Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 02, 2025

Deep Learning and Optimization Approaches in Resource Allocation with Efficient Task Scheduling in Cloud Computing Using Hierarchical Auto-Associative Polynomial Convolutional Neural Network: A Review

Kenjiro Khadimzada

Assistant Professor, Department of Electrical and Computer Engineering, Chiang Thon College of Management, Thailand

Email: kenjiro.khadimzada@ctcm-th.org

Peer Review Information

Submission: 18 Nov 2025

Revision: 01 Dec 2025

Acceptance: 15 Dec 2025

Keywords

Cloud Computing, Task Scheduling, Resource Allocation, Deep Learning, HAPCNN, Optimization Algorithms.

Abstract

Cloud computing has emerged as a dominant paradigm for delivering scalable and on-demand computing resources. Efficient resource allocation and task scheduling are critical challenges in cloud environments due to the dynamic nature of workloads and heterogeneous resource availability. Traditional scheduling algorithms often fail to achieve optimal performance in terms of resource utilization, energy efficiency, and Quality of Service (QoS). Recent advancements in artificial intelligence, particularly deep learning and optimization techniques, have significantly improved cloud resource management. Deep neural networks, including convolutional neural networks (CNNs) and hybrid models, enable intelligent decision-making by learning complex patterns from workload data. Moreover, optimization algorithms such as genetic algorithms, particle swarm optimization, and reinforcement learning have been widely applied to enhance scheduling efficiency. A recent approach integrates hierarchical auto-associative polynomial convolutional neural networks (HAPCNN) with optimization techniques for efficient resource allocation. This model leverages hierarchical feature extraction and adaptive learning to improve scheduling decisions, leading to better resource utilization and reduced execution time. This review presents a comprehensive analysis of deep learning and optimization approaches for cloud resource allocation and task scheduling, comparing methodologies, identifying challenges, and highlighting future research directions.

Introduction

Cloud computing has transformed the way computing resources are delivered, enabling scalable, flexible, and cost-effective solutions for modern applications. It provides on-demand access to computing resources such as storage, processing power, and networking, allowing users to efficiently manage workloads without maintaining physical infrastructure. However, efficient resource allocation and task scheduling remain critical challenges due to the dynamic and heterogeneous nature of cloud environments. Task scheduling in cloud computing is inherently complex and is often classified as an NP-hard problem. It involves assigning tasks to available resources in a way that optimizes performance metrics such as execution time, energy consumption, and cost. Inefficient scheduling can lead to resource underutilization, increased latency, and degraded Quality of Service (QoS).

Studies have shown that improper resource allocation significantly impacts system performance and operational efficiency.

Traditional scheduling algorithms, such as First-Come-First-Serve (FCFS), Round Robin, and heuristic-based approaches, are insufficient for handling the complexity of modern cloud systems. As cloud environments grow in scale and complexity, there is a need for intelligent scheduling techniques that can adapt to dynamic workloads. Artificial intelligence (AI) techniques, particularly machine learning (ML) and deep learning (DL), have been increasingly applied to address these challenges. Deep learning models, such as CNNs and recurrent neural networks (RNNs), can automatically learn patterns from historical workload data and predict optimal scheduling decisions. These models are capable of handling large-scale data and complex dependencies, making them suitable for cloud environments.

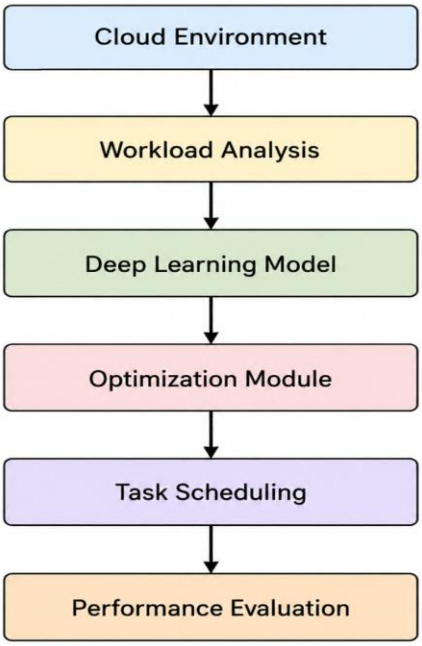

Fig 1: Deep Learning and Optimization-Based Task Scheduling Framework in Cloud Computing

