Machine Learning Driven Task Scheduling for Sustainable and Energy Efficient Cloud Infrastructure

Perbhat Singh [1], Dr. Yogesh[2]
[1]Research Scholar, Department of Computer Science, NIILM University, Kaithal, Haryana
[2]Supervisor, Department of Computer Science, NIILM University, Kaithal, Haryana

Abstract: The rapid expansion of cloud computing has transformed the global digital ecosystem by enabling scalable, flexible, and on-demand computing services for governments, industries, academic institutions, and individual users. However, the large-scale growth of cloud infrastructure has also increased energy consumption, carbon footprint, heat generation, and operational complexity in data centers. In this context, task scheduling has emerged as one of the most critical mechanisms for improving resource utilization, service quality, and energy efficiency in cloud environments. Conventional scheduling methods are often rule-based, static, or reactive, and therefore they struggle to adapt to dynamic workloads, heterogeneous infrastructure, fluctuating user demand, and sustainability goals. This research article explores the role of machine learning-driven task scheduling as an intelligent and adaptive framework for building sustainable and energy-efficient cloud infrastructure. The study adopts an exploratory and conceptual research approach to examine how machine learning models can predict workload patterns, classify task behavior, optimize task placement, reduce idle resource consumption, and support green cloud operations. The article discusses the limitations of traditional scheduling methods, explains the relevance of supervised, unsupervised, and reinforcement learning techniques, and proposes an innovative framework for energy-aware task scheduling in cloud systems. It further evaluates the potential of machine learning to improve makespan, throughput, SLA compliance, load balancing, and power efficiency simultaneously. The findings suggest that machine learning-driven scheduling is not merely a technical enhancement but a strategic pathway toward sustainable cloud infrastructure. The article concludes that intelligent task scheduling can significantly contribute to the future of green computing by enabling adaptive, predictive, and context-aware resource management in cloud data centers.

Keywords: Cloud computing, task scheduling, machine learning, sustainable infrastructure, energy efficiency, green cloud computing, resource optimization, intelligent scheduling.

I. INTRODUCTION

Cloud computing has become one of the most influential technological paradigms of the modern digital age. It supports a wide range of applications, including e-commerce platforms, online education, smart healthcare, digital finance, scientific computing, enterprise systems, and large-scale data analytics. The cloud model provides computing resources as services and allows users to access processing power, storage, software, and platforms on demand. This capability has made cloud infrastructure essential for digital transformation across sectors. Yet, the success of cloud computing has also produced a serious challenge: rising energy consumption in data centers and cloud environments.

Cloud data centers operate continuously and consume significant electricity for servers, processors, storage devices, cooling systems, and network equipment. A large portion of this energy is wasted due to underutilized servers, poor task distribution, inefficient resource allocation, and outdated scheduling practices. In many cloud systems, tasks are assigned to virtual machines or physical hosts without sufficiently considering energy usage patterns, workload behavior, or sustainability objectives. As a result, cloud providers face increasing operational costs, thermal management issues, performance bottlenecks, and environmental concerns. The question is no longer whether cloud systems are efficient in terms of performance alone, but whether they are sustainable in the long term.

Task scheduling plays a central role in this issue. In cloud infrastructure, task scheduling refers to the process of assigning incoming computational tasks to available resources in a way that satisfies performance goals, user demands, and service-level agreements. An effective scheduling mechanism must manage multiple objectives simultaneously, such as minimizing execution time, balancing system load, improving throughput, reducing response time, and lowering energy consumption. Traditional scheduling algorithms like First Come First Serve, Round Robin, Min-Min, Max-Min, and heuristic-based approaches have been useful in early cloud environments, but they often fail to adapt intelligently to changing system conditions. These techniques generally do not learn from historical data, cannot anticipate future workloads, and are limited in handling complex multi-objective optimization.

Machine learning offers a promising solution to this challenge. By learning from workload traces, system logs, resource usage histories, and performance feedback, machine learning models can make informed scheduling decisions that are predictive rather than merely reactive. They can forecast resource demand, classify tasks according to priority and energy impact, detect workload anomalies, and optimize task placement dynamically. In this sense, machine learning-driven task scheduling can support not only better performance but also greener and more sustainable cloud operations. This article explores this possibility in depth and presents an innovative and exploratory discussion on how machine learning can reshape task scheduling for sustainable cloud infrastructure.

