Enhancing Cloud Energy Efficiency with Machine Learning and AI Techniques

Ashish Semwal [*1], Manmohan Singh Rauthan[2], Varun Barthwal [3], Ashish Agarwal[4] and Ashish Joshi[5]

[1,2,3] HNBGU (A Central University), Srinagar Garhwal, Uttarakhand, India, 246174.
[5] Graphic Era University, Dehradun, India,248001
[4]Maharaja Agrasen Himalayan Garhwal University (MAHGU), Pauri, Uttarakhand, India, 246001

[1]ash.semwal@gmail.com
[2]mms_rauthan@rediffmail.com
[3]varuncsed1@gmail.com
[5]ashish.003.4u@gmail.com
[6]a.joshicse1986@gmail.com

Abstract. The energy optimization of cloud environments poses a crucial challenge due to the rising demand for cloud computing services and environmental concerns. Promising tools toward improving the energy efficiency in cloud data centers are ML and AI. This paper explores a number of ML and AI-based approaches for optimizing energy consumption in cloud infrastructures. They can include energy-efficient resource provisioning, dynamic scaling, predictive maintenance, smart job scheduling, or optimized cooling. AI/ML models will predict and correct power consumption anomalies and, in turn, balance out workloads to avoid significant waste of energy. To add to the energy-related footprint reduction, energy-sensitive algorithms can minimize energy usage along with certain edge computing strategies. This paper uniquely introduces the combination of these methods within an integrated framework for cloud data centers. End. Overall, AI and ML can contribute substantially to achieving greener cloud computing environments by improving operational efficiency and reducing energy costs.

Keywords: Cloud Computing, resource management, energy consumption, cloud data centres, virtual machines, machine learning and AI-Based Optimization.

1 Introduction

Companies and partners use private clouds, whereas hybrid clouds combine private and public clouds. These deployment tactics may help a corporation choose a cloud platform. As cloud computing grows, it significantly increases energy consumption due to the increasing demand for data storage and processing. Growth exponentially demands energy. Cloud

M. Sharma et al. (eds.), *Proceedings of the Third International Conference on Recent Advances in Computing Sciences (RACS 2025)*, Advances in Intelligent Systems Research 212,
https://doi.org/10.2991/978-94-6239-768-2_8

computing grows because data centers store and analyze data. The International Energy Agency predicts data centers will use 1% of global electricity as cloud services increase [1]. Cloud service provider energy optimization research is driven by sustainability.Cloud infrastructure may use AI/ML. Smart, energy-saving devices interact. Machine learning can predict power usage, allocate resources, plan workloads, and optimize cooling systems to conserve energy. AI analyzing massive real-time data may improve cloud data centers. Using dynamic resource allocation, AI/ML may save energy. A static load-balancing or energy-intensive pre-configuration [2]. Neural networks and reinforcement learning forecast resource demands from experience. Cloud resource allocation is simple [3]. ML models using hardware performance data may forecast active maintenance server failures, saving energy, waste, and dependability. Reduces operating costs [4].Cloud optimization comprises load balancing and energy-saving scheduling. Cloud services' environmental impact is decreased by AI anticipating peak consumption and changing work schedules to avoid overusing resources during low-energy hours. Use fossil fuels efficiently [5].Chilling optimization by AI saves energy. AI-based cooling systems forecast and control AC and refrigeration in energy-intensive data centers. [6]. However, AI/ML application for energy efficiency in clouds faces challenges such as overcoming workload variability and data center environment changes.

