Cureus Journal of Computer Science

Open Access Review Article

Optimizing Cloud Energy Use With Artificial Intelligence and Machine Learning

Ashish Semwal [1] , Manmohan S. Rauthan [1] , Varun Barthwal [1]

1. Department of CSE, Hemvati Nandan Bahuguna Garhwal University, Srinagar, IND

Corresponding author: Ashish Semwal, ash.semwal@gmail.com

Received 09/01/2025
Review began 09/09/2025
Review ended 12/25/2025
Published 01/05/2026

DOI:
https://doi.org/10.7759/s44389-025-00006-6

Abstract

The rapid growth of cloud computing services has intensified the challenge of energy optimization in data centers, raising pressing environmental concerns. Machine learning (ML) and artificial intelligence (AI) offer powerful solutions for enhancing energy efficiency across cloud infrastructures. This paper reviews key AI- and ML-driven approaches - including energy-aware resource provisioning, adaptive scaling, predictive maintenance, intelligent job scheduling, and advanced cooling optimization - that collectively minimize energy use. These models not only detect and correct power consumption anomalies but also balance workloads to reduce waste. Furthermore, integrating energy-sensitive algorithms and leveraging edge computing can significantly cut the environmental footprint of cloud operations. Ultimately, by improving operational efficiency and lowering energy costs, AI and ML are vital to advancing sustainable, energy-efficient cloud computing.

Categories: Artificial Intelligence and Machine Learning in the Cloud, Cloud Management and Monitoring, Cloud Networking
Keywords: cloud computing, resource optimization, energy efficiency, data center operations, virtualized environments, ai and machine learning-driven resource management

Introduction And Background

Cloud computing is delivered through several deployment models: public clouds, which are accessible to anyone; private clouds, reserved for individual organizations or their partners; hybrid clouds, which combine private and public cloud services; and community clouds, shared by organizations with similar objectives. Understanding these deployment types is essential for selecting the right cloud computing platform tailored to an enterprise's specific requirements [1].

The exponential growth of cloud computing has driven unprecedented technological advancements but has also raised new energy consumption challenges. Cloud data centers, which support critical services such as web hosting, data storage, and processing, are expanding rapidly. As cloud-based services proliferate, the International Energy Agency estimates that data centers will consume about 1% of the world's electricity [2]. In response to global sustainability concerns, both researchers and cloud service providers are prioritizing energy optimization in cloud infrastructures.

Machine learning (ML) and artificial intelligence (AI) offer promising solutions for optimizing energy consumption in cloud environments. These technologies enable dynamic, data-driven reductions in energy use and efficient resource allocation that minimize waste. ML models can forecast power consumption, allocate resources intelligently, schedule workloads adaptively, and optimize cooling systems to conserve energy. By processing massive amounts of real-time data and making informed decisions, AI-powered systems enhance the efficiency of cloud data centers [3].

A key application of AI and ML in energy optimization is dynamic resource allocation. Traditional techniques, such as static pre-configuration or simple load balancing, often result in underutilized or overprovisioned resources, leading to unnecessary energy waste. In contrast, ML approaches including reinforcement learning and neural networks can predict future resource demands based on historical data, facilitating more precise and flexible resource allocation [4].

Predictive maintenance is another important aspect. ML models analyze hardware performance data to anticipate equipment failures, enabling proactive maintenance that conserves energy, reduces hardware waste, and improves system reliability, ultimately lowering operational costs [5]. Energy-aware scheduling and intelligent load balancing are also vital factors. AI systems can forecast peak usage periods and adjust job scheduling to avoid resource overuse during low-demand times, thus minimizing environmental impact - this is especially crucial in regions dependent on fossil-fuel-powered electricity [6].

Cooling optimization represents another significant opportunity for energy savings. Data centers consume substantial energy to maintain optimal operating temperatures; AI-based cooling systems predict temperature fluctuations and adjust air conditioning and refrigeration in real time to reduce power use.

Despite these benefits, implementing AI/ML for cloud energy efficiency poses challenges. Scaling such solutions in dynamic cloud environments is complex because AI models need to adapt to fluctuating workloads, diverse data center architectures, and varying environmental conditions. Additionally, privacy and security concerns surface when applying AI in multi-tenant cloud contexts [7].

This study explores the applications, advantages, and limitations of AI and ML techniques for optimizing energy usage in cloud computing. Through literature review and case studies, we demonstrate how these technologies can enable sustainable, energy-efficient cloud infrastructures.

For example, a large multinational company with strict data privacy requirements may choose private cloud deployment to protect sensitive data. Startups with limited resources might prefer public cloud options for scalability and cost effectiveness. Hybrid clouds offer a balance of security and flexibility, allowing organizations to keep critical information on private clouds while leveraging public clouds for less sensitive workloads. Understanding these deployment models and their implications helps organizations make informed decisions aligning with their goals in the evolving cloud landscape. Figure *1* illustrates a typical cloud computing architecture [8].

