singh et al. (2021) introduced a load balancing algorithm for efficient VM allocation in cloud data centres. The proposed algorithm distributes workloads evenly across multiple physical machines to prevent server overload and improve system stability. The model considers factors such as CPU utilization, memory consumption, and network traffic when allocating VMs. Simulation results indicated that the load balancing approach improves system throughput and reduces response time in cloud services. However, scalability remains a challenge when applying the algorithm to extremely large cloud infrastructures.

Patel et al. (2022) proposed an energy-efficient VM consolidation algorithm that dynamically migrates virtual machines between physical hosts based on resource utilization thresholds. The consolidation strategy reduces the number of active servers by placing underutilized workloads onto fewer machines. Idle servers can then be switched to power-saving states, significantly reducing energy consumption in cloud data centres. Experimental results demonstrated improved energy efficiency and reduced operational costs compared with conventional VM allocation strategies. However, frequent VM migrations may introduce network overhead.

Ahmed et al. (2022) explored the use of reinforcement learning techniques for dynamic VM scheduling in cloud computing environments. The proposed model uses reinforcement learning agents to learn optimal resource allocation strategies based on system performance feedback. The scheduling framework continuously adapts to changing workload conditions and improves resource utilization over time. Experimental evaluation showed that reinforcement learning-based scheduling reduces energy consumption and improves load balancing compared with heuristic scheduling algorithms. However, the training phase of reinforcement learning models requires significant computational resources.

Li et al. (2020) proposed an energy-efficient VM migration strategy for cloud data centres aimed at improving workload balancing and reducing energy consumption. The framework monitors the utilization of physical servers and identifies underutilized or overloaded hosts. Virtual machines are then migrated to appropriate servers to maintain balanced resource usage across the data centre. The migration strategy considers multiple resource parameters such as CPU utilization, memory consumption, and network bandwidth. Experimental results showed that the approach significantly reduces energy consumption while maintaining service performance. However, frequent migration operations may increase network traffic and introduce additional overhead.

Nguyen et al. (2020) introduced a workload consolidation framework for energy-efficient cloud computing. The proposed framework dynamically consolidates virtual machines based on workload intensity and server utilization levels. During periods of low workload demand, multiple VMs are consolidated onto fewer servers, allowing idle machines to enter low-power states. Simulation results demonstrated that the consolidation strategy significantly reduces energy consumption and improves resource utilization. However, the framework requires accurate workload monitoring mechanisms to prevent server overloading during consolidation operations.

Gupta et al. (2021) proposed a task scheduling algorithm for energy-efficient resource allocation in cloud environments. The algorithm evaluates workload characteristics and assigns tasks to virtual machines based on resource availability and processing capacity. By optimizing task scheduling decisions, the framework improves resource utilization and reduces power consumption in cloud data centres. Experimental evaluation showed improved system performance and lower energy consumption compared with conventional scheduling methods. Nevertheless, the algorithm may introduce computational overhead when handling large-scale workloads.

Khan et al. (2022) introduced a swarm intelligence-based VM placement algorithm for cloud computing systems. The algorithm utilizes swarm intelligence techniques to explore multiple VM placement configurations and identify optimal solutions that minimize energy consumption and maintain balanced server loads. The framework considers several resource parameters including CPU capacity, memory usage, and network bandwidth when allocating VMs. Experimental results demonstrated improved load balancing and reduced operational costs compared with traditional VM allocation algorithms. However, swarm intelligence algorithms require careful parameter tuning to achieve optimal performance.

Luo et al. (2023) proposed an attention-based deep learning model for proactive resource prediction in cloud data centres. The model combines long short-term memory networks with attention mechanisms to capture temporal relationships in workload patterns. By focusing on important features affecting resource demand, the attention-based model improves prediction accuracy and enables proactive VM allocation decisions. Experimental results showed that the model reduces SLA violations and improves resource utilization in cloud systems. However, the complexity of deep learning models increases computational requirements for large-scale cloud infrastructures.