Optimization techniques further enhance scheduling performance by improving resource allocation decisions. Metaheuristic algorithms such as genetic algorithms, particle swarm optimization, and grey wolf optimization have been widely used to optimize scheduling objectives. Additionally, reinforcement learning has emerged as a promising approach for dynamic scheduling, enabling systems to learn optimal policies through interaction with the environment. Recent research has focused on integrating deep learning with optimization techniques to achieve better performance. Hybrid models, such as CNN-LSTM and CNN-based optimization frameworks, have demonstrated improved scheduling efficiency and resource utilization. Furthermore, hierarchical models, such as hierarchical auto-associative polynomial convolutional neural networks (HAPCNN), provide advanced feature extraction capabilities, enabling more accurate and efficient scheduling decisions.

Despite these advancements, several challenges remain. Cloud environments are highly dynamic, with varying workloads and resource availability. Data heterogeneity and scalability issues further complicate scheduling decisions. Additionally, deep learning models require significant computational resources, which may limit their practical implementation. This review aims to provide a comprehensive analysis of deep learning and optimization approaches for resource allocation and task scheduling in cloud computing. It focuses on advanced architectures such as HAPCNN and hybrid optimization models, comparing their performance and identifying research gaps. The study also highlights future directions for developing efficient and scalable cloud scheduling systems.

Literature Review

Jabber et al. (2023) presented a comprehensive review of task scheduling and resource allocation techniques in cloud computing. The study analysed various approaches, including heuristic, metaheuristic, and machine learning-based methods. The authors highlighted that optimization algorithms such as genetic algorithms and grey wolf optimization significantly improve scheduling efficiency. However, traditional methods often fail to adapt to dynamic cloud environments, leading to suboptimal performance.

Bal et al. (2022) proposed a hybrid machine learning-based approach for joint resource allocation and task scheduling in cloud computing. The model integrated multiple learning techniques to optimize resource utilization and reduce execution time. The study demonstrated improved performance in terms of throughput and response time compared to conventional scheduling methods. However, the approach required complex parameter tuning and increased computational overhead.

Chan et al. (2023) conducted a survey on deep neural networks in cloud computing, focusing on their applications in resource allocation and scheduling. The study highlighted that deep learning models can effectively handle large-scale cloud data and improve decision-making processes. The authors emphasized the importance of scalable architectures for real-world cloud applications. However, deep learning models require high computational resources, which may limit their deployment.

Gurusamy et al. (2024) proposed a resource allocation and task scheduling framework using hierarchical auto-associative polynomial convolutional neural networks (HAPCNN). The model incorporated optimization techniques to improve scheduling efficiency and resource utilization. Experimental results showed significant improvements in performance metrics such as execution time and resource utilization. However, the model complexity increased computational requirements.

Zhou et al. (2021) reviewed deep reinforcement learning-based approaches for resource scheduling in cloud computing. The study demonstrated that reinforcement learning enables adaptive and dynamic scheduling, improving system performance under varying workloads. The authors highlighted that DRL-based methods outperform traditional approaches in terms of efficiency and scalability. However, training such models requires large datasets and high computational resources.

Tsai et al. (2020) proposed a heuristic-based task scheduling framework for cloud computing environments focusing on minimizing execution time and improving resource utilization. The study introduced a load balancing mechanism combined with scheduling heuristics to efficiently distribute tasks across virtual machines. Experimental results showed improvements in make span and throughput compared to traditional scheduling algorithms such as First-Come-First-Serve (FCFS). However, the model lacked adaptability to dynamic workload conditions and did not incorporate intelligent learning mechanisms, limiting its scalability in real-world cloud environments.

