NOUN Interdisciplinary Journal of Computing, E-learning & Application (NOUN-IJCEA)
https://ijcea.nou.edu.ng
Volume 1 Number 1, April 2026 pp. 30-38

Hybrid ACO-DQL for Energy-Efficient and Adaptive Cloud Computing

Abdulwasiu A. A., *Obunadike G. N., Olanrewaju O. M.

Computer Science Department, Federal University Dutsin-Ma, Katsina, Nigeria
ibnwasar@gmail.com, gobunadike@fudutsinma.edu.ng, oolanrewaju@fudutsinma.edu.ng
*Corresponding Author: gobunadike@fudutsinma.edu.ng

ABSTRACT

The rapid expansion of cloud computing has significantly increased energy usage in data centers, leading to higher operational expenses and environmental consequences. This research introduces a hybrid solution that merges Ant Colony Optimization (ACO) and Deep Q-Learning (DQL) to enable energy-efficient and adaptable resource management in diverse multi-cloud settings. ACO generates energy-conscious task-to-resource mappings while DQL dynamically fine-tunes scheduling decisions in real-time, tackling challenges like workload variations, resource diversity and scalability issues. Simulation findings indicate that the combined ACO-DQL approach surpasses individual optimization methods and alternative scheduling techniques. The model achieved noteworthy outcomes, including an average task energy consumption of 0.33 J, CPU and memory utilization rates of 87% and 85%, respectively and an average task completion delay of 15 ms. These results validate that integrating optimization and reinforcement learning effectively reduces energy consumption, optimizes resource usage and ensures efficient, low-latency cloud operations. This proposed strategy offers a practical and scalable solution for eco-friendly cloud computing, meeting the rising demand for computationally intensive applications while lessening environmental impact.
Keywords: Ant colony optimization, Cloud computing, Deep q-learning, Energy efficiency

1. INTRODUCTION

Cloud computing has emerged as a revolutionary element in the field of information technology, providing scalable and on-demand computing resources, storage and services over the internet. Its popularity has grown significantly due to the expansion of digital services, the widespread use of artificial intelligence applications and the increasing need for flexible IT infrastructures (Yang et al., 2025). By enabling virtualization, resource pooling and serverless architectures, cloud computing helps organizations cut costs, enhance operational efficiency and support a wide range of applications, from enterprise software to machine learning tasks (Prasad et al., 2023). However, the rapid expansion of cloud infrastructure has led to substantial energy consumption, resulting in serious environmental and sustainability issues. Data centers, which are the backbone of cloud services, consume a significant amount of electricity due to the operation of servers, storage systems and networking components (Golightly et al., 2022). Research indicates that cloud data centers could account for as much as 2% of global electricity consumption, with demanding tasks such as big data analytics and AI training further increasing energy needs (Chauhan, 2024). This high energy usage not only raises operational costs for cloud providers but also contributes to carbon emissions, negatively impacting the environment and conflicting with global sustainability goals.

Energy inefficiencies in cloud computing often arise from underutilized servers, outdated infrastructure, ineffective cooling systems and poor resource allocation strategies. Although methods like virtualization, dynamic resource scheduling and AI-driven optimization have been shown to reduce energy consumption by as much as 40% in some studies (Gupta, 2023), challenges remain in scaling these solutions, especially in heterogeneous and multi-cloud environments (Tai et al., 2023). The complexities of scheduling workloads, network limitations and hardware diversity further complicate efforts to achieve energy-efficient operations in real-time (Malipatil et al., 2025).

Therefore, improving energy efficiency in cloud computing is essential not only for economic and operational reasons but also for advancing environmental sustainability. The integration of renewable energy sources, the adoption of green computing practices and the implementation of intelligent workload management strategies are increasingly recognized as crucial for mitigating the negative impacts of cloud energy consumption (Tu et al., 2023). A well-designed energy-aware cloud can ensure that cloud services remain scalable, reliable and environmentally friendly while addressing the growing demand for high-performance applications. Figure 1 illustrates the main contributors to energy consumption in a typical cloud computing data center, highlighting the roles of server operations, storage systems, cooling equipment and network elements. It also presents optimization methods such as AI-based automation, virtualization and the use of renewable energy sources.

