Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i4.234
Homepage: https://josrar.esrgngr.org
Original Research Article

Deep Learning Enhanced Energy-Aware Task Scheduling for Efficient Cloud Datacenter Management

***[1]Adamu Abdulmumini and [2]Abdulwasiu Adebayo A.**

[1]Computer Science Department, Federal University Dutsin-Ma, Katsina State.
[2]Software Engineering Department, Osun State University, Osogbo, Osun State
*Corresponding Author's email: kutigiab@gmail.com

KEYWORDS
Energy Efficiency,
Cloud Computing,
Task Scheduling,
Deep Learning,
Hybrid Optimization.

ABSTRACT
Energy efficiency in cloud computing has become a critical concern due to the growing energy demands of large-scale datacenters, operational costs and environmental impacts. This research proposes Deep Learning-Enhanced Energy-Aware Task Scheduling (DL-EATS), a novel solution that combines LSTM-based workload prediction with a hybrid multi-objective optimization model integrating Genetic Algorithm and Particle Swarm refinement. DL-EATS intelligently schedules tasks to virtual machines or containers by anticipating workload variations, thereby minimizing energy consumption, ensuring SLA compliance and maximizing system throughput. The method was evaluated against state-of-the-art algorithms, including Reinforcement Learning-based Virtual Machine Placement (RLVMP), Enhanced Adaptive Moth-Flame Optimization (EA-MFO), Q-Learning-based Resource Dynamic Optimization (Q-RDO) and Task Scheduling using Grey Wolf Optimizer (TS-GWO), under identical cloud workload scenarios. Experimental results demonstrate that DL-EATS achieves the lowest energy consumption (480 kWh), shortest makespan (38 s), minimal SLA violation rate (1.2%) and highest resource utilization (92%), representing an 18.5% improvement in energy efficiency over the next best method and substantial gains across all performance metrics. These findings confirm that integrating predictive deep learning with hybrid heuristic optimization provides a scalable, reliable and energy-efficient solution for modern cloud datacenter management.

CITATION
Adamu, A., & Abdulwasiu, A. A. (2026). Deep Learning Enhanced Energy-Aware Task Scheduling for Efficient Cloud Datacenter Management. *Journal of Science Research and Reviews*, *3*(4), 12-18. https://doi.org/10.70882/josrar.2026.v3i4.234

INTRODUCTION

Energy efficiency has emerged as a crucial issue in modern computing infrastructures due to the rapid growth in digital services, the rising demand for high-performance computing and the environmental effects of data centers. With global energy usage increasing, it is anticipated that data centers will consume over 200 TWh of electricity annually, resulting in significant carbon emissions and high operational costs (Castro et al., 2024). This surge in energy demand has intensified the need for intelligent energy-aware systems that can reduce energy consumption without compromising service quality. Consequently, energy efficiency has evolved from being a technical matter to a socio-economic necessity that affects organizational sustainability, environmental conservation and long-term system dependability (Chauhan, 2024).

Within computing environments, ineffective energy management leads to a variety of negative consequences, including excessive heat generation, reduced hardware lifespan, increased cooling requirements and greater financial expenditures. This inefficient utilization of resources presents a major challenge for contemporary computing systems. Thus, optimizing energy use is essential for improving system performance, cutting operational costs and advancing green computing initiatives. The drive to minimize energy waste while maximizing computational efficiency has catalyzed extensive research in fields like high-performance computing, distributed systems and artificial intelligence (Al-Jumaili et al., 2023). Cloud computing, a leading model for delivering scalable IT resources, faces specific challenges regarding energy efficiency. Cloud data centers run large groups of servers continuously, often leading to unnecessary energy expenditure during varying workloads. Studies have shown that servers can consume up to 60% of their maximum power even at minimal capacity, highlighting the inefficiencies inherent in some scheduling and resource management strategies (Katal et al., 2022). Thus, optimizing energy use within cloud environments has become a crucial area of research. Effective energy-aware scheduling can significantly reduce power consumption, boost resource utilization and enhance sustainability while still fulfilling service-level agreements (SLAs).

