International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

Artificial Intelligence for Energy-Efficient Green Cloud Data Centres: A Review of Workload Prediction, Intelligent Scheduling and Cooling Optimization

Dhavale Priyal Sandesh[1] Dhavale Pranjal Sandesh[2] Gaikar Aditi Bhausaheb[3] Dr. V. M. Dhede[4] Prof. M. S. Nalawade[5]

[1,2,3] Department Of Electronics And Telecommunication Jaihind College Of Engineering Kuran, Pune, Maharashtra, India

[4]**Seminar** Guide & Head of Department Department of Electronics and Telecommunication Jaihind College of Engineering Kuran, Pune, Maharashtra, India

[5]**Seminar** Coordinator Department of Electronics and Telecommunication Jaihind College of Engineering Kuran, Pune, Maharashtra, India

Abstract—The rapid growth of online services, digital education and large-scale artificial intelligence has increased the electricity demand of cloud data centres, where servers, storage, networks and cooling equipment run around the clock. Rule-based management policies, which react to fixed thresholds, often leave servers underused and cooling plants over-provisioned. Artificial intelligence (AI) offers a different approach: models learn from operational data, anticipate demand and choose actions that save energy while protecting service quality. This review examines how AI is applied to four tasks in a green data centre: forecasting workload, allocating and consolidating virtual machines, scheduling tasks with awareness of energy and carbon, and controlling cooling systems. Statistical forecasting, recurrent neural networks, reinforcement learning and data-driven cooling control are explained and compared through representative studies. An AI-based green-cloud framework is then described that combines a learning-based demand predictor, an energy- and carbon-aware scheduler, a cooling optimiser and a renewable-supply monitor in a closed control loop. A simulation-based evaluation plan using energy, response time, utilization and carbon emission as metrics is outlined. Challenges including data quality, model interpretability, safety of autonomous control and the energy cost of the AI models themselves are discussed.

Index Terms—Green cloud computing, artificial intelligence, machine learning, workload prediction, deep reinforcement learning, cooling optimization, carbon-aware scheduling, data centre.

I. INTRODUCTION

Cloud computing supports a wide range of daily activities, from online classes and banking to video streaming and the training of large AI models. The facilities that host these services are energy-intensive. A data centre must power thousands of servers, keep them connected through high-speed networks and remove the heat they produce. Estimates of global consumption indicate that efficiency gains have been essential in limiting growth so far, but further demand, particularly from AI workloads, puts that balance at risk [1].

Many data centres still manage resources with static rules. For example, a host is considered overloaded when CPU utilization crosses a fixed value, and cooling set-points are kept conservatively low to avoid any risk of hot spots. Such rules are simple, yet they cannot anticipate changes in demand and they ignore the complex interaction between IT load, outside weather and cooling equipment [2]. As a result, energy is wasted on idle servers and unnecessary cooling.

Artificial intelligence, and machine learning (ML) in particular, can learn these relationships from the large volume of telemetry that data centres already record. Learned models can predict demand minutes or hours ahead, choose placements that keep servers well utilized, and adjust cooling based on predicted conditions rather than worst-case assumptions [3]. The purpose of this paper is to review these AI techniques in an organised way and to describe a framework that combines them.

The contributions of the paper are:

- A review of AI methods for workload prediction, resource allocation, task scheduling and cooling control in cloud data centres.
- A comparison of the reviewed approaches by objective, technique, data needs and limitations.
- An AI-based green-cloud framework with a closed-loop structure and an evaluation plan.
- A discussion of open challenges for practical deployment.

© 2026 IJSRET

Section II gives the background, Section III reviews AI techniques, Section IV presents the framework, Section V compares approaches, Section VI describes evaluation, Section VII discusses challenges, and Section VIII concludes.

II. BACKGROUND

1 Energy Profile Of A Data Centre

Energy in a data centre is consumed by IT equipment and by supporting infrastructure, mainly cooling and power conversion [2]. Power usage effectiveness (PUE) is the ratio of total facility energy to IT energy; values close to 1.0 indicate low overhead. Within the IT load, servers are not energy-proportional: an idle server still draws a large fraction of its peak power [4]. Two broad opportunities follow. The first is to reduce the number of active servers by placing work efficiently.