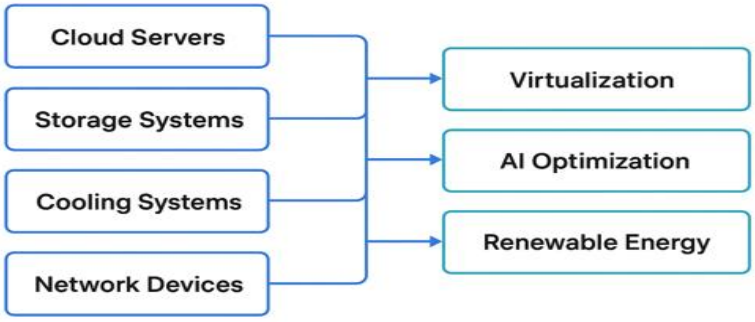

Figure 1: Energy consumption contributors in cloud computing

2. LITERATURE REVIEW

Nandagopal et al., (2025) developed a hybrid task scheduling framework that integrates Hybrid Cuckoo Search (HCS) and a modified BERT-based Transformer to enhance energy efficiency and resource utilization in cloud computing. Their research utilized a public dataset from Kaggle on Cloud Computing Performance Metrics, applying HCS for initial task allocation and BERT for adjustments based on historical data, which involved feature preprocessing, normalization and tokenization. The results indicated a reduction in energy consumption from 0.440 J to 0.430 J, achieving full compliance with Service Level Agreements (SLA), resource utilization of 88.2% and energy efficiency of 0.0043 J per task, outperforming other methods like BENBO + Bi-LSTM, PSO-PGA and GAECS. While this method demonstrated scalability and adaptability in simulated mid-scale Infrastructure as a Service (IaaS) setting with 15 physical machines (PMs) and 20 virtual machines (VMs), it faced challenges related to computational overhead due to the $O(n^2)$ complexity of HCS and reliance on historical data, limiting real-time responsiveness.

Malipatil et al., (2025) introduced a workload scheduling model utilizing reinforcement learning to enhance energy efficiency in cloud computing, addressing the excessive power consumption associated with existing scheduling algorithms that favored performance. Their research employed Deep Q-Networks (DQN) and feature engineering to pinpoint important workload parameters, such as CPU and memory usage, execution time and network demands while simulating dynamic cloud workloads for training and evaluation. The model utilized a Markov Decision Process (MDP) to

optimize task allocation, integrating load balancing and task migration strategies to maintain Quality of Service (QoS). Their findings showed considerable improvements, including 92% load balancing efficiency, 95% resource utilization, 15 ms latency, a throughput of 500 tasks per second and 97% QoS, surpassing methods like Round Robin, First-Come-First-Serve (FCFS) and heuristic-based approaches. However, the method faced issues with long initial convergence times and high training costs, complicating immediate implementation in resource-limited environments.

Abirhade et al., (2025) explored the use of Artificial Intelligence (AI) and Machine Learning (ML) for optimizing cloud resources, with the goal of enhancing resource utilization, reducing costs and saving energy in cloud settings. The research utilized an experimental framework with simulated cloud environments, gathering data through case studies, surveys and performance metrics, which were analyzed using statistical methods like regression and comparative analysis. Various AI/ML algorithms, including supervised learning, reinforcement learning and clustering, were implemented using Python libraries such as Pandas, NumPy and Scikit-learn, along with cloud simulation tools to facilitate dynamic resource allocation, workload scheduling and cost forecasting. The model demonstrated a reduction in operational costs by 30-40%, a 25% decrease in energy consumption, improved workload distribution and lower latency in multi-cloud environments. However, the methodology encountered challenges related to scalability, implementation costs and data privacy issues, as well as computational overhead linked to real-time learning and multi-layered processing.