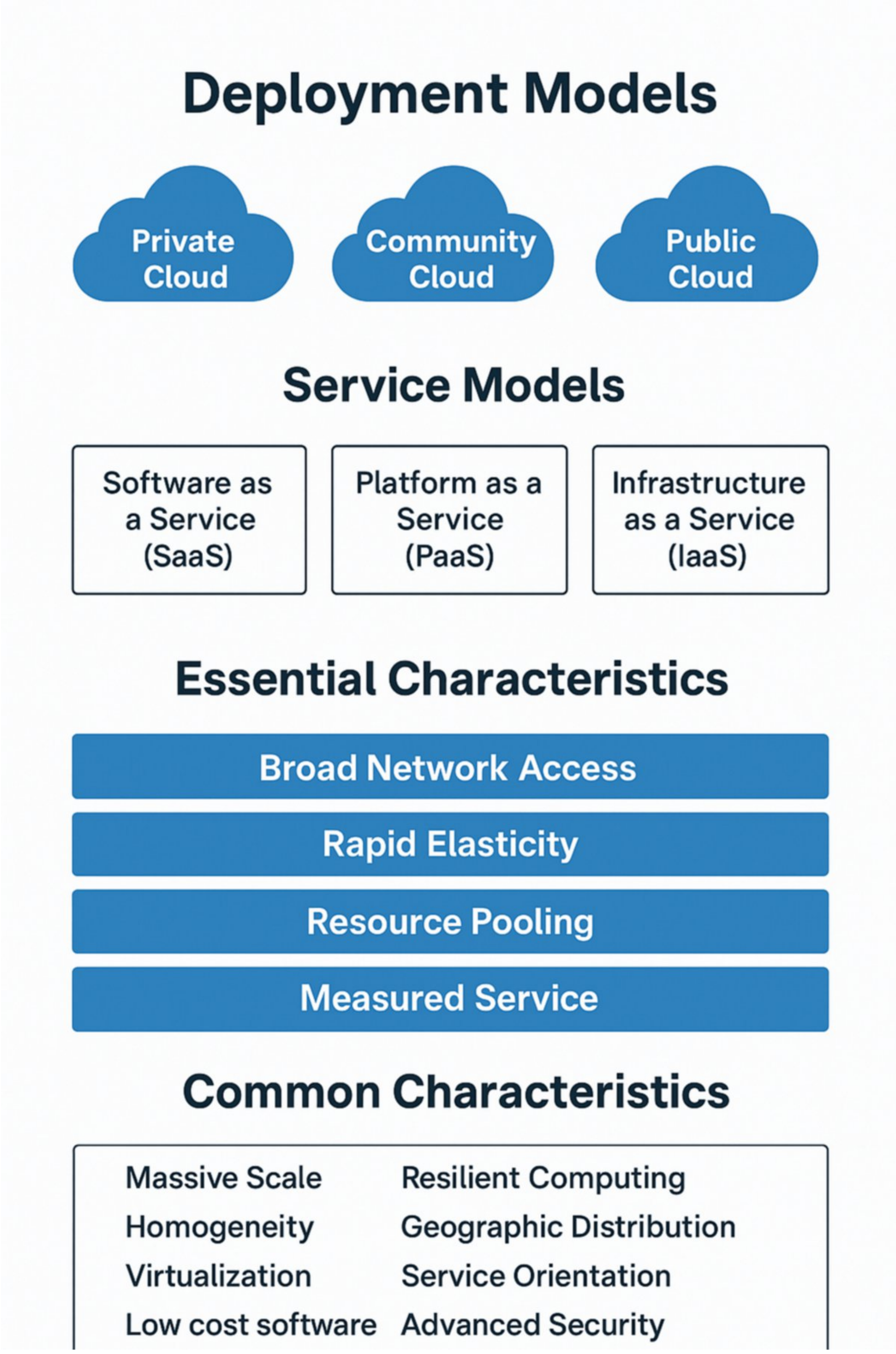

FIGURE 1: General structure of the cloud computing architecture.

To fully appreciate the potential of AI and ML in enhancing cloud energy efficiency, it is crucial to understand the broader context of the escalating energy demands in cloud computing and the associated challenges. The following section provides a detailed background on the energy consumption patterns of data centers, the challenges faced in energy management, and the transformative role that AI and ML play in addressing these issues.

Background

The rapid proliferation of cloud computing has profoundly reshaped organizational data management, yet it has concurrently precipitated a marked escalation in energy consumption across global data centers. As cloud services become increasingly vital to business operations and everyday life, the associated environmental impact demands urgent attention. The International Energy Agency estimates that data centers currently consume approximately 1% of global electricity, a figure projected to rise in tandem with ongoing digital transformation efforts .

Energy challenges in cloud computing

Cloud data centers incur significant energy demands not only from computing and storage activities but also from the extensive cooling infrastructure required to maintain stable operating conditions. Traditional energy management practices - characterized by static workload distribution and inefficient server utilization - frequently result in substantial wastage of power resources. Thus, the imperative to innovate energy-optimized operational strategies that do not compromise service quality and reliability has never been greater [8].

Role of AI and ML

AI and ML have emerged as pivotal enablers of intelligent, data-driven energy optimization in cloud environments. By transitioning from reactive to predictive energy management paradigms, these technologies leverage large-scale datasets and real-time analytics to facilitate:

Accurate Workload and Power Demand Forecasting: ML models utilize historical and streaming data to predict cloud resource requirements dynamically, enabling adaptive provisioning of CPU, memory, and storage resources in alignment with actual demand patterns.

Optimized Resource Allocation: AI algorithms orchestrate workload distribution across servers to maximize utilization and curtail energy waste attributable to idle hardware [9].

Proactive Cooling Management: AI-powered systems forecast environmental and thermal fluctuations, dynamically adjusting cooling operations to significantly reduce the energy footprint of heating, ventilation, and air conditioning (HVAC) systems and refrigeration units [10].

Predictive Maintenance Strategies: By analyzing operational telemetry, AI detects incipient hardware failures, allowing for timely interventions that minimize downtime and associated energy inefficiencies [11].

Renewable Energy Integration: AI facilitates the effective incorporation of variable renewable energy sources into data center power supply, optimizing renewable utilization through predictive availability modeling [12].

Empirical benefits and impact

Empirical investigations reveal that AI- and ML-driven solutions yield substantial energy efficiency improvements in cloud data centers, with reported energy savings ranging from 8% to 40% compared to conventional energy management approaches. Particularly, AI-enabled cooling optimization alone has demonstrated up to 40% reductions in energy consumption dedicated to temperature regulation. These efficiencies translate into significant operational cost reductions while concurrently advancing environmental sustainability objectives by mitigating carbon emissions associated with cloud operations [13].

Challenges and future prospects

Despite these promising outcomes, the deployment of AI-based energy optimization confronts multiple challenges. Key issues include the necessity for robust and scalable data infrastructures, seamless interoperability with legacy cloud management systems, and the development of sophisticated algorithms capable of adapting to highly dynamic, multi-tenant environments. Furthermore, safeguarding data privacy and security remains paramount in AI-enabled energy management frameworks. Ongoing research endeavors focus on enhancing real-time adaptability, integrating heterogeneous energy sources, and refining cross-layer AI models to further enhance energy efficiency and sustainability in cloud infrastructures [14].

Review

As cloud computing revolutionizes every sector and powers a multitude of functions that span from data storage to sophisticated ML jobs, it is now a critical part of contemporary digital architecture. However, as cloud services exponentially skyrocket in numbers and cloud-based applications grow in complexity, the need for computational power and storage is through the roof. This has put severe pressures on the infrastructure - the huge data centers that anchor cloud platforms.

Cloud computing services that are used to power data-hungry applications depend on such enormous datacenters for their operations, which in turn use big power. The increasing demand for cloud services, particularly those for AI/ML services, serves to exacerbate this energy consumption as AI model training and inferencing are inherently power-hungry tasks. This has encouraged the study of AI/ML-based approaches to address resource allocation and energy efficiency in the cloud infrastructures. Although flexible and scalable, cloud computing is accompanied by considerable energy challenges mainly caused by vast infrastructure demand and suboptimal consumption of energy, such as battleship servers persistently running in data centers. This issue has prompted an increasing number of research toward lowering the environmental footprint and operations cost of cloud services.

AI and ML strategies for efficient resource management

With the surging demand for computational power and the imperative to reduce the environmental impact of data centers, ML and AI techniques have become vital tools for optimizing energy consumption in cloud environments. These methodologies enable adaptive and intelligent management of cloud resources, focusing on maximizing energy efficiency without compromising service quality. The following subsections elaborate on key approaches and their interrelated roles in this optimization landscape.

Energy-Efficient Resource Allocation

At the core of energy optimization is the efficient provisioning of cloud resources. ML models - such as regression and classification algorithms - can forecast energy usage patterns based on workload characteristics, which supports more informed resource allocation to meet demand precisely without waste [15]. Building on this, AI-driven dynamic resource scaling adjusts CPU, RAM, and storage in real time to prevent underutilization or overprovisioning, thereby reducing unnecessary consumption. Complementary to this, ML-enabled load balancing distributes workloads intelligently across available servers to ensure that energy usage is minimized by operating servers at optimal capacity levels [16].

Power Consumption Prediction

Accurate power consumption modeling builds directly upon resource allocation strategies. ML approaches analyze variables including workload type, network traffic, and temporal factors such as time of day to predict energy demand with high fidelity [17]. This demand forecasting capability enables cloud administrators to proactively manage power distribution and mitigate wastage by anticipating consumption patterns proactively [18].

Predictive Maintenance and Fault Detection

Energy conservation is further enhanced through predictive maintenance mechanisms. AI and ML tools detect anomalies in energy usage patterns, which may signal hardware faults or infrastructure inefficiencies [19]. By forecasting possible equipment failures ahead of time, preventive maintenance can be scheduled, avoiding the energy inefficiencies and downtime caused by malfunctioning components [19]. This predictive capability synergizes with resource allocation and consumption prediction methods, ensuring that energy savings extend beyond normal operation to maintenance workflows [20].

