International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

A Comprehensive Review of Energy Optimization Strategies for Sustainable Green Computing

Chandana H M[1], *Ramesh B*[2]
[1,2]*Department of CSE, Malnad College of Engineering, Hassan, Karnataka 573202, India.*
Emails: *chandana.hsn@gmail.com* [1], *sanchara@gmail.com* [2]

Abstract
Rapid growth of cloud platforms and large-scale data centers has placed extreme pressure on global power grids and has heavily amplified carbon footprints globally. This literature review delivers a step by evaluation of energy-conservation strategy for green computing, organized across five complementary domains: (1) containerized workload management lever-aging Kubernetes combined with nature-inspired solvers such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Dynamic Voltage and Frequency Scaling (DVFS); (2) data-driven energy forecasting using Long Short-Term Memory (LSTM), Transformer, and ensemble neural architectures; (3) thermodynamic optimization of cooling subsystems via metaheuristic search, multi-agent reinforcement learning, and intelligent thermal control; (4) power management at edge data centers through demand aggregation, workload migration, and online Lyapunov-based decision-making; and (5) clean energy sourcing via combined photovoltaic–wind–hydrogen storage topologies. The contribution of this review is synthesized from ten IEEE papers the survey identifies recurring methodological patterns, outstanding technical constraints, and pathways aiming for a cohesive structure for sustainability, autonomously managed infrastructure, self-managing cloud and cloud frameworks. Our comprehensive literature review indicates a lack of research combining high fidelity energy modelling with adaptive control within a single closed - loop framework. To mitigate this limitation, we propose a hybrid architecture combining an LSTM network for multi horizon energy demand forecasting with a Deep Reinforcement Learning agent for real time resource optimization. Overall, the examined works report energy reductions of 9.2% to 45%, quantifiable Power Usage Effectiveness (PUE) improvements, and significant CO2 cuts, validating the impactful potential of algorithmic and systems level innovations for environmental sustainability.

Keywords: *Containerization; DVFS; Green Computing; Kubernetes; PSO;*

1. Introduction

Global reliance on digital infrastructure has propelled data center construction at a pace that strains energy supply net-works. Estimates place current data center electricity consumption at roughly 2% of global demand, and that share is widely expected to grow sharply as artificial intelligence, cloud-native applications, large-scale analytics, and Internet-of-Things (IoT) services mature [1, 5]. AI model training is a particularly acute driver: projections indicate that AI-related power demand within data centers may climb by 160% by 2030 [7], with individual frontier-model training campaigns already exceeding hundreds of megawatt-hours. Green computing addresses this challenge by embedding environmental responsibility into every phase of a computing system's life cycle, from design and fabrication through operation and eventual decommissioning. When applied to data center infrastructure, the discipline converges on three mutually reinforcing objectives: curtailing the absolute energy drawn by IT hardware, raising the efficiency of ancillary systems (especially cooling, which routinely exceeds 40% of total facility consumption), and displacing fossil-derived power with renewable alternatives [5, 10]. The Power Usage Effectiveness (PUE) metric, defined as the ratio of total facility power to IT equipment power, has become the de facto industry benchmark for data center en-ergy efficiency [10]. A PUE of 1.0 represents ideal efficiency, whereas values exceeding

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

2.0 indicate that more than half of all consumed energy is dedicated to non-IT overhead. State-of-the-art hyper scale facilities have pushed PUE toward 1.1, while the global average remains closer to 1.5 [5, 9]. The energy landscape is further complicated by the rapid adoption of containerized workloads and micro service architectures. Container orchestration platforms such as Kubernetes enable fine-grained scheduling flexibility but simultaneously require sophisticated energy-aware policies to prevent resource fragmentation and idle-power waste [1]. Complementary to scheduling strategies, predictive energy modeling using deep neural networks offers the ability to anticipate demand shifts before they occur, enabling pre-emptive reconfiguration rather than reactive correction [2, 3]. The integration of renewable energy sources introduces additional complexity. Solar and wind generation are inherently intermittent, requiring storage buffers and intelligent dispatch strategies to ensure uninterrupted service availability [8, 7]. Hydrogen-based long-duration storage, combined with photovoltaic and wind generation, represents a promising pathway to near-zero-carbon data center operation, though significant optimization challenges remain in system sizing and energy management [8]. The remainder of this paper is structured as follows. Section 2 presents the Literature Survey. Section 3 introduces a taxonomy of energy optimization domains. Sections 4–9 review the surveyed literature in detail. Section 10 draws cross-cutting insights. Section 11 presents the Research Gap and Gap Analysis. Section 12 details the proposed method-ology. Section 13 concludes with a future outlook.