II. NEED AND SIGNIFICANCE OF THE STUDY

The need for sustainable cloud infrastructure has become increasingly urgent in the face of growing digital dependency and environmental pressure. Every year, more organizations migrate services to the cloud, increase data traffic, and rely on real-time applications that demand constant availability. While cloud computing improves accessibility and scalability, it also intensifies the pressure on data centers, which in turn consume more energy and generate higher carbon emissions. Sustainable computing is therefore no longer an optional concern; it is a strategic necessity.

Task scheduling is one of the most manageable and impactful points of intervention in cloud operations. Unlike hardware replacement or large-scale structural redesign, scheduling policies can be improved through software intelligence, adaptive algorithms, and predictive analytics. This makes machine learning-driven scheduling an especially attractive area for research. If scheduling mechanisms can be trained to assign tasks more intelligently, then cloud systems can reduce wasted energy, improve resource utilization, and deliver higher quality service without proportional increases in power consumption.

The significance of this study also lies in its interdisciplinary value. It brings together cloud computing, machine learning, operational efficiency, and environmental sustainability. It advances the idea that cloud optimization should not be judged only by speed or throughput, but also by how responsibly infrastructure consumes energy. The present article contributes to this shift by examining task scheduling as both a technical and sustainability-oriented problem.

III. OBJECTIVES OF THE ARTICLE

This research article is guided by the following objectives:

1. To examine the role of task scheduling in achieving energy efficiency in cloud infrastructure.
2. To analyze the limitations of traditional task scheduling techniques in dynamic cloud environments.
3. To explore the applicability of machine learning techniques for intelligent and adaptive task scheduling.
4. To propose an innovative conceptual framework for machine learning-driven sustainable cloud scheduling.
5. To assess how machine learning-based scheduling can improve energy usage, load balancing, QoS, and resource utilization.
6. To identify major challenges, research opportunities, and future directions in this domain.

IV. RESEARCH QUESTIONS

The article addresses the following exploratory questions:

1. Why are conventional task scheduling techniques insufficient for sustainable cloud infrastructure?
2. How can machine learning improve the intelligence and adaptability of scheduling decisions?
3. Which machine learning paradigms are most suitable for energy-aware task scheduling?
4. What are the practical benefits of machine learning-driven scheduling in terms of sustainability and system performance?
5. What challenges must be addressed for real-world deployment of such models in cloud data centers?

V. CONCEPTUAL BACKGROUND

5.1 Cloud Infrastructure and Scheduling:

Cloud infrastructure consists of interconnected physical servers, virtual machines, storage units, network devices, middleware systems, and application layers. The scheduler acts as a decision-making engine that determines where and when each task should execute. In virtualized environments, this becomes especially complex because a single physical host may run multiple virtual machines, each serving different users or applications. The scheduler must therefore consider not only processor availability but also memory usage, task urgency, host load, migration costs, bandwidth conditions, and energy implications.

5.2 Sustainability in Cloud Computing:

Sustainability in cloud computing refers to the ability to deliver digital services while minimizing energy waste, thermal burden, environmental damage, and unnecessary infrastructure strain. A sustainable cloud environment aims for optimal performance with minimal ecological cost. This involves reducing idle server power, consolidating workloads effectively, improving cooling efficiency, and applying intelligent resource management strategies. Since scheduling directly affects which resources remain active and how intensively they are used, it is strongly linked to sustainability outcomes.

5.3 Energy Efficiency as a Scheduling Goal

Energy efficiency in cloud systems does not mean sacrificing performance for power savings. Rather, it means achieving an intelligent balance where computational work is completed with the least possible resource waste. An energy-aware scheduler may consolidate tasks onto fewer active machines during low-demand periods, avoid overloading particular hosts, and distribute workloads in ways that reduce both execution delay and power draw. The challenge lies in managing these trade-offs without violating service-level agreements.

VI. LIMITATIONS OF TRADITIONAL TASK SCHEDULING METHODS

Traditional task scheduling techniques in cloud environments are usually based on deterministic or heuristic logic. Although these methods are computationally simple and easy to implement, they often perform poorly in highly dynamic and heterogeneous environments.