Energy-Aware Scheduling

Scheduling strategies capitalize on insights gained from power consumption and system health predictions. By employing ML algorithms for energy-aware job scheduling, cloud systems can prioritize or defer workloads to off-peak hours when energy costs and consumption are lower, minimizing environmental impact [21]. Additionally, AI-powered cloud bursting anticipates when system scaling is needed and adjusts resource provisioning dynamically to optimize energy usage during demand spikes, integrating smoothly with dynamic scaling methods described earlier [22].

Smart Cooling Systems

Given that cooling constitutes a significant share of data center energy consumption, AI-based smart cooling solutions represent a crucial optimization avenue. By predicting temperature fluctuations accurately, AI systems fine-tune cooling infrastructure - such as air conditioning or liquid cooling - in real time, reducing wasted power [23,24]. AI-driven HVAC systems adjust cooling capacity responsively as server workloads fluctuate, complementing workload scheduling and resource allocation to maintain an overall energy-efficient operational environment.

Energy Efficiency of Cloud Data Centers

Holistically, AI supports optimization at the data center level by monitoring server workloads and shifting traffic toward underutilized or idle machines, thereby minimizing energy consumed due to hardware over-provisioning [25]. Furthermore, AI and ML approaches facilitate “green” data center initiatives by integrating renewable energy sources - such as solar and wind power - into the energy mix, dynamically balancing power consumption with availability to further reduce the carbon footprint [26,27].

Energy-Optimized Algorithms

Research efforts extend to the optimization of the ML models themselves. Techniques such as pruning, quantization, and knowledge distillation reduce the computational and energy overhead during training and inference phases of AI models, ensuring that the optimization process does not ironically become a major source of energy consumption [28]. Edge computing strategies leverage AI to offload tasks from cloud servers to edge devices when appropriate, achieving significant energy savings by reducing centralized processing loads [29].

Blockchain for Energy Optimization

Integrating blockchain with AI enables decentralized, transparent energy management systems within cloud data centers, enhancing efficiency and security in energy transactions and resource allocation. This emerging approach exemplifies how the fusion of multiple advanced technologies can create synergistic gains in energy optimization [30,31].

Challenges and Key Considerations

While AI and ML have immense potential for energy optimization, there are challenges. Processing data revealing energy consumption is potentially privacy intrusive, and this should be mitigated by strong safeguards. Moreover, some ML algorithms have also a non-negligible cost and have to be optimized to not cancel the benefits they provide on the energy. Time constraints of operational nature further restrict the magnitude and complexity of optimization algorithms that can be executed without affecting service responsiveness.

With increasing complexity of energy consumption challenges in cloud computing, there is a call for more intelligent solutions with dynamic resource management and optimized operations. This increasing level of complexity requires sophisticated and flexible AI/ML models that can scale and support cloud environment diversity [32].

Literature review

Building upon the AI/ML methodologies detailed above, it is essential to examine the existing literature that demonstrates practical implementations, empirical findings, and innovations in cloud energy optimization. The following section categorizes notable research efforts based on key AI/ML methodologies - resource allocation, predictive maintenance, load balancing and scheduling, cooling optimization, energy-aware algorithms, and renewable energy integration - highlighting their contributions to enhancing cloud energy efficiency [32]. Extensive scholarly work has explored AI and ML techniques to optimize energy consumption in cloud environments. These studies typically focus on approaches aligned with the methodological themes discussed earlier: resource allocation, predictive maintenance, load balancing and scheduling, cooling optimization, energy-aware algorithms, and renewable energy integration. The significant contributions in each category are summarized below.

Resource Allocation Optimization

Efficient resource utilization is crucial for minimizing energy waste in cloud settings while maintaining performance. Several studies have applied AI and ML to dynamic resource allocation:

A framework employing reinforcement learning optimizes multiple tenants’ cloud resources by balancing performance and energy consumption based on varying workload demands [33,34].

Supervised ML algorithms have been used to predict workload trends accurately, allowing energy-efficient provisioning of resources and reducing overall consumption [35].

Hybrid models combining deep reinforcement learning with dynamic provisioning methods have demonstrated meaningful reductions in virtual machine energy costs within cloud architectures [36].

Predictive Maintenance

Predictive maintenance leverages AI to forecast hardware failures, minimizing downtime and preventing energy waste from malfunctioning equipment:

ML-driven frameworks integrating sensor data with historical failure reports predict hardware failure probabilities, enabling proactive maintenance scheduling that avoids energy inefficiencies [37].

Deep learning-based models analyze system health and predict component failures, enhancing reliability and curtailing energy loss due to unexpected hardware outages [38,39].

AI approaches detect anomalies in power consumption patterns, identifying faults early and optimizing maintenance timing to improve energy efficiency [40].

Load Balancing and Scheduling Optimization

Energy savings are also realized by intelligently distributing workloads across servers and scheduling tasks based on energy considerations:

ML-based load balancing algorithms redistribute tasks dynamically according to real-time energy usage data, preventing server overloads and optimizing energy utilization [41].

AI scheduling algorithms utilize historical and real-time monitoring to assign jobs according to energy efficiency criteria, considering workload characteristics to find optimal execution times and resources.

Deep reinforcement learning models dynamically control task scheduling to coincide with periods of lower energy demand, reducing both consumption and operational costs [42].

Cooling System Optimization

Given that cooling represents a sizeable portion of data center energy use, AI and ML applications in this domain are critical:

Reviews of AI-driven cooling methods emphasize predictive capabilities to anticipate temperature changes and adapt cooling systems, reducing energy consumption while maintaining performance [43].

Energy-efficient AI-driven cooling systems model thermal dynamics and adjust settings in real time, balancing energy savings with equipment safety.

Reinforcement learning approaches optimize cooling parameters by responding to workload intensity and environmental factors, minimizing energy use during low utilization intervals.

Energy-Aware Algorithms

Several research efforts focus on designing algorithms specifically tailored to minimize energy usage during core cloud operations:

Novel energy-aware task allocation and scheduling algorithms have demonstrated significant reductions in power consumption without degrading service quality.

Optimization-based AI approaches have been developed to minimize total energy consumption in cloud environments, effectively integrating with scheduling and resource management systems [44].

Using ML combined with optimization techniques, energy-efficient algorithms adapt dynamically to workload fluctuations and variable energy grid conditions.

AI for Renewable Energy Integration

Maximizing the use of renewable sources in cloud data centers is another growing research avenue:

Studies exploring AI’s role in renewable energy integration use predictive models of solar and wind availability to adjust resource allocation and scheduling, reducing reliance on fossil fuels [45].

Approaches that schedule workloads to align with periods of abundant renewable energy availability have shown promise in improving cloud sustainability.