2. Literature Survey

The body of research addressing energy optimization in data centers and green computing spans multiple disciplines, including distributed systems, machine learning, thermal engineering, and power systems. This section reviews the ten IEEE-published works selected for this survey, presenting their motivations, methodologies, and principal findings. The reviewed contributions are organized thematically across five domains: cloud orchestration and containerization, predictive energy modeling, cooling system optimization, edge data center management, AI workload scheduling, and renewable energy integration

2.1. Cloud Orchestration and Containerization

Beena et al. [1] present a holistic orchestration frame-work that merges Docker-based containerization with Kubernetes cluster management and three complementary optimization algorithms: ACO, PSO, and DVFS. The work is motivated by the observation that conventional static re-source allocation leads to significant idle-power waste in multi-tenant cloud environments. Two novel components underpin the proposed architecture. The Adaptive Resource Adjustment Algorithm (ARAA) employs the Kubernetes Metrics API to continuously sample CPU and memory occupancy at sub-second granularity, enabling reactive scaling decisions. The Automated Resource-Aware Kubernetes Life-cycle Management (ARKLM) component orchestrates container lifecycle events—including spawning, migration, and termination—while incorporating real-time carbon-intensity signals and economic cost constraints. The ACO algorithm guides task-to-node assignment by treating energy consumption as pheromone-weighted path costs, while PSO optimizes resource allocation across nodes by treating each node con-figuration as a particle in search space. DVFS further reduces dynamic power by modulating supply voltage and clock frequency in proportion to instantaneous computational demand. Experimental results conducted on a representative Kubernetes cluster demonstrate that DVFS achieves the lowest average CPU occupancy at 0.94%, while ACO attains the fastest task completion time of 31.4 seconds. The authors acknowledge that evaluation on a single node cluster limits generalizability to multi-cloud and heterogeneous infrastructure deployments. Badhan et al. [5] provide a comprehensive survey of energy efficiency strategies in cloud data centers, consolidating findings related to virtual machine (VM) placement, consolidation policies, DVFS mechanisms, and renewable energy coupling. The survey synthesizes evidence that coordinated VM consolidation paired with DVFS or energy-aware schedulers typically yields power reductions in the range of 10–40%. The authors additionally propose DRACON, a conceptual cloud-edge hybrid architecture that projects a 45% energy reduction and

IRJAEH
International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

a 98% task completion rate through the combination of demand-response signals, adaptive consolidation, and renewable supply integration. The authors explicitly note that DRACON remains a simulation-derived proposal without empirical validation on physical infrastructure, and they identify this as a priority for future investigation.

2.2. Predictive Energy Modeling via Deep Learning

Zhang et al. [2] develop a combined Attentional-LSTM (A-LSTM) and Cutting Plane Method (CPM) pipeline for data center energy consumption prediction and allocation optimization. The motivation arises from the inadequacy of conventional LSTM models in distinguishing between time steps of differing energetic relevance within a long historical window. The A-LSTM extends standard LSTM gating with a soft-attention mechanism that assigns scalar weights to each hidden state, enabling selective amplification of the most informative temporal context within a 48-step (4-hour) prediction horizon. The CPM handles nonlinear feasibility constraints in the allocation subproblem through iterative linearization, reaching near-optimal solutions 23% faster than conventional LP solvers. The framework is validated on three datasets: the ASDC-EnergyTrace benchmark, Microsoft Azure VM traces, and a physics-informed synthetic benchmark. The A-LSTM attains a mean absolute error of 10.9 kWh and delivers a 10.3% reduction in operational energy cost. The authors identify limited interpretability of the attention weights and single data center validation as the principal limitations. Zhang et al. [3] argue that heterogeneous power signals in large-scale data centers resist adequate representation by any single neural architecture, motivating a multi-model ensemble that assigns complementary roles to three architectures. A deep neural network extracts nonlinear feature interactions from static configuration and workload metadata. An LSTM module tracks short-horizon temporal dynamics in power telemetry. A Transformer module captures broad temporal dependencies through multi-head self-attention over extended historical windows. The three outputs are fused through a learned weighted combination. When paired with a reinforcement learning scheduler, the system achieves an R2 of 0.966, reduces PUE from 1.52 to 1.35 (a 14.7% energy saving), and accumulates a 3,961 kg CO2 reduction over a 10-day operational test.