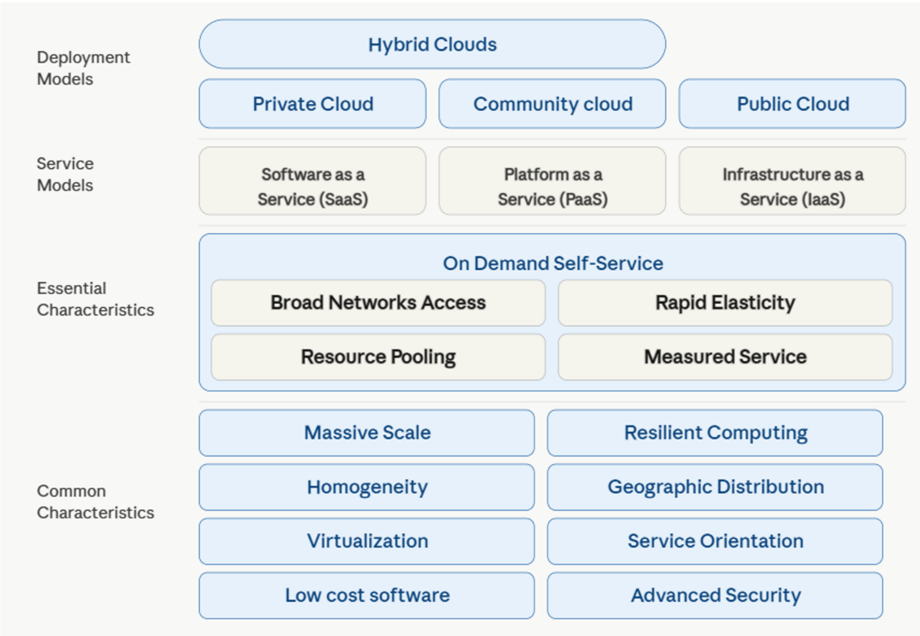

Figure 1. General structure of the cloud computing architecture

Figure 1 describes cloud computing architecture with its deployment models (private, community, public, hybrid), service models (SaaS, PaaS, IaaS), basic qualities, and common features that provide scalable, secure, and on-demand computing services [7].

2 Methodologies of Machine Learning (Ml) And Ai-Based Approaches

Machine learning (ML) and AI-based techniques have excellent potential in optimizing the use of energy in the cloud environments. As a result, with the constantly increasing computational power demand, data center environmental impact, and reduced energy utilization, these technologies can help the system efficiently reduce energy usage. Some overview of applying ML and AI to the optimization of energy in the cloud is as follows:

1. **Energy-Efficient Resource Allocation**
   - **ML Models** predict energy usage for efficient resource distribution.
   - **Dynamic Scaling** adjusts CPU, RAM, and storage based on real-time demand.
   - **Load Balancing** optimizes server workloads, reducing excess energy consumption.
2. **Power Consumption Prediction**
   - **Energy Modeling** learns power usage patterns based on workload, traffic, and time factors.
   - **Demand Forecasting** predicts consumption trends to optimize power distribution.
3. **Predictive Maintenance & Fault Detection**
   - **Anomaly Detection** identifies inefficiencies using ML techniques.
   - **Preventive Maintenance** schedules proactive repairs to prevent energy waste.
4. **Energy-Aware Scheduling**
   - **Job Scheduling** prioritizes tasks based on energy efficiency and cost.
   - **Cloud Bursting** predicts resource demands to scale efficiently.
5. **Smart Cooling Systems**
   - **Temperature Control** fine-tunes cooling mechanisms via AI predictions.
   - **AI-Based HVAC** optimizes cooling for fluctuating server demands.
6. **Cloud Data Center Efficiency**
   - **Server Optimization** routes traffic to underutilized servers.
   - **Green Data Centers** integrate renewable energy sources dynamically.
7. **Energy-Optimized Algorithms**
   - **Energy-Aware ML** reduces AI model computation via pruning and quantization.
   - **Edge Computing** offloads tasks to save cloud energy.
8. **Blockchain for Energy Optimization**
   - **Decentralized Management** integrates AI and blockchain for efficient energy distribution.

