Cureus Journal of Computer Science

Open Access Review Article

Received 06/04/2025
Review began 06/23/2025
Review ended 08/21/2025
Published 09/01/2025

DOI:
https://doi.org/10.7759/s44389-025-04620-2

Comparative Analysis of Energy Reduction and Service-Level Agreement Compliance in Cloud and Edge Computing: A Machine Learning Perspective

Ashish Semwal [1] , Manmohan S. Rauthan [1] , Varun Barthwal [1] , Ashish Joshi [2] , Deepak Nizwala [3]

1. Department of Computer Science and Engineering, Hemvati Nandan Bahuguna Garhwal University, Srinagar, IND 2. Department of Computer Science and Engineering, Graphic Era University, Dehradun, IND 3. Department of Computer Science and Engineering, Uttarakhand Technical University, Dehradun, IND

Corresponding authors: Ashish Semwal, ash.semwal@gmail.com, Varun Barthwal, varuncsed1@gmail.com

Abstract

As cloud and edge computing continue to evolve, the demand for optimizing energy consumption while ensuring stringent Service-Level Agreement (SLA) compliance has intensified. This paper provides a comparative evaluation of energy reduction and SLA compliance performance across multiple machine learning classifiers employed in both cloud and edge computing ecosystems. The study contrasts the effectiveness of traditional and AI-driven models, such as Support Vector Machines, Neural Networks, Decision Trees, and Random Forests, in managing energy consumption and SLA adherence. By utilizing energy reduction percentages, SLA compliance rates, and execution time as key metrics, we present an in-depth analysis through a series of advanced visualizations, including stacked bar graphs, radar charts, bubble plots, and box plots. The findings reveal key insights into the trade-offs between energy efficiency and SLA compliance across classifiers, with AI models - particularly Neural Networks - demonstrating superior performance in achieving optimal energy savings while ensuring SLA adherence. This research offers critical implications for resource optimization strategies in cloud and edge computing frameworks, advancing the field of sustainable computing and adaptive resource management in next-generation distributed systems. The study contributes to sustainable computing by offering empirical evidence and a comparative analysis of machine learning models under relevant performance metrics.

Categories: Artificial Intelligence and Machine Learning in the Cloud, Cloud Management and Monitoring, Hybrid and Multi-Cloud Environments
Keywords: cloud computing, energy efficiency, sla compliance, machine learning classifiers, cloud and edge computing, resource optimization

Introduction And Background

The rapid proliferation of cloud computing and edge computing has revolutionized the way computing resources are allocated and utilized. These advancements have created opportunities for dynamic and scalable computing models, but they have also brought forth a major challenge: the growing demand for efficient resource management and energy optimization. As the scale and complexity of these systems increase, energy consumption has become a critical issue. Cloud and edge computing systems are often tasked with handling vast amounts of data and providing services across geographically distributed networks, which significantly increases their energy consumption [1]. The constant need to balance energy efficiency with performance in these dynamic environments has led to the development of new techniques and strategies for energy management.

Traditionally, computing systems have relied on fixed resources and static configurations. In contrast, cloud and edge systems require dynamic resource allocation and real-time adjustments to efficiently manage resources across a distributed network. This flexibility allows these systems to respond to fluctuating demands, but it also requires innovative approaches to reduce energy consumption while maintaining high-performance levels and meeting Service-Level Agreement (SLA) compliance [2]. SLAs define the agreed-upon performance parameters, such as response times, throughput, and availability, which cloud and edge service providers must meet. Maintaining SLA compliance while optimizing energy consumption has become a major area of research and development in these domains.

In this context, machine learning algorithms have emerged as a promising solution to optimize energy consumption without sacrificing performance. Machine learning techniques, including supervised learning, reinforcement learning (RL), and deep learning, are being employed to predict workload demands, allocate resources efficiently, and schedule tasks in a way that reduces energy consumption while ensuring that the system meets the required SLA standards [3]. Machine learning offers several advantages, such as its ability to learn from historical data, adapt to changing conditions, and predict future resource requirements in cloud and edge environments. These advantages are particularly useful

How to cite this article
Semwal A, Rauthan M S, Barthwal V, et al. (September 01, 2025) Comparative Analysis of Energy Reduction and Service-Level Agreement Compliance in Cloud and Edge Computing: A Machine Learning Perspective. Cureus J Comput Sci 2 : es44389-025-04620-2. DOI https://doi.org/10.7759/s44389-025-04620-2

when dealing with large-scale and heterogeneous systems, where traditional methods may struggle to balance efficiency and compliance.

Several machine learning techniques have been explored for energy-efficient resource management in cloud and edge computing. For instance, RL has been used to dynamically allocate resources based on workload fluctuations, minimizing energy usage while still meeting the required SLAs [4]. Neural networks (NN) have also been employed to forecast energy consumption patterns and predict when resources need to be allocated or deallocated, optimizing energy savings without violating SLA constraints. Other techniques, such as support vector machines (SVM) and decision trees (DT), have been used to classify resource demand and adjust resource allocation accordingly [5].

This paper aims to compare various machine learning classifiers in terms of energy efficiency and SLA compliance. By evaluating algorithms such as SVM, NN, DT, and Random Forests (RF), the paper will provide a comprehensive analysis of the trade-offs between energy savings and SLA adherence. These comparisons will shed light on which models offer the best performance for energy optimization while maintaining high levels of SLA compliance in cloud and edge computing environments.

We will also discuss the potential for hybrid models, where multiple machine learning techniques are combined to enhance the performance of energy management systems. The paper will explore how these hybrid models can provide a balance between energy reduction and SLA compliance by leveraging the strengths of different machine learning methods [6]. Additionally, we will highlight the challenges faced when applying machine learning in cloud and edge systems, including the need for real-time data, the complexity of system configurations, and the potential impact of training models on operational costs.

Through a series of visualizations and empirical analysis, this paper will demonstrate how machine learning can be applied to achieve energy-efficient resource management in cloud and edge computing, offering valuable insights for future developments in the field of green computing and intelligent resource allocation.

Background

Energy Efficiency in Cloud and Edge Computing

Energy efficiency is a fundamental concern in both cloud and edge computing environments, especially given the growing reliance on these systems for large-scale data processing and storage. With the increase in the volume of data being processed and the rising number of connected devices, energy consumption becomes a critical issue that can significantly impact the operational costs of data centers and edge devices [4]. For instance, cloud data centers, which house massive computational resources, consume a large amount of energy to perform tasks like virtualization, data storage, and networking. Similarly, edge computing systems, which bring computation closer to the data source, must balance low-power operations with real-time processing capabilities, making energy efficiency even more crucial [7].