2.3. Cooling System Optimization

Jiang et al. [6] address the energy burden imposed by Computer Room Air Handler (CRAH) systems by constructing a whole-chain thermodynamic model that captures power draw across all cooling subsystem components, including CRAH fans, chilled-water pumps, chillers, cooling-tower fans, and condenser-water pumps. The Enhanced Whale Optimization Algorithm (E-WOA) is proposed as the solver, modifying baseline WOA with a nonlinear time-adaptive inertia weight and a differential-evolution mutation step to improve convergence speed and escape local optima. Deployed on operational data collected from a Shanghai-based data center facility, E-WOA reduces aggregate cooling power from 1,206.6 kW to 1,095.85 kW, achieving a 9.2% hourly energy saving. Li and Cen [9] survey multi-agent deep reinforcement learning (MADRL) applied to data center HVAC governance, reviewing architectures including MA-CWSC, MADDPG, MAPPO, and LLM-assisted control. The MA-CWSC architecture deploys five autonomous DQN agents that govern cooling-tower fan speed, condenser-water pump frequency, and chiller load concurrently, with a shared reward signal defined by the Coefficient of Performance. Comparative results demonstrate that MA-CWSC outperforms PID control while achieving approximately five times faster convergence per training iteration. Naganandhini et al. [10] deliver a comparative examination of cooling paradigms including air-based, direct-to-chip liquid, immersion, free cooling, evaporative, and AI-guided dynamic cooling management. Direct-to-chip liquid cooling, phase-change materials (PCMs), and AI-guided dynamic control emerge as most capable of pushing PUE toward the theoretical lower bound of 1.0.

2.4. Edge Data Center Energy Management

Liu et al. [4] investigate energy management complexities at edge data centers, where high service-demand volatility, physical proximity to end users, limited footprint, and tight electrothermal coupling create distinct operational challenges. The

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

proposed approach combines load aggregation, workload migration, and online Lyapunov-based decision-making. A temperature-aware migration heuristic relocates in-flight workloads when host CPU temperature surpasses a threshold of 75°C, using an RC thermal model to predict temperature trajectories. The resulting online algorithm, solved via Generalized Benders Decomposition (GBD), reduces operating expenditure by 20% compared to migration alone and by 40% compared to aggregation alone over 500 scheduling intervals using Google cluster traces.

2.5. AI Workload Scheduling

Melnatami Krishnaram [7] investigates the energy challenges posed by large-scale AI workloads, characterizing training energy footprints of Llama-2 (147 MWh), GPT-3 (1,287 MWh), and Bloom (433 MWh). The proposed scheduler profiles each job along four dimensions—numerical precision, data modality, software framework, and resource footprint—and assigns workloads to low, medium, or high-energy tiers. High-intensity computations are deferred to off-peak grid windows with 15-minute checkpoint transitions. At a 0.69% peak utilization target, the framework realizes a 12.8% wall-clock efficiency gain and avoids 364 MWh of peak-period consumption.

2.6. Renewable Energy Integration

Ettahri et al. [8] optimize a hybrid energy system comprising PV arrays, wind turbines, lead-acid battery banks, and a hydrogen subsystem including an electrolyzer, storage tanks, and a PEM fuel cell. PSO is applied to determine four decision variables: PV rated power, wind turbine count, battery autonomy duration, and hydrogen tank count, with the objective of minimizing Total Annualized Cost using the Capital Recovery Factor. The system is calibrated on 8,760 hours of meteorological data from Meknes, Morocco. The optimized configuration achieves a levelized cost of energy of $0.906/kWh and a 70.75% renewable energy fraction.

3. Taxonomy of Energy Optimization Domains

3.1. Cloud Orchestration and Containerization

This domain addresses software-driven governance of workloads on virtualized or containerized infrastructure [1, 5]. The primary levers include real-time resource provisioning, horizontal and vertical autoscaling, workload consolidation, and scheduling guided by bio-inspired or machine-learning algorithms.

3.2. Predictive Energy Modeling via Deep Learning

Reliable prediction of near-future energy demand enables pre-emptive resource reconfiguration [2, 3]. The domain covers LSTM, Gated Recurrent Unit (GRU), Transformer, attention-augmented networks, and hybrid ensembles applied to multivariate time-series power telemetry.

3.3. Cooling System Optimization

Thermal management represents the dominant non-IT energy expenditure in most facilities, routinely consuming 30–40% of total power [6, 9, 10]. Optimization efforts target chiller operating points, airflow distribution, precision air-conditioning control variables, and hot/cold-aisle contain-ment geometries.

