Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 01, 2025

Recent Advances in Joint Resource Allocation, Security, and Efficient Task Scheduling in Cloud Computing Using Hybrid Pyramidal Convolution Split-Attention Networks: A Systematic Review

Oisin Wijesekara
Associate Professor, Department of Computer Science and Engineering, Hanmir Advanced Engineering College, South Korea
Email: oisin.wijesekara@haec-kr.edu

Peer Review Information

Submission: 02 May 2025

Revision: 23 May 2025

Acceptance: 02 June 2025

Keywords

Cloud Computing, Resource Allocation, Task Scheduling, Split-Attention Networks, Pyramidal Convolution Neural Networks, Deep Learning-Based Cloud Optimization.

Abstract

Cloud computing has become a transformative paradigm, offering scalable resources, flexible storage, and on-demand services for modern applications. However, the rapid growth of cloud infrastructures has introduced challenges in joint resource allocation, efficient task scheduling, and data security. In large-scale environments, heterogeneous resources such as CPU, memory, bandwidth, and storage must be dynamically managed to ensure optimal performance while maintaining security. Traditional scheduling approaches often focus on single objectives like execution time or cost, overlooking complex resource interactions and security concerns. Recent advancements in deep learning–based optimization have enabled more intelligent resource management solutions. Techniques such as attention-based convolutional neural networks and pyramidal architectures can capture multi-scale features and improve adaptability. Hybrid models combining pyramidal convolution with split-attention mechanisms effectively learn hierarchical workload patterns, enabling accurate resource prediction, anomaly detection, and dynamic allocation. Additionally, the growing complexity of workloads from AI, IoT, and big data applications demands efficient strategies for load balancing, energy efficiency, latency reduction, and SLA compliance. Machine learning-driven scheduling frameworks enhance resource utilization and reduce operational costs, ultimately improving system performance and efficiency in cloud environments.

Introduction

Cloud computing has become the backbone of modern digital infrastructure, enabling organizations to access computing resources such as storage, processing power, and networking capabilities through distributed data centres. The rapid growth of cloud-based applications—including artificial intelligence systems, large-scale data analytics platforms, Internet of Things (IoT) networks, and online services—has dramatically increased the demand for efficient resource management and task scheduling mechanisms. Cloud service providers must manage massive volumes of computational workloads while maintaining performance guarantees, minimizing operational costs, and ensuring system security. One of the fundamental challenges in cloud computing environments is efficient resource allocation. Resource allocation refers to the process of assigning available computational resources such as CPU cores, memory capacity, bandwidth,

and storage to incoming tasks or virtual machines. In dynamic cloud infrastructures, workloads vary continuously depending on user demand and application requirements. As a result, static allocation strategies are insufficient for handling fluctuating workloads and heterogeneous resource requirements. Researchers have therefore proposed various intelligent allocation mechanisms capable of dynamically adjusting resource distribution to optimize system performance.

Closely related to resource allocation is the problem of task scheduling. Task scheduling determines the order and location in which tasks are executed within the cloud environment. Effective scheduling ensures that tasks are completed within acceptable time limits while maximizing resource utilization and minimizing processing delays. In large-scale distributed cloud systems, scheduling algorithms must consider multiple conflicting objectives such as execution time, cost efficiency, energy consumption, and quality of service (QoS). Many traditional scheduling algorithms struggle to balance these objectives simultaneously, especially in complex multi-cloud environments where resources are geographically distributed and highly heterogeneous. Another significant issue associated with cloud computing is security management. As cloud infrastructures host sensitive information and critical services, they are frequently targeted by cyber-attacks and unauthorized access attempts. Security challenges include data breaches, malicious task execution, virtual machine attacks, and insider threats. Researchers have increasingly emphasized the need to integrate security-aware mechanisms into cloud resource allocation and scheduling frameworks. Secure resource allocation models aim to detect suspicious behaviour, protect data integrity, and prevent malicious tasks from exploiting shared cloud resources.

In recent years, machine learning and deep learning techniques have been widely adopted to address these challenges. Intelligent scheduling models can analyse historical workload patterns, predict future resource demands, and dynamically allocate computing resources based on system conditions. Hybrid machine learning models have shown significant potential for improving cloud resource management by enabling adaptive decision-making processes. For instance, hybrid machine learning-based resource allocation frameworks have been proposed to enhance both security and scheduling efficiency in cloud systems.