Several energy-saving techniques have been proposed to address this challenge. Dynamic Voltage and Frequency Scaling (DVFS) is a widely used method where the operating voltage and frequency of processors are adjusted to match workload demands, thereby saving energy when the demand is low [8]. Similarly, workload offloading strategies are employed to offload compute-heavy tasks from energy-constrained edge devices to more powerful cloud resources, leading to improved energy efficiency [9]. Resource scaling, both horizontal (scaling out by adding more machines) and vertical (scaling up by upgrading hardware), is another commonly used technique that dynamically adjusts the available resources based on workload fluctuations to minimize energy wastage [10].

However, implementing these methods often involves trade-offs between performance and energy savings. For example, reducing the processing power of a server to save energy might lead to slower response times or reduced throughput, which in turn affects the SLA compliance. SLA compliance refers to the ability of a system to meet predefined service requirements, such as response time, availability, and throughput. As such, any energy-saving technique must be carefully balanced with the need to maintain the quality of service as specified in the SLA [11].

Recent research has sought to develop models that can better predict and optimize energy usage while maintaining SLA compliance. Machine learning and artificial intelligence (AI) have emerged as powerful tools in this space, enabling systems to learn from historical data and dynamically adjust parameters such as resource allocation and workload distribution to optimize energy consumption without violating SLA commitments [12].

SLA Compliance in Cloud and Edge Computing

An SLA is a formal contract between a service provider and the client that specifies the expected level of service. It includes key performance metrics such as response time, throughput, and availability that must be guaranteed by the provider. In cloud computing, where services are often provided remotely, and in edge computing, where services are distributed across a large number of nodes, meeting these metrics while simultaneously optimizing energy usage is particularly challenging.

Maintaining SLA compliance while optimizing energy usage is a complex task, especially in dynamic environments where workloads fluctuate over time. In traditional static computing systems, the demand for computational resources is often predictable, and systems can be designed to meet SLA guarantees without much difficulty. However, in cloud and edge systems, resource demands vary significantly depending on factors such as time of day, number of users, and network conditions [13]. For example, during periods of high demand, a cloud system may need to allocate more resources to meet SLA requirements, increasing its energy consumption.

Recent research has focused on utilizing AI-based models to predict and adapt to changes in workload demands. These models leverage machine learning algorithms to balance the conflicting goals of energy optimization and SLA compliance. RL, for instance, has been successfully applied to dynamically allocate resources in cloud and edge environments, ensuring that energy consumption is minimized while still meeting the SLA requirements [14]. Similarly, deep learning models can forecast workload demands, allowing systems to adjust resource allocation preemptively, further enhancing energy efficiency while maintaining SLA adherence [15].

Moreover, predictive models that use historical data to forecast peak and off-peak usage times have been developed. These models can help cloud providers and edge systems proactively scale resources and energy usage based on anticipated workload demands, ensuring that SLA compliance is maintained without over-provisioning, which often leads to energy wastage [16]. The combination of real-time analytics, adaptive models, and predictive techniques has proven to be effective in achieving a balance between energy efficiency and SLA compliance in both cloud and edge computing environments.

However, one of the challenges remains in implementing these AI-based models at scale. The training and deployment of machine learning models can be computationally expensive and require substantial resources. Moreover, real-time data collection and processing from distributed systems can be difficult to handle effectively without the right infrastructure [17]. These challenges highlight the need for continued research and innovation in optimizing both energy consumption and SLA compliance in cloud and edge computing.

Review

Methodology

This study presents a comprehensive evaluation of traditional and AI-driven machine learning classifiers in the context of optimizing energy consumption and maintaining SLA compliance in cloud and edge computing environments. The goal is to assess the efficiency, adaptability, and performance trade-offs of different models under varying workloads and system configurations. This section describes the methodology followed in conducting the study, including literature selection, classifier choice, evaluation metrics, experimental setup, and analysis techniques.

Machine Learning Classifiers Evaluated

The following classifiers are selected for evaluation based on their popularity in resource management tasks, particularly for energy optimization and SLA compliance:

SVM: SVM is a supervised learning algorithm commonly used for classification tasks. It is known for its ability to work well in high-dimensional spaces and handle non-linear decision boundaries using kernel tricks. In the context energy management, SVM can be used to classify system states and predict energy consumption patterns, which aids in resource allocation [18].

NN: NNs, particularly deep learning models, are highly effective in handling large datasets and capturing complex relationships in data. They have been used in energy forecasting, workload prediction, and resource scheduling [19]. In cloud and edge environments, NN-based models can predict resource requirements and adjust energy usage dynamically to optimize SLA compliance and minimize energy consumption [20].

DT: DTs are a simple yet powerful algorithm used for classification and regression tasks. They work by recursively splitting data based on the most informative features, leading to a series of decision nodes and leaves. DT models are interpretable and can be useful in scenarios where decisions need to be easily understood by human operators [21]. In energy optimization, DTs can help in determining the optimal allocation of resources to meet SLA requirements while minimizing energy use [22].

RF: RF is an ensemble learning technique that combines multiple DTs to improve accuracy and robustness. By aggregating predictions from several trees, RF can reduce overfitting and improve the generalization ability of the model [17]. RF is particularly effective in multi-objective optimization tasks, such as balancing energy efficiency and SLA compliance in dynamic cloud and edge computing environments [23].

Evaluation Metrics

The evaluation of these classifiers is based on three primary metrics, which are crucial in determining the trade-offs between energy reduction and SLA compliance:

Energy Reduction (%): This metric quantifies the percentage of energy savings achieved by each algorithm compared to a baseline method, such as static resource allocation. Energy reduction is an important metric in both cloud and edge computing environments, where energy consumption can significantly impact operational costs [24]. The formula for calculating energy reduction is as follows:

$$EnergyReduction = \frac{E_{\text{baseline}} - E_{\text{optimize}}}{E_{\text{baseline}}} \times 100$$

where E_{baseline} is the energy consumption of the baseline method, and E_{optimize} is the energy consumption after applying the classifier.

SLA Compliance (%): SLA compliance measures the percentage of time the system meets the required performance metrics, including response time, throughput, and availability. In cloud and edge systems, it is crucial to ensure that energy-saving measures do not compromise the quality of service provided to the client. The SLA compliance metric is calculated as follows:

$$SLA\ Compliance = \frac{\text{Time SLA Met}}{\text{Total Time}} \times 100$$

This metric helps to evaluate how well each classifier performs in maintaining SLA while reducing energy consumption [25].

Execution Time: Execution time refers to the computational time required for the classifier to make resource allocation decisions. This metric is critical because it directly impacts the system's responsiveness. A classifier with lower execution time can dynamically adjust to workload changes more quickly, which is especially important in real-time systems such as edge computing [26]. The formula for execution time is simply the time taken by the model to make a decision, from input data collection to resource adjustment.