3.4. Edge Data Center Energy Management

Edge deployments introduce distinct challenges: high service-demand volatility, physical proximity to end users, limited footprint, and tight electrothermal coupling [4]. This domain encompasses demand-side aggregation, inter-host workload relocation, and real-time stochastic optimization under uncertainty.

3.5. Renewable Energy Integration

Reducing the carbon intensity of data center power re-quires direct coupling with photovoltaic, wind, and storage subsystems [8, 7]. The inherent intermittency of renewable sources demands intelligent energy management systems capable of balancing supply uncertainty against firm service-level commitments.

4. Cloud Orchestration and Con-Tainerization

4.1. Adaptive Energy Optimization via Kubernetes

Beena et al. [1] present a holistic orchestration framework that merges Docker-based containerization with Kubernetes cluster management and three complementary optimization algorithms: ACO, PSO, and DVFS. Two novel components MAE of 10.9 kWh and delivers a 10.3% reduction in operational energy cost.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

4.2. Multi-Model Fusion: DNN–LSTM–Transformer

Zhang et al. [3] argue that heterogeneous power signals resist adequate representation by any single neural architecture. Their ensemble assigns complementary roles: a DNN extracts nonlinear feature interactions; an LSTM tracks short-horizon dynamics; and a Transformer captures broad temporal dependencies through multi-head self-attention: The fused architecture achieves RMSE of 4.99 kW, MAE of 3.03 kW, and R2 = 0.966. When paired with a reinforcement learning scheduler, average PUE improves from 1.52 to 1.35—a 14.7% energy saving with a cumulative 3,961 kg CO2 reduction over a 10-day operational test COOLING SYSTEM OPTIMIZATION

4.2.1. Precision Air Conditioning via Enhanced Whale Optimization

Jiang et al. [6] tackle the energy burden imposed by CRAH systems by constructing a whole-chain model capturing power draw across all cooling components. The unified objective is:

$$\min P = \sum P_{fan,CRAH} + \sum P_{p,CW} + \sum P_{chiller} + \sum P_{p,CG} + \sum P_{fan,tower} \quad (5)$$

E-WOA reduces aggregate cooling power from 1,206.6 kW to 1,095.85 kW—a 9.2% hourly saving.

4.2.2. Multi-Agent Reinforcement Learning for HVAC

Li and Cen [9] survey MADRL in data center HVAC governance. The MA-CWSC architecture deploys five autonomous DQN agents governing cooling-tower fan speed, condenser-water pump frequency, and chiller load concurrently, with a shared reward maximizing overall Coefficient of Performance:

4.2.3. The MA-CWSC architecture

outperforms conventional PID control and achieves approximately five times faster convergence per training iteration Comprehensive Cooling Strategy Analysis. Naganandhini et al. [10] deliver a comparative examination of cooling paradigms, evaluating each against Power Usage Effectiveness: Direct-to-chip liquid cooling, phase-change materials (PCMs), and AI-guided dynamic cooling management emerge as most capable of pushing PUE toward 1.0. emerge as most capable of pushing PUE toward 1.0.

5. Edge Data Center Energy Manage-Ment

Liu et al. [4] investigate energy management at edge data centers via a temperature-aware migration heuristic. When host CPU temperature Ticpu surpasses K = 75 ◦C, in-flight workloads are relocated to cooler peers. Processor tempera-ture evolves according to an RC thermal model: from the comparative analysis.

5.1. Algorithm–Orchestration Integration

Every cloud-oriented contribution concludes that optimization intelligence must be embedded within orchestration layers. The ARAA/ARKLM framework [1], the A-LSTM+CPM pipeline [2], and the RL-augmented scheduler [3] each exemplify tight, closed-loop coupling between prediction and the control plane.

T cpu(t) = Pi(t)R2 +T air(t)+ T ini −Pi(t)R2 −T air(t) e−Δt/(R2C2)B. Multi-Objective Balancing (8)

The long-horizon cost minimization is decomposed via Lyapunov optimization into tractable per-slot sub-problems requiring no knowledge of future states. The resulting online algorithm, solved via Generalized Benders Decomposition (GBD), reduces operating expenditure by 20% versus migration alone and 40% versus aggregation alone.