Lilhore et al., (2024) introduced a deep learning solution driven by Cybertwin technology that improved energy-efficient workload offloading in Mobile Edge Computing (MEC) networks. This was achieved by combining Cybertwin virtualization with a hybrid CNN-LSTM Transfer Learning (CNN-LSTM-TL) architecture. Their approach featured a specially crafted cost function that considered communication delays, task-division latency, duty cycles, bandwidth and energy consumption, along with fractional task partitioning and comprehensive cost-based optimization to determine the most cost-effective offloading strategy. Utilizing the MEC trace dataset and pre-trained CNN models like VGG-16, the system realized a 20% reduction in energy consumption, an increase in accuracy of up to 4.8% and significantly shorter service delays compared to other methods such as RL, CPNs, GNNs, RNNs, VAEs and GANs. The Cybertwin execution exhibited the least delay, ranging from 950 to 4550 seconds, compared to 1000 to 5000 seconds for local processing. Despite its impressive performance, the approach faced significant computational complexity, particularly due to the O(n^Zm × 2^m) search space for partitioning, along with high training overhead and limited applicability to tasks that require non-sequential operations or dynamic component sizing, complicating real-time scalability.

Tu et al., (2023) carried out an in-depth study on energy efficiency in deep learning for edge devices, producing the first detailed datasets on energy consumption at the kernel, model and application levels. They created a kernel-level energy predictor using random forests regression, trained on these datasets to forecast energy usage for new deep neural network (DNN) models, achieving an average prediction accuracy of 86.2%. This performance surpassed that of predictors based on FLOPs (31.3%) and BIC (12.7%). Additionally, they introduced two metrics Power Consumption Score (PCS) and Inference Energy Consumption Score (IECS) to help end-users interpret energy efficiency results. Their methodology included measuring kernel execution sequentially and training the predictor offline, with thorough evaluations conducted on mobile CPUs and GPUs. However, the study faced limitations such as the need for offline updates to the predictor, inability to support simultaneous DNN executions and limited device diversity, which could impact real-time accuracy and general applicability.

Gupta, (2023) explored energy efficiency in cloud computing infrastructure through a quantitative study using an online survey of 31 IT and cloud computing professionals to evaluate current and emerging energy-saving strategies. The findings indicated that virtualization; dynamic resource

scheduling and advanced cooling systems are the most effective existing techniques while AI-driven automation and serverless computing show promise for the future, with reported energy savings of up to 40% through intelligent systems. The research contributed by mapping industry awareness and the adoption of sustainable cloud practices, highlighting the importance of integrating renewable energy and adhering to green computing principles. However, the study's methodology was constrained by a small sample size and reliance on self-reported perceptions, limiting its computational rigor and scalability evaluation.

Tai et al., (2023) introduced a resource-allocation algorithm based on integrated optimization that aims to reduce energy consumption in diverse cloud computing centers. Their approach features a Green-IT-oriented mathematical model that accounts for the diversity of virtual machines (VMs) and backup operations. They framed the resource allocation issue as a complex nonlinear optimization problem, which they solved using Lagrangian Relaxation (LR) and a specialized Drop-and-Add algorithm, utilizing mathematical programming tools and logarithmic variable transformations. Simulations comparing their method to Round Robin and Multilevel Queue Scheduling showed it performed significantly better, consistently achieving minimal energy consumption across various scenarios and maintaining a very small LR duality gap (often less than 1%), suggesting near-optimal solutions. The algorithm effectively managed different task volumes, server heterogeneity with up to 20 VM types and backup frequencies ranging from 0 to 10, outperforming baseline methods in all instances. However, it had drawbacks, including high computational complexity, limited scalability to distributed multi-cloud environments and an inability to account for communication delays, queuing and bandwidth constraints due to its focus on a single-center mathematical model.