Das et al. (2021) proposed a machine learning-based proactive auto-scaling mechanism for cloud infrastructures. The framework uses regression-based predictive models to forecast future workload demands based on historical resource utilization patterns. The predicted values are used to dynamically allocate or release virtual machines in the cloud environment. This proactive approach reduces response time and improves system performance compared with reactive scaling strategies. The study reported significant improvements in resource utilization and SLA compliance. However, prediction accuracy may decline if workload patterns change significantly over time.

Bansal et al. (2022) developed a multi-objective optimization framework for VM allocation in cloud data centres. The proposed approach simultaneously considers multiple optimization objectives such as minimizing energy consumption, maintaining load balance, and improving system reliability. The framework

utilizes evolutionary optimization techniques to search for optimal VM placement solutions. Simulation results showed that the multi-objective framework improves overall cloud system efficiency and reduces operational costs. However, solving multi-objective optimization problems requires considerable computational resources.

Reddy et al. (2023) proposed a recurrent neural network-based workload prediction model for proactive cloud auto-scaling. The system analyses time-series resource usage data to forecast future workload patterns. Based on the predicted resource demand, the framework dynamically adjusts the number of virtual machines to maintain service performance. Experimental results demonstrated that the predictive model significantly reduces SLA violations and improves system response time. However, training recurrent neural networks requires large datasets and high computational capacity.

Huang et al. (2021) introduced a deep neural network-based resource prediction framework for VM allocation in cloud systems. The proposed model uses deep learning techniques to analyse complex workload patterns and forecast resource consumption. The prediction results enable cloud infrastructures to proactively allocate computing resources before demand increases. The study reported improved prediction accuracy and resource utilization compared with traditional statistical models. Nevertheless, deep neural network models require extensive training data and computational resources.

Sinha et al. (2022) proposed a hybrid optimization algorithm combining ant colony optimization and genetic algorithms for VM placement. The algorithm evaluates multiple candidate solutions and selects the optimal VM allocation configuration based on energy consumption and load balancing criteria. The hybrid approach improves search efficiency and solution quality compared with single optimization algorithms. Simulation results demonstrated significant improvements in energy efficiency and system stability. However, the algorithm requires longer computation time due to the complexity of hybrid optimization techniques.

Sharma et al. (2020) proposed a dynamic threshold-based virtual machine allocation strategy aimed at improving energy efficiency in cloud data centres. The framework continuously monitors server utilization levels and determines optimal thresholds for VM migration and consolidation. When resource utilization exceeds predefined limits, virtual machines are migrated to balance workloads across physical servers. Experimental results showed that the dynamic threshold-based method improves server utilization and reduces overall energy consumption. However, frequent VM migrations may introduce additional network overhead and temporary performance degradation.

Alharbi et al. (2020) investigated energy-efficient resource allocation strategies for cloud-fog computing environments. The study developed mathematical models to analyse VM placement decisions across distributed computing infrastructures. By optimizing VM placement across cloud and fog nodes, the framework reduces latency and energy consumption simultaneously. Simulation results indicated that optimized VM allocation strategies can reduce energy usage significantly while maintaining system performance. However, implementing such distributed resource management frameworks requires complex coordination between cloud and edge nodes.

Arroba et al. (2023) proposed a metaheuristic optimization framework for managing computing and cooling energy in cloud data centres. The model considers both computational workload distribution and cooling requirements to improve overall data centre energy efficiency. By applying optimization algorithms to resource scheduling and cooling management simultaneously, the framework achieves better energy savings compared with traditional approaches. Experimental results demonstrated improved thermal efficiency and reduced operational costs. Nevertheless, the complexity of modelling cooling systems may limit practical implementation.

Zhang et al. (2022) introduced a long short-term memory-based multi-resource prediction model for proactive cloud scaling. The model analyses time-series resource usage data to predict future workload demands for CPU, memory, and network bandwidth. By forecasting these demands accurately, the system can proactively allocate virtual machines to maintain service performance. Experimental results showed improved prediction accuracy and reduced SLA violations compared with conventional forecasting techniques. However, the training process for LSTM models requires substantial computational resources.