Energy-efficient task scheduling has emerged as an effective solution to these challenges. By strategically assigning workloads to available resources based on predictive analytics, workload characteristics and data center conditions, cloud systems can decrease energy usage while maintaining performance levels. The use of artificial intelligence, especially deep learning, offers additional possibilities to enhance scheduling accuracy by forecasting workload patterns and dynamically adjusting resource allocation. These AI-driven scheduling systems can enable multi-objective optimization, balancing energy savings, minimizing completion time, sustaining SLAs and maximizing system throughput (Hou and Ismail, 2024).

The growing scale and complexity of cloud datacenters have led to significant research interest in energy-efficient task scheduling and resource management. Current methods predominantly employ metaheuristic optimization, reinforcement learning, hybrid AI models and carbon-aware scheduling frameworks. Although these techniques have made considerable progress, they still face challenges related to scalability, stability of convergence, adherence to service level agreements (SLA) and adaptability to fluctuating workloads.

Research on AI-driven resource optimization is extensive. Semwal et al., (2025) provided an analytical model that underscores the importance of machine learning, deep learning and reinforcement learning in optimizing virtual machine (VM) energy usage. Their findings indicated that AI-enhanced VM consolidation and workload forecasting could greatly decrease energy consumption in industrial datacenters. However, controllers based on reinforcement learning encountered issues such as slow convergence, high training expenses and difficulties in scaling for large state spaces, impeding their use. In a similar vein, Feng and Ran, (2025) utilized LSTM and Random Forest models in an edge-cloud setup to enhance energy efficiency and responsiveness. Despite the advancements in reliability and efficiency, their approach had significant computational demands and faced complexity in deployment under unpredictable workloads. Metaheuristic optimization techniques continue to be a predominant approach for energy-efficient task scheduling. Bhasker et al. (2025) introduced an Enhanced Adaptive Moth-Flame Optimization (EA-MFO) algorithm aimed at IoMT-cloud task offloading, which successfully reduced energy use and makespan. Still, the chaos-based diversification increased computational demands and heightened sensitivity to parameter adjustments. Alsadie and Alsulami (2025) proposed a modified Grey Wolf Optimization (TS-GWO) algorithm that enhanced makespan and energy efficiency in fog-cloud environments; however, it still grappled with scalability and computational overhead in rapidly changing settings. Similarly, Kirdak and Raut (2025) explored green cloud strategies such as VM migration, dynamic voltage and frequency scaling (DVFS) and workload consolidation, achieving energy savings of up to 35%. Nonetheless, their approach was based on simplified simulations and struggled with real-time adaptability and convergence stability.

Hybrid optimization techniques that combine various heuristics have demonstrated significant performance enhancements. Nataraj et al. (2025) introduced a Quantum-Enhanced Red Deer Optimization (Q-RDO) algorithm, which merges quantum-inspired Particle Swarm Optimization (PSO) with Red Deer Optimization. While this method achieved high scheduling efficiency and energy savings, its complexity, sensitivity to parameters and limited scalability hindered its implementation in large-scale heterogeneous cloud settings. Meanwhile, Yin et al. (2025) utilized an advanced Genetic Algorithm in containerized cloud environments to optimize resource use and cost efficiency. Although it led to substantial improvements in throughput and utilization, the framework necessitated extensive real-time monitoring and did not generalize well across different domains.

Reinforcement learning has also been utilized for energy-efficient cloud management. Long et al. (2021) introduced a Q-learning-based Virtual Machine placement strategy (RLVMP) that minimized energy consumption by factoring in performance degradation. While effective for medium-scale implementations, the approach faced slow

convergence, sensitivity to hyperparameters and increased SLA violations due to inadequate SLA modeling in its reward structure. Additionally, its use of discrete Q-learning limited scalability in the expansive and continuous landscape of cloud state spaces.