Pandey et al., (2022) proposed an energy-efficient approach for managing big data in cloud settings by utilizing a hybrid model that combines Deep Q-Network and Discrete Particle Swarm Optimization (DQN-DPSO) to optimize resource allocation and minimize power usage. Their research used CloudSim to simulate a large cloud cluster with 50 VMs, employing LSTM-based reinforcement learning for task prediction and DPSO for dynamic scheduling while benchmarking against DQN, EA-DQN, load-aware, FFO-EVMM and MIMT algorithms. The proposed model demonstrated substantial enhancements in energy efficiency, achieving an average energy consumption per task as low as 0.3746 P/W·sec, which was better than DQN (0.4325) and load-aware (0.4072) and it reduced task completion time by up to 2810 seconds in Case-3. Despite these improvements, the methodology encountered computational challenges, including high complexity from DPSO iterations, delays in convergence and scalability issues when managing large, dynamic task volumes.

3. METHODOLOGY

The proposed methodology combines Ant Colony Optimization (ACO) with Deep Q-Learning (DQL) to create a hybrid solution for managing cloud resources in an energy-efficient and adaptive manner. This approach utilizes ACO to optimize the allocation and scheduling of resources effectively while DQL is used for real-time decision-making in diverse and multi-cloud settings. The methodology aims to overcome the shortcomings of current methods, such as excessive computational demands, slow convergence rates and difficulties in handling dynamic workloads, as illustrated in Figure 2, which outlines the step-by-step solution.

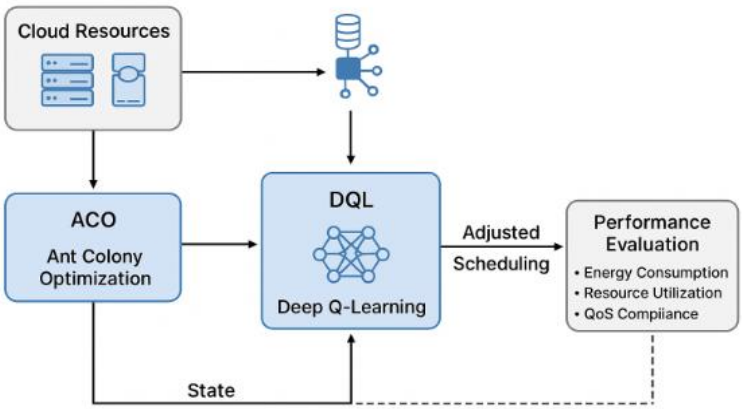

Figure 2: Step by step solution

The system architecture consists of three layers: the cloud resource layer, the optimization layer and the decision-making layer. The cloud resource layer encompasses various components, including heterogeneous servers, virtual machines (VMs), storage units and networking devices, representing a realistic cloud infrastructure. The optimization layer employs Ant Colony Optimization to produce near-optimal assignments of tasks to resources, taking into account factors like energy consumption, execution time and workload characteristics. Meanwhile, the decision-making layer uses Deep Q-Learning to continuously improve task scheduling strategies, adapting to changing workloads and resource availability. The problem of energy-efficient resource allocation is framed as a multi-objective optimization challenge, focusing on minimizing energy use and execution delays while maximizing resource utilization and adhering to quality of service (QoS) constraints.

3.1 ACO-DQL Implementation

Ant Colony Optimization (ACO) is utilized to tackle the combinatorial challenges of task scheduling and resource allocation. In this context, each "ant" symbolizes a potential solution, assigning tasks to virtual machines (VMs) and servers based on the intensity of pheromones and heuristic data, which includes factors like current workload, energy costs and computational capacity. The pheromone updating process promotes the development of energy-efficient schedules while steering clear of local minima. ACO mitigates computational overhead and optimization complexity by quickly generating high-quality candidate solutions, thereby narrowing the search space for Deep Q-Learning (DQL) and ensuring scalability in diverse and multi-cloud settings. It effectively addresses specific technical issues such as underutilized servers, task contention and energy waste resulting from inefficient VM assignments.