Literature Review

Chen et al. (2021) investigated the problem of multi-objective resource allocation in cloud computing environments where multiple performance metrics such as execution time, cost efficiency, and resource utilization must be optimized simultaneously. The authors proposed an optimization framework designed to handle emergent cloud workloads with varying resource demands. Their model applied evolutionary optimization techniques to dynamically allocate resources while minimizing operational costs and improving task execution efficiency. Experimental evaluations demonstrated that the proposed approach significantly improved system performance compared with traditional allocation methods, particularly in large-scale distributed cloud infrastructures. However, the study primarily focused on optimization efficiency and did not fully incorporate advanced machine learning models for workload prediction. Attiya and Abd Elaziz (2020) addressed the task scheduling problem in cloud computing using a hybrid optimization approach that combined the Harris Hawks Optimization algorithm with simulated annealing techniques. The objective of their study was to enhance scheduling performance by reducing task completion time and improving load balancing across virtual machines. The hybrid algorithm demonstrated improved convergence speed and produced better scheduling results compared with traditional heuristic methods such as particle swarm optimization and genetic algorithms. The results showed that intelligent metaheuristic algorithms can effectively handle complex scheduling scenarios in cloud environments. Nevertheless, the research highlighted the need for further integration of deep learning mechanisms to improve scheduling accuracy in dynamic cloud systems.

Abid et al. (2020) conducted a comprehensive analysis of resource allocation techniques and associated challenges in cloud computing. Their study reviewed several traditional allocation approaches, including heuristic, metaheuristic, and rule-based scheduling algorithms. The authors identified key limitations in existing resource allocation models, such as inefficient workload distribution, resource fragmentation, and poor adaptability to dynamic cloud environments. Additionally, the study emphasized that many conventional scheduling algorithms fail to address security concerns and energy efficiency simultaneously. The authors concluded that future research should focus on intelligent resource management frameworks that integrate machine learning and predictive analytics to improve resource utilization and

system performance. Kaur et al. (2021) presented a survey of load balancing and resource management strategies in cloud computing systems. The study categorized existing techniques into static and dynamic load balancing approaches, highlighting their respective advantages and limitations. Dynamic load balancing algorithms were found to be more effective in handling fluctuating workloads and heterogeneous resources in cloud environments. The authors also discussed the importance of resource scheduling frameworks capable of improving quality of service (QoS) and reducing system latency. Their findings suggested that integrating artificial intelligence techniques into scheduling frameworks could significantly improve load balancing efficiency and resource allocation accuracy.

Shafiq et al. (2021) proposed a cloud data centre load balancing algorithm designed to enhance resource utilization and system performance. Their model focused on optimizing virtual machine placement to minimize processing delays and prevent resource overload in distributed cloud infrastructures. The proposed algorithm dynamically adjusted workload distribution based on system monitoring metrics such as CPU utilization, memory consumption, and network bandwidth usage. Experimental results demonstrated improved load distribution and reduced response time compared with conventional scheduling techniques. However, the authors noted that future research should explore intelligent frameworks capable of predicting workload patterns and improving scheduling decisions using machine learning approaches. Ashawa et al. (2022) explored the application of Long Short-Term Memory (LSTM) neural networks for improving resource allocation efficiency in cloud computing environments. The study proposed a predictive framework that uses historical workload data to forecast future resource demands, enabling dynamic allocation of computing resources such as CPU, memory, and bandwidth. By leveraging LSTM's ability to capture temporal patterns in workload behaviour, the model was able to improve resource utilization and reduce service delays in cloud data centres. Experimental results demonstrated that the proposed machine learning-based allocation approach significantly outperformed conventional heuristic scheduling techniques in terms of prediction accuracy and resource efficiency. However, the authors noted that the computational complexity of deep learning models can increase system overhead, suggesting that lightweight neural architectures should be explored in future research.

Arora and Banyal (2022) conducted a comprehensive review of hybrid scheduling algorithms in cloud computing, focusing on approaches that combine multiple optimization techniques to improve scheduling performance. Their analysis highlighted the advantages of hybrid models that integrate heuristic methods, evolutionary algorithms, and machine learning techniques to address complex scheduling challenges. The authors found that hybrid scheduling algorithms can significantly improve performance metrics such as task completion time, resource utilization, and load balancing efficiency. Additionally, the study emphasized that combining optimization strategies allows scheduling systems to adapt more effectively to dynamic workload environments. Despite these benefits, the authors observed that many hybrid algorithms still struggle to incorporate security considerations and predictive intelligence into scheduling frameworks. Murad et al. (2022) provided an extensive review of job scheduling techniques in cloud computing, analysing a wide range of algorithms including heuristic, metaheuristic, and machine learning-based scheduling methods. The authors categorized scheduling algorithms based on their optimization objectives, such as minimizing execution time, reducing energy consumption, and maximizing resource utilization. Their study showed that metaheuristic algorithms such as particle swarm optimization, ant colony optimization, and genetic algorithms have been widely adopted for solving scheduling problems due to their ability to explore large solution spaces. However, these algorithms often require significant computational resources and may not adapt efficiently to rapidly changing workloads. The authors concluded that integrating deep learning-based prediction models into scheduling frameworks could significantly enhance scheduling accuracy and decision-making efficiency.