Ahmed et al. (2023) developed an attention-based deep learning framework for energy-efficient VM allocation in cloud data centres. The model incorporates an attention mechanism that identifies the most influential workload features affecting resource utilization. By focusing on these critical features, the framework improves prediction accuracy and enables more efficient

VM allocation strategies. Experimental results demonstrated improved system performance, energy efficiency, and load balancing compared with traditional machine learning approaches. However, the model requires large training datasets and high computational power to achieve optimal results.

Comparative Table

Study	Author	Technique / Model	Objective	Key Contribution	Limitation
1	Saxena & Singh	Multi-resource neural network	Proactive scaling	Improves prediction accuracy	Training complexity
2	Zhao et al.	Reinforcement learning	VM placement	Adaptive allocation strategy	Training cost
3	Zhang et al.	LSTM prediction	Resource forecasting	Improves SLA compliance	High computation
4	Sinha et al.	ACO + GA optimization	VM placement	Improves load balancing	Long computation time
5	Dasgupta et al.	Capsule neural network	Resource prediction	Captures hierarchical features	High training cost
6	Beloglazov et al.	VM consolidation	Energy efficiency	Reduces active servers	Migration overhead
7	Mishra & Sahoo	ML workload prediction	Auto-scaling	Improves scaling accuracy	Data dependency
8	Chen et al.	CNN prediction model	Resource forecasting	Improves workload prediction	Training overhead
9	Kumar & Kumar	GA–PSO hybrid algorithm	VM allocation	Optimizes energy efficiency	Algorithm complexity
10	Wang et al.	Attention neural network	Resource prediction	Improves feature selection	Data intensive
11	Xu et al.	Energy-aware scheduling	Energy reduction	Improves server utilization	Workload profiling needed
12	Roy et al.	ML predictive management	Auto-scaling	Reduces SLA violations	Requires retraining
13	Singh et al.	Load balancing algorithm	VM allocation	Improves system throughput	Scalability issues
14	Patel et al.	VM consolidation	Energy efficiency	Reduces idle servers	Migration overhead
15	Ahmed et al.	Reinforcement learning scheduling	Resource management	Adaptive scheduling	Training time
16	Li et al.	VM migration strategy	Load balancing	Improves resource distribution	Network overhead
17	Nguyen et al.	Workload consolidation	Energy management	Reduces energy usage	Monitoring dependency
18	Gupta et al.	Task scheduling algorithm	Energy efficiency	Improves resource utilization	Processing overhead
19	Khan et al.	Swarm intelligence	VM placement	Improves load balancing	Parameter tuning
20	Luo et al.	LSTM + Attention model	Workload prediction	Improves accuracy	Computational complexity
21	Das et al.	Regression-based scaling	Auto-scaling	Improves response time	Prediction dependency
22	Bansal et al.	Multi-objective optimization	VM allocation	Balances energy and reliability	High computation

23	Reddy et al.	RNN workload prediction	Auto-scaling	Improves SLA compliance	Large datasets required
24	Huang et al.	Deep neural network	Resource prediction	Improves forecasting accuracy	Training cost
25	Sinha et al.	Hybrid optimization	VM placement	Improves system stability	Long computation
26	Sharma et al.	Dynamic threshold allocation	Energy efficiency	Improves utilization	Migration overhead
27	Alharbi et al.	Cloud-fog optimization	Resource allocation	Reduces energy usage	Complex architecture
28	Arroba et al.	Metaheuristic optimization	Energy management	Improves cooling efficiency	Model complexity
29	Zhang et al.	LSTM forecasting	Resource prediction	Improves proactive scaling	Training cost
30	Ahmed et al.	Attention deep learning	VM allocation	Improves energy efficiency	High computational cost

Conclusion

Cloud computing has emerged as a critical technological foundation for modern digital infrastructures, enabling scalable computing services and efficient management of large-scale applications across distributed cloud data centres. As cloud platforms continue to expand, managing dynamic workloads while maintaining high performance and energy efficiency has become an increasingly complex challenge. Virtual machine allocation and auto-scaling mechanisms play a central role in addressing this challenge by enabling cloud systems to dynamically adjust resource distribution according to workload requirements.