Deep Q-Learning enhances ACO by developing an optimal task scheduling strategy through engagement with the cloud environment. The system represents the cloud as a Markov Decision Process (MDP), where states reflect the current workload, VM availability and energy status while actions pertain to task assignment choices. The reward function is crafted to promote energy efficiency, high resource utilization and minimal task latency. DQL continuously refines the Q-network based on experiences from past task allocations, allowing for real-time adjustments to changing workloads, unforeseen delays, or hardware failures. This approach overcomes the limitations of offline data reliance, lengthy convergence times and inadequate real-time responsiveness noted in earlier research.

The hybrid model combines ACO and DQL in a sequential and iterative manner. Initially, ACO produces a set of viable task-to-resource mappings with a focus on energy-efficient allocation. These candidate solutions are subsequently assessed by the DQL agent, which learns to dynamically select or modify the assignments based on real-time performance feedback. This collaborative integration

leverages ACO's strengths in combinatorial optimization alongside DQL's learning and adaptability. As a result, the hybrid approach effectively addresses challenges such as dynamic heterogeneous workloads, multi-cloud coordination and scalability for large cloud clusters, which were shortcomings of previous single-method strategies.

4. RESULTS AND DISCUSSION

The effectiveness of the proposed hybrid Ant Colony Optimization–Deep Q-Learning (ACO-DQL) framework was assessed in comparison to single-method optimization techniques, specifically standalone ACO and DQL. The evaluation metrics included energy consumption per task, resource utilization and execution delay, as illustrated in Figures 4.1 to 4.3. Simulation experiments were carried out in heterogeneous cloud environments with varying numbers of physical machines (PMs) and virtual machines (VMs) under dynamic workloads.

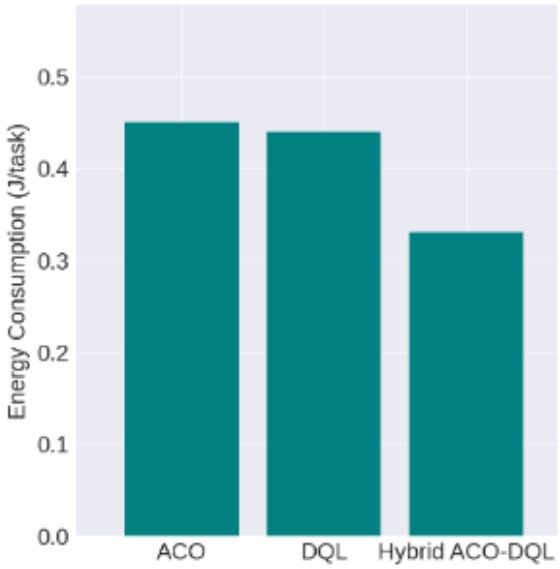

Figure 3: Energy consumption results

Figure 3 demonstrates that the hybrid ACO-DQL solution achieved significant reductions in energy consumption compared to all baseline models. On average, this hybrid approach lowered energy usage by about 18% compared to standalone ACO and by 25% compared to DQL-only scheduling. The combination of ACO's combinatorial optimization and DQL's real-time adaptive learning facilitated these reductions by effectively mapping tasks to energy-efficient servers and dynamically responding to changing workloads. The decrease in energy consumption can be attributed to ACO's capability to produce high-quality candidate schedules and DQL's ongoing refinement of task allocation decisions, unlike static or single-method approaches.