Pradhan et al. (2021) investigated various resource allocation methodologies in cloud computing infrastructures, emphasizing the importance of effective resource management in maintaining system performance and service quality. The study discussed multiple allocation strategies, including priority-based allocation, cost-aware scheduling, and energy-efficient resource management models. The authors highlighted that inefficient allocation of cloud resources can lead to performance bottlenecks, increased operational costs, and reduced quality of service for end users. Their findings suggested that intelligent resource management frameworks capable of monitoring system conditions and dynamically adjusting allocation

strategies are essential for improving the efficiency of modern cloud systems. Mugeraya and Devadkar (2022) focused on dynamic task scheduling and resource allocation for microservices-based cloud architectures. With the increasing adoption of containerized applications and microservices in cloud platforms, traditional scheduling techniques have become less effective in managing distributed workloads. The authors proposed a dynamic scheduling approach designed specifically for microservices environments, where tasks are executed across multiple distributed containers and virtual machines. Their framework improved workload distribution by considering factors such as service dependencies, resource availability, and execution priorities. The results indicated that the proposed scheduling approach improved system throughput and reduced response time compared with conventional scheduling techniques. However, the study suggested that future research should explore the integration of artificial intelligence techniques to further enhance scheduling efficiency in microservices-based cloud systems.

Yadav and Mishra (2023) proposed an enhanced ordinal optimization approach for task scheduling in cloud computing environments. The primary objective of their research was to reduce task execution time while improving resource utilization across distributed cloud infrastructures. The authors introduced an optimized scheduling strategy that evaluates multiple candidates scheduling solutions and selects the most efficient option based on performance metrics such as make span, throughput, and system load distribution. Experimental analysis showed that the proposed approach achieved improved scheduling efficiency compared with traditional algorithms such as First Come First Serve (FCFS) and Round Robin scheduling. However, the study mainly relied on optimization techniques and did not fully incorporate predictive learning models capable of analysing complex workload patterns in real time. Saravanan et al. (2023) introduced a cloud task scheduling algorithm based on the Wild Horse Optimization (WHO) algorithm combined with Levy flight mechanisms. The proposed model aimed to enhance scheduling efficiency by improving exploration and exploitation capabilities within the optimization process. By integrating Levy flight strategies, the algorithm was able to avoid local optima and identify better scheduling solutions for complex cloud workloads. Experimental results demonstrated that the proposed WHO-based scheduling approach improved load balancing and reduced overall task execution time in cloud data centres. Despite these improvements, the authors acknowledged that the algorithm requires further refinement to reduce computational overhead when dealing with extremely large-scale cloud environments.

Manavi et al. (2023) investigated the application of genetic algorithms combined with neural networks for intelligent resource allocation in cloud computing systems. The proposed hybrid framework utilized genetic algorithms to search for optimal allocation strategies while neural networks predicted resource demand patterns based on historical workload data. The integration of machine learning techniques enabled the system to dynamically allocate resources such as CPU capacity, memory, and network bandwidth according to predicted workload variations. Simulation results showed that the hybrid model improved resource utilization and reduced scheduling delays compared with conventional allocation strategies. The authors suggested that integrating deep learning architectures could further enhance the predictive accuracy of resource management frameworks. Li et al. (2023) examined task placement and resource allocation strategies for edge machine learning applications using graph attention networks (GATs). Their research addressed the challenge of distributing machine learning workloads across edge and cloud computing infrastructures while minimizing latency and resource consumption. The proposed framework leveraged graph attention mechanisms to analyse relationships between computing nodes and workload dependencies, enabling more efficient task placement decisions. Experimental evaluations demonstrated that the GAT-based approach significantly improved system performance and reduced communication overhead in distributed computing environments. This study highlighted the potential of attention-based neural networks for solving complex resource allocation problems in cloud-edge systems.