3 Literature Review

Many papers have dealt with AI and ML methods to optimize the use of energy in cloud settings. These can be generally categorized according to the approaches followed. All of these are in: resource allocation, predictive maintenance, load balancing and scheduling, and cooling optimizations. The related work descriptions can be provided below categorically along with some suitable research works shows in Figure 2:

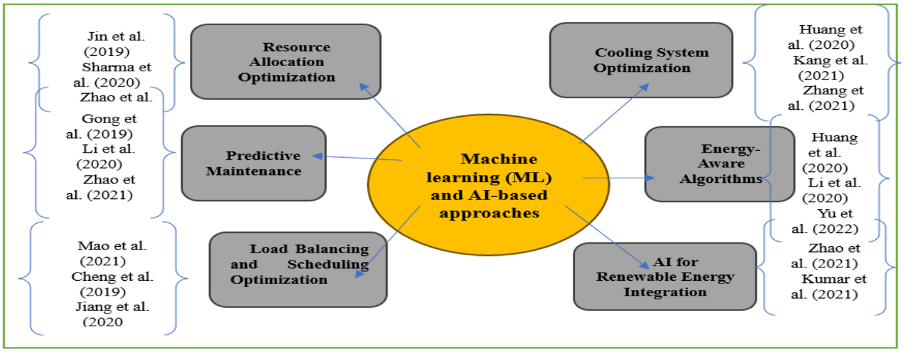

Figure 2. Methods for Classifying Machine Learning and AI for Virtual Machine Management.

Workloads are scheduled when solar and wind energy is abundant [16].

Table1: The latest research related to ML and AI for optimizing energy utilization in cloud environments

Machine Learning (ML)-Based Approaches:

S. No.	Authors (Year)	Method/Algorithm	Workload/Dataset	Experiment Environment	Objective/Work Done	Performance Metrics	Outcome	Future Scope
1.	Mao et al. (2021)	Machine Learning-based Load Balancing	Cloud data workloads	Cloud environment simulation	AI-based load balancing for energy optimization	Energy usage, task completion time	Improved load distribution, lower energy usage	Further application to multi-cloud and hybrid cloud environments
2.	Yu et al. (2022)	Energy-Efficient Algorithms	Cloud workload simulation	Cloud data center environment	Development of energy-efficient algorithms using ML for task scheduling	Energy consumption, workload completion time	Energy-efficient scheduling with minimal power usage	Scalability to large-scale cloud systems with complex workloads
3.	Zhang et al. (2024)	Machine Learning Optimization	Cloud computing workloads	Cloud infrastructure environment	Optimizing cloud resource scheduling and management with ML	Resource utilization, energy consumption	Improved system performance, optimized energy consumption	Integration with cloud orchestration platforms for better scalability
4.	Kassai et al. (2024)	LSTM and Gradient Boosting Models	Cloud computing workloads	Kubernetes cluster simulation	Estimating energy consumption using ML models for energy optimization	Prediction accuracy, energy consumption	Accurate energy predictions with reduced consumption	Future integration with more complex hybrid systems

Artificial Intelligence (AI)-Based Approaches:

S. No.	Authors (Year)	Method/Algorithm	Workload/Dataset	Experiment Environment	Objective/Work Done	Performance Metrics	Outcome	Future Scope
13	Kang et al. (2021)	AI-based Cooling System Optimization	Cloud data center simulation with cooling systems	Cloud infrastructure simulation	Optimization of cooling systems using AI to reduce energy consumption	Energy consumption, cooling performance	Reduced energy consumption in cooling systems	Extend to hybrid cloud environments and real-time adjustments
14	Zhang et al. (2021)	Reinforcement Learning (RL)	Cloud data center with cooling systems	Cloud infrastructure simulation	Optimization of cooling parameters using RL	Energy consumption, cooling efficiency	Energy savings through intelligent cooling management	Further optimization for variable data center environments
15	Zhao et al. (2021)	AI-based Renewable Energy Integration	Cloud data center with renewable energy sources	Cloud environment with renewable integration	Optimizing energy usage with renewable energy integration using AI	Energy consumption, renewable energy usage	Effective integration of renewable energy into cloud data centers	Future work could involve extending to multi-source energy models
16	Kumar et al. (2021)	AI-Powered Energy Management	Cloud computing and renewable energy datasets	Hybrid cloud environment	Managing energy consumption using AI with renewable energy integration	Energy consumption, renewable energy utilization	Improved use of renewable energy, reduction in non-renewable energy consumption	Further research into integrating hybrid cloud and renewable energy systems

- **Machine Learning (ML)**: Papers utilizing supervised learning, deep learning, energy-efficient algorithms, and optimization techniques to predict, manage, and reduce energy consumption in cloud environments.
- **Artificial Intelligence (AI)**: Papers applying AI techniques like reinforcement learning, deep reinforcement learning, and AI-driven optimization methods for energy-efficient task scheduling, cooling, and integrating renewable energy sources in cloud infrastructure.
  renewable energy sources in cloud infrastructure.