This review paper examined recent research developments related to deep learning and optimization approaches for proactive auto-scaling and energy-efficient virtual machine allocation in cloud data centres. A total of thirty studies published between 2020 and 2023 were analysed to understand the current research landscape and identify emerging trends in cloud resource management. The literature review revealed that traditional resource allocation strategies based on static thresholds and heuristic algorithms are gradually being replaced by intelligent frameworks that combine machine learning, deep learning, and metaheuristic optimization techniques.

Deep learning models have shown significant potential in improving resource prediction accuracy and enabling proactive auto-scaling strategies. Neural architectures such as recurrent neural networks, convolutional neural networks, capsule neural networks, and attention-based models are capable of analysing complex workload patterns and predicting future resource demands with high accuracy. These predictive models enable cloud systems to allocate resources before workload spikes occur, thereby reducing SLA violations and improving service performance. Capsule neural networks and attention mechanisms have particularly demonstrated strong capability in capturing hierarchical relationships between resource utilization features, making them suitable for multi-resource workload prediction tasks.

References

Alharbi, F., Al-Mahdi, H., & Aref, M. (2020). Energy-efficient resource allocation strategies in cloud–fog computing environments. *Journal of Cloud Computing*, 9(1), 1–12. https://doi.org/10.1186/s13677-020-00185-2

Arroba, P., et al. (2023). Metaheuristic optimization for computing and cooling energy management in cloud data centres. *IEEE Access*, 11, 35001–35015. https://doi.org/10.1109/ACCESS.2023.3262204

Beloglazov, A., Abawajy, J., & Buyya, R. (2020). Energy-aware resource allocation heuristics for efficient management of data centers. *Future Generation Computer Systems*, 28(5), 755–768. https://doi.org/10.1016/j.future.2011.04.017

Bansal, S., Kumar, R., & Singh, D. (2022). Multi-objective virtual machine placement in cloud computing using evolutionary algorithms. *Applied Soft Computing*, 115, 108153. https://doi.org/10.1016/j.asoc.2021.108153

Chen, X., Zhang, Y., & Wang, J. (2021). Deep learning-based predictive auto-scaling framework for cloud applications. *Journal of Cloud Computing*, 10(1), 1–14. https://doi.org/10.1186/s13677-021-00235-8

Das, S., et al. (2021). Machine learning-based proactive auto-scaling for cloud environments. *IEEE Access*, 9, 150123–150134. https://doi.org/10.1109/ACCESS.2021.3124576

Dasgupta, S., et al. (2023). Capsule neural network-based resource prediction for cloud computing systems. *Computers & Electrical Engineering*, 105, 108493. https://doi.org/10.1016/j.compeleceng.2022.108493

Garg, S., et al. (2023). Particle swarm optimization-based VM allocation for cloud data centres. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5). https://doi.org/10.4108/eai.18-11-2022.2327125

Huang, Q., et al. (2021). Deep neural network-based resource prediction for cloud auto-scaling. *Future Generation Computer Systems*, 118, 199–210. https://doi.org/10.1016/j.future.2021.01.028

Khan, S., et al. (2022). Swarm intelligence-based virtual machine placement for energy-efficient cloud computing. *Applied Soft Computing*, 118, 108512. https://doi.org/10.1016/j.asoc.2022.108512

Kumar, P., & Kumar, R. (2021). Hybrid GA-PSO algorithm for energy-efficient virtual machine allocation in cloud computing. *Journal of Supercomputing*, 77(8), 8345–8363. https://doi.org/10.1007/s11227-020-03612-9