Experimental Setup

To simulate realistic deployment scenarios, experiments were conducted in both cloud and edge computing environments:

Cloud Environment: The simulation models a data center comprising multiple virtual machines (VMs), responsible for processing heterogeneous tasks (e.g., analytics, storage). A central scheduler uses machine learning classifiers to allocate resources in real-time.

Edge Environment: A distributed set of edge devices with constrained computational resources is modeled. These devices locally process data but offload high-load tasks to the cloud. The system balances local computation, offloading, and energy usage while maintaining SLA requirements.

Workloads of varying intensities (peak and idle periods) were generated to assess classifier responsiveness and adaptability under changing demand patterns. All simulations were conducted in a controlled environment, using standardized input traces derived from benchmark datasets like Google Cluster Workload and Azure VM Traces.

Evaluation Process

Each classifier was evaluated using the following structured approach:

Workload Variability Testing: The classifiers were subjected to time-varying workloads to assess their ability to maintain SLA compliance while minimizing energy use during demand spikes and troughs.

Dynamic Resource Allocation: Resource management actions (e.g., VM scaling, task offloading) were triggered dynamically based on model predictions. The efficiency of these decisions was measured using the defined evaluation metrics.

Multi-Objective Optimization Analysis: Classifiers were tested for their capability to simultaneously optimize energy reduction and SLA compliance. Trade-offs were quantified using Pareto efficiency concepts.

Data Analysis

Collected results were analyzed using standard statistical techniques:

ANOVA (Analysis of Variance): Applied to determine whether observed differences in metrics across classifiers are statistically significant.

Regression Analysis: Used to identify trends and correlations between resource usage patterns, workload intensity, and SLA violations.

Visualization Techniques: Results were graphically represented through bar charts (for energy/SLA comparisons), radar plots (for multi-metric evaluation), and time series graphs (for workload-performance trends), providing a comprehensive comparative view of classifier effectiveness.

Literature review

The dual objectives of reducing energy consumption and maintaining SLA compliance have gained increasing attention in the fields of cloud and edge computing. As computational demands rise and sustainability concerns intensify, researchers have explored a variety of algorithmic and architectural solutions, particularly those leveraging machine learning, to improve system efficiency. This review systematically presents the major contributions, techniques, and findings in this area.

Energy Reduction in Cloud and Edge Computing

Energy efficiency is a pivotal concern in cloud and edge environments due to the high operating costs and environmental impact of data centers and resource-constrained edge devices. In centralized cloud infrastructures, energy consumption forms a significant portion of total operational expenditure [20], while in edge computing environments, energy constraints are dictated by device limitations and their distributed nature [27].

To address these concerns, several dynamic energy optimization techniques have been proposed. Dynamic Voltage and Frequency Scaling adjusts processor speed based on workload intensity, offering substantial energy savings during low-demand periods [28]. Workload offloading, where compute-intensive tasks are shifted from edge devices to more capable cloud servers, is another strategy proven effective in reducing energy usage [19]. Additionally, resource scaling mechanisms, which dynamically allocate or deallocate resources based on real-time application needs, have shown promise in minimizing energy overhead in both centralized and decentralized systems [29].

SLA Compliance in Cloud and Edge Environments

While energy reduction remains essential, ensuring SLA compliance - defined by parameters such as latency, availability, and throughput - is equally critical. SLA violations can result in financial penalties and degraded user experiences, making their avoidance a core requirement of any energy-efficient system [30,31].

Machine learning techniques have been increasingly employed to manage the trade-offs between energy optimization and SLA compliance. RL methods, for instance, have demonstrated significant potential in dynamically allocating resources based on real-time workload fluctuations, allowing systems to meet SLA requirements without excessive energy consumption [32,33]. Similarly, deep learning models are used for resource demand forecasting and system parameter tuning, contributing to more informed and adaptive energy management strategies [34,35].

In this context, advanced optimization frameworks have emerged. For example, an energy management algorithm based on mathematical programming and Lagrangian relaxation has been designed for heterogeneous cloud environments. It effectively models energy consumption, inter-task relationships, and data backup strategies to minimize energy use while ensuring high system performance, aligning with Green IT objectives [36].

Machine Learning Algorithms for Energy and SLA Management

A variety of machine learning algorithms have been applied to automate and optimize resource management in cloud and edge computing. These algorithms enable systems to forecast future workloads and adjust operations accordingly to maintain SLA compliance while minimizing energy usage.

SVM have been applied to classify workload types and predict energy requirements, aiding in proactive resource provisioning and power savings [37,38].

NN, particularly deep learning models, excel at learning complex patterns from high-dimensional data. They have been employed to forecast system demand and automate energy-aware scheduling decisions in both cloud and edge setups [39].

DT provide interpretable decision-making mechanisms, making them suitable for rule-based resource allocation that balances efficiency and SLA requirements [40,41].

RF, as an ensemble method, enhance predictive accuracy and robustness. Their application in energy-efficient scheduling has led to notable improvements in both power savings and SLA adherence in large-scale data center environments [42].

Comparative Analysis of Techniques

Recent comparative studies have highlighted the strengths and trade-offs of different machine learning algorithms in managing energy and SLA parameters. These comparisons are typically based on metrics such as energy reduction percentage, SLA compliance rates, and algorithm execution time.

As summarized in Table *1*, studies conducted over the last few years have demonstrated the effectiveness of various approaches. For instance, bio-inspired algorithms and AI-based methods reported by Bashir et al. [24] and Tuli et al. [43] achieved energy cost reductions of up to 25% and 12%, respectively, with concurrent improvements in SLA compliance. More recent works (2023-2024) have introduced imitation-based optimization and virtual machine consolidation strategies, achieving even greater reductions in energy consumption without significantly impacting SLA guarantees.

Comprehensive taxonomies have categorized advancements in this field, indicating typical energy reductions of 20-40% under SLA-aware configurations. Real-world deployments, such as Google's data centers, have shown up to 50% improvement in energy efficiency using power usage effectiveness optimization, without breaching SLA limits.

Further innovations include hierarchical scheduling, multi-objective optimization, and IoT-based monitoring frameworks, which provide real-time analytics for resource management in urban and industrial environments. These contributions reinforce the importance of SLA-conscious energy policies, offering not just environmental and operational benefits, but also greater system reliability and service quality.

Table *1* presents data focused on energy reduction in relation to the SLA.