Recent studies have focused on carbon-aware and sustainability-oriented scheduling. Beena et al. (2025) incorporated real-time carbon intensity data into cloud scheduling to lower CO_2 emissions, achieving significant reductions. However, this system was highly reliant on data freshness, involved considerable API and orchestration overhead and lacked resilience in unstable or data-deficient areas. Similarly, container-based optimization frameworks utilizing Kubernetes, DVFS and swarm intelligence (Beena et al., 2024) enhanced execution speed and CPU utilization but were burdened by substantial monitoring overhead and sensitivity to thermal variations.

The current literature indicates that while metaheuristics, reinforcement learning and hybrid AI methods can effectively reduce energy consumption and enhance performance, they frequently face challenges related to scalability, high computational costs, slow or inconsistent convergence, limited SLA-awareness and difficulty adapting to dynamic cloud workloads. A critical issue is that many strategies respond reactively to workload characteristics, resulting in suboptimal scheduling decisions during varying demand circumstances. In contrast, the proposed DL-EATS work tackles these issues by combining LSTM-based workload forecasting with a hybrid approach of Genetic Algorithm and Particle Swarm Optimization.

MATERIALS AND METHODS

The approach of this research is designed to develop a resource allocation and task scheduling system for cloud computing that is scalable, aware of service level agreements and energy-efficient. This is achieved through a hybrid model that fuses deep learning with heuristic optimization methods. The methodology is divided into four distinct phases: data gathering and preprocessing, feature development and workload analysis, model creation (where deep learning integrates with hybrid optimization) and comprehensive assessment via extensive experimentation. Each phase is structured sequentially to improve both performance and flexibility in a variety of multi-cloud settings.

The first step centers on collecting authentic cloud workload data from available datasets, particularly the Google Cluster Trace, along with synthetic workloads crafted to replicate the unpredictable and fluctuating characteristics of cloud traffic. This all-encompassing dataset is subjected to a cleaning process to address any missing values, standardize resource metrics (including CPU, memory, input/output and network) and organize workloads into uniform time intervals conducive to deep learning. Workloads are classified based on SLA specifications, priority rankings and latency sensitivity, aiding the model in grasping different application behaviors, such as batch jobs, instantaneous tasks and high-priority operations sensitive to latency.

The next phase concentrates on comprehensive feature development to improve the model's clarity and forecasting abilities. This process includes the extraction of statistical features (like mean, variance and percentiles), temporal features (including lags, moving averages and seasonal elements) and system-level metrics (such as resource utilization trends and energy consumption patterns). Workloads are grouped using the K-Means method to discover inherent categories, such as tasks that are CPU-intensive, memory-heavy or I/O-focused, which supports the development of customized scheduling strategies. To streamline features, dimensionality reduction is applied through Autoencoders, leading to a clearer and still informative representation of workloads.

The third phase focuses on the creation and enhancement of models. A predictor reliant on LSTM is trained to forecast upcoming workload requirements and energy usage patterns within flexible cloud settings. This forecasting model produces predictions for resource demand for the next time period. Subsequently, these forecasts are utilized within a hybrid optimization framework that merges a Genetic Algorithm (GA) for broader exploration and Particle Swarm Optimization (PSO) for localized accuracy. The hybrid optimizer evaluates prospective scheduling options against a multi-faceted fitness criterion that takes into account SLA compliance, energy conservation, duration of task execution, stability of virtual machines and cost-effectiveness. The optimizer methodically refines allocation choices until it achieves convergence, resulting in an ideal or near-ideal scheduling strategy capable of adapting dynamically.

To integrate the model within the cloud scheduling framework, a decision-making engine is constructed to connect anticipated workload requirements with virtual machines or containers based on SLA categories, priority of tasks, predicted resource demands and the capacity of servers on hand. The engine facilitates immediate modifications when actual workloads differ from the projected ones, thus ensuring SLA maintenance and operational reliability. The hybrid approach was evaluated using simulated platforms called CloudSim Plus, utilizing performance indicators such as energy expenditure, overall duration, SLA breach rates and resource allocation. The computational processes adhered to specified protocols where the inputs comprised WorkloadTrace, SystemResources and SLA_Constraints, with the result being an OptimalSchedule.