The second is to reduce cooling energy by operating the cooling plant closer to what the actual heat load requires.

2 Why Learning Helps

Both opportunities depend on knowing the future. Consolidating virtual machines safely requires knowing whether demand will rise in the next interval; adjusting cooling safely requires predicting how temperatures will respond. Workload traces from large providers show strong daily and weekly patterns as well as recurring behaviour of individual customers, which makes them suitable for learning [5]. AI therefore acts as the *prediction and decision* engine of a green data centre.

3. Main Families Of Methods

Statistical time-series models such as the autoregressive integrated moving average (ARIMA) capture trend and seasonality in a compact form [6]. Neural networks, including long short-term memory (LSTM) networks designed to retain information over long sequences [7], can model non-linear patterns.

Reinforcement learning (RL) learns a policy by trial and reward and is suited to sequential decisions such as scheduling [8]. Data-driven control combines prediction models with optimisation to choose set-points for physical systems such as chillers and cooling towers [3].

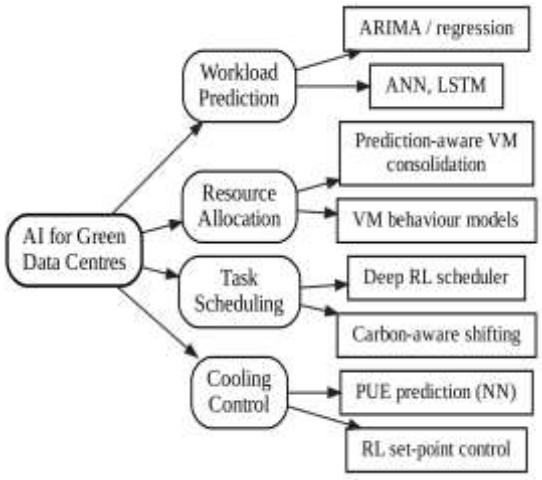


Fig. 1. Taxonomy of AI applications for energy efficiency in green cloud data centres.

III. REVIEW OF AI TECHNIQUES

1. Workload Prediction

Calheiros *et al.* applied an ARIMA model to forecast web request rates and showed that accurate prediction improves the quality of service delivered by provisioning resources ahead of demand [6]. Kumar and Singh trained an artificial neural network whose weights were tuned with an adaptive differential evolution algorithm, reporting lower prediction error than conventional gradient training on benchmark traces [9]. Cortez *et al.* described Resource Central, a system deployed at a large cloud provider that learns from past behaviour of virtual machines to predict their lifetime, utilization and workload class, and makes these predictions available to the resource manager [5]. Recurrent models such as LSTM are commonly used for this task because they handle long sequences with daily and weekly cycles [7].

2. Intelligent Resource Allocation And Consolidation

Heuristic consolidation, such as the adaptive threshold and best-fit methods of Beloglazov and Buyya, already saves considerable energy [10]. AI improves on these by replacing reactive thresholds with predicted utilization, so that a host is relieved before it becomes overloaded and an underloaded host is emptied only when its demand is expected to stay low. Prediction-aware consolidation reduces both unnecessary migrations and SLA violations.

3. Reinforcement-Learning-Based Scheduling

Mao *et al.* framed multi-resource job scheduling as a reinforcement learning problem and trained a deep policy network, DeepRM, that learned to reduce average job slowdown and performed comparably with or better than hand-tuned heuristics [8]. The same formulation can include

© 2026 IJSRET

energy in the reward, so that the agent learns to pack jobs onto fewer servers or to delay flexible jobs until energy is cheaper or cleaner. Radovanović *et al*. demonstrated carbon-aware scheduling in production, using day-ahead forecasts of grid carbon intensity to shift flexible workloads in time [11].

4. Cooling Optimization

Cooling is one of the largest sources of overhead. Gao described the use of neural networks trained on sensor data to predict PUE from operating variables, allowing operators to test configuration changes before applying them [3]. Evans and Gao later reported that an ML system recommending cooling actions reduced the energy used for cooling in a Google data centre by up to 40% [12]. Li *et al.* applied deep reinforcement learning to cooling control and reported energy savings while keeping temperatures within limits [13].