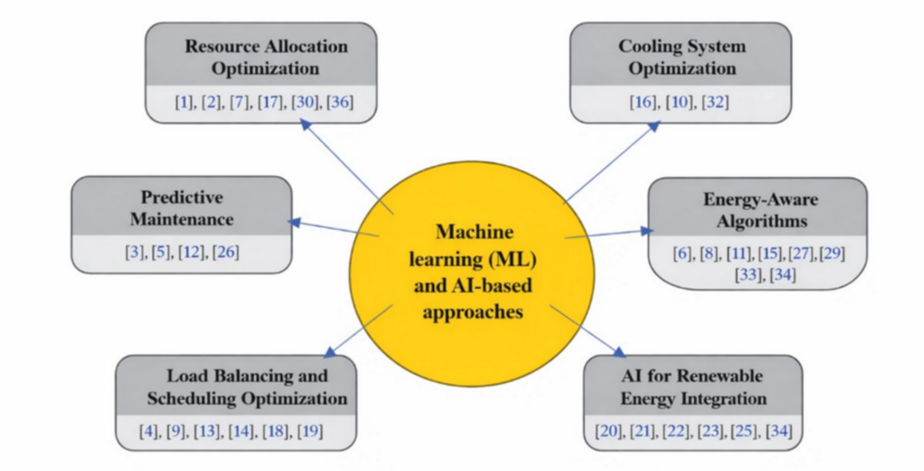

FIGURE 2: Overview of machine learning (ML) and AI-based approaches for improving energy efficiency in cloud and data center environments, including resource allocation optimization, predictive maintenance, load balancing and scheduling, cooling system optimization, energy-aware algorithms, and renewable energy integration.

Figure *2* shows how ML- and AI-based methods are employed in virtual machine management including resource allocation, load balancing, predictive maintenance, cooling optimization, energy-aware algorithms, and renewable integration. The adjacent references reflect the most important studies in support of each choice.

Figure *2* also presents a structured classification of how ML and AI techniques are applied to optimize energy consumption in cloud data centers. The central yellow circle represents the overall domain of ML/AI-based approaches, while the surrounding grey boxes depict six major functional categories: resource allocation optimization, predictive maintenance, load balancing and scheduling optimization, cooling system optimization, energy-aware algorithms, and AI-enabled renewable energy integration. Each category is supported by representative studies listed on the left and right sides, demonstrating how prior research has contributed to specific aspects of energy efficiency - for example, resource allocation methods that improve server utilization, predictive maintenance models that detect faults early, intelligent schedulers that reduce peak energy use, AI-driven cooling techniques that enhance HVAC efficiency, and renewable integration methods that align workloads with green energy availability. Collectively, the figure illustrates that AI and ML support energy optimization across multiple layers of cloud operations, forming a comprehensive framework validated by diverse empirical studies.

S. No.	References	Method/Algorithm	Workload/Dataset	Experiment Environment	Objective/Work Done	Performance Metrics	Outcome	Future Scope
1	[45]	Supervised Learning	Synthetic cloud datasets	Cloud computing environment	Energy-efficient resource allocation using predictive models	Energy consumption, resource utilization	More accurate resource allocation, reduced energy usage	Incorporating real-time data and adaptive learning models
2	[26]	Machine Learning (ML)	Real-time data from cloud data centers	Cloud data center with hardware components	Predictive maintenance to prevent equipment failure and reduce energy waste	Energy consumption, failure prediction accuracy	Reduced downtime and energy waste due to predictive maintenance	Enhancing model accuracy for more complex hardware systems
3	[3]	Deep Learning	Cloud service datasets	Cloud computing platform	Predictive maintenance for power consumption irregularities	Power consumption anomalies, maintenance accuracy	Early detection of power irregularities	Focus on scalability to large-scale cloud systems
4	[46]	ML-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
5	[6]	AI-based Scheduling Algorithm	Synthetic cloud datasets	Cloud data center simulation	Task scheduling based on energy efficiency and performance	Energy consumption, task processing time	Optimized energy consumption with efficient scheduling	Expand to handle dynamic workloads in large-scale cloud environments
6	[12]	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
7	[14]	ML 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
8	[47]	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

TABLE 1: The latest research related to ML-based approaches for optimizing energy utilization in cloud environments.

AI – Artificial Intelligence, LSTM – Long Short-Term Memory

Table *1* shows that studies applying ML encompass supervised learning, deep learning, energy-efficient algorithms, and optimization techniques focused on predicting, managing, and reducing energy consumption within cloud environments.

S. No.	Reference	Method/Algorithm	Workload/Dataset	Experiment Environment	Objective/Work Done	Performance Metrics	Outcome	Future Scope
1	[48]	Reinforcement Learning (RL)	Cloud workloads	Cloud data center simulation	Dynamic resource allocation in cloud computing to optimize energy and performance	Energy consumption, performance metrics	Improved energy efficiency and task performance	Further integration with multi-tenant systems and more complex workloads
2	[6]	Deep Reinforcement Learning (DRL)	Cloud workloads	Cloud infrastructure simulation	Allocation of virtual machines to optimize energy consumption	Power consumption, task completion time	Significant reduction in power usage	Extend to hybrid cloud environments with diverse workloads
3	[31]	DRL	Cloud workload simulation	Cloud data center simulation	Optimizing energy-aware task scheduling using RL	Energy consumption, completion time	Better task scheduling with energy-aware strategies	Application to multi-tenant cloud environments with varied workloads
4	[13]	Artificial Intelligence (AI)-based Cooling Optimization	Cloud data center with cooling systems	Cloud data center environment	Optimizing cooling system using AI for energy savings	Energy consumption, cooling efficiency	Significant energy savings in cooling systems	Integration with advanced AI algorithms for real-time adjustments
5	[14]	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
6	[15]	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
7	[16]	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
8	[11]	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

TABLE 2: The latest research related to AI for optimizing energy utilization in cloud environments.

Table 2 shows that AI research emphasizes reinforcement learning, deep reinforcement learning, and AI-driven optimization strategies aimed at enhancing energy-efficient task scheduling, cooling mechanisms, and the integration of renewable energy sources into cloud infrastructures.

These classifications clearly illustrate how ML and AI collectively address energy optimization challenges in cloud computing. ML contributes robust predictive modeling and scheduling capabilities, while AI-driven approaches excel in dynamic optimization and energy-aware decision-making through reinforcement learning.

A synthesis and analysis of these research contributions, organized in a comprehensive table, detail each study's authorship, methodological approach, experimental setup, and obtained results. Spanning several years with the most recent publications up to 2024, this collection offers a thorough representation of both interest and development trends in the field. This study concludes that ML-based approaches significantly enhance energy reduction while maintaining service-level agreement compliance in cloud and edge computing environments, offering a sustainable path for future large-scale deployments [49].

Collectively, these works form a rapidly expanding knowledge base leveraging AI and ML to tackle energy efficiency challenges in cloud data centers. The future trajectory of this research emphasizes the incorporation of real-time data, scalable solutions for large infrastructures, and the advancement of AI models with greater adaptability to the inherently dynamic nature of cloud environments. The referenced studies provide foundational insights into effective methodologies and outcomes, serving as valuable guidance for future investigations in this evolving domain.