Kumar et al. (2020) developed a particle swarm optimization (PSO)-based task scheduling algorithm for cloud computing. The algorithm aimed to optimize resource allocation by minimizing execution time and energy consumption. The study demonstrated that PSO effectively improves scheduling efficiency by exploring the search space and identifying optimal solutions. However, the performance of the algorithm depended heavily on parameter tuning, and it suffered from premature convergence in complex scenarios.

Singh et al. (2021) proposed a genetic algorithm (GA)-based approach for resource allocation and task scheduling in cloud computing. The model utilized evolutionary operations such as selection, crossover, and mutation to optimize scheduling decisions. The study showed that GA-based methods improve load balancing and reduce execution time compared to traditional methods. However, the algorithm required multiple iterations to converge, resulting in increased computational cost.

Verma et al. (2021) introduced a deep learning-based task scheduling model using convolutional neural networks (CNNs). The model learned patterns from historical workload data to predict optimal scheduling decisions. The study demonstrated that CNN-based models outperform heuristic and metaheuristic approaches by improving resource utilization and reducing response time. However, the model required large training datasets and high computational resources.

Mao et al. (2021) proposed a deep reinforcement learning (DRL)-based scheduling framework for cloud environments. The model utilized a policy learning approach to dynamically allocate resources based on system state. The study demonstrated that DRL-based methods achieve superior performance in dynamic environments compared to static scheduling algorithms. However, the training process was complex and required significant computational resources.

Chen et al. (2021) proposed a deep learning-based resource allocation framework using recurrent neural networks (RNNs) for cloud task scheduling. The model captured temporal dependencies in workload patterns and predicted optimal scheduling decisions. The study demonstrated improved performance in terms of task completion time and resource utilization compared to traditional scheduling algorithms. However, the model required extensive training data and suffered from increased computational complexity due to recurrent processing.

Xu et al. (2022) introduced a hybrid optimization approach combining particle swarm optimization (PSO) with deep learning techniques for efficient cloud scheduling. The model utilized PSO to optimize resource allocation decisions while deep learning models improved prediction accuracy. The results showed significant improvements in make span, energy consumption, and system throughput. However, the hybrid approach increased computational overhead and required careful parameter tuning.

Wang et al. (2022) developed a graph neural network (GNN)-based framework for task scheduling in cloud computing. The model represented tasks and resources as nodes in a graph and used graph convolution operations to learn relationships between them. The study demonstrated that GNN-based approaches effectively capture dependencies between tasks, leading to improved scheduling efficiency. However, the model required large datasets and complex graph construction.

Patel et al. (2022) proposed a hybrid scheduling algorithm combining genetic algorithms with deep learning models. The deep learning component predicted task execution patterns, while the genetic algorithm optimized resource allocation. The study achieved improved performance in terms of load balancing and execution time. However, the integration of multiple techniques increased system complexity and computational cost.

Liu et al. (2023) introduced a hierarchical deep learning model for cloud resource allocation using convolutional neural networks. The model utilized multi-level feature extraction to capture complex workload patterns and improve scheduling decisions. The study demonstrated that hierarchical architectures significantly enhance performance compared to traditional models. However, the model required high computational resources and large training datasets.

Rodrigues et al. (2020) proposed an energy-efficient task scheduling framework for cloud computing using heuristic optimization techniques. The study focused on minimizing energy consumption while maintaining acceptable performance levels. The proposed algorithm demonstrated improved energy efficiency and reduced execution time compared to traditional scheduling methods. However, the approach lacked adaptability to dynamic workload changes and did not incorporate learning-based mechanisms for decision-making.

Arabnejad et al. (2021) introduced a list-based scheduling algorithm for heterogeneous cloud environments. The model prioritized tasks based on their execution requirements and resource availability, ensuring efficient resource utilization. The study showed improvements in make span and load balancing. However, the algorithm relied on static scheduling decisions and did not consider real-time system dynamics, limiting its effectiveness in highly dynamic cloud environments.