6. AI Workload Scheduling for Sustain-Ability

Melnatami Krishnaram [7] investigates energy challenges posed by large-scale AI workloads, characterizing training footprints of Llama-2 (147 MWh), GPT-3 (1,287 MWh), and Bloom (433 MWh). The proposed scheduler profiles each job by numerical precision, data modality, software frame-work, and resource footprint, assigning workloads to low, medium, or high-energy tiers. High-intensity computations are deferred to off-peak grid windows with 15-minute check-point transitions. At a 0.69% peak utilization target, the framework realizes a 12.8% wall-clock efficiency gain and avoids 364 MWh of peak-period consumption.

7. Renewable Energy Integration

Ettahri et al. [8] optimize a hybrid energy system comprising PV arrays, wind turbines, lead-acid battery banks, and a hydrogen subsystem

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4825-4833
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0635

(electrolyzer, storage tanks, PEM fuel cell). PSO determines four decision variables by minimizing Total Annualized Cost using Capital Recovery Factor:

$$CRF(i, N) = \frac{i(1+i)N}{(1+i)N - 1} \quad (9)$$

with standard PSO update rules:

$$v_i(t+1) = wv_i(t) + c_1r_1\,P_{best,i} - x_i(t) + c_2r_2\,G_{best} - x_i(t) \quad (10)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (11)$$

Calibrated on 8,760 h of meteorological data from Meknes, Morocco, the optimized configuration achieves an LCOE of \$0.906/kWh and a 70.75% renewable energy fraction.

8. Synthesis and Comparative Analysis

Table 1 consolidates the principal attributes and outcomes of the ten reviewed works. Five recurring themes emerge. None of the surveyed works optimizes energy in isolation. Each framework explicitly balances computational performance, thermal constraints, and economic objectives. The Lyapunov-based formulation in [4] formalizes this trade-off through a tunable coefficient V that interpolates between cost aggressiveness and constraint conservatism.

8.1. Cooling as a First-Class Optimization Target

A persistent finding is that cooling infrastructure consumes upwards of 40% of total facility power [5, 6, 9, 10]. The 9.2% saving reported by [6] and the PID-outperforming MADRL agents surveyed in [9] collectively reinforce cooling as a high-leverage intervention point.

8.2. Predictive–Prescriptive Pipelines

Works [2] and [3] establish a reusable two-stage pattern: high-fidelity forecasting (LSTM- or Transformer-based) feeding into a prescriptive optimizer (Cutting Plane or reinforcement learning). The modular separation simplifies maintenance and permits independent component upgrades as model architectures evolve.

8.3. Sustainability Metrics Beyond PUE

More recent contributions demonstrate a matured perspective on sustainability accounting. Carbon emission tracking [3], renewable energy fraction [8], grid-carbon-intensity signaling [1], and Water Usage Effectiveness (WUE) [3] are increasingly reported alongside PUE.

9. Research Gap and Gap Analysis

Although substantial progress has been reported, the existing body of literature still exhibits several critical unresolved limitations that collectively hinder the realization of fully autonomous, end-to-end sustainable green computing infrastructures [1–10]. Absence of Closed-Loop Prediction and Control Integration (G1). A fundamental deficiency is the disconnection between energy forecasting and real-time control decision-making [2, 1, 5]. Predictive models developed by Zhang et al. [2] and Zhang et al. [3] produce accurate multi-horizon energy forecasts but operate as standalone modules whose outputs are not fed back into adaptive control loops. Similarly, the Kubernetes-based orchestration framework of Beena et al. [1] and the VM consolidation strategies of Badhan et apply scheduling heuristics without continuous refinement driven by forecasting accuracy metrics. No existing contribution establishes a unified closed-loop architecture where prediction and optimization mutually reinforce each other at every scheduling interval.

9.1. Reactive Rather Than Anticipatory Resource Management (G2)

Resource management decisions in the majority of surveyed works are event-triggered rather than forecast-driven [1, 4]. The orchestration framework of Beena et al. [1] initiates scaling actions when CPU or memory thresholds are breached, and the temperature-aware migration heuristic of Liu et al. [4] relocates workloads only after thermal limits are exceeded. Although [2] and [3] produce forward-looking demand forecasts, these predictions are not embedded as anticipatory state features within any scheduling or resource control agent.

9.2. Disjoint Optimization of Compute and Cooling Systems (G3)

Cooling infrastructure consistently accounts for 30–40% of total facility energy consumption [6, 9, 10], yet compute-layer and cooling-layer optimization remain entirely decoupled [6, 9, 10, 4]. The E-WOA-based CRAH optimizer of Jiang et al. [6] minimizes cooling power against static thermal loads without awareness of concurrent workload placement decisions. The MADRL-based HVAC controllers surveyed by Li and Cen [9] treat cooling as an isolated control problem independent of IT scheduling.