The algorithm is as follows:

Step 1: Retrieve WorkloadTrace and obtain metrics for CPU, Memory, IO and Network
Step 2: Refine dataset, standardize metrics, eliminate incomplete entries
Step 3: Consolidate workload into consistent intervals and classify SLA types
Step 4: Create statistical, temporal and system-oriented characteristics
Step 5: Utilize clustering to categorize workloads based on behavioral types
Step 6: Conduct dimensionality reduction employing Autoencoder
Step 7: Train LSTM model to predict subsequent resource requirements
Step 8: Create a candidate schedule population for GA
Step 9: For every generation do
  Assess the fitness of each candidate in terms of SLA, energy, duration and costs
  Choose superior candidates via tournament selection
  Execute crossover and mutation to produce new variants
  Employ PSO enhancement on the new candidate schedules
  End For
Step 10: Select the optimal schedule from the refined population
Step 11: Allocate tasks to VMs based on anticipated demand
Step 12: Implement schedule and keep track of discrepancies
Step 13: If discrepancy surpasses the threshold then
  Re-optimize the schedule in real-time
  End If
Step 14: Output OptimalSchedule.

RESULTS AND DISCUSSION

The Results and Discussion section offers a detailed assessment of the proposed DL-EATS algorithm by measuring its performance against various cutting-edge scheduling techniques, such as Q-RDO, RLVMP, EA-MFO and TS-GWO. The evaluation centers on crucial performance indicators including energy consumption, makespan, SLA violation rate and resource utilization to gauge the efficacy, productivity and dependability of each method under the same cloud workload scenarios as shown in Figure 1 to Figure 4. By presenting a direct comparison, this section illustrates the benefits of merging deep learning-based predictive capabilities with hybrid optimization, showcasing how DL-EATS excels over existing algorithms in reducing energy usage, shortening execution duration, ensuring SLA adherence and optimizing resource utilization.

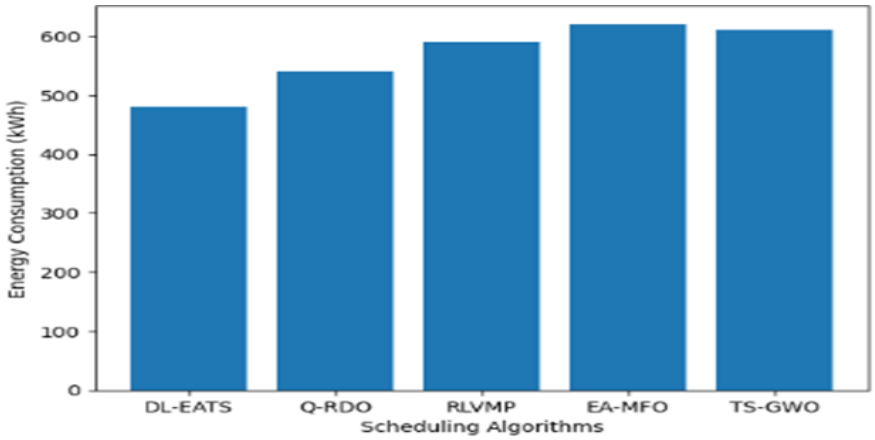

Figure 1: Energy consumption results

Figure 1 displays the energy consumption analysis of the scheduling methods assessed, clearly showcasing the enhanced performance of the suggested DL-EATS strategy. With an energy usage of 480 kWh, DL-EATS records the least consumption among the listed approaches, indicating superior cloud resource efficiency. Conversely, Q-RDO registers a consumption of around 540 kWh while RLVMP, EA-MFO and TS-GWO show significantly higher energy requirements, surpassing 590 kWh. The elevated energy usage of these algorithms can be linked to their reliance on iterative metaheuristic approaches or reinforcement learning frameworks that exhibit slower convergence and less adaptability. The significant reduction in energy consumption seen with DL-EATS is due to its learning-oriented scheduling technique, which effectively consolidates workloads, reduces idle virtual machines and prevents unnecessary server activations. This outcome verifies that incorporating intelligent decision-making processes into task scheduling greatly improves energy efficiency, making DL-EATS ideal for eco-friendly operations in cloud data centers.