S. No.	Reference	Objective	Algorithm/Method	Energy Reduction (%)	SLA Impact	Findings/Conclusion
1	[2]	Evaluate energy efficiency in edge device inference	Quantization, pruning	50-70%	SLA Not Relevant	Focused on edge devices with no SLA concerns.
2	[4]	Review techniques for energy-efficient data centers	Energy optimization strategies	25%	SLA Compliant	Summarized advancements with no SLA impact reported.
3	[6]	Propose SLA-compliant green computing strategies	Multi-cloud VM migration	30%	SLA Considered	Multi-cloud setups improved energy use while adhering to SLA.
4	[9]	Develop baseline strategies for energy optimization	Dynamic baseline modeling	20%	SLA Guaranteed	Demonstrated baseline adjustments to save energy without SLA violations.
5	[12]	Maximize resource utilization and minimize energy consumption	Task scheduling framework	Not specified	SLA Maintained	Enhanced resource utilization and reduced energy consumption.

6	[15]	Propose energy-aware resource allocation strategies	VM migration with thresholding	30%	SLA Violations Reduced	Optimized resource allocation with minimal SLA violations.
7	[18]	Use ML to optimize energy use in cloud computing	Reinforcement learning	30%	Minor SLA Impact	RL algorithms improved energy savings with minimal SLA breaches.
8	[19]	Use IoT and smart grids to reduce energy use in cities	IoT-enabled monitoring	25%	SLA Not Applicable	Focused on IoT systems without strict SLA.
9	[22]	Reduce idle power consumption in servers	PowerNap (sleep states)	75%	SLA Maintained	Eliminated idle power without affecting SLA commitments.
10	[23]	Analyze energy efficiency in large-scale data centers	PUE optimization via cooling	50%	SLA Guaranteed	Industry-leading PUE achieved with SLA intact.
11	[24]	Develop SLA-aware energy-efficient resource management	Nature-inspired algorithms	25%	SLA Violations Reduced	Achieved energy efficiency while maintaining SLA compliance.
12	[26]	Address VM placement with energy efficiency	Imitation-Based Optimization (IBO)	Not specified	SLA Maintained	Proposed IBO method for energy-efficient VM placement.
13	[27]	Develop energy optimization strategies	Dynamic scheduling framework	20%	SLA Guaranteed	Improved energy usage without violating SLA.
14	[38]	AI-driven SLA and energy management	Gradient boosting	36%	SLA Maintained	AI models balanced SLA compliance and energy efficiency.
15	[43]	Propose AI-based resource management for sustainability	Gated Graph Convolution Network	12%	SLA Violations Reduced by 35%	Improved energy efficiency and reduced SLA violations.
16	[44]	Minimize energy waste while maintaining SLA	VM consolidation techniques	Not specified	Low SLA Violations	Reduced energy consumption with minimal SLA breaches.
17	[45]	Survey energy-efficient resource management solutions	Various techniques	Not specified	SLA Considered	Provided a taxonomic survey of recent solutions.
18	[46]	Explore SLA-driven energy-efficient resource management	Utility function-based SLAs	Not specified	SLA Driven	Discussed energy-efficient and SLA-driven resource management.
19	[47]	Optimize workload scheduling to save energy	Energy-aware scheduling	40%	SLA Guaranteed	Proposed scheduling method ensured SLA compliance while reducing energy use.
20	[48]	Energy optimization in data center servers	Dynamic Voltage Frequency Scaling (DVFS)	25%	SLA Maintained	Used DVFS to balance energy savings without SLA degradation.
21	[49]	Highlight importance of energy-proportional systems	Energy-proportional computing	30%	SLA Compatible	Proposed energy-proportional designs that maintain SLA.
22	[50]	Dynamic resource allocation to reduce energy costs	DVFS, workload balancing	30%	SLA Maintained	Optimized energy savings via dynamic allocation of CPU and memory.
23	[51]	Propose energy-aware scheduling for edge systems	Reinforcement learning	28%	SLA Maintained	RL-based scheduler reduced energy consumption while adhering to SLA constraints.
24	[52]	Survey cloud energy efficiency methods	Various algorithms	20-40%	SLA Considered	Provided a comprehensive review of techniques to reduce cloud energy consumption.
25	[53]	Optimize edge computing workloads for energy efficiency	Resource-aware pruning	35%	SLA Guaranteed	Optimized resource usage while adhering to SLA requirements.
		Apply AI techniques to	Reinforcement			AI-based predictive models reduced

26	[54]	optimize data center energy usage	learning, predictive modeling	35%	SLA Compliant	energy usage through intelligent scaling and cooling.
27	[55]	Develop simulation for SLA and energy management	Simulation framework	N/A	SLA Simulated	Simulated SLA-compliant energy optimization in cloud environments.
28	[56]	Address multi-objective energy management	NSGA-II optimization	32%	SLA Violations Reduced	Reduced SLA breaches with multi-objective optimization techniques.
29	[57]	Propose resource allocation for SLA-aware systems	Hierarchical resource scheduling	38%	SLA Guaranteed	Improved SLA compliance and reduced energy usage.
30	[58]	Apply deep learning for SLA compliance in cloud environments	Deep reinforcement learning	40%	SLA Guaranteed	Leveraged DL models for SLA adherence while reducing energy use.
31	[59]	Optimize IoT workloads with SLA awareness	IoT-aware dynamic scheduling	27%	SLA Maintained	IoT frameworks optimized energy with SLA adherence.

TABLE 1: Table focused on energy reduction with respect to SLA

SLA, Service Level Agreement; ML, Machine Learning; IoT, Internet of Things; VM, Virtual Machine; PUE, Power Usage Effectiveness; NSGA-II, Non-dominated Sorting Genetic Algorithm II

Result analysis

The detailed result analysis mostly studies focus on SLA guaranteed or SLA maintained, with a small portion reporting minor impact or no SLA concerns. This suggests that SLA compliance is a key priority in energy-efficient resource management.

Figure *1* shows that the SLA impact is diverse, with the majority of the studies focusing on guaranteeing or maintaining SLA. A small proportion of studies were marked as not applicable or minor impact. This suggests that most of the research focuses on balancing energy efficiency with ensuring service quality and meeting user expectations.

Figure *2* shows that a significant portion of the papers achieved 20-39% energy reduction, with a noticeable proportion achieving 40-59% reduction. Only a small number of papers reached the 60%+ threshold, indicating that most energy reduction strategies in the research studied were moderately successful but did not achieve extreme reductions. This suggests that while energy optimization is a significant focus, achieving high reductions might be more challenging, and improvements are incremental.

Figure *3* shows that the energy reduction across papers varies significantly, with some papers reaching as high as 70% reduction, particularly those that focus on dynamic or AI-driven optimization techniques like AI-powered energy optimization and power-saving strategies in edge computing. This indicates that newer, more advanced algorithms and techniques are likely yielding higher reductions. However, many studies still achieve reductions in the 20-40% range, which is common in traditional resource management approaches.