First Come First Serve is straightforward but ignores task characteristics and energy cost. It can lead to long waiting times and poor resource matching.

Round Robin ensures fairness but often neglects the actual resource demands of different tasks.

Min-Min and Max-Min attempt to optimize execution order, yet they remain limited by static assumptions and incomplete system awareness.

Priority-based scheduling can support urgent tasks but may cause starvation of lower-priority jobs and inefficient energy use.

Heuristic and metaheuristic methods such as Genetic Algorithms or Ant Colony Optimization improve search quality but may still depend heavily on manually designed fitness functions and may not learn continuously from system behavior.

The central weakness of these traditional techniques is that they are not truly adaptive. They do not learn patterns from historical execution data, cannot generalize effectively to unseen workload combinations, and do not improve their decision quality over time. In modern cloud systems, where workloads fluctuate rapidly and user expectations are high, this is a serious limitation. Sustainable infrastructure requires scheduling mechanisms that can anticipate future states, not just react to the current one.

VII. MACHINE LEARNING AS A TRANSFORMATIVE SCHEDULING PARADIGM

Machine learning transforms scheduling from a rule-following process into a data-informed decision-making system. Instead of assigning tasks solely on predefined logic, machine learning allows the scheduler to infer patterns, predict outcomes, and optimize actions based on past and present data. This is especially valuable in cloud systems, where workload intensity, task diversity, and resource status change continuously.

A machine learning-driven scheduler can answer critical operational questions such as:

- Which host is likely to execute this task with minimum energy cost?
- Which virtual machine is at risk of overload in the next time interval?
- When should a task be delayed, migrated, or redirected to avoid unnecessary power consumption?
- What combination of allocation decisions will produce the best long-term energy-performance balance?

By addressing such questions intelligently, machine learning supports a proactive scheduling model. It helps the cloud environment become self-adjusting, context-aware, and sustainable.

VIII. MACHINE LEARNING TECHNIQUES FOR TASK SCHEDULING

8.1 Supervised Learning

Supervised learning models can be trained on labeled historical data to predict execution time, power usage, response delay, or resource demand. Algorithms such as Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting, and Artificial Neural Networks can classify tasks and predict which host or VM is most suitable for a specific workload.

For example, a supervised learning scheduler may predict the expected energy consumption of assigning a task to different machines and then select the assignment with the lowest cost while maintaining acceptable performance. This is particularly useful in recurring workload scenarios where patterns can be learned reliably over time.

8.2 Unsupervised Learning

Unsupervised learning is useful when labeled data is unavailable or incomplete. Techniques such as K-means clustering, hierarchical clustering, and principal component analysis can group tasks with similar characteristics, identify workload categories, and discover hidden behavioral patterns. This can improve scheduling by enabling the system to treat similar tasks collectively and allocate them to optimized resource pools.

For instance, compute-intensive tasks, memory-heavy tasks, and latency-sensitive tasks can be clustered separately, allowing the scheduler to handle them using differentiated energy-aware strategies.

8.3 Reinforcement Learning

Reinforcement learning is one of the most promising approaches for intelligent scheduling in cloud systems. In this model, the scheduler acts as an agent that interacts with the cloud environment and learns which actions produce the best long-term rewards. Rewards may be defined in terms of lower energy consumption, reduced SLA violations, improved load balance, or better throughput.

Over time, the reinforcement learning agent learns scheduling policies that balance multiple objectives dynamically. This makes it suitable for real-time environments where future conditions are uncertain. Deep Reinforcement Learning further expands this capacity by combining reinforcement learning with neural networks, allowing the scheduler to handle complex state spaces and multi-dimensional optimization.

8.4 Deep Learning for Workload Prediction

Deep learning models such as Long Short-Term Memory networks and recurrent neural architectures are useful for predicting workload trends over time. These models can capture temporal dependencies in cloud traffic and help the scheduler anticipate future demand. When workload surges can be predicted in advance, tasks can be scheduled more efficiently, servers can be prepared proactively, and unnecessary power cycling can be reduced.

IX.PROPOSED INNOVATIVE FRAMEWORK: ML-DRIVEN SUSTAINABLE TASK SCHEDULING MODEL

This article proposes an exploratory framework for machine learning-driven sustainable task scheduling in cloud infrastructure. The framework is conceptual in nature but designed to be practical and extensible.