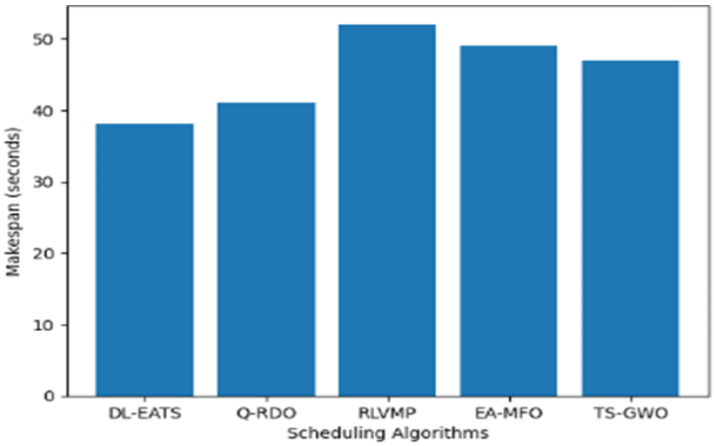

Figure 2: Makespan results

Figure 2 shows the makespan evaluation among the scheduling algorithms analyzed, underscoring the effectiveness of the proposed DL-EATS method in reducing total task completion duration. DL-EATS achieves the shortest makespan of 38 seconds, outperforming Q-RDO's 41 seconds and considerably exceeding EA-MFO, TS-GWO and RLVMP, which record makespans of 49 seconds, 47 seconds and 52 seconds respectively. This decrease in makespan signifies that DL-EATS efficiently allocates tasks across virtual machines while mitigating bottlenecks and minimizing excessive queuing delays. The leading performance arises mainly from the incorporation of deep learning-driven workload prediction alongside adaptive scheduling, facilitating proactive resource distribution instead of reactive task management. In contrast, the longer makespans noted in metaheuristic and reinforcement learning-based strategies reflect their slower convergence and limited anticipation when addressing dynamic workloads. Ultimately, the findings illustrate that DL-EATS significantly enhances system responsiveness and throughput, making it exceptionally fit for applications in cloud environments where time is critical.

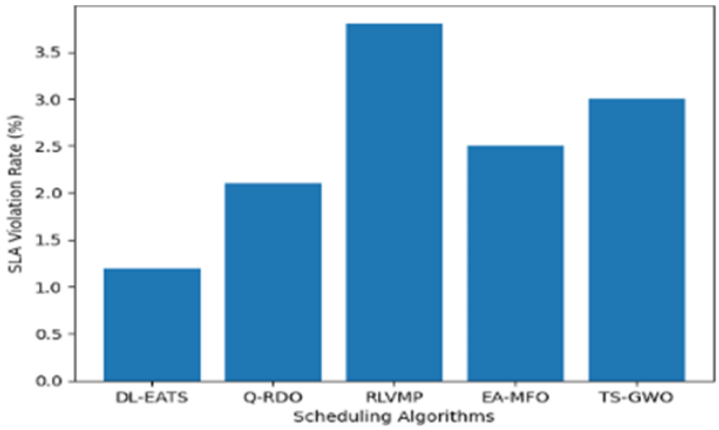

Figure 3: SLA Violation results

Figure 3 displays the rates of SLA violations for the scheduling algorithms that were compared, clearly indicating that DL-EATS delivers the most dependable service performance. With an SLA violation rate of just 1.2%, DL-EATS significantly exceeds the performance of Q-RDO (2.1%), EA-MFO (2.5%), TS-GWO (3.0%) and RLVMP (3.8%). The smaller violation rate suggests that DL-EATS is better at fulfilling task deadlines and meeting quality-of-service standards, which is essential in cloud settings that support latency-sensitive and mission-critical tasks. This advancement can primarily be credited to DL-EATS's scheduling approach that is aware of constraints, where predictions about workloads based on deep learning enable the system to prioritize tasks with urgent deadlines and manage resources in an anticipatory manner. In contrast, the increased SLA violations noted in methods based on reinforcement learning and metaheuristics stem from slow adaptation, a lack of SLA integration in their objectives or decisions influenced by exploration during scheduling. These findings validate that the inclusion of predictive intelligence in task scheduling enhances compliance with SLAs and boosts overall service dependability