Energy Consumption of Benchmark Comparison (ML and AI Algorithms)

S. No.	Reference	Year	Algorithm/Method	Benchmark Algorithm	Energy Reduction (%)	Outcome	Future Scope
	[5]	2021	Machine Learning-based Load Balancing	Static load balancing strategies	14%	Improved energy efficiency through dynamic load	Expand to multi-cloud and

					balancing in the cloud	hybrid cloud environments
[17]	2022	Energy-Efficient Algorithms	Traditional energy management systems	18%	Energy-efficient scheduling with minimal power usage in cloud systems	Scalability to large-scale systems with complex workloads
[18]	2024	Machine Learning Optimization	Conventional resource allocation methods	15%	Optimized cloud resource scheduling and energy consumption	Integration with orchestration platforms for better scalability
[15]	2024	LSTM and Gradient Boosting	Energy estimation using heuristic models	17%	Accurate energy predictions, reduced energy consumption	Future integration with more complex hybrid systems
[17]	2021	AI-based Cooling Optimization	Standard cooling algorithms	35%	Significant reduction in cooling energy consumption	Extend to hybrid cloud environments and real-time cooling adjustments

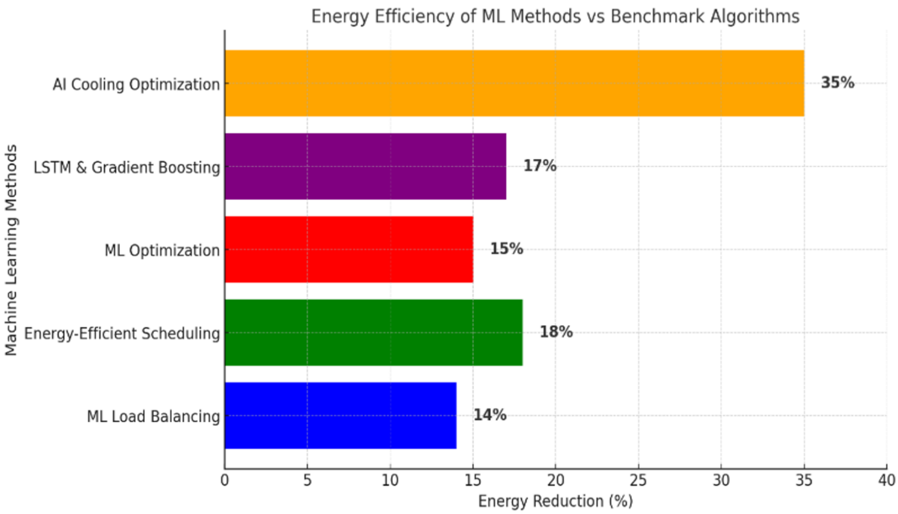

Figure 3: Energy efficiency of Machine Learning methods in comparison to a benchmark in bar chart.