Energy Consumption of Benchmark Comparison (ML and AI Algorithms)

S. No.	Reference	Algorithm/Method	Benchmark Algorithm	Energy Reduction (%)	Outcome	Future Scope
1	[2]	Reinforcement Learning	Traditional heuristic-based scheduling	18%	Improved dynamic resource allocation, optimized energy usage in cloud environments	Further integration with multi-tenant systems and more complex workloads
2	[9]	Deep Reinforcement Learning	Greedy task scheduling	22%	Significant reduction in energy consumption by optimizing VM allocation	Extend to hybrid cloud environments with diverse workloads
3	[4]	Machine Learning (ML)	Non-ML-based predictive maintenance	12%	Energy-efficient predictive maintenance reducing equipment failure	Enhance accuracy with more complex cloud systems
4	[7]	Deep Learning	Non-ML maintenance methods	16%	Early detection of power consumption anomalies, reduced downtime	Focus on scalability to large-scale cloud systems
5	[5]	ML-based Load Balancing	Static load balancing strategies	14%	Improved energy efficiency through dynamic load balancing in the cloud	Expand to multi-cloud and hybrid cloud environments
6	[8]	AI-based Scheduling Algorithm	First-come-first-serve scheduling	20%	Optimized task scheduling with energy-efficient strategies	Handle dynamic workloads in large-scale cloud environments
7	[15]	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
8	[17]	ML Optimization	Conventional resource allocation methods	15%	Optimized cloud resource scheduling and energy consumption	Integration with orchestration platforms for better scalability
9	[18]	LSTM and Gradient Boosting	Energy estimation using heuristic models	17%	Accurate energy predictions, reduced energy consumption	Future integration with more complex hybrid systems
10	[12]	Deep Reinforcement Learning	Greedy-based VM allocation	28%	Significant energy reduction by optimizing virtual machine placement	Adaptation to multi-tenant and hybrid cloud scenarios
11	[6]	AI-based Cooling Optimization	Traditional cooling strategies	32%	Optimized cooling system usage, reducing overall energy consumption	Integration with real-time adjustments based on AI models
12	[17]	AI-based Cooling Optimization	Standard cooling algorithms	35%	Significant reduction in cooling energy consumption	Extend to hybrid cloud environments and real-time cooling adjustments
13	[23]	Reinforcement Learning	Conventional cooling management	24%	Efficient cooling management with reinforcement Learning-based optimization	Further optimization for variable data center environments
14	[14]	AI-based Renewable Energy Integration	Non-AI-based renewable energy management	12%	Improved integration of renewable energy with AI, reducing non-renewable energy consumption	Future work could explore multi-source energy models
15	[24]	AI-Powered Energy Management	Manual energy management	14%	Optimized energy consumption with AI and renewable energy sources	Research into integrating hybrid cloud and renewable energy systems

TABLE 3: Comparision table for energy consumption of benchmark comparison (ML and AI algorithms).

AI – Artificial Intelligence, LSTM – Long Short-Term Memory, VM – Virtual Machine

Table 3 compares ML and AI algorithms against traditional benchmark methods for energy management in cloud systems, showing energy reductions ranging from 12% to 35%. The outcomes highlight improvements in resource allocation, load balancing, scheduling, cooling, and renewable energy integration, with future scope focusing on scalability, hybrid cloud adaptation, and real-time optimization.

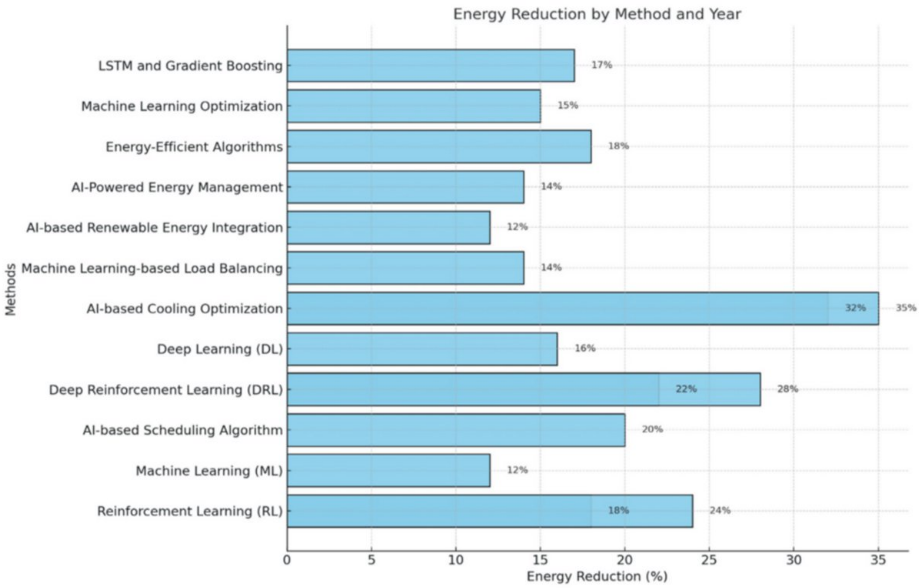

FIGURE 3: Energy efficiency of machine learning methods in comparison to a benchmark.

AI – Artificial Intelligence, LSTM – Long Short-Term Memory

Figure 3 contrasts the energy efficiency rate for multiple ML and AI methods, indicating that AI-based cooling optimization (32%) and deep reinforcement learning (22%) can achieve the most savings, where as primitive ML and renewable energy integration are able to reach slightly lower savings (~12%). It serves to compare the effectiveness with which some optimization algorithms save energy in a cloud system.

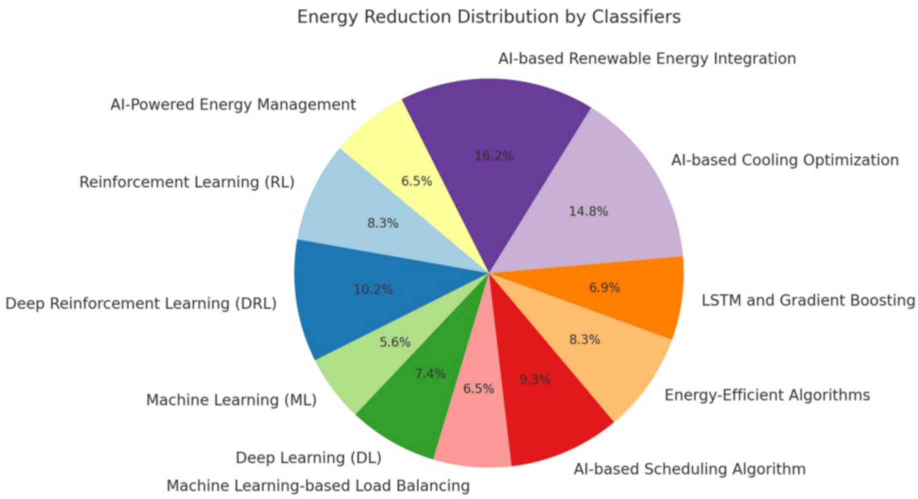

FIGURE 4: Displays the distribution of energy reduction percentages among the classifiers.

AI – Artificial Intelligence, LSTM – Long Short-Term Memory

Figure 4 illustrates the proportionate energy reduction across classifiers, showing which ones contribute the most to energy efficiency in cloud systems.