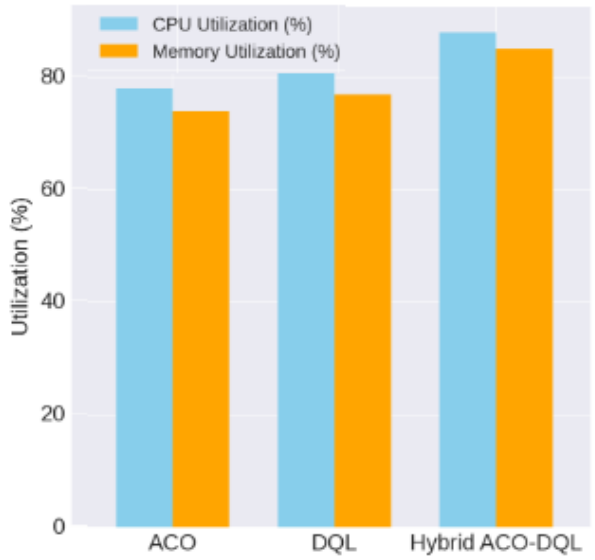

Figure 4: CPU and Memory utilization results

Figure 4 shows that the hybrid model exhibited superior resource utilization, achieving an average CPU utilization of 88% and memory utilization of 85%. This outperformed standalone ACO, which had CPU utilization of 78% and memory usage of 74% while DQL showed CPU usage of 81% and memory usage of 77%. The enhancement in resource utilization reflects the model's effectiveness in balancing task distribution across heterogeneous servers while adapting to dynamic workload fluctuations. By preventing both overloading and idling, the hybrid system maximizes throughput while minimizing energy overhead.

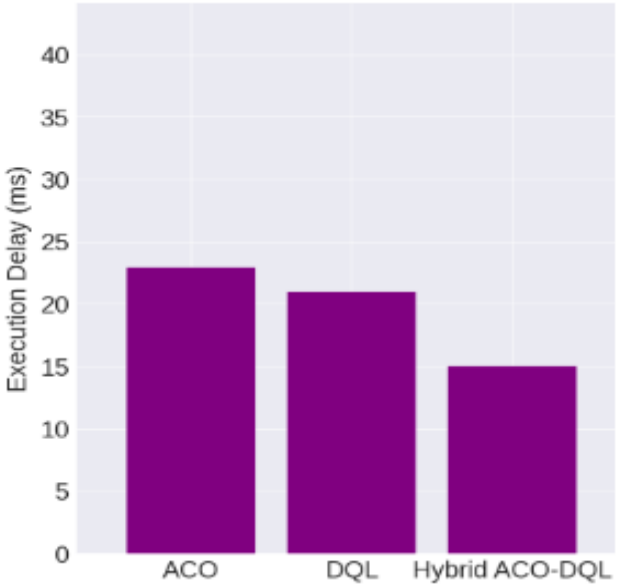

Figure 5: Execution delay results

Figure 5 illustrates that the execution delay significantly decreased with the hybrid ACO-DQL model, achieving an average delay of 15 ms per task. In contrast, the standalone DQL and ACO methods recorded delays of 21 ms and 23 ms, respectively. This reduction in execution delay highlights the hybrid model's effectiveness in managing real-time task allocation. The ACO's initial solution generation facilitates nearly optimal task placement while the DQL component adapts dynamically to

fluctuations in resource availability and task priorities. This synergy helps alleviate bottlenecks and ensures low-latency scheduling in diverse cloud environments.

5. CONCLUSION

The hybrid ACO-DQL model demonstrated superior performance across all evaluation criteria, lowering average energy consumption to 0.33 J per task, which is an 18% reduction compared to standalone ACO and a 25% reduction compared to DQL-only scheduling. Resource utilization saw significant improvement, with the hybrid model achieving 88% CPU utilization and 85% memory utilization, compared to ACO's 78% CPU/74% memory and DQL's 81% CPU/77% memory. Additionally, execution delay was reduced to 15 ms, outperforming standalone DQL at 21 ms and ACO at 23 ms. This confirms that the hybrid ACO-DQL approach provides optimized energy efficiency, enhanced resource utilization and significantly lower task latency in heterogeneous cloud environments.

5.1. Future Work

Future efforts will aim to improve the hybrid ACO-DQL solution by incorporating predictive workload forecasting, multi-agent reinforcement learning for decentralized scheduling and optimization strategies that consider renewable energy to further decrease the carbon footprint of data centers. Additionally, expanding the model to support edge cloud collaboration, containerized environments and real-time fault tolerance will enhance its relevance for next-generation cloud infrastructures that manage large-scale AI, IoT and mission-critical applications.