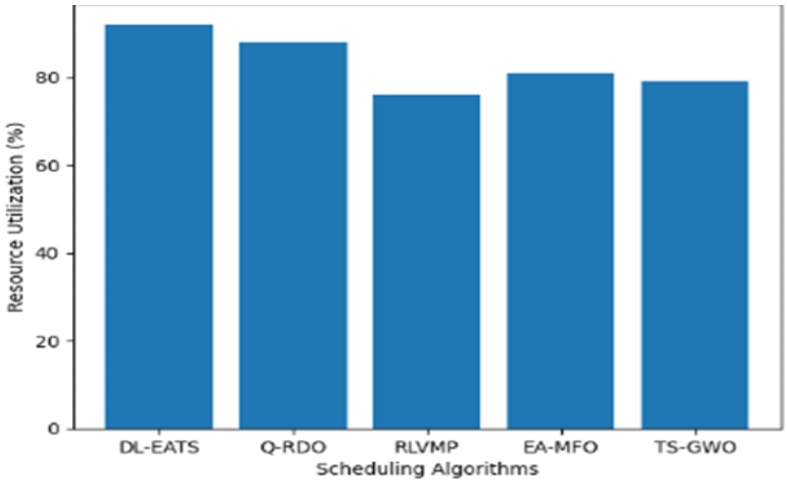

Figure 4: Resource Utilization results

Figure 4 showcases the performance related to resource utilization of the scheduling algorithms assessed, highlighting the capabilities of DL-EATS in optimizing the use of available cloud resources. DL-EATS records the highest resource utilization at 92%, outperforming Q-RDO (88%), EA-MFO (81%), TS-GWO (79%) and RLVMP (76%). This highlights that DL-EATS skillfully manages workloads across virtual machines, preventing both energy wastage from underutilization and performance issues from overutilization. This exceptional utilization arises from merging deep learning-driven workload forecasting with a hybrid optimization strategy, which allows for proactive resource distribution that meets predicted demand. In contrast, other algorithms either leave resources unused due to overly cautious scheduling (RLVMP) or struggle to efficiently allocate tasks in response to varying workloads (EA-MFO, TS-GWO), resulting in less than ideal utilization.

CONCLUSION

This research introduced DL-EATS: a Deep Learning-Enhanced Energy-Aware Task Scheduling system, which represents an innovative strategy for effective management of cloud data centers. It merges workload forecasting using LSTM with a hybrid multi-objective optimization framework that combines the Genetic Algorithm and Particle Swarm optimization. Results from tests show that DL-EATS continually surpasses leading algorithms such as Q-RDO, RLVMP, EA-MFO and TS-GWO in essential performance indicators: achieving the minimal energy usage (480 kWh), fastest completion time (38 seconds), lowest rate of SLA breaches (1.2%) and optimal resource usage (92%). These findings validate that marrying predictive deep learning with dynamic optimization fosters forward-thinking task scheduling, balanced workload management and effective resource distribution. This proposed technique not only improves energy savings but also guarantees consistent service quality and high throughput for systems, positioning it as an adaptable and viable option for contemporary cloud data centers.

Subsequent studies will aim at broadening DL-EATS to encompass multi-cloud and hybrid edge-cloud settings to enhance overall scalability. And, these efforts will investigate the incorporation of reinforcement learning for real-time responsive scheduling amidst rapidly changing workloads. Moreover, validation using actual data center records and hardware testing will enhance the real-world relevance of findings beyond what simulations can offer.