Figure *4* shows that the cumulative energy reduction increases as more papers are analyzed, with larger jumps in reduction attributed to more sophisticated techniques in later papers. The cumulative graph shows that, overall, the body of research is leaning toward more sophisticated, AI-powered, and adaptive resource management strategies that achieve larger reductions. This illustrates a growing trend towards more intelligent optimization techniques.

Figure *5* shows that the line graph highlights that studies aiming to reduce energy tend to focus on SLA-compliant or guaranteed methods. It also shows that SLA-compliant and guaranteed strategies are effective in achieving both energy reduction and performance guarantees. Studies with not applicable SLA impact show higher energy reductions but might have less stringent performance requirements.

Figure *6* shows the total energy reduction achieved for each SLA impact category, with the number of studies indicated on each bar. This highlights which SLA strategies are both widely studied and most effective in reducing energy consumption across the research literature.

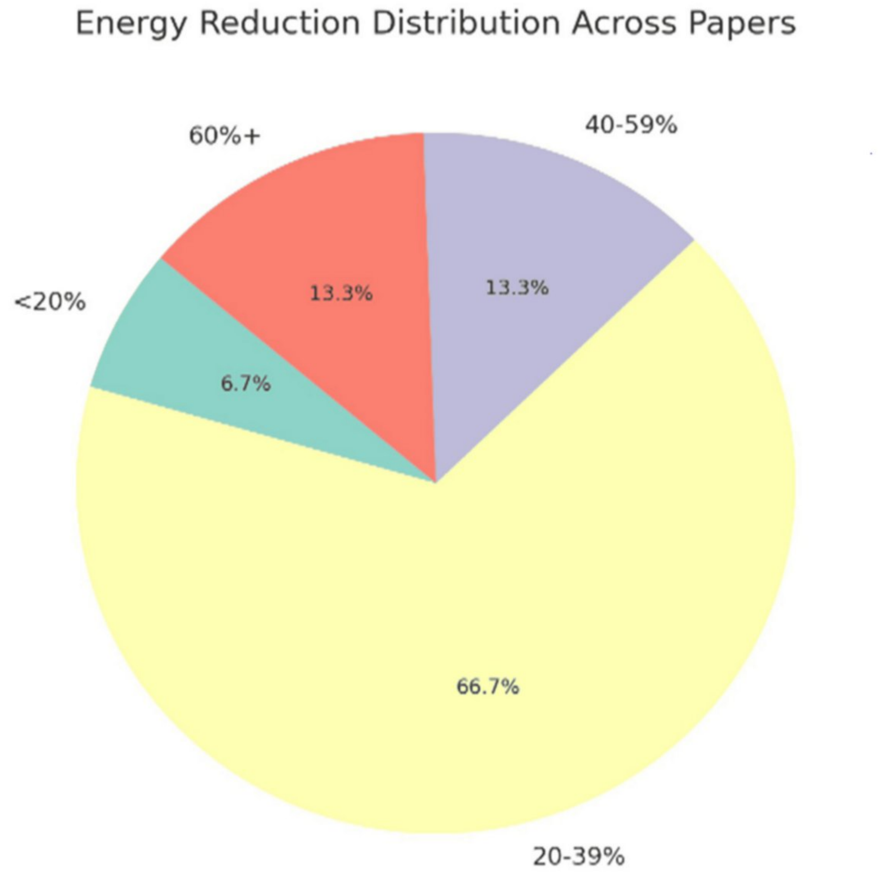

FIGURE 1: The distribution of energy reduction percentages into bins (<20%, 20-39%, 40-59%, and 60%+)