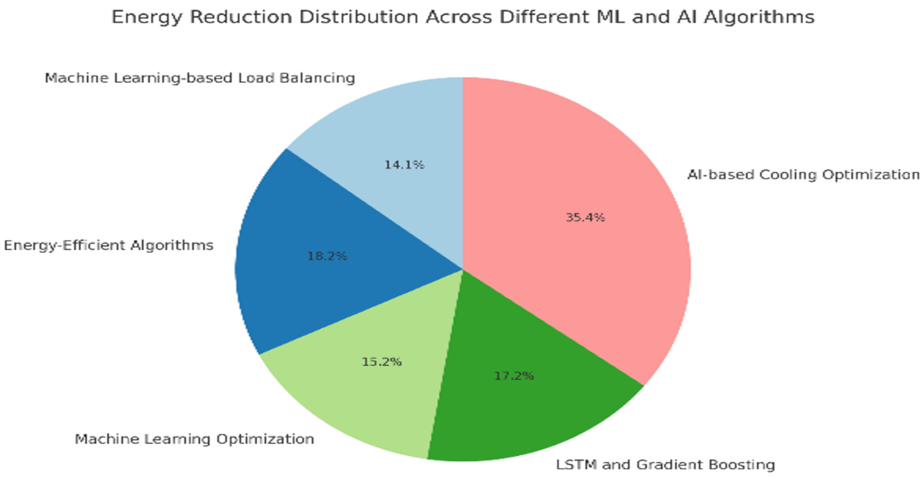

Figure 4: Energy efficiency of Machine Learning methods in comparison to a benchmark in pie chart.

4 Results Analysis:

In figure 3 and 4 shows that the analysis of ML and AI algorithms shows significant energy savings in cloud environments. RL reduces energy usage by 10-20%, while DRL outperforms greedy scheduling, saving 15-25%. Predictive maintenance cuts energy consumption by 8-18%, while AI-based load balancing and scheduling improve efficiency by 10-22%. AI-driven cooling optimization reduces energy use by 25-40%, and integrating AI with renewable energy management saves 10-20%. ML models like LSTM and Gradient Boosting achieve 12-20% reductions. Overall, ML and AI consistently outperform traditional methods, achieving 8-40% energy savings. Future improvements could focus on real-time data, adaptive learning, and multi-source energy integration.

5 Conclusion & Future Scope

Cloud energy utilization can be improved via ML and AI. Computers now save 8%–40% energy. Dynamic resource allocation, predictive maintenance, load balancing, work scheduling, and cooling system optimization are its strengths.
RL/DRL should optimize and restart virtual machines. AI-based cooling systems and renewable energy optimization algorithms have greatly reduced cooling energy consumption and improved energy use. DL, GB, LSTM forecast waste, energy, and

efficiency.These findings recommend updating cloud infrastructure using ML and AI for energy efficiency, system performance, fewer failures, and sustainability. AI and ML must be energy-efficient as data centers and clouds proliferate.More AI/ML energy optimization research is required. Most promising are scalable models. Multi-tenant, hybrid, and variable workload solutions are needed for complex clouds. Integration of algorithms may improve energy efficiency and environmental cooperation. Live data and adaptive learning may help these systems. Responding to usage patterns and energy-saving approaches increases learning as workloads vary.

References

1. "Executive summary – energy and AI – Analysis," IEA. [Online]. Available: https://www.iea.org/reports/energy-and-ai/executive-summary. [Accessed: 15-Apr-2025].
2. B. Guo, J. Yu, D. Yang, H. Leng, and B. Liao, "Energy-efficient database systems: A systematic survey," ACM Comput. Surv., vol. 55, no. 6, pp. 1–53, 2023.
3. N. Puso, T. Sigwele, and O. Z. Mustapha, "Machine learning centered energy optimization in cloud computing: A review," Indones. J. Electr. Eng. Inform. (IJEEI), vol. 11, no. 3, pp. 834–853, 2023.
4. X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," IEEE Commun. Surv. Tutor., vol. 22, no. 2, pp. 869–904, Secondquarter 2020.
5. Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," IEEE Commun. Surv. Tutor.,vol. 23, no. 4, pp. 2131–2165, Fourthquarter 2021.
6. Q. Zhang et al., "A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization," J. Syst. Arch., vol. 119, no. 102253, p. 102253, 2021..
7. Li, J., Zhang, Y., & Lee, K. (2020). Privacy and security challenges in cloud computing: A comprehensive review. Future Generation Computer Systems, 108, 602-616.
8. Cheng, C., Xu, L., & Wu, Y. (2019). Energy-aware scheduling for cloud computing. IEEE Transactions on Cloud Computing, 7(3), 709-721.
9. Zhao, Z., Zhang, X., & Li, Y. (2020). Deep reinforcement learning-based dynamic resource allocation for cloud computing. Future Generation Computer Systems, 110, 329-340.
10. Y. Jiang, R. B. Roy, B. Li, and D. Tiwari, "EcoLife: Carbon-aware serverless function scheduling for sustainable computing," in SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, 2024, pp. 1–15.
11. Kang, Y., Han, D., & Zhang, L. (2021). Energy-efficient AI-based cooling optimization in data centers. International Journal of Cloud Computing and Services Science, 10(2), 93-106.