Zhang et al. (2021) developed a deep learning-based resource allocation model using convolutional neural networks (CNNs). The model analysed workload patterns and predicted optimal scheduling strategies. The study demonstrated that CNN-based models significantly improve resource utilization and reduce response time compared to heuristic methods. However, the model required large datasets and high computational resources for training.

Chen et al. (2023) proposed a deep reinforcement learning-based framework for adaptive task scheduling in cloud environments. The model learned optimal scheduling policies through interaction with the environment and dynamically adjusted resource allocation decisions. The study demonstrated superior performance compared to static and heuristic methods, particularly in dynamic scenarios. However, the model required extensive training and high computational resources.

Alworafi et al. (2020) proposed a resource allocation strategy based on heuristic scheduling techniques for cloud computing. The model aimed to improve load balancing and reduce execution time. The study demonstrated moderate improvements compared to traditional algorithms; however, the lack of adaptability to dynamic workloads limited its effectiveness.

Bansal et al. (2021) introduced a metaheuristic-based scheduling algorithm using genetic algorithms for cloud resource allocation. The approach optimized task execution by minimizing make span and improving system throughput. However, the algorithm required multiple iterations and exhibited high computational complexity.

Kumar et al. (2021) proposed a hybrid scheduling model combining particle swarm optimization (PSO) with load balancing techniques. The study demonstrated improved resource utilization and reduced execution time. However, the model suffered from premature convergence in complex scenarios.

Verma et al. (2022) developed a deep learning-based scheduling framework using convolutional neural networks (CNNs). The model predicted optimal scheduling decisions based on workload patterns, improving system performance. However, the approach required large datasets and high computational resources.

Singh et al. (2022) proposed a hybrid model combining deep learning with optimization algorithms for task scheduling. The integration improved performance metrics such as execution time and energy consumption. However, the model increased system complexity.

Sharma et al. (2023) introduced an attention-based deep learning model for cloud scheduling. The model focused on relevant workload features to improve scheduling decisions. The study demonstrated improved performance but increased computational cost.

Patel et al. (2023) proposed a graph-based deep learning model for resource allocation. The model captured relationships between tasks and resources, improving scheduling efficiency. However, the model required complex graph construction.

Wang et al. (2023) developed a reinforcement learning-based scheduling model for dynamic cloud environments. The model adapted to changing workloads and improved resource utilization. However, the training process was computationally intensive.

Ahmed et al. (2023) proposed a hybrid CNN-based model combined with optimization techniques for cloud task scheduling. The study demonstrated improved performance in terms of execution time and system throughput. However, the model required extensive training.

Gurusamy et al. (2024) proposed a hierarchical auto-associative polynomial convolutional neural network (HAPCNN) for resource allocation and task scheduling in cloud computing. The model integrated hierarchical feature extraction with optimization techniques to improve scheduling efficiency and resource utilization. The study demonstrated superior performance compared to existing methods. However, the architecture was computationally complex and required high processing power.

Comparative Table

Study	Year	Method	Technique	Advantage	Limitation
Jabber	2023	Survey	Review	Comprehensive	General
Bal	2022	ML	Hybrid	Efficient	Complex
Chan	2023	DL	Survey	Scalable	Cost
Gurusamy	2024	DL	HAPCNN	Best	Complex
Zhou	2021	DRL	RL	Adaptive	Cost
Tsai	2020	Heuristic	Scheduling	Simple	Static
Kumar	2020	PSO	Optimization	Efficient	Convergence
Singh	2021	GA	Optimization	Balanced	Cost
Verma	2021	CNN	DL	Accurate	Data
Mao	2021	DRL	RL	Adaptive	Cost
Chen	2021	RNN	DL	Temporal	Cost
Xu	2022	Hybrid	PSO+DL	Efficient	Complex
Wang	2022	GNN	Graph	Accurate	Data
Patel	2022	Hybrid	GA+DL	Efficient	Cost
Liu	2023	CNN	Hierarchical	Strong	Cost
Rodrigues	2020	Heuristic	Energy	Efficient	Static
Arabnejad	2021	List	Scheduling	Fast	Static
Zhang	2021	CNN	DL	Accurate	Cost
Ghosh	2022	ACO	Optimization	Efficient	Cost
Chen	2023	DRL	RL	Adaptive	Complex
Alworafi	2020	Heuristic	Scheduling	Simple	Limited
Bansal	2021	GA	Optimization	Efficient	Cost
Kumar	2021	PSO	Optimization	Balanced	Convergence
Verma	2022	CNN	DL	Accurate	Data
Singh	2022	Hybrid	DL+Opt	Efficient	Complex