Table I Summary Of Representative Ai Studies For Green Data Centres

Ref.	Task	AI Technique	Reported Outcome	Limitation
[6]	Workload prediction	ARIMA time-series model	Better QoS through proactive provisioning	Linear model; limited for bursts
[9]	Workload prediction	Neural network with adaptive differential evolution	Lower error than standard training	Training cost; trace-specific tuning
[5]	VM behaviour prediction	Learned models on VM history	Predictions used in production manager	Needs large provider-scale data
[8]	Job scheduling	Deep reinforcement learning (DeepRM)	Lower job slowdown than heuristics	Simplified simulated cluster
[11]	Carbon-aware scheduling	Carbon-intensity forecasting with load shifting	Flexible load moved to cleaner hours	Only delay-tolerant workloads
[12]	Cooling	ML recommendations for cooling	Up to 40% cooling energy reduction reported	Proprietary; facility-specific
[13]	Cooling	Deep reinforcement learning control	Energy saved within thermal limits	Safety of exploration

IV. PROPOSED AI-BASED GREEN CLOUD FRAMEWORK

Drawing on the reviewed methods, an AI-based green-cloud framework is described here. Its structure is shown in Fig. 2.

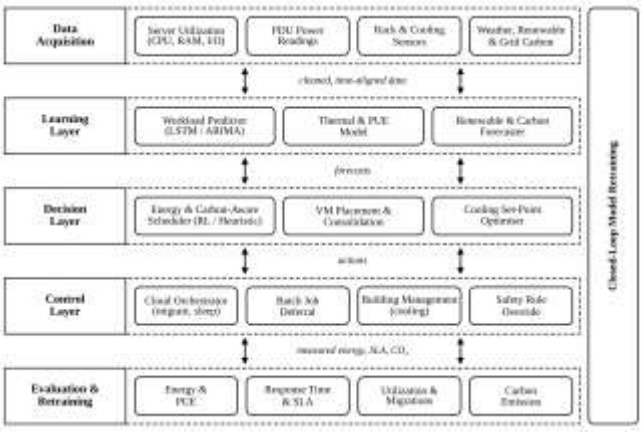


Fig. 2. Architecture of the AI-based green cloud framework with prediction, decision and control layers.

1 Data Acquisition

The data layer collects server utilization (CPU, memory, disk, network), power readings from power-distribution units, rack inlet and outlet temperatures, cooling-plant variables, outside weather, and the carbon intensity and availability of renewable supply. Data are cleaned, aligned to a common time step and stored.

2 Learning Layer

The learning layer contains three models. A workload predictor, such as an LSTM network, forecasts demand per VM and per cluster for the next intervals [7]. A thermal model predicts rack temperatures and cooling power for a given IT load and set-point [3]. A renewable and carbon forecaster estimates solar or wind output and grid carbon intensity [11].

3 Decision Layer

An energy-aware scheduler uses the predictions to place VMs and tasks. Interactive services are placed to meet latency limits with the fewest active hosts, while deferrable batch tasks are scheduled in periods of high renewable supply or low carbon intensity. The scheduler can be a reinforcement learning agent whose reward combines energy, SLA penalty and carbon emission [8], or a heuristic that uses predicted

© 2026 IJSRET

rather than current utilization [10]. A cooling optimiser selects set-points that satisfy temperature limits with the least cooling power.

4. Control And Feedback

The control layer executes decisions through the cloud orchestrator and the building management system, switching idle hosts to low-power states and changing cooling set-points. Safety rules override the AI in case of abnormal temperature or overload. Measured results are fed back to retrain the models periodically, forming a closed loop.

5. Comparative Analysis

Table II compares the main AI families used for green data centres in terms of what they need and what they offer.