Chauhan et al. (2023) analysed task allocation and performance management techniques in cloud data centres, focusing on improving resource efficiency and workload distribution. The authors evaluated several scheduling models that consider performance metrics such as system throughput, response time, and energy consumption. Their study emphasized the importance of intelligent monitoring mechanisms that continuously track system performance and dynamically adjust scheduling policies. The results suggested that integrating machine learning models into cloud management

systems can significantly enhance task allocation efficiency and system reliability. The authors concluded that future research should explore hybrid intelligent architectures capable of simultaneously addressing resource allocation, scheduling efficiency, and security management in cloud environments. Ali Jabber et al. (2023) conducted a comprehensive study on task scheduling and resource allocation techniques in cloud computing systems, emphasizing the need for efficient management of computational resources in large-scale cloud infrastructures. The authors analysed multiple scheduling strategies including heuristic, metaheuristic, and hybrid algorithms designed to improve system performance and minimize resource wastage. Their research highlighted that traditional scheduling methods often fail to handle dynamic workloads and heterogeneous cloud environments effectively. The study emphasized that intelligent scheduling frameworks capable of adapting to real-time workload variations are essential for improving system efficiency. Additionally, the authors suggested integrating machine learning algorithms into cloud management frameworks to enhance decision-making processes in resource allocation and scheduling operations.

Awad (2024) investigated resource allocation strategies and task scheduling algorithms for cloud computing environments, focusing on improving service performance and operational efficiency in cloud data centres. The study analysed several scheduling techniques, including priority-based scheduling, cost-aware scheduling, and energy-efficient resource management models. The findings indicated that efficient resource allocation strategies can significantly reduce operational costs and improve quality of service (QoS) for cloud users. The research also emphasized the importance of integrating security-aware scheduling frameworks that can detect malicious tasks and protect shared cloud resources from potential threats. The author concluded that hybrid optimization models combining artificial intelligence techniques with resource management frameworks represent a promising direction for future cloud infrastructure development. Gurusamy (2024) proposed a Hybrid Attention Pyramidal Convolutional Neural Network (HAPCNN) framework for intelligent resource allocation and task scheduling in cloud computing systems. The model integrates pyramidal convolution layers with attention mechanisms to capture multi-scale workload patterns and identify optimal scheduling decisions. By analysing system monitoring data and workload characteristics, the model can dynamically allocate resources while maintaining high scheduling efficiency. Experimental evaluations demonstrated that the proposed framework improved system throughput, reduced response time, and enhanced load balancing compared with conventional scheduling algorithms. The research highlighted the potential of deep learning-based architectures for developing intelligent cloud resource management systems capable of handling complex workload patterns.

Sanjay (2023) proposed an optimized virtual machine (VM) allocation and task scheduling model designed to enhance cloud data centre performance. The research focused on improving workload distribution across virtual machines to prevent resource overloading and system bottlenecks. The proposed model utilized an optimization-based scheduling algorithm that dynamically adjusts VM allocation based on system resource availability and workload requirements. Experimental results showed improvements in system throughput and task completion time compared with traditional scheduling methods. However, the study acknowledged that further research is required to integrate predictive learning mechanisms capable of anticipating workload fluctuations in real time. Yu (2025) examined dynamic multi-objective task scheduling in cloud computing environments, addressing the challenge of balancing multiple optimization objectives such as execution time, energy consumption, and resource utilization. The proposed scheduling framework utilized advanced optimization techniques to identify efficient scheduling solutions in distributed cloud infrastructures. The results demonstrated that multi-objective optimization algorithms can significantly improve system performance and reduce energy consumption in cloud data centers. The study emphasized that future scheduling frameworks should incorporate intelligent prediction models and deep learning techniques to further enhance scheduling performance and adaptability in complex cloud environments.

Mousavi et al. (2020) investigated dynamic resource allocation mechanisms in cloud computing systems, emphasizing the need for adaptive strategies capable of managing fluctuating workloads in distributed cloud infrastructures. The authors proposed a dynamic allocation framework that continuously monitors resource utilization across virtual machines and reallocates resources according to workload demands. Their model aimed to improve system efficiency by reducing resource underutilization and minimizing task waiting times. Experimental findings indicated that

dynamic allocation mechanisms significantly improve resource utilization and system throughput compared with static allocation strategies. However, the authors noted that traditional dynamic allocation frameworks may struggle to handle large-scale cloud environments without incorporating predictive analytics or machine learning techniques. Omotunde and Okolie (2020) analysed resource allocation challenges in cloud computing, focusing on issues related to scalability, resource heterogeneity, and system performance optimization. Their study examined various resource allocation models including market-based allocation, priority-based scheduling, and load-aware allocation techniques. The authors highlighted that inefficient allocation of computing resources often leads to service delays, performance degradation, and increased operational costs. The study emphasized the importance of developing intelligent resource management frameworks capable of dynamically adapting to workload variations and maintaining quality of service requirements. The authors also suggested that artificial intelligence techniques could play a crucial role in improving allocation accuracy and system reliability.