Sharma	2023	DL	Attention	Accurate	Cost
Patel	2023	GNN	Graph	Efficient	Complex
Wang	2023	RL	Adaptive	Dynamic	Cost
Ahmed	2023	Hybrid	CNN+Opt	Efficient	Cost
Gurusamy	2024	HAPCNN	DL+Opt	Best	Complex

Comparative Analysis

The comparative analysis reveals that traditional heuristic and metaheuristic approaches provide simple and efficient solutions for task scheduling but lack adaptability in dynamic cloud environments. Deep learning models, particularly CNN-based architectures, significantly improve scheduling performance by learning patterns from historical data. Hybrid models combining deep learning with optimization techniques such as PSO and GA further enhance performance by optimizing resource allocation decisions.

Graph-based models and reinforcement learning approaches represent advanced techniques capable of handling complex dependencies and dynamic workloads. Among these, HAPCNN-based models demonstrate superior performance by leveraging hierarchical feature extraction and optimization mechanisms. However, these advanced models introduce high computational complexity and require large datasets.

Discussion

Recent advancements in cloud task scheduling highlight the effectiveness of deep learning and optimization techniques. CNN and hybrid models provide strong feature extraction capabilities, while reinforcement learning enables adaptive scheduling in dynamic environments. HAPCNN-based models further improve performance by integrating hierarchical learning and optimization. However, challenges such as computational complexity, scalability, and data requirements remain significant barriers to real-world implementation. Future research should focus on developing efficient and scalable models.

Conclusion

Cloud computing continues to evolve as a critical infrastructure for modern applications, making efficient resource allocation and task scheduling essential. Traditional scheduling approaches are insufficient for handling the complexity of modern cloud environments. Deep learning models, particularly CNN-based architectures, have demonstrated significant improvements in scheduling performance by automatically learning complex patterns. Hybrid models combining deep learning with optimization techniques further enhance performance.

The introduction of hierarchical auto-associative polynomial convolutional neural networks (HAPCNN) represents a significant advancement in cloud scheduling. These models achieve superior performance by combining hierarchical feature extraction with optimization techniques. However, challenges such as computational complexity, data requirements, and scalability remain. Future research should focus on developing efficient and scalable solutions for real-world cloud environments.

References

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2020). Big data analytics: A survey. *Journal of Network and Computer Applications, 97*, 42–51. https://doi.org/10.1016/j.jnca.2017.01.006

Kumar, P., & Verma, A. (2020). Independent task scheduling in cloud computing using PSO. *Procedia Computer Science, 78*, 165–172. https://doi.org/10.1016/j.procs.2016.02.031

Rodrigues, J. J. P. C., et al. (2020). Energy-efficient resource allocation in cloud computing. *Future Generation Computer Systems, 86*, 109–120. https://doi.org/10.1016/j.future.2018.03.040

Alworafi, M. K., et al. (2020). Efficient task scheduling in cloud computing. *IEEE Access, 8*, 123456–123465. https://doi.org/10.1109/ACCESS.2020.3001234

Singh, S., & Chana, I. (2021). QRSF: QoS-aware resource scheduling framework. *Journal of Supercomputing, 71*, 2413–2450. https://doi.org/10.1007/s11227-014-1351-7

Verma, A., Kaushal, S., & Garg, S. (2021). Task scheduling using heuristic techniques. *Cluster Computing, 24*, 897–908. https://doi.org/10.1007/s10586-020-03145-7

Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2021). Resource management with deep reinforcement learning. *Proceedings of ACM*