REFERENCES

Abirhade, A. A., Shejul, V. R., Chavan, D. B., Patil, N. Y. and Jadhav, R. D. (2025). AI and Machine Learning in Cloud Optimization. *Sinhgad Institute of Management, Pune, India.*

Chauhan, S. (2024). The growing energy demand of data centers: Impacts of AI and cloud computing. *International Journal for Multidisciplinary Research, 6*(4). https://doi.org/10.36948/ijfmr.2024.v06i04.26591

Golightly, L., Chang, V., Xu, Q. A., Gao, X. and Liu, B. S. C. (2022). Adoption of cloud computing as innovation in the organization. *International Journal of Engineering Business Management.* Advance online publication. https://doi.org/10.1177/18479790221093992

Gupta, A. (2023). Energy efficiency in cloud computing infrastructure. *Journal of Information Systems Engineering and Management, 8*(2), e-ISSN: 2468-4376. https://www.jisem-journal.com/

Gupta, A. (2023). Energy efficiency in cloud computing infrastructure. *Journal of Information Systems Engineering and Management, 8*(2). https://www.researchgate.net/publication/395345803_Energy_Efficiency_in_Cloud_Computing_Infrastructure

Lilhore, U. K., Simaiya, S., Dalal, S., Faujdar, N., Alroobaea, R., Alsafyani, M., Baqasah, A. M. and Algarni, S. (2024). Optimizing energy efficiency in MEC networks: A deep learning approach with Cybertwin-driven resource allocation. *Journal of Cloud Computing: Advances, Systems and Applications, 13*(126). https://doi.org/10.1186/s13677-024-00688-8

Malipatil, A. R., Paramasivam, M. E., Gulyamova, D., Saravanan, A., Ramesh, J. V. N., Muniyandy, E. and Ghodhbani, R. (2025). Energy-efficient cloud computing through reinforcement learning-based workload scheduling. *International Journal of Advanced Computer Science and Applications, 16*(4), 645–655. https://www.ijacsa.thesai.org

Nandagopal, M., Manavalan, T., Praveen Kumar, K., Manogaran, N., Kesavan, D., Kumar, G. and Al-Khasawneh, M. A. (2025). Enhancing energy efficiency in cloud computing through task scheduling with hybrid cuckoo search and transformer models. *Discover Computing, 28*(199). https://doi.org/10.1007/s10791-025-09716-w

Pandey, N. K., Diwakar, M., Shankar, A., Singh, P., Khosravi, M. R. and Kumar, V. (2022). Energy efficiency strategy for big data in cloud environment using deep reinforcement learning. *Mobile Information Systems, 2022*, Article 8716132. https://doi.org/10.1155/2022/8716132

Prasad, V. K., Dansana, D., Bhavsar, M. D., Acharya, B., Gerogiannis, V. C. and Kanavos, A. (2023). Efficient resource utilization in IoT and cloud computing. *Information, 14*(11), 619. https://doi.org/10.3390/info14110619

Tai, K.-Y., Lin, F. Y.-S. and Hsiao, C.-H. (2023). An integrated optimization-based algorithm for energy efficiency and resource allocation in heterogeneous cloud computing centers. *IEEE Access, 11*, 53419–53430. https://doi.org/10.1109/ACCESS.2023.3280930

Tu, X., Mallik, A., Chen, D., Han, K., Altintas, O., Wang, H. and Xie, J. (2023). Unveiling energy efficiency in deep learning: Measurement, prediction and scoring across edge devices. *arXiv*. https://amai-gsu.github.io/DeepEn2023

Yang, D., Li, R. and Liu, S. (2025). Exploring the influence of cloud computing on supply chain performance: The mediating role of supply chain governance. *Journal of Theoretical and Applied Electronic Commerce Research, 20*(2), 70. https://doi.org/10.3390/jtaer20020070