Table Ii Comparison Of Ai Method Families For Data-Centre Energy Management

Criterion	Statistical Models	Neural Networks (LSTM)	Reinforcement Learning	Data-Driven Cooling Control
Main use	Short-term load forecast	Non-linear load forecast	Scheduling and placement	Set-point selection
Data needed	Small	Large history	Simulator or long interaction	Dense sensor data
Interpretability	High	Low	Low	Medium
Adaptivity	Low	Medium	High	Medium to high
Risk in operation	Low	Low	Exploration risk	Thermal safety risk
Example references	[6]	[7], [9]	[8]	[12], [13]

V. EVALUATION METHODOLOGY

The framework can be evaluated through simulation. CloudSim provides models of hosts, VMs, power consumption and allocation policies and can be extended with prediction modules written in Java or connected to Python models [14]. Real workload traces should be used to train and test the predictors. Experiments should compare the AI-based policy with conventional baselines such as static threshold consolidation and round-robin placement. The metrics should include total energy consumption, average response time, host utilization, number of migrations, SLA violation rate and estimated carbon emission obtained by multiplying hourly energy by hourly carbon intensity [11]. For cooling, a simplified thermal model can relate IT load and set-point to cooling power, allowing the optimiser to be tested without physical equipment.

VI. CHALLENGES AND FUTURE DIRECTIONS

Several challenges limit the deployment of AI in data centres. Data quality is the first: missing sensor readings and inconsistent logs reduce model accuracy. Second, many learned models, especially deep networks and RL policies, are hard to interpret, which makes operators hesitant to let them control critical equipment [12]. Third, RL agents learn by exploration, and unsafe exploration in a real facility could cause overheating or service failure; safe RL and training in accurate simulators are therefore important [13].

Fourth, workloads change over time, so models must be retrained and monitored for drift. Fifth, the AI models themselves consume energy for training and inference, and this cost must be smaller than the savings they produce. Future work is likely to focus on lightweight models, transfer learning between facilities, explainable decision-making, and joint optimisation of computing, cooling and on-site renewable energy.

VII. CONCLUSION

This review has shown how artificial intelligence can improve the efficiency of cloud data centres. Statistical and neural predictors forecast workload, learned and heuristic schedulers place work on fewer servers and at cleaner times, and data-driven controllers reduce cooling energy. Reported results, including large cooling savings in production facilities, show that the approach is practical. The proposed framework combines these elements in a closed loop with safety rules and feedback. Its evaluation through simulation, using energy, response time, utilization and carbon emission as metrics, offers a clear route for further academic study of AI-driven green cloud computing.

REFERENCES

1. E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," Science, vol. 367, no. 6481, pp. 984–986, 2020.

© 2026 IJSRET

2. M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 732–794, 2016.
3. J. Gao, "Machine learning applications for data center optimization," Google, Mountain View, CA, USA, White Paper, 2014.
4. L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," Computer, vol. 40, no. 12, pp. 33–37, 2007.
5. E. Cortez, A. Bonde, A. Muzio, M. Russinovich, M. Fontoura, and R. Bianchini, "Resource Central: Understanding and predicting workloads for improved resource management in large cloud platforms," in Proc. 26th ACM Symp. Operating Systems Principles (SOSP), 2017, pp 153–167.
6. R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload prediction using ARIMA model and its impact on cloud applications' QoS," IEEE Transactions on Cloud Computing, vol. 3, no. 4, pp. 449–458, 2015.
7. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
8. H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in Proc. 15th ACM Workshop on Hot Topics in Networks (HotNets), 2016, pp. 50–56.
9. J. Kumar and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," Future Generation Computer Systems, vol. 81, pp. 41–52, 2018.
10. A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience, vol. 24, no. 13, pp. 1397–1420, 2012.
11. A. Radovanović et al., "Carbon-aware computing for datacenters," IEEE Transactions on Power Systems, vol. 38, no. 2, pp. 1270–1280, 2023.
12. R. Evans and J. Gao, "DeepMind AI reduces Google data centre cooling bill by 40%," DeepMind Blog, Jul. 2016.
13. Y. Li, Y. Wen, D. Tao, and K. Guan, "Transforming cooling optimization for green data center via deep reinforcement learning," IEEE Transactions on Cybernetics, vol. 50, no. 5, pp. 2002–2013, 2020.
14. R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and Experience, vol. 41, no. 1, pp. 23–50, 2011.

© 2026 IJSRET