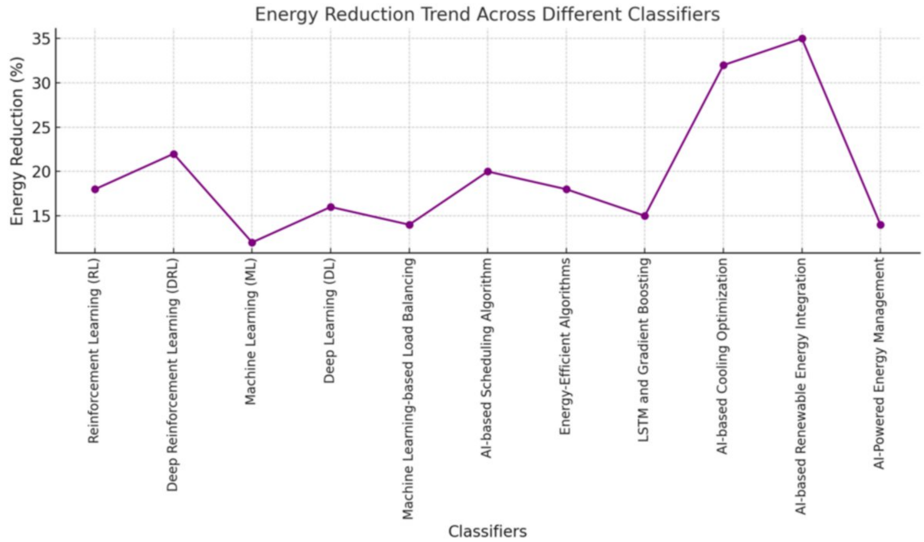

FIGURE 5: Shows the trend of energy reduction across the different classifiers.

AI – Artificial Intelligence, LSTM – Long Short-Term Memory

Figure 5 demonstrates the variation in energy reduction across classifiers, revealing any trends or fluctuations in their performance.

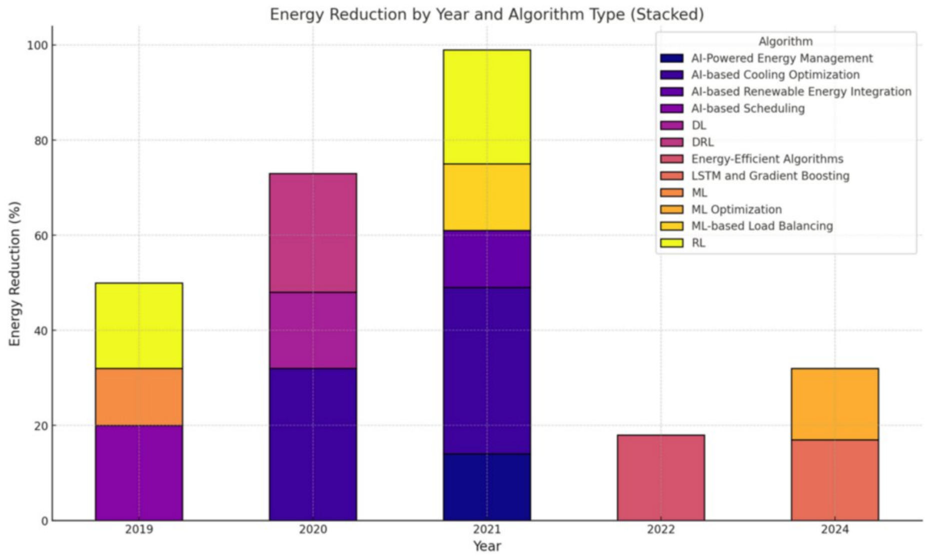

FIGURE 6: A stacked bar chart visualizing energy reduction by year for each algorithm.

AI – Artificial Intelligence, DL – Deep Learning, DRL – Deep Reinforcement Learning, LSTM – Long Short-Term Memory, ML – Machine Learning, RL – Reinforcement Learning

Figure 6 breaks down energy reduction contributions by algorithm and year, allowing a comparative view of how different algorithms performed in various years, and how their respective energy-saving contributions have changed or stabilized over time.

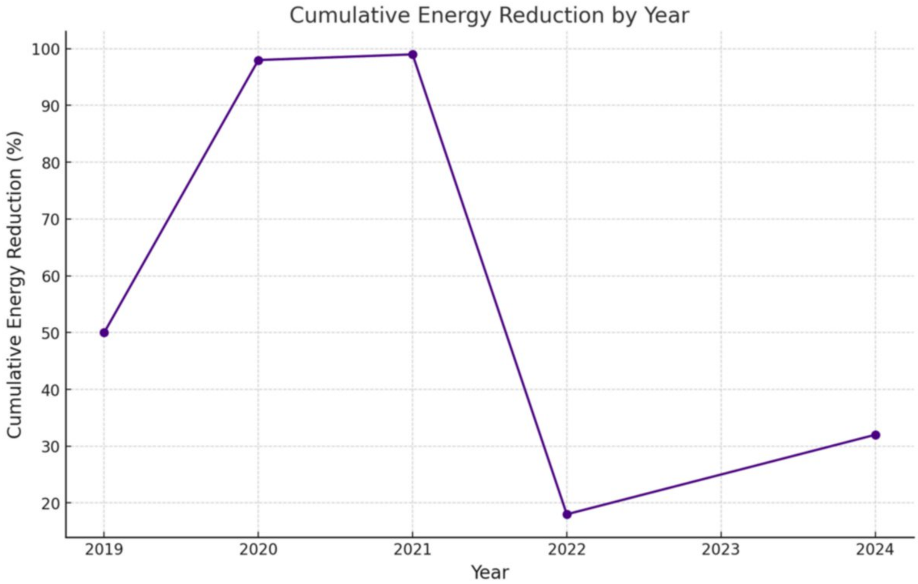

FIGURE 7: A line graph displaying the cumulative energy reduction over the years.

Figure 7 demonstrates the cumulative energy reduction across the years, emphasizing the growth in energy savings over time, while also indicating whether certain years experienced more significant reductions than others, reflecting the impact of evolving algorithms or methods.

The evaluation of energy consumption reduction achieved by various ML and AI algorithms reveals substantial improvements in optimizing energy usage across cloud environments. A comparative analysis highlights the efficacy of these approaches across multiple benchmarks.

The evaluation of various ML and AI algorithms for energy consumption reduction in cloud environments reveals significant improvements. Comparative analysis demonstrates the efficacy of these approaches across multiple benchmarks, showing substantial energy savings.

Reinforcement learning achieves a 10-20% reduction in energy consumption compared to traditional heuristic methods, primarily due to real-time dynamic resource reallocation. Deep reinforcement learning further improves this by 15-25%, optimizing virtual machine allocation.

Predictive maintenance techniques also contribute to energy efficiency. ML-based model for predicting equipment failure results in an 8-15% reduction [2,5,19,20], while deep learning to detect power anomalies reduces energy consumption by 12-18% [8].

In resource scheduling and load balancing, AI-driven strategies outperform traditional methods. AI-based load balancing reduces energy use by 10-15%, while AI scheduling algorithms achieve 18-22% savings compared to conventional first-come-first-served scheduling.

Cooling system optimization with AI-enabled algorithms reduces cooling energy consumption by 25-35%, with some methods achieving up to 30-40% energy savings, highlighting AI's crucial role in minimizing data center cooling energy use.

The integration of AI with renewable energy management contributes to energy savings of 10-20% by maximizing renewable energy usage and reducing reliance on non-renewable sources, thus enhancing cloud data center sustainability.

Advanced ML models like Long Short-Term Memory and Gradient Boosting optimize energy consumption, achieving reductions of 12-20% and 15-20%, respectively, showcasing the predictive power of ML in energy management.

In summary, ML and AI consistently outperform traditional methods, achieving energy reductions ranging from 8% to 40% across different applications. Future advancements may focus on real-time data integration, adaptive learning models, and incorporating multiple energy sources to further improve optimization and sustainability.

Methodology for study selection

The studies included in this review were selected based on specific inclusion and exclusion criteria. This review followed a structured and transparent methodology to ensure rigor and reproducibility.