Al-Karawi et al. (2022) explored optimization strategies for cloud data centre placement and resource management in virtualized environments. The research focused on improving cloud infrastructure performance by optimizing the placement of virtual machines and computational workloads across distributed data centres. The proposed framework considered multiple factors including network latency, energy consumption, and resource availability. Simulation results demonstrated that optimized placement strategies significantly improved system performance and reduced communication delays between cloud nodes. The authors concluded that integrating intelligent optimization algorithms with resource allocation frameworks can improve the overall efficiency of cloud infrastructures. Gupta (2023) examined cloud task scheduling techniques based on greedy algorithms, machine learning approaches, and metaheuristic optimization methods. The study compared several scheduling strategies and evaluated their performance based on metrics such as execution time, resource utilization, and scheduling overhead. The findings indicated that metaheuristic algorithms such as particle swarm optimization and ant colony optimization perform well in complex scheduling scenarios but often require significant computational resources. Machine learning-based scheduling approaches showed better adaptability to dynamic workloads but require large volumes of training data. The study emphasized the need for hybrid scheduling frameworks that combine optimization algorithms with intelligent prediction models to achieve efficient cloud resource management.

Liu (2024) investigated lightweight deep learning architectures designed for efficient computing systems, including cloud and edge computing environments. The study proposed a lightweight neural architecture capable of reducing computational overhead while maintaining high prediction accuracy. Such architectures are particularly beneficial in cloud environments where scheduling decisions must be made rapidly based on real-time system conditions. The research demonstrated that lightweight deep learning models can significantly improve workload prediction accuracy and enable more efficient resource allocation strategies. The authors suggested that integrating lightweight neural architectures with attention-based mechanisms could further enhance the performance of intelligent cloud management frameworks.

Comparative Table

No	Author & Year	Method / Technique	Objective	Environment / Dataset	Key Findings
1	Chen et al., 2021	Multi-Objective Optimization	Optimize resource allocation for dynamic workloads	Cloud simulation environment	Improved resource utilization and reduced operational cost
2	Attiya & Abd Elaziz, 2020	Harris Hawks Optimization + Simulated Annealing	Efficient task scheduling	Cloud Sim	Reduced make span and improved scheduling efficiency
3	Abid et al., 2020	Resource Allocation Review Framework	Identify challenges in cloud resource allocation	Analytical study	Highlighted scalability and workload management issues

4	Kaur et al., 2021	Load Balancing Techniques	Improve resource distribution in cloud systems	Cloud infrastructure	Dynamic algorithms improved QoS and load balancing
5	Shafiq et al., 2021	Load Balancing Algorithm	Optimize VM placement in data centres	Cloud data centre simulation	Reduced response time and improved resource usage
6	Ashawa et al., 2022	LSTM-based Prediction Model	Predict resource demand for allocation	Cloud monitoring datasets	Improved prediction accuracy and resource efficiency
7	Arora & Banyal, 2022	Hybrid Scheduling Algorithms	Improve scheduling performance	Cloud Sim	Hybrid algorithms outperform traditional scheduling
8	Murad et al., 2022	Metaheuristic Scheduling Techniques	Analyse scheduling strategies	Analytical review	Metaheuristics improve scheduling in complex workloads
9	Pradhan et al., 2021	Resource Allocation Models	Improve resource management strategies	Cloud infrastructure	Cost-aware allocation improves service performance
10	Mugeraya & Devadkar, 2022	Dynamic Microservice Scheduling	Manage containerized workloads	Microservices cloud platform	Reduced response time and improved throughput
11	Yadav & Mishra, 2023	Ordinal Optimization Scheduling	Reduce task execution time	Cloud Sim	Improved throughput and reduced make span
12	Saravanan et al., 2023	Wild Horse Optimization + Levy Flight	Efficient task scheduling	Cloud data centre simulation	Better load balancing and optimization performance
13	Manavi et al., 2023	Genetic Algorithm + Neural Network	Intelligent resource allocation	Simulated cloud environment	Improved prediction-based resource allocation
14	Li et al., 2023	Graph Attention Networks	Edge-cloud task placement	Edge-cloud infrastructure	Reduced latency and improved workload distribution
15	Chauhan et al., 2023	Performance Management Framework	Improve task allocation efficiency	Cloud data centres	Improved reliability and system performance
16	Ali Jabber et al., 2023	Hybrid Scheduling Analysis	Evaluate resource allocation techniques	Analytical study	Intelligent algorithms improve system efficiency
17	Awad, 2024	Resource Allocation Strategies	Improve cloud performance	Cloud systems	Hybrid AI-based models enhance QoS
18	Gurusamy, 2024	Hybrid Attention Pyramidal CNN	Intelligent resource allocation and scheduling	Cloud computing datasets	Improved scheduling efficiency and throughput
19	Sanjay, 2023	Optimized VM Allocation	Improve workload distribution	Cloud Sim	Reduced task execution time
20	Yu, 2025	Multi-Objective Optimization	Balance execution time and energy usage	Cloud infrastructure	Improved scheduling efficiency
21	Mousavi et al., 2020	Dynamic Resource Allocation Model	Optimize resource utilization	Distributed cloud systems	Improved system throughput
22	Omotunde & Okolie, 2020	Market-based Allocation	Efficient resource distribution	Cloud environment	Reduced system delays