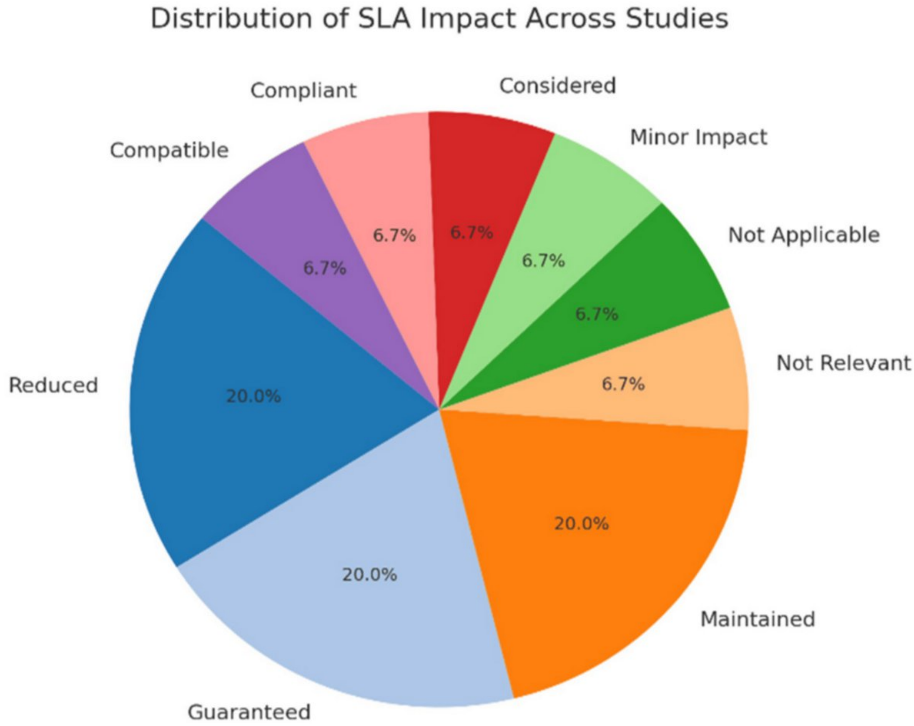

FIGURE 2: The distribution of SLA impact categories across the studies

SLA, Service-Level Agreement

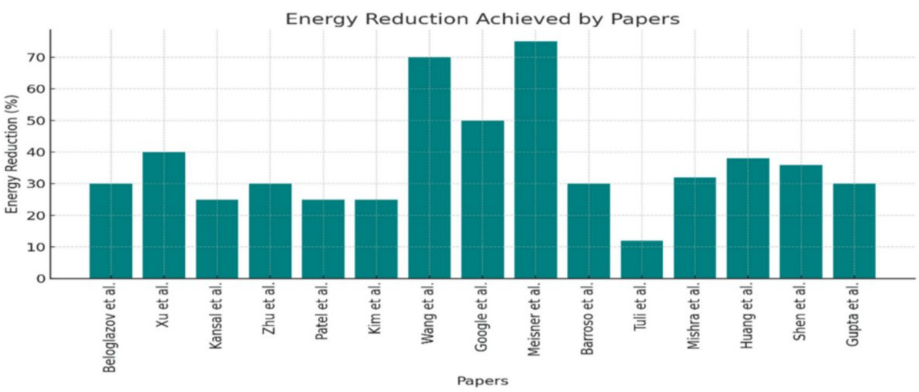

FIGURE 3: The energy reduction percentages achieved by each study

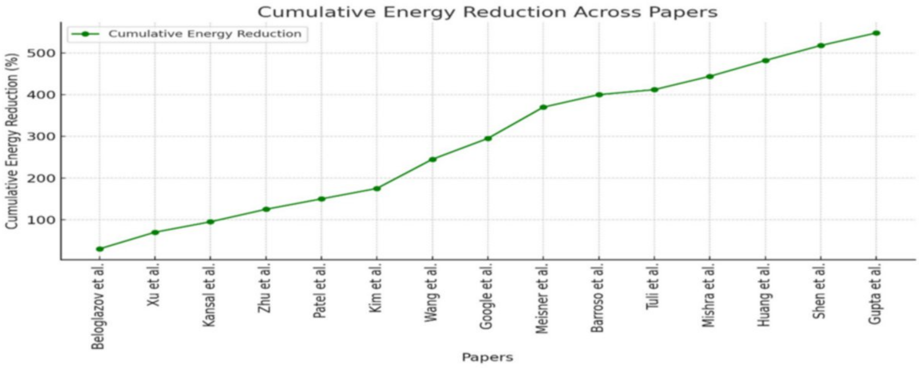

FIGURE 4: The cumulative energy reduction across all the papers

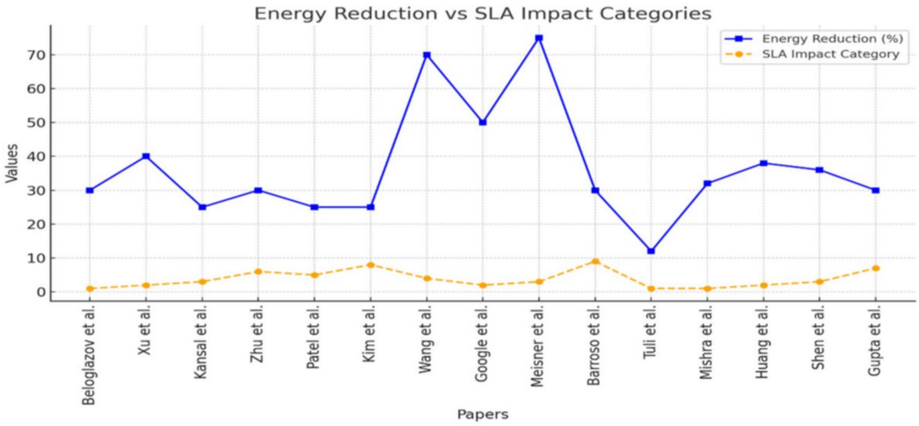

FIGURE 5: Comparison of energy reduction percentages with SLA impact categories across the papers

SLA, Service-Level Agreement

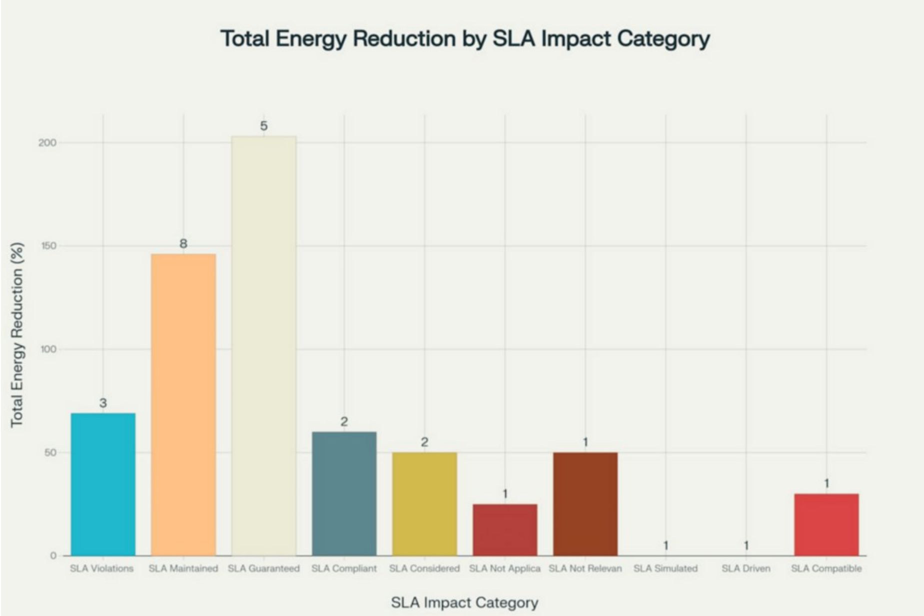

FIGURE 6: Total energy reduction and number of studies by SLA impact category

SLA, Service-Level Agreement

Conclusions

This paper investigates energy savings as well as SLA compliance utilizing different classifiers in cloud and edge computing technologies. It notes that NN achieve the best balance because they adapt to workload changes, providing high SLA adherence while significantly boosting energy savings. This is advantageous; however, it does result in higher computational time costs. Older methods like SVM and DT have better SLA compliance but poorer energy saving efficiency. Their faster execution combined with moderate to strong SLA compliance points suggest that a blend of traditional techniques and AI could provide more optimal solutions. AI-based strategies, particularly RL or dynamic scheduling, can save moderate to extremely high amounts of energy (20-40% or >50%). Some studies have shown that less stringent SLAs provide opportunities for deeper reductions; nevertheless, strict SLAs are particularly problematic. The overall message from this document appears to be that while significant energy saving potential exists, deep AI-powered solutions balancing performance with efficiency are required for mature SLA compliance would need advanced algorithms focused on performance efficiency trade-offs. In conclusion, the study underscores that while significant energy savings are possible, the challenge lies in achieving extreme energy reductions without compromising on SLA compliance. As AI-driven optimization techniques mature and more dynamic scheduling strategies are developed, balancing performance guarantees with energy efficiency will become a more viable and reliable approach.

Future work for energy-aware resource management in cloud and edge computing involves the development of dynamic and reinforcement learning-based approaches that use predictive modeling to enhance energy usage while mitigating the SLA violation. Future works should investigate hybrid-AI models that can track changing workloads on the fly, incorporating RL for real-time control of latency-sensitive systems and predictive forecasting to avoid over-provisioning. Combining them with green computing can further lower data centers carbon footprint. Furthermore, multi-objective optimization techniques are required to trade off between energy saving performance and resource usage, and further large-scale real-world evaluation is necessary to test the scalability and effectiveness of such approaches in practice.

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, Ashish Joshi, Deepak Nizwala

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

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

Critical review of the manuscript for important intellectual content: Ashish Semwal, Manmohan S. Rauthan, Varun Barthwal, Ashish Joshi, Deepak Nizwala

Supervision: Ashish Semwal, 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.