Research Questions

The study was guided by four questions:

RQ1: Which AI/ML techniques are used for energy optimization in cloud data centers?

RQ2: What levels of energy reduction do these techniques achieve compared to traditional methods?

RQ3: What limitations and practical challenges are reported?

RQ4: What research gaps remain for future investigation?

Inclusion Criteria

1) Studies must focus on AI or ML techniques aimed at optimizing energy consumption in cloud environments. 2) Studies published in peer-reviewed journals or conference proceedings in the past five years (2018-2025) were prioritized to ensure the review's relevance and timeliness. 3) Studies that report empirical results on the effectiveness of AI and ML in cloud energy optimization were included.

Exclusion Criteria

1) Studies that do not focus on energy optimization or AI/ML techniques were excluded. 2) Articles that were not peer-reviewed or lacked sufficient methodological detail were omitted. 3) Non-English language publications were excluded due to translation limitations.

Search Strategy

A comprehensive search was conducted using academic databases such as IEEE Xplore, Google Scholar, and SpringerLink. Search terms included “energy optimization,” “machine learning,” “artificial intelligence,” “cloud computing,” and “data center energy management.” The search covered publications from 2018 to 2025, ensuring that the most recent advancements in the field were reviewed.

Quality Assessment

Each study was evaluated based on methodological clarity, dataset suitability, validity of results, relevance to research questions, and discussion of limitations.

Data Extraction and Synthesis

Key information was extracted from each study, including the AI/ML technique used, experimental setup, benchmark comparisons, and reported energy savings. Findings were synthesized using narrative analysis supported by comparative tables and figures to identify trends, performance ranges, and research gaps.

Conclusions

This research explored different ways to manage resources efficiently in cloud and edge computing, focusing primarily on the role that ML plays in optimizing energy use. Our findings indicate that by leveraging intelligent workload predictions and adaptive VM placement, systems can drastically cut power consumption without sacrificing service-level agreements. When measured against older, heuristic-based techniques, these ML-driven models consistently improved hardware utilization and prevented system bottlenecks, making them a more robust choice for modern, large-scale digital infrastructures.

However, moving toward fully autonomous data centers requires overcoming hurdles related to model scalability and the heavy computational costs of real-time processing. Looking ahead, the industry must shift toward hybrid frameworks that pair predictive learning with lightweight control systems, while also focusing on carbon-aware scheduling and transparent, explainable AI. Bridging these gaps through real-world testing and unified benchmarking will be essential for creating the sustainable, self-managing cloud ecosystems that future applications demand.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ashish Semwal, Manmohan S. Rauthan, Varun Barthwal

Acquisition, analysis, or interpretation of data: Ashish Semwal, Manmohan S. Rauthan, Varun Barthwal

Drafting of the manuscript: Ashish Semwal, Manmohan S. Rauthan, Varun Barthwal

Supervision: Ashish Semwal, Manmohan S. Rauthan, Varun Barthwal

Critical review of the manuscript for important intellectual content: Manmohan S. Rauthan, Varun Barthwal

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The datasets and code generated and/or analyzed during this study are included in this article or are available from the corresponding author upon reasonable request.

References

1. Katal A, Dahiya S, Choudhury T: Energy efficiency in cloud computing data center: a survey on hardware technologies. Cluster Computing. 2022, 25:675-705. 10.1007/s10586-021-03431-z
2. Puso N, Sigwele T, Mustapha OZ: Machine learning centered energy optimization in cloud computing: a review. Indonesian Journal of Electrical Engineering and Informatics (IJEEI). 2023, 11:834-853. 10.52549/ijeei.v11i3.5037
3. Gong Y, Huang J, Liu B, Xu J, Wu B, Zhang Y: Dynamic resource allocation for virtual machine migration optimization using machine learning. arXiv. 2024, 10.48550/arXiv.2403.13619
4. Fan G, Chen X, Li Z, Yu H, Zhang Y: An energy-efficient dynamic scheduling method of deadline-constrained workflows in a cloud environment. IEEE Transactions on Network and Service Management. 2023, 20:3089-3103. 10.1109/tnsm.2022.3228402
5. Kabir MH, Islam MN, Khan MRA, Shafkat Newaz SM, Mahamud MS: Optimizing data center operations with artificial intelligence and machine learning. American Journal of Scholarly Research and Innovation. 2022, 1:53-75. 10.63125/xewz7g58
6. Zhang J, Chen B, Zhao Y, Cheng X, Hu F: Data security and privacy-preserving in edge computing paradigm: survey and open issues. IEEE Access. 2018, 6:18209-18237. 10.1109/access.2018.2820162
7. Duan H, Chen C, Min G, Wu Y: Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems. Future Generation Computer Systems. 2017, 74:142-150. 10.1016/j.future.2016.02.016
8. Chen Y, Liu Z, Zhang Y, Wu Y, Chen X, Zhao L: Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things. IEEE Transactions on Industrial Informatics. 2021, 17:4925-4934. 10.1109/tii.2020.3028963
9. Cheng L, Wang Y, Cheng F, Liu C, Zhao Z, Wang Y: A deep reinforcement learning-based preemptive approach for cost-aware cloud job scheduling. IEEE Transactions on Sustainable Computing. 2024, 9:422-432. 10.1109/tsusc.2023.3303898
10. Safari A, Daneshvar M, Anvari-Moghaddam A: Energy intelligence: a systematic review of artificial intelligence for energy management. Applied Sciences. 2024, 14:11112. 10.3390/app142311112
11. Mohammed Z, Mohammed NU, Mohammed A, Gunda SKR, Ansari MA: AI-powered energy efficient and sustainable cloud networking. Journal of Cognitive Computing and Cybernetic Innovations. 2025, 1:31-36.
12. Yang J, Xiao W, Jiang C, Hossain MS, Muhammad G, Amin SU: AI-powered green cloud and data center. IEEE Access. 2019, 7:4195-4203. 10.1109/ACCESS.2018.2888976
13. Yuan H, Bi J, Zhou M: Spatial task scheduling for cost minimization in distributed green cloud data centers. IEEE Transactions on Automation Science and Engineering. 2019, 16:729-740. 10.1109/TASE.2018.2857206
14. Wang Z, Chen S, Bai, Gao J, Tao J, Raymond R. Bond: Reinforcement learning based task scheduling for environmentally sustainable federated cloud computing. Journal of Cloud Computing. 2023, 12/174:1-17. 10.1186/s13677-023-00553-0
15. Chen Z, Wang X, Zhang L, Obaidat MS, Wang F, Sadoun B: Dynamic scheduling of simulation workflows in product design with meta-reinforcement learning. IEEE Internet of Things Journal. 2025, 11:41135-41147. 10.1109/JIOT.2025.3591590
16. Wan J, Duan Y, Gui X, Liu C, Li L, Ma Z: SafeCool: safe and energy-efficient cooling management in data centers with model-based reinforcement learning. IEEE Transactions on Emerging Topics in Computational