9.1 Input Layer

The system continuously collects data from the cloud environment, including CPU usage, memory demand, task length, task priority, VM state, host utilization, historical execution time, energy consumption patterns, and SLA metrics.

9.2 Data Processing and Feature Engineering

Raw cloud logs are cleaned and transformed into meaningful features. These may include average power draw per task category, estimated completion time, host thermal load, queue length, workload burst intensity, and migration cost indicators. Good feature engineering improves the intelligence of the scheduler and ensures that energy-related decisions are based on relevant information.

9.3 Prediction and Classification Engine

A supervised learning component predicts task execution cost and energy impact. An unsupervised component clusters similar tasks and identifies stable workload segments. A deep learning module forecasts short-term workload demand. Together, these sub-systems create a rich decision profile for every incoming task.

9.4 Decision and Scheduling Layer

A reinforcement learning agent uses the predictions and classifications to decide task placement. It selects the most suitable VM or host while balancing performance and sustainability goals. The reward function encourages energy-efficient execution, reduced response time, low SLA violation, and improved utilization.

9.5 Feedback and Continuous Learning

After each scheduling action, the system observes the actual outcomes and feeds them back into the model. This allows continuous improvement. The scheduler gradually becomes more accurate, more context-aware, and more effective in balancing energy with service quality.

X. EXPLORATORY ADVANTAGES OF THE PROPOSED APPROACH

10.1 Reduction in Energy Consumption

Machine learning can help identify underutilized resources and direct tasks toward more efficient execution paths. By predicting load behavior and minimizing unnecessary host activation, the scheduler can reduce power waste significantly.

10.2 Improved Resource Utilization

A well-trained scheduler distributes workloads more intelligently across virtual and physical resources. This prevents both underuse and overload. Better utilization means fewer idle servers and more effective use of active infrastructure.

10.3 Better SLA Compliance

Because machine learning models can estimate task completion patterns and host performance more accurately, they reduce the risk of poor placement decisions that lead to SLA violations. This improves trust and service quality for cloud users.

10.4 Adaptive Decision-Making

Unlike static methods, machine learning models can adapt to changing workload behaviors, newly emerging patterns, and evolving infrastructure conditions. This makes them suitable for large-scale and real-time cloud systems.

10.5 Support for Green Computing Goals

Sustainable cloud infrastructure requires operational intelligence. Machine learning-driven scheduling contributes to green computing by aligning system-level decisions with energy-conscious principles, not just raw computational output.

XI.KEY PERFORMANCE INDICATORS FOR EVALUATION

To assess the effectiveness of machine learning-driven scheduling, the following performance indicators are especially relevant:

- Energy Consumption: Total power used during task execution.
- Makespan: Total completion time for a set of scheduled tasks.
- Throughput: Number of tasks completed per time unit.
- Resource Utilization: Degree of effective use of CPU, memory, and storage.
- SLA Violations: Frequency of performance or service breaches.
- Response Time: Time taken to begin or complete task processing.
- Migration Overhead: Cost associated with moving tasks or VMs between hosts.
- Carbon-Aware Efficiency: Emerging metric combining energy usage with environmental impact.

A sustainable scheduler should optimize several of these simultaneously, not just one.

XII. CHALLENGES IN REAL-WORLD IMPLEMENTATION

12.1 Data Quality and Availability:

Machine learning models depend on good-quality data. In cloud systems, logs may be noisy, incomplete, inconsistent, or distributed across platforms. Poor data quality reduces the reliability of predictions.

12.2 Scalability

Large cloud environments generate huge volumes of data and involve thousands of scheduling decisions per second. A machine learning scheduler must scale efficiently without becoming a performance bottleneck itself.

12.3 Model Interpretability

In critical infrastructure, operators often hesitate to trust black-box models. If the scheduler makes unexpected decisions without understandable reasoning, it may face resistance in practical deployment.

12.4 Dynamic Workload Variability

Cloud workloads are highly volatile. Models trained on past data may lose relevance if workload characteristics shift substantially. This requires continual learning and retraining strategies.