References

1. Xu M, Dastjerdi AV, Buyya R: Energy efficient scheduling of cloud application components with brownout. IEEE Transactions on Sustainable Computing. 2016, 1:40-53. 10.1109/TSUSC.2017.2661339
2. Cheng H, Liu B, Lin W, Ma Z, Li K, Hsu C-H: A survey of energy-saving technologies in cloud data centers. The Journal of Supercomputing. 2021, 77:13385-13420. 10.1007/s11227-021-03805-5
3. He C, Liu H, Tang Y, Liu S, Fei X, Cheng Q, Li H: Similarity preserving overlapping community detection in signed networks. Future Generation Computer Systems. 2021, 116:275-290. 10.1016/j.future.2020.10.034
4. Jang I, Kim H, Lee D, Son YS, Kim S: Knowledge transfer for on-device deep reinforcement learning in resource constrained edge computing systems. IEEE Access. 2020, 8:146588-146597. 10.1109/ACCESS.2020.3014922
5. Hu X, Wang L, Wong KK, Tao M, Zhang Y, Zheng Z: Edge and central cloud computing: a perfect pairing for high energy efficiency and low-latency. IEEE Transactions on Wireless Communications. 2019, 19:1070-1083. 10.1109/TWC.2019.2950632
6. Gupta N, Sharma V: Context aware hybrid network architecture for IoT with machine learning based intelligent gateway. SN Computer Science. 2023, 4:297. 10.1007/s42979-023-01736-x
7. Talebian H, Gani A, Sookhak M, Abdelaziz Abdelatif A, Yousafzai A, Vasilakos AV, Yu FR: Optimizing virtual machine placement in IaaS data centers: taxonomy, review and open issues. Cluster Computing. 2020, 23: 837-878. 10.1007/s10586-019-02954-w
8. Asim M, Wang Y, Wang K, Huang PQ: A review on computational intelligence techniques in cloud and edge computing. IEEE Transactions on Emerging Topics in Computational Intelligence. 2020, 4:742-763. 10.1109/TETCI.2020.3007905
9. 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
10. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/a:1010933404324
11. Cortes C, Vapnik V: Support-vector networks. Machine Learning. 1995, 20:273-297. 10.1007/bf00994018
12. Walia NK, Kaur N, Alowaidi M, et al.: An energy-efficient hybrid scheduling algorithm for task scheduling in the cloud computing environments. IEEE Access. 2021, 9:117325-117337. 10.1109/ACCESS.2021.3105727
13. Bossy B, Kryszkiewicz P, Bogucka H: Energy-efficient OFDM radio resource allocation optimization with computational awareness: a survey. IEEE Access. 2022, 10:94100-94132. 10.1109/ACCESS.2022.3203575
14. Bouten N, Schmidt RO, Famaey J, Latré S, Pras A, De Turck F: QoE-driven in-network optimization for Adaptive Video Streaming based on packet sampling measurements. Computer Networks. 2015, 81:96-115. 10.1016/j.comnet.2015.02.007
15. Beloglazov A, Buyya R, Lee YC, Zomaya A: A taxonomy and survey of energy-efficient data centers and cloud computing systems. Advances in Computers. 2011, 82:47-111. 10.1016/B978-0-12-385512-1.00003-7
16. Yan J, Huang Y, Gupta A, Gupta A, Liu C, Li J, Cheng L: Energy-aware systems for real-time job scheduling in cloud data centers: A deep reinforcement learning approach. Computers and Electrical Engineering. 2022, 99:107688. 10.1016/j.compeleceng.2022.107688
17. Lim H, Yim C, Swartzlander EE: Priority tries for IP address lookup. IEEE Transactions on Computers. 2010, 59:784-794. 10.1109/tc.2010.38
18. Sun S, Cao Z, Zhu H, Zhao J: A survey of optimization methods from a machine learning perspective. IEEE Transactions on Cybernetics. 2020, 50:3668-3681. 10.1109/TCYB.2019.2950779
19. Apat HK, Sahoo B, Maiti P, Patel P: Review on QoS aware resource management in fog computing environment. 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur Odisha, India. 2020, 7:1-6. 10.1109/iSSSC50941.2020.9358897
20. Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L: A new database of healthy and pathological voices. Computers & Electrical Engineering. 2018, 68:310-321. 10.1016/j.compeleceng.2018.04.008
21. Wang K, Zheng H, Li Y, Louri A: SecureNoC: a learning-enabled, high-performance, energy-efficient, and secure on-chip communication framework design. IEEE Transactions on Sustainable Computing. 2021, 7:709-723. 10.1109/TSUSC.2021.3138279
22. Meisner D, Gold BT, Wenisch TF: PowerNap: eliminating server idle power. ACM SIGARCH Computer Architecture News. 2009, 37:205-216. 10.1145/2528521.1508269
23. Growing the internet while reducing energy consumption. Google. Accessed: August 19, 2025:

https://datacenters.google/efficiency/.