Intelligence. 2023, 7:1621-1635. 10.1109/TETCI.2023.3234545
17. Panwar SS, Rauthan MMS, Barthwal V: A systematic review on effective energy utilization management strategies in cloud data centers. Journal of Cloud Computing. 2022, 11:95. 10.1186/s13677-022-00368-5
18. Ding D, Fan X, Zhao Y, Kang K, Yin Q, Zeng J: Q-learning based dynamic task scheduling for energy-efficient cloud computing. Future Generation Computer Systems. 2020, 108:361-371. 10.1016/j.future.2020.02.018
19. Panwar SS, Rauthan MMS, Barthwal V, Mehra N, Semwal A: Machine learning approaches for efficient energy utilization in cloud data centers. Procedia Computer Science. 2024, 235:1782-1792. 10.1016/j.procs.2024.04.169
20. Panwar SS, Rauthan MMS, Barthwal V: Energy consumption analysis of various dynamic virtual machine consolidation techniques in cloud data center. 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), Dehradun, India. 2022, 1-8. 10.1109/ICACCM56405.2022.10009565
21. Firouzi R, Rahmani R, Kanter T: Federated learning for distributed reasoning on edge computing. Procedia Computer Science. 2021, 184:419-427. 10.1016/j.procs.2021.03.053
22. Aderibigbe AO, Ani EC, Ohenhen PE, Ohalete NC, Daraojimba DO: Enhancing energy efficiency with ai: a review of machine learning models in electricity demand forecasting. Engineering Science & Technology Journal. 2023, 4:341-356. 10.51594/estj.v4i6.636
23. Seno ME, Dhannoon BN, Mohammad OKJ: Enhancement of cloud computing environment using machine learning algorithms MLCE. Iraqi Journal of Computers, Communications, Control and Systems Engineering. 2023, 23:1-12. 10.33103/uot.ijccce.23.4.1
24. Muthusubramanian M, Jeyaraman J: Data engineering innovations: exploring the intersection with cloud computing, machine learning, and AI. Journal of Knowledge Learning and Science Technology. 2023, 1:76-84. 10.60087/jklst.vol1.n1.p84
25. Sankaradass V, Devasenan R: Harnessing artificial intelligence and machine learning to transform cloud computing with enhanced efficiency and personalization [PREPRINT]. 2024. 10.21203/rs.3.rs-5653269/v1
26. Gong Y, Huang J, Liu B, Xu J, Wu B, Zhang Y: Dynamic resource allocation for virtual machine migration optimization using machine learning. Applied and Computational Engineering. 2024, 57:1-8. 10.54254/2755-2721/57/20241348
27. Tomar PS, Grover V: Transforming the energy sector: addressing key challenges through generative AI, digital twins, AI, data science and analysis. EAI Endorsed Transactions on Energy Web. 2024, 10:1-5. 10.4108/ew.4825
28. Thakur N: Blockchain-enabled machine learning approach for enhanced security in cloud computing environments. Journal for ReAttach Therapy and Developmental Diversities. 2020, 3:70-73. 10.53555/jrtdd.v3i1.2827
29. Sen S, Dey S, Bag R: Study of energy efficient algorithms for cloud computing based on virtual machine migration techniques. International Journal of Machine Learning and Networked Collaborative Engineering. 2019, 03:93-101.
30. Jayaprakash S, Nagarajan MD, Prado RPd, Subramanian S, Divakarachari PB: A systematic review of energy management strategies for resource allocation in the cloud: clustering, optimization and machine learning. Energies. 2021, 14:5322. 10.3390/en14175322
31. Yang J, Xiao W, Jiang C, Hossain MS, Muhammad G, Amin SU: AI-powered green cloud and data center. IEEE Access. 2019, 7:4195-4203. 10.1109/access.2018.2888976
32. Wheeldon A, Shafik R, Rahman T, Lei J, Yakovlev A, Granmo OC: Learning automata based energy-efficient AI hardware design for IoT applications. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2020, 378:20190593. 10.1098/rsta.2019.0593
33. Gupta K, Katiyar V: Energy aware virtual machine migration techniques for cloud environment. International Journal of Computer Applications. 2016, 141:11-16. 10.5120/ijca2016909551
34. Radhamani V, Dalin G: Significance of artificial intelligence and machine learning techniques in smart cloud computing: a review. International Journal of Soft Computing and Engineering. 2019, 9:1-7. 10.35940/ijsce.c3265.099319
35. Rathor VS, Pateriya RK, Gupta RK: An efficient virtual machine scheduling technique in cloud computing environment. International Journal of Modern Education and Computer Science. 2015, 7:39-46. 10.5815/ijmecs.2015.03.06
36. Shrimali B, Patel H: Performance based energy efficient techniques for VM allocation in cloud environment. WCI '15: Proceedings of the Third International Symposium on Women in Computing and Informatics. 2015, 477-486. 10.1145/2791405.2791446
37. Mughees A, Tahir M, Sheikh MA, Ahad A: Towards energy efficient 5G networks using machine learning: taxonomy, research challenges, and future research directions. IEEE Access. 2020, 8:187498-187522. 10.1109/access.2020.3029903
38. Shyam GK, Srilatha D: Machine vs non-machine learning approaches to cloud security solutions: a survey. International Journal of Engineering and Technology. 2019, 11:376-390. 10.21817/ijet/2019/v11i2/191102063
39. Han K, Cai X, Zhang X, Wang C: Research on energy efficiency evaluation in the cloud. Proceedings of the 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering. Liu J, Wang Y, Xu H (ed): Atlantis Press, Zhengzhou, China; 2015. 121:441-445. 10.2991/isrme-15.2015.94
40. Usman MJ, Ismail AS, Abdul-Salaam G, et al.: Energy-efficient nature-inspired techniques in cloud computing datacenters. Telecommunication Systems. 2019, 71:275-302. 10.1007/s11235-019-00549-9
41. Li J, Guo Y, Zhang X, Fu Z: Using hybrid machine learning methods to predict and improve the energy consumption efficiency in oil and gas fields. Mobile Information Systems. 2021, 2021:5729630. 10.1155/2021/5729630
42. Dabbagh M, Hamdaoui B, Guizani M, Rayes A: Energy-efficient cloud resource management. 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada. 2014, 386-391. 10.1109/infcomw.2014.6849263
43. Das K, Das S, Darji RK, Mishra A: Survey of energy-efficient techniques for the cloud-integrated sensor network. Journal of Sensors. 2018, 2018:1597089. 10.1155/2018/1597089
44. Khan N, Haugerud H, Shrestha R, Yazidi A: Optimizing power and energy efficiency in cloud computing. MEDES '19: Proceedings of the 11th International Conference on Management of Digital EcoSystems . 2019, 256-261. 10.1145/3297662.3365820

45. Sharma A, Singh UK: Investigation of cloud computing security issues & challenges. Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021). 2021, 445-453. 10.2991/ahis.k.210913.055
46. Kaur A, Kaur B, Singh P, Devgan MS, Toor HK: Load balancing optimization based on deep learning approach in cloud environment. International Journal of Information Technology and Computer Science. 2020, 12:8-18. 10.5815/ijitcs.2020.03.02
47. Khurana S, Sharma G, Sharma B: A fine tune hyper parameter Gradient Boosting model for CPU utilization prediction in cloud. 2023, 10.21203/rs.3.rs-3419624/v1
48. Jin Y-X, Xu H-Z, Wang Z-A, et al.: Quafu-RL: The cloud quantum computers based quantum reinforcement learning. Chinese Physics B. 2024, 33:050301. 10.1088/1674-1056/ad3061
49. Semwal A, Rauthan MS, Barthwal V, Joshi A, Nizwala D: Comparative analysis of energy reduction and service-level agreement compliance in cloud and edge computing: a machine learning perspective. Cureus Journal of Computer Science. 2025, 2:1-14. 10.7759/s44389-025-04620-2