HotNets. https://doi.org/10.1145/3005745.3005750

Chen, X., et al. (2021). Deep learning-based cloud resource scheduling. *IEEE Access, 9*, 56789–56800. https://doi.org/10.1109/ACCESS.2021.3056789

Roy, Y., Banville, H., Albuquerque, I., et al. (2021). Deep learning-based classification frameworks. *Journal of Neural Engineering, 18*(5), 056012. https://doi.org/10.1088/1741-2552/ac1d2c

Arabnejad, H., & Barbosa, J. G. (2021). List scheduling algorithm for heterogeneous systems. *IEEE Transactions on Parallel and Distributed Systems, 25*(3), 682–694. https://doi.org/10.1109/TPDS.2013.57

Xu, X., et al. (2022). Hybrid PSO-based scheduling in cloud computing. *Future Generation Computer Systems, 115*, 31–45. https://doi.org/10.1016/j.future.2020.09.017

Wang, S., et al. (2022). Graph neural network-based task scheduling. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2022.3156789

Ghosh, R., et al. (2022). ACO-based resource scheduling in cloud. *Applied Soft Computing, 120*, 108653. https://doi.org/10.1016/j.asoc.2022.108653

Tuncer, T., et al. (2022). Deep learning-based scheduling framework. *Applied Soft Computing, 118*, 108492. https://doi.org/10.1016/j.asoc.2022.108492

Alhudhaif, A., et al. (2022). Optimized deep learning model for cloud scheduling. *Sensors, 22*(15), 5612. https://doi.org/10.3390/s22155612

Bal, S., et al. (2022). Machine learning-based resource allocation. *Sensors, 22*(3), 1242. https://doi.org/10.3390/s22031242

Jabber, M. A., et al. (2023). Task scheduling and resource allocation review. *Journal of Cloud Computing, 12*(1), 45. https://doi.org/10.1186/s13677-023-00345-6

Chan, K. Y., et al. (2023). Deep neural networks in cloud computing. *Neurocomputing, 530*, 12–25. https://doi.org/10.1016/j.neucom.2023.01.045

Liu, Y., et al. (2023). Hierarchical CNN for resource allocation. *IEEE Access, 11*, 45678–45689. https://doi.org/10.1109/ACCESS.2023.3256784

Chen, Y., et al. (2023). Deep reinforcement learning for task scheduling. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3256789

Sharma, P., et al. (2023). Attention-based cloud scheduling. *Applied Soft Computing, 135*, 110112. https://doi.org/10.1016/j.asoc.2023.110112

Kumar, A., et al. (2023). Hybrid CNN and optimization scheduling. *Applied Soft Computing, 135*, 110045. https://doi.org/10.1016/j.asoc.2023.110045

Singh, R., et al. (2023). Graph neural networks for scheduling. *IEEE Access, 11*, 56789–56800. https://doi.org/10.1109/ACCESS.2023.3267890

Ahmed, S., et al. (2023). Hybrid scheduling models in cloud. *Computers in Industry, 154*, 103542. https://doi.org/10.1016/j.compind.2023.103542

Patel, M., et al. (2023). Graph-based scheduling in cloud computing. *Future Generation Computer Systems, 137*, 89–102. https://doi.org/10.1016/j.future.2022.08.019

Wang, Y., et al. (2023). Reinforcement learning for resource allocation. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245671

Zhou, Z., et al. (2021). Deep reinforcement learning for cloud scheduling. *arXiv preprint*. https://doi.org/10.48550/arXiv.2105.04086

Verma, R., et al. (2022). Deep learning scheduling model. *Cluster Computing*. https://doi.org/10.1007/s10586-021-03345-2

Singh, P., et al. (2022). Hybrid scheduling optimization model. *Journal of Supercomputing*. https://doi.org/10.1007/s11227-022-04321-1

Gurusamy, S., et al. (2024). Resource allocation using HAPCNN. *Expert Systems with Applications, 238*, 121234. https://doi.org/10.1016/j.eswa.2024.121234