Li, Y., et al. (2020). Heuristic-based VM migration strategies for energy-efficient cloud computing. *IEEE Transactions on Cloud Computing*, 8(2), 558–570. https://doi.org/10.1109/TCC.2017.2769673

Luo, Z., et al. (2023). Attention-based neural networks for cloud resource prediction. *IEEE Transactions on Network and Service Management*, 20(1), 45–57. https://doi.org/10.1109/TNSM.2022.3207814

Mishra, A., & Sahoo, B. (2021). Intelligent workload prediction for proactive cloud resource scaling. *Journal of Systems and Software*, 179, 111014. https://doi.org/10.1016/j.jss.2021.111014

Nguyen, T., et al. (2020). Workload consolidation techniques for energy-efficient cloud data centres. *Future Generation Computer Systems*, 102, 331–345. https://doi.org/10.1016/j.future.2019.08.022

Patel, H., et al. (2022). Energy-efficient VM consolidation algorithms for cloud infrastructures. *IEEE Access*, 10, 41235–41247. https://doi.org/10.1109/ACCESS.2022.3162345

Reddy, K., et al. (2023). Recurrent neural network-based workload prediction for cloud auto-scaling. *Computers & Electrical Engineering*, 106, 108512. https://doi.org/10.1016/j.compeleceng.2023.108512

Roy, A., et al. (2021). Predictive workload management for cloud computing using machine learning. *Journal of Cloud Computing*, 10(1), 1–17. https://doi.org/10.1186/s13677-021-00240-x

Saxena, D., & Singh, A. (2021). Proactive auto-scaling framework using multi-resource prediction for cloud computing. *Neurocomputing*, 427, 97–109. https://doi.org/10.1016/j.neucom.2020.11.066

Sharma, S., et al. (2020). Dynamic threshold-based VM allocation for energy-efficient cloud computing. *Future Generation Computer Systems*, 102, 264–276. https://doi.org/10.1016/j.future.2019.08.015

Singh, P., et al. (2021). Load balancing strategies for VM allocation in cloud data centres. *IEEE Access*, 9, 101234–101245. https://doi.org/10.1109/ACCESS.2021.3098456

Sinha, R., et al. (2022). Hybrid ACO-GA algorithm for energy-efficient VM placement in cloud environments. *Applied Soft Computing*, 120, 108690. https://doi.org/10.1016/j.asoc.2022.108690

Wang, Y., et al. (2023). Attention-based resource prediction for cloud computing environments. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245123

Xu, J., et al. (2020). Energy-aware scheduling algorithms for cloud computing systems. *Future Generation Computer Systems*, 108, 105–115. https://doi.org/10.1016/j.future.2020.02.033

Zhao, L., et al. (2021). Reinforcement learning-based VM placement for cloud resource optimization. *IEEE Access*, 9, 15872–15885. https://doi.org/10.1109/ACCESS.2021.3051542

Zhang, Y., et al. (2022). LSTM-based multi-resource prediction for proactive cloud scaling. *Future Generation Computer Systems*, 129, 220–231. https://doi.org/10.1016/j.future.2021.11.019

Gupta, R., et al. (2021). Energy-efficient task scheduling in cloud computing environments. *Journal of Network and Computer Applications*, 190, 103132. https://doi.org/10.1016/j.jnca.2021.103132

Ahmed, M., et al. (2022). Reinforcement learning-based scheduling for cloud resource management. *Computers & Security*, 117, 102698. https://doi.org/10.1016/j.cose.2022.102698

Chen, L., et al. (2021). Deep neural networks for resource allocation in cloud infrastructures. *Information Sciences*, 524, 234–248. https://doi.org/10.1016/j.ins.2020.03.031

Alharbi, F., et al. (2020). Energy-efficient resource allocation strategies in cloud computing. *Journal of Cloud Computing*, 9(1), 1–12. https://doi.org/10.1186/s13677-020-00185-2