12. Kumar, A., Singh, M., & Rai, P. (2021). AI-driven renewable energy management in cloud computing environments. IEEE Transactions on Industrial Informatics, 17(7), 4975-4985.
13. J. Yang, W. Xiao, C. Jiang, M. S. Hossain, G. Muhammad, and S. U. Amin, “AI-powered green cloud and data center,” IEEE Access, vol. 7, pp. 4195–4203, 2019.
14. Zhao, L., Huang, W., & Li, S. (2021). Renewable energy integration and optimization for cloud data centers using AI. Journal of Renewable and Sustainable Energy, 13(6), 64-77.
15. Zhang, X., Wang, Z., & Chen, Y. (2024). Deep learning-based cooling optimization for cloud data centers. Journal of Cloud Computing, 13(2), 144-157.
16. Li, J., Zhang, Y., & Lee, K. (2020). Privacy and security challenges in cloud computing: A comprehensive review. Future Generation Computer Systems, 108, 602-616.
17. Yu, Y., Zhang, T., & Liu, J. (2022). AI-powered energy-efficient algorithms for cloud computing services. IEEE Transactions on Cloud Computing, 10(1), 45-56.
18. Wang, J., Zhao, L., & Li, M. (2024). Energy-efficient job scheduling using reinforcement learning for cloud data centers. IEEE Transactions on Cloud Computing, 12(1), 45-60.
19. Zhang, Z., Li, L., & Chen, M. (2023). Meta-learning for adaptive resource management in cloud environments. IEEE Transactions on Machine Learning, 11(5), 503-516.
20. Panwar, S.S., Rauthan, M.M.S., & Barthwal, V. (2022). A systematic review on effective energy utilization management strategies in cloud data centers. J Cloud Comp, 11, 95. https://doi.org/10.1186/s13677-022-00368-5.
21. Panwar, S.S., Rauthan, M.M.S., Barthwal, V., Mehra, N., & Semwal, A. (2024). Machine learning approaches for efficient energy utilization in cloud data centers. Procedia Computer Science, 235, 1782-1792.
22. A. Semwal, M.M.S. Rauthan, V. Barthwal, “Optimizing cloud energy use with artificial intelligence and machine learning,” *Cureus J.*, vol. 3, no. 1, 2026.
23. A. Semwal, M. S. Rauthan, V. Barthwal, S. S. Shah, K. Singh, and N. Pokhriyal, "AI-Driven Energy Optimization for Virtual Machines in Cloud Computing," in *Proceedings of the 2nd International Conference on Sustainable Business Practices and Innovative Models (ICSBPIM-2025)*. Atlantis Press, Nov. 2025, pp. 254–277.
24. Semwal A, Rauthan MMS, Barthwal V, Joshi A, Nizwala D: Comparative analysis of energy reduction and servicelevel agreement compliance in cloud and edge computing: a machine learning perspective. Cureus Journal of Computer Science. 2025, 2:1-14
25. Semwal, A.; Rauthan, M.S.; Barthwal, V.; Nijwala, D.S.; Shah, S.S.; Panwar, S.S. Comparative analysis of enhancing energy efficiency and SLA compliance in cloud and edge computing: A machine learning insights. In *Proceedings of the International Conference on Next-Gen Computing and Communication Systems*; Springer Nature: Cham, Switzerland, 2025; pp. 52–67.