12.5 Multi-Objective Conflict

Energy minimization, response time reduction, and SLA preservation may not always align. The scheduler must make intelligent trade-offs, which is challenging even for advanced learning systems.

12.6 Security and Privacy

Cloud logs and workload traces may contain sensitive operational information. Machine learning frameworks must be designed in ways that protect data confidentiality and system integrity.

XIII. DISCUSSION

The exploratory analysis presented in this article indicates that machine learning-driven task scheduling is a highly promising direction for sustainable cloud infrastructure. The strength of this approach lies not only in computational optimization but also in its ability to turn scheduling into a self-improving process. In traditional systems, the scheduler executes a fixed logic. In an intelligent cloud, the scheduler becomes an adaptive agent capable of learning, predicting, and refining its behavior continuously.

This shift is important because the future of cloud infrastructure will be shaped by complexity. Workloads will become more diverse, applications more latency-sensitive, users more demanding, and sustainability regulations more serious. Static or semi-dynamic scheduling approaches will become increasingly inadequate. Cloud providers will need systems that can understand workload context, anticipate demand, and manage infrastructure in a way that is both economically viable and environmentally responsible.

The article also suggests that sustainable scheduling should not be treated as a single-algorithm problem. Rather, it should be viewed as a layered intelligence system where prediction, classification, optimization, and feedback work together. The most effective future schedulers are likely to be hybrid systems that combine supervised learning, reinforcement learning, and deep workload forecasting within a unified decision framework.

At the same time, caution is necessary. Machine learning is not a magical solution. Its success depends on careful design, transparent metrics, continuous evaluation, and alignment with operational realities. The scheduler itself must remain lightweight enough to justify its energy savings. The goal is not to add intelligence for its own sake, but to produce measurable improvements in sustainability and service quality.

XIV.FUTURE RESEARCH DIRECTIONS

Several pathways remain open for future exploration in this area.

One important direction is the development of carbon-aware scheduling, where the scheduler considers not just electricity usage but also the carbon intensity of energy sources at different times and locations. Another promising direction is federated or distributed learning for multi-cloud or edge-cloud environments, where scheduling intelligence can be shared without exposing raw operational data. Further research is also needed on explainable AI for cloud scheduling, so that intelligent decisions become more transparent and trustworthy.

There is also strong potential in combining machine learning-driven scheduling with container orchestration platforms, serverless computing, and green software engineering practices. Similarly, integrating scheduling intelligence with real-time thermal management and renewable energy availability could significantly improve sustainable cloud design. These areas offer fertile ground for innovative doctoral-level and publication-oriented research.

XV. CONCLUSION

The present research article has explored the emerging role of machine learning-driven task scheduling in developing sustainable and energy-efficient cloud infrastructure. It has shown that task scheduling is far more than an operational mechanism; it is a strategic instrument through which cloud systems can improve performance, reduce energy waste, and move toward environmentally responsible computing. Traditional scheduling techniques, although useful in simpler settings, are increasingly inadequate for the dynamic and heterogeneous nature of modern cloud environments. Their inability to learn from data, anticipate workload shifts, and handle complex optimization trade-offs limits their value in sustainability-oriented cloud systems.

Machine learning introduces a more intelligent and adaptive paradigm. Through supervised learning, unsupervised learning, reinforcement learning, and deep predictive models, cloud schedulers can become proactive, responsive, and context-aware. They can make smarter placement decisions, improve resource utilization, reduce idle energy consumption, and support better SLA outcomes. The exploratory framework proposed in this article demonstrates how a multi-layered machine learning scheduler can combine prediction, classification, optimization, and feedback to support sustainable cloud operations.

The broader contribution of this article lies in reframing cloud scheduling as a sustainability challenge as well as a computational one. In the coming years, cloud infrastructure will need to be not only faster and larger, but also greener and more efficient. Machine learning-driven task scheduling offers a strong path forward in that transformation. With continued research, careful implementation, and real-world validation, it can become one of the foundational pillars of sustainable cloud computing.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50–58.

[2] Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data

centers for cloud computing. *Future Generation Computer Systems, 28*(5), 755–768.

[3] Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience, 24*(13), 1397–1420.

[4] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems, 25*(6), 599–616.

[5] Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A. F., & Buyya, R. (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience, 41*(1), 23–50.