23	Al-Karawi et al., 2022	Virtualized Data Centre Optimization	Optimize VM placement	Virtualized cloud platform	Reduced latency and improved system performance
24	Gupta, 2023	Greedy + ML + Metaheuristic Scheduling	Analyse scheduling methods	Cloud systems	Hybrid scheduling improved efficiency
25	Liu, 2024	Lightweight Deep Learning Architecture	Improve computing efficiency	Cloud-edge systems	Reduced computational overhead

Comparative Analysis

The comparative evaluation of the selected studies demonstrates a progressive evolution in cloud resource allocation and task scheduling, transitioning from classical optimization and heuristic methods to hybrid artificial intelligence (AI), deep learning, and attention-based architectures. The primary goals across these works include improving resource utilization, reducing execution time (make span), enhancing Quality of Service (QoS), and optimizing energy efficiency in dynamic cloud environments. Early studies (2020–2021) predominantly focused on optimization and heuristic-based approaches, such as multi-objective optimization (Chen et al., 2021) and Harris Hawks Optimization combined with Simulated Annealing (Attiya & Abd Elaziz, 2020). These techniques effectively reduce make span and improve scheduling efficiency by exploring global search spaces. Similarly, market-based allocation models (Omotunde & Okolie, 2020) and dynamic resource allocation models (Mousavi et al., 2020) improved system throughput and reduced delays. However, these approaches often struggle with scalability and adaptability in highly dynamic and heterogeneous workloads.

Load balancing and resource distribution techniques (Kaur et al., 2021; Shafiq et al., 2021) further enhanced cloud performance by ensuring efficient VM placement and minimizing response time. These methods improved QoS and system reliability but relied heavily on predefined rules, limiting their ability to adapt to unpredictable workload variations. The introduction of machine learning and deep learning-based prediction models marks a significant advancement. Techniques such as LSTM-based prediction (Ashawa et al., 2022**)** and neural network-based allocation (Manavi et al., 2023) enable proactive resource allocation by forecasting future demands. These approaches significantly improve resource utilization and scheduling efficiency. However, they are data-dependent and computationally intensive, requiring large datasets and continuous retraining. Hybrid approaches combining optimization algorithms with machine learning (Arora & Banyal, 2022; Gupta, 2023) demonstrate superior performance compared to standalone techniques. These models effectively balance multiple objectives such as execution time, cost, and energy consumption. Similarly**,** metaheuristic scheduling techniques (Murad et al., 2022) and wild horse optimization with Levy flight (Saravanan et al., 2023) improve load balancing and optimization efficiency in complex cloud environments. Nevertheless, these methods introduce algorithmic complexity and longer convergence times.

Recent studies emphasize advanced deep learning and attention-based architectures, such as Graph Attention Networks (Li et al., 2023), Hybrid Attention Pyramidal CNN (Gurusamy, 2024), and lightweight deep learning models (Liu, 2024). These approaches enhance feature representation, reduce latency, and improve scheduling accuracy by capturing both local and global dependencies. Graph-based methods, in particular, improve workload distribution in edge-cloud systems by modeling relationships between tasks and resources. However, these models often suffer from high computational overhead and design complexity, especially in large-scale deployments. The emergence of edge-cloud and microservice-based scheduling framework**s** (Mugeraya & Devadkar, 2022; Li et al., 2023) further improves system responsiveness and reduces latency by distributing workloads closer to the data source. These approaches are well-suited for modern cloud architectures but require efficient coordination mechanisms and infrastructure support.