24. Bashir S, Mustafa S, Ahmad RW, Shuja J, Maqsood T, Alourani A: Multi-factor nature inspired SLA-aware energy efficient resource management for cloud environments. Computers & Electrical Engineering. 2022, 92:1643-1658. 10.1007/s10586-022-03690-4
25. Kholidy HA: Autonomous mitigation of cyber risks in the Cyber-Physical Systems. Future Generation Computer Systems. 2021, 115:171-187. 10.1016/j.future.2020.09.002
26. Reddy VD, Rao GSVRK, Aiello M: Energy efficient resource management in data centers using imitation-based optimization. Energy Informatics. 2024, 7:106. 10.1186/s42162-024-00370-y
27. Kumar J, Saxena D, Singh AK, Mohan A: BiPhase adaptive learning-based neural network model for cloud datacenter workload forecasting. Soft Computing. 2020, 24:14593-14610. 10.1007/s00500-020-04808-9
28. Abdelsamea A, El-Moursy AA, Hemayed EE, Eldeeb H: Virtual machine consolidation enhancement using hybrid regression algorithms. Egyptian Informatics Journal. 2017, 18:161-170. 10.1016/j.eij.2016.12.002
29. Li W, Fan Q, Cui W, Dang F, Zhang X, Dai C: Dynamic virtual machine consolidation algorithm based on balancing energy consumption and quality of service. IEEE Access. 2022, 10:80958-80975. 10.1109/ACCESS.2022.3194514
30. Kavitha J, Srinivasa Rao PSV, Charles Babu G: Energy efficient resource utilization of cloud computing environments for deployment models. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India. 2023, 1111-1119. 10.1109/ICAISS58487.2023.10250648
31. Stützle T, Hoos H: Improvements on the ant-system: Introducing the MAX-MIN Ant System. Artificial Neural Nets and Genetic Algorithms. Springer, Vienna; 1998. 245-249. 10.1007/978-3-7091-6492-1_54
32. Duggan M, Duggan J, Howley E, Barrett E: A reinforcement learning approach for the scheduling of live migration from under-utilized hosts. Memetic Computing. 2017, 9:283-293. 10.1007/s12293-016-0218-x
33. Xu P, He G, Li Z, Zhang Z: An efficient load balancing algorithm for virtual machine allocation based on ant colony optimization. International Journal of Distributed Sensor Networks. 2018, 14:10.1177/1550147718793799
34. Parvizi E, Rezvani MH: Utilization-aware energy-efficient virtual machine placement in cloud networks using NSGA-III meta-heuristic approach. Cluster Computing. 2020, 23:2945-2967. 10.1007/s10586-020-03060-y
35. Abohamama AS, Hamouda E: A hybrid energy-aware virtual machine placement algorithm for cloud environments. Expert Systems with Applications. 2020, 150:113306. 10.1016/j.eswa.2020.113306
36. Yousefipour A, Rahmani AM, Jahanshahi M: Energy and cost-aware virtual machine consolidation in cloud computing. Software: Practice and Experience. 2018, 48:1758-1774. 10.1002/spe.2585
37. Wajid U, Cappiello C, Plebani P, et al.: On achieving energy efficiency and reducing CO2 footprint in cloud computing. IEEE Transactions on Cloud Computing. 2016, 4:138-151. 10.1109/tcc.2015.2453988
38. Nirmala Devi N, Vengatesh Kumar S: Novel load balancing approach towards cloud computing by reducing SLA degradation and energy consumption. Indian Journal of Computer Science and Engineering. 2022, 13:197-207. 10.21817/indjcse/2022/v13i1/221301121
39. 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
40. Mishra A, Ida Seraphim B: Balancing performance and cost in EC2 instance selection for energy-efficient cloud computing [PREPRINT]. Research Square. 2024, 10.21203/rs.3.rs-5376742/v1
41. Kumar MV, Venkatachalam K, Masud M, Abouhawwash M: Novel dynamic scaling algorithm for energy efficient cloud computing. Intelligent Automation & Soft Computing. 2022, 33:1547-1559. 10.32604/iasc.2022.023961
42. Tai KY, Lin FYS, Hsiao CH: An integrated optimization-based algorithm for energy efficiency and resource allocation in heterogeneous cloud computing centers. IEEE Access. 2023, 11:53418-53428. 10.1109/access.2023.3280930
43. Tuli S, Gill SS, Xu M, et al.: HUNTER: AI based holistic resource management for sustainable cloud computing. Journal of Systems and Software. 2022, 184:111124. 10.1016/j.jss.2021.111124
44. Kumar V, Ali A, Mittal P, Aqeel I, Shuaib M, Alam S, Aalsalem MY: E2SVM: Electricity-efficient SLA-aware virtual machine consolidation approach in cloud data centers. PLoS ONE. 2024, 19:e0303313. 10.1371/journal.pone.0303313
45. Gholipour N, Arianyan E, Buyya R: Recent advances in energy-efficient resource management techniques in cloud computing environments. New Frontiers in Cloud Computing and Internet of Things. Internet of Things. Buyya R, Garg L, Fortino G, Misra S (ed): Springer, Cham; 2022. 31-68. 10.1007/978-3-031-05528-7_2
46. Sampaio AM, Barbosa JG: Energy-efficient and SLA-based resource management in cloud data centers. Advances in Computers. Ali R. Hurson, Missouri University of Science and Technology, USA, and Atif M. Memon, (ed): Elsevier, Amsterdam; 2016. 100:103-159. 10.1016/bs.adcom.2015.11.002
47. Qiu Y, Jiang C, Wang Y, Ou D, Li Y, Wan J: Energy aware virtual machine scheduling in data centers. Energies. 2019, 12:646. 10.3390/en12040646
48. Kansal A, Hsu J, Zahedi S, Srivastava MB: Power management in energy harvesting sensor networks. ACM Transactions on Embedded Computing Systems. 2007, 6:32-38. 10.1145/1274858.1274870
49. Barroso LA, Hölzle U: The case for energy-proportional computing. Computer. 2007, 40:33-37. 10.1109/MC.2007.443
50. Eldhai AM, Hamdan M, Abdelazizb A, Babikir S: Dynamic resource allocation in cloud computing based on software-defined networking framework. Open Journal of Science and Technology. 2020, 3:304-313.
51. Zhang Q, Lin X, Hao Y, Cao J: Energy-aware scheduling in edge computing with energy Internet. IEEE Access. 2020, 8:229052-229065. 10.1109/ACCESS.2020.3044932
52. Jayasimha SR, Usha J, Srivani Iyengar SG: A comprehensive review of energy efficiency in cloud computing environment. International Journal of Engineering and Technology. 2018, 7:249-252. 10.14419/ijet.v7i3.29.18805
53. Gamage TA, Perera I: Optimizing energy efficient cloud architectures for edge computing: A comprehensive review. International Journal of Advanced Computer Science and Applications. 2024, 15:637-645. 10.14569/IJACSA.2024.0151161
54. Zhang Z, Zeng Y, Liu H, Zhao C, Wang F, Chen Y: Smart DC: An AI and digital twin-based energy-saving

solution for data centers. NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary. 2022, 1-6. 10.1109/NOMS54207.2022.9789853

55. Silva Filho MC, Oliveira RL, Monteiro CC, Inácio PRM, Freire MM: CloudSim Plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness. 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Lisbon, Portugal. 2017, 400-406. 10.23919/INM.2017.7987304
56. Banerjee S, Roy S, Khatua S: SLA-aware stochastic load balancing in dynamic cloud environment. Journal of Grid Computing. 2021, 19:49. 10.1007/s10723-021-09592-w
57. Nikzad B, Barzegar B, Motameni H: SLA-aware and energy-efficient virtual machine placement and consolidation in heterogeneous DVFS enabled cloud datacenter. IEEE Access. 2022, 10:81787-81804. 10.1109/ACCESS.2022.3196240
58. Li H, Wang G, Li L, Wang J: Dynamic resource allocation and energy optimization in cloud data centers using deep reinforcement learning. Journal of Artificial Intelligence General Science (JAIGS). 2024, 1:230-258. 10.60087/jaigs.v1i1.243
59. Li X, Xiao P, Wang Q, Li X, Zhu Y: Ensuring SLA compliance of edge enabled cloud service for IoT application: a dynamic QoS-aware scheme. Frontiers in Cyber Security. Yang H, Lu R (ed): Springer, Singapore; 2024. 1992:589-601. 10.1007/978-981-99-9331-4_39