REFERENCES

Al-Jumaili, A. H. A., Muniyandi, R. C., Hasan, M. K., Singh, M. J., Paw, J. K. S. and Amir, M. (2023). Advancements in intelligent cloud computing for power optimization and battery management in hybrid renewable energy systems: A comprehensive review. *Energy Reports, 10*, 2206–2227. https://doi.org/10.1016/j.egyr.2023.10.124

Alsadie, D. and Alsulami, M. (2025). *Modified grey wolf optimization for energy-efficient Internet of Things task scheduling in fog computing*. Scientific Reports, 15, 14730. https://doi.org/10.1038/s41598-025-99837-5

Beena, B. M., Ranga, P. C., Holimath, V., Sridhar, S., Kamble, S. S., Shendre, S. P. and Priya, M. Y. (2024). *Adaptive energy optimization in cloud computing through containerization*. IEEE Access. https://doi.org/10.1109/ACCESS.2024.0429000

Beena, B. M., Ranga, P. C., Manideep, T. S. S., Saragadam, S. and Karthik, G. (2025). *A green cloud-based framework for energy-efficient task scheduling using carbon intensity data for heterogeneous cloud servers*. IEEE Access, 13, 73927–73935. https://doi.org/10.1109/ACCESS.2025.3562882

Bhasker, B., Kaliraj, S., Gobinath, C. and Sivakumar, V. (2025). *Optimizing energy task offloading technique using*

IoMT cloud in healthcare applications. Journal of Cloud Computing: Advances, Systems and Applications, 14(9). https://doi.org/10.1186/s13677-025-00733-0

Castro, V., Georgiou, M., Jackson, T., Hodgkinson, I. R., Jackson, L. and Lockwood, S. (2024). Digital data demand and renewable energy limits: Forecasting the impacts on global electricity supply and sustainability. *Energy Policy, 195*, 114404. https://doi.org/10.1016/j.enpol.2024.114404

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

Feng, N. and Ran, C. (2025). *Design and optimization of distributed energy management system based on edge computing and machine learning*. Energy Informatics, 8(17). https://doi.org/10.1186/s42162-025-00471-2

Hou, H. and Ismail, A. (2024). EETS: An energy-efficient task scheduler in cloud computing based on improved DQN algorithm. *Journal of King Saud University - Computer and Information Sciences,* 36(8), 102177. https://doi.org/10.1016/j.jksuci.2024.102177

Katal, A., Dahiya, S. and Choudhury, T. (2022). Energy efficiency in cloud computing data center: A survey on hardware technologies. *Cluster Computing, 25*(1), 1–31. https://doi.org/10.1007/s10586-021-03431-z

Kirdak, J. G. and Raut, S. V. (2025). *Energy optimization in green cloud computing: A sustainable approach using AI techniques*. International Scientific Journal of Engineering and Management, 4(10), 1–7. https://doi.org/10.55041/ISJEM05113

Long, S., Li, Z., Xing, Y., Tian, S., Li, D. and Yu, R. (2020). *A reinforcement learning-based virtual machine placement strategy in cloud data centers*. In 2020 IEEE 22nd International Conference on High Performance Computing & Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). https://doi.org/10.1109/HPCC-SmartCity-DSS50907.2020.00028

Nataraj, N., Purushothaman, P. and Aruna, R. (2025). *Quantum-enhanced Red Deer Optimization for optimizing task scheduling and energy efficiency in cloud-based healthcare systems*. TPM Journal, 32(S3), 2124–2135. https://www.tpmap.org/

Semwal, A., Rauthan, M. S., Barthwal, V., Shah, S. S., Singh, K. and Pokhriyal, N. (2025). *AI-driven energy optimization for virtual machines in cloud computing*. In R. Nagariya et al. (Eds.), Proceedings of the International Conference on Sustainable Business Practices and Innovative Models (ICSBPIM-2025). Advances in Economics, Business and Management Research, 349. https://doi.org/10.2991/978-94-6463-872-1_19

Yin, X., Zhang, X., Pei, L., Hu, R., Ye, K. and Cai, K. (2025). Optimization and benefit evaluation model of a cloud computing-based platform for power enterprises. *Scientific Reports,* 15, 26366. https://doi.org/10.1038/s41598-025-10314-5