[6] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.

[7] Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2016). Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (pp. 50–56).

[8] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

[9] Xu, J., & Fortes, J. A. B. (2010). Multi-objective virtual machine placement in virtualized data center environments. In *2010 IEEE/ACM International Conference on Green Computing and Communications* (pp. 179–188).

[10] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*..

[11] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50–58.

[12] Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems, 28*(5), 755–768.

[13] Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience, 24*(13), 1397–1420.

[14] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems, 25*(6), 599–616.

[15] Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A. F., & Buyya, R. (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience, 41*(1), 23–50.

[16] Ding, W., Fan, C., Song, X., & Liu, F. (2020). Adaptive virtual machine consolidation framework based on resource utilization and power-performance ratio in cloud data centers. *Future Generation Computer Systems, 109*, 589–602.

[17] Farahnakian, F., Ashraf, A., Pahikkala, T., Liljeberg, P., Plosila, J., Tenhunen, H., & Hänninen, H. (2015). Using ant colony system to consolidate VMs for green cloud computing. *IEEE Transactions on Services Computing, 8*(2), 187–198.

[18] Ghribi, C., Hadji, M., & Zeghlache, D. (2013). Energy efficient VM scheduling for cloud data centers using linear programming. In *2013 IEEE 5th International Conference on Cloud Computing Technology and Science* (pp. 1–8).

[19] Kansal, A., & Bahl, V. (2010). Joulemeter: Complete energy management for virtualized data centers. In *Proceedings of the 7th International Conference on Autonomic Computing* (pp. 175–184).

[20] Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2016). Resource management with deep reinforcement learning. In *Proceedings of*

the 15th ACM Workshop on Hot Topics in Networks (pp. 50–56).
[21] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.
[22] Sharma, O., Kanaujia, B. K., & Abraham, A. (2016). VM consolidation for cloud data center using median and standard deviation based threshold approach. *Procedia Computer Science, 85*, 610–616.
[23] Singh, J., Chana, I., & Buyya, R. (2020). Comparative analysis of VM consolidation algorithms for energy efficiency in cloud data centers. *Procedia Computer Science, 171*, 249–258.
[24] Svard, M., Tordsson, J., & Elmroth, E. (2011). A load balancing and energy efficiency approach for cloud computing environments. In *Proceedings of the 2011 IEEE International Conference on Cloud Computing Technology and Science* (pp. 1–8).
[25] Torre, E., Delimitrou, C., & Bianchini, R. (2020). A dynamic evolutionary multi-objective virtual machine placement approach for cloud data centers. *Information and Software Technology, 123*, 106310.
[26] Usmani, Z., & Singh, S. (2016). A survey of virtual machine placement techniques in cloud data centers. *Procedia Computer Science, 78*, 491–498.
[27] Verma, A., Ahuja, P., & Neogi, A. (2008). pMapper: Power and migration cost aware application placement in virtualized systems. In *Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware* (pp. 243–264).
[28] Wang, Z., Wang, Y., & Liu, X. (2023). Reinforcement learning based task scheduling for cloud computing. *The Journal of Supercomputing, 79*, 1–29.
[29] Xu, J., & Fortes, J. A. B. (2010). Multi-objective virtual machine placement in virtualized data center environments. In *2010 IEEE/ACM International Conference on Green Computing and Communications* (pp. 179–188).
[30] Yazdanov, K., & Fetzer, C. (2013). A decision tree and reinforcement learning approach for adaptive VM scaling in cloud environments. In *Proceedings of the 6th International Conference on Cloud Computing* (pp. 1–8).
[31] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing* (pp. 10–10).
[32] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications, 1*(1), 7–18.
[33] Zhang, Y., Wang, Y., & Qian, Z. (2011). Energy-aware resource provisioning in cloud computing environments. *IEEE Transactions on Cloud Computing, 1*(2), 109–118.
[34] Nandagopal, M., et al. (2025). Enhancing energy efficiency in cloud computing through task scheduling using hybrid intelligent optimization. *Journal of Intelligent & Fuzzy Systems*. Advance online publication.
[35] Yadav, M., et al. (2025). Energy-efficient workflow scheduling using dynamic task clustering in cloud environments. *Journal of Grid Computing*. Advance online publication.