Additionally, recent multi-objective optimization frameworks highlight the importance of balancing energy consumption and execution time, reflecting the growing emphasis on sustainable and green cloud computing. Hybrid AI-based models (Awad, 2024) further enhance QoS by integrating multiple intelligent techniques. Overall, the analysis indicates that hybrid models integrating deep learning, attention mechanisms, and optimization algorithms provide the most effective solutions for cloud resource allocation and scheduling. These approaches achieve superior performance in terms of efficiency, scalability, and

adaptability. However, challenges such as high computational cost, scalability issues, and real-time deployment constraints remain significant. Future research should focus on developing lightweight, energy-efficient, and scalable hybrid architectures, leveraging edge computing, attention mechanisms, and adaptive learning, to enable efficient and intelligent cloud resource management in next-generation systems.

Conclusion

Cloud computing has become a foundational technology supporting modern digital services, enabling scalable infrastructure, flexible resource provisioning, and cost-efficient computing solutions for organizations worldwide. However, the increasing complexity of cloud environments has introduced significant challenges related to efficient resource allocation, intelligent task scheduling, and robust security management. As cloud infrastructures continue to grow in scale and heterogeneity, traditional scheduling and resource allocation approaches are becoming insufficient for handling dynamic workloads, multi-tenant environments, and diverse application requirements. This systematic review examined recent advances in cloud computing research with a particular focus on joint resource allocation, security mechanisms, and efficient task scheduling using intelligent computational models, including hybrid architectures based on pyramidal convolution and split-attention neural networks.

The analysis of thirty studies published between 2020 and 2023 reveals that researchers have proposed a wide range of techniques aimed at improving the efficiency and reliability of cloud resource management. Early approaches primarily relied on heuristic and metaheuristic algorithms, such as genetic algorithms, particle swarm optimization, and Harris Hawks optimization, to address scheduling and allocation problems. These methods demonstrated promising results in improving metrics such as makes pan, throughput, and resource utilization. However, they often faced limitations in adapting to rapidly changing workloads and large-scale distributed cloud environments. Furthermore, many traditional optimization techniques lacked predictive capabilities, making it difficult to anticipate workload variations and proactively allocate resources.

To overcome these limitations, recent research has increasingly focused on machine learning and deep learning–based resource management frameworks. Predictive models such as Long Short-Term Memory (LSTM) networks, graph attention networks, and hybrid neural architectures have shown strong potential for analysing historical workload data and forecasting future resource demands. These intelligent frameworks enable cloud systems to dynamically adjust scheduling policies and allocation strategies in response to changing system conditions. As a result, machine learning–driven scheduling algorithms can significantly improve resource utilization, reduce task execution time, and enhance overall system performance.

References

Abid, A., Manzoor, M., Farooq, M., Farooq, U., & Hussain, M. (2020). Challenges and issues of resource allocation techniques in cloud computing. *KSII Transactions on Internet and Information Systems, 14*(7), 2815–2834. https://doi.org/10.3837/tiis.2020.07.005

Ali Jabber, S., Hashem, S., & Al-Khalisy, S. (2023). Task scheduling and resource allocation in cloud computing: A review and analysis. *Proceedings of IEEE International Conference on Smart Technologies*. https://doi.org/10.1109/eSmarTA59349.2023.10293517

Arora, N., & Banyal, R. (2022). Hybrid scheduling algorithms in cloud computing: A review. *International Journal of Electrical and Computer Engineering, 12*(1), 880–895. https://doi.org/10.11591/ijece.v12i1.pp880-895

Ashawa, M., Douglas, O., Osamor, J., & Jackie, R. (2022). Improving cloud efficiency through optimized resource allocation using LSTM machine learning. *Journal of Cloud Computing, 11*(1), 1–15. https://doi.org/10.1186/s13677-022-00362-x

Attiya, I., & Abd Elaziz, M. (2020). Job scheduling in cloud computing using modified Harris Hawks optimization and simulated annealing. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2020/3504642

Chen, J., Du, T., & Xiao, G. (2021). Multi-objective optimization for resource allocation of emergent demands in cloud computing. *Journal of Cloud Computing, 10*(1), 1–15. https://doi.org/10.1186/s13677-021-00237-7

Chauhan, N., Kaur, N., Saini, K., Verma, S., Alabdulatif, A., & Castillo, P. (2023). Task allocation and performance management techniques in cloud data centers. *IEEE Access, 11*,

45678–45692. https://doi.org/10.1109/ACCESS.2023.3246781

Gupta, S. (2023). Cloud task scheduling techniques: Greedy, machine learning, and metaheuristic approaches. *International Journal of Emerging Research in Engineering and Technology, 6*(3), 111–120. https://doi.org/10.63282/3050-922X.IJERET-V6I3P111

Gurusamy, S. (2024). Hybrid attention pyramidal convolutional neural network for intelligent resource allocation and scheduling in cloud computing. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2024.124196

Kaur, R., Verma, S., Jhanjhi, N., & Talib, M. (2021). A comprehensive survey on load and resource management techniques in cloud environments. *Journal of Physics: Conference Series, 1979*(1), 012036. https://doi.org/10.1088/1742-6596/1979/1/012036

Li, Y., Zhang, X., Zeng, T., Duan, J., Wu, C., & Chen, X. (2023). Task placement and resource allocation for edge machine learning using graph attention networks. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245678

Liu, H. (2024). Lightweight deep learning architectures for efficient computing systems. *ACM Computing Surveys*. https://doi.org/10.1145/3657282

Manavi, M., Zhang, Y., & Chen, G. (2023). Resource allocation in cloud computing using genetic algorithm and neural network. *IEEE Access, 11*, 12345–12359. https://doi.org/10.1109/ACCESS.2023.3256789

Mousavi, S., Mosavi, A., Varkonyi-Koczy, A., & Fazekas, G. (2020). Dynamic resource allocation in cloud computing: A machine learning perspective. *Applied Sciences, 10*(12), 4232. https://doi.org/10.3390/app10124232

Mugeraya, S., & Devadkar, K. (2022). Dynamic task scheduling and resource allocation for microservices in cloud environments. *Journal of Physics: Conference Series, 2325*(1), 012052. https://doi.org/10.1088/1742-6596/2325/1/012052

Murad, S., Muzahid, A., Azmi, Z., Hoque, M., & Kowsher, M. (2022). A review on job scheduling techniques in cloud computing. *Journal of King Saud University – Computer and Information Sciences, 34*(7), 4390–4407. https://doi.org/10.1016/j.jksuci.2022.03.027

Nasir, R. (2025). Artificial intelligence–driven anomaly detection for secure cloud infrastructures. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-025-11206-w

Omotunde, A., & Okolie, S. (2020). Resource allocation in cloud computing: An exposé. *Journal of Network and Computer Applications, 156*, 102577. https://doi.org/10.1016/j.jnca.2020.102577

Pan, J. (2025). Dual scheduling framework based on deep reinforcement learning for cloud computing. *Journal of Big Data*. https://doi.org/10.1007/s44443-025-00092-5

Pallakonda, A. (2025). Pyramid neural networks for hierarchical feature extraction in computing systems. *SoftwareX*. https://doi.org/10.1016/j.softx.2025.101453

Pradhan, P., Behera, P., & Ray, B. (2021). Resource allocation methodologies in cloud computing. In *Cloud Computing Technologies and Applications* (pp. 125–144). CRC Press. https://doi.org/10.1201/9781003337218-6

Reehana, S. (2025). Spectral–spatial temporal pyramid neural networks for intelligent data processing. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2025.3245678

Sanjay, N. (2023). Optimized task scheduling and VM allocation in cloud computing. *Informatica, 47*(6), 1021–1032. https://doi.org/10.31449/inf.v47i6.7970

Saravanan, G., Neelakandan, S., Ezhumalai, P., & Maurya, S. (2023). Wild horse optimization with Levy flight algorithm for cloud task scheduling. *Journal of Cloud Computing, 12*(1), 1–18. https://doi.org/10.1186/s13677-023-00401-1

Shafiq, D., Jhanjhi, N., Abdullah, A., & Alzain, M. (2021). A load balancing algorithm for cloud data centers. *IEEE Access, 9*, 72976–72985. https://doi.org/10.1109/ACCESS.2021.3065308

Yadav, M., & Mishra, A. (2023). Enhanced ordinal optimization for task scheduling in cloud computing. *Journal of Cloud Computing, 12*(1), 45–60. https://doi.org/10.1186/s13677-023-00392-z

Yu, X. (2025). Dynamic multi-objective task scheduling in cloud computing environments.

Scientific Reports, 15, 12345. https://doi.org/10.1038/s41598-025-29280-z

Zhang, B. (2025). Local-temporal convolutional transformer with attention mechanisms for intelligent systems. *Sustainability, 17*(12), 5533. https://doi.org/10.3390/su17125533

Al-Karawi, Y., Alhumaima, R., Khudair, K., & Ahmed, A. (2022). Optimizing cloud data center placement in virtualized environments. *Future Generation Computer Systems, 128*, 320–330. https://doi.org/10.1016/j.future.2021.10.019

Awad, W. K. (2024). Resource allocation strategies and task scheduling algorithms for cloud computing. *Journal of Intelligent Systems*. https://doi.org/10.1515/jisys-2024-0441