9.3. Limited Renewable-Aware Real-Time Workload Dispatch (G4)

Renewable energy integration operates exclusively at the infrastructure planning and static sizing level rather than as a dynamic operational decision [8, 7]. Ettahri et al. [8] apply PSO to optimize rated capacities based on annual meteorological averages without providing a real-time energy management layer. Melnatami Krishnaram [7] incorporates grid availability as a coarse scheduling constraint but does not couple decisions with live renewable surplus or battery state-of-charge signals.

9.4. Scalability Limitations and Lack of Multi-Environment Validation (G5)

Most proposed frameworks have been validated under controlled or single-site conditions [1, 5, 4]. Beena et al. [1] acknowledge that their single-node cluster evaluation may not generalize to multi-cloud configurations. The DRACON architecture of Badhan et al. [5] is simulation-derived with-out empirical validation on physical infrastructure.

9.5. Absence of Standardized Multi-Metric Bench-marking (G6)

A recurring methodological weakness is the use of heterogeneous and incompatible evaluation metrics [1–10]. Energy savings are variously reported as absolute kWh reductions [2], percentage improvements [1, 5, 4], PUE improvements [3, 6], carbon mass reductions [3], levelized cost of energy [8], and wall-clock efficiency gains [7]. Incomplete Environmental Sustainability Assessment (G7). Environmental evaluation is predominantly confined to operational energy metrics, most commonly PUE [10, 3, 8, 1]. Broader sustainability indicators including embodied carbon, electronic waste, freshwater consumption [10], water usage effectiveness, and renewable penetration ratios [3, 8] are inconsistently incorporated.

9.6. Structured Gap Analysis Table

Table 2 summarizes the seven identified research gaps and their proposed remediation’s.

9.7. Thematic Clustering and Severity Assessment

The seven identified gaps cluster into three thematic groups. Prediction and control gaps (G1, G2) reflect the inability to tightly couple temporal demand forecasting with proactive resource control. Integration gaps (G3, G4) ex-pose the fragmented treatment of compute, cooling, and renewable systems as isolated problems despite strong physical interdependencies [4, 6, 8]. Evaluation and sustainability gaps (G5, G6, G7) reveal methodological weaknesses that undermine reproducibility and holistic environmental accountability [10, 5, 3]. G1 and G3 carry the highest potential impact. The con-sistent finding that cooling accounts for 30–40% of facility energy [6, 9, 10, 5], combined with demonstrated feasibility of LSTM-based demand prediction [2, 3] and DRL-based control [9, 3], indicates that a unified closed-loop architecture could yield compounded savings substantially beyond any single-domain work. G4 is of growing strategic urgency as carbon-reporting obligations tighten globally [8, 7].

10. Proposed Methodology

10.1. System Architecture Overview

The proposed framework couples two ML models in a closed-loop pipeline: (1) a stacked LSTM that continuously forecasts energy demand and thermal load, and (2) a Deep Reinforcement Learning (DRL) agent that translates those forecasts into real-time resource control actions. Operating at 15-minute scheduling intervals, the LSTM feeds its predictions as state features to the DRL agent, which selects control actions spanning VM scheduling, DVFS configuration, cooling setpoints, and renewable dispatch. Observed outcomes update both models: the LSTM via online fine-tuning and the DRL agent via experience replay. This closed-loop design directly addresses G1–G5.

10.2. LSTM-Based Energy Prediction

A two-layer stacked LSTM network is trained on historical server telemetry to forecast energy consumption and aggregate cooling load over a 1-hour horizon. The input feature vector includes server CPU utilization, memory utilization, network I/O, aggregate IT power draw, inlet air temperature, and renewable generation availability, collected at 1-

minute granularity over a sliding window of 60 steps. The first LSTM layer (128 units) captures short-term work-load spikes while the second layer (64 units) captures slower trending behaviour. A fully connected output head produces predicted energy demand and cooling load at horizons of 1 h and 3 h ahead. The model is trained using Mean Squared Error (MSE) loss with the Adam optimizer and is fine-tuned online on a rolling window of the 48 most recent observations.

10.3. DRL-Based Resource Optimization

The DRL agent learns a control policy mapping the observed system state to resource allocation actions to minimize energy expenditure while satisfying thermal constraints and service-level agreements, formulated as a Markov Decision Process.

State, Action, and Reward

The state vector concatenates current IT utilization metrics (per-server CPU, memory, and network load), thermal telemetry (CPU inlet temperatures and current PUE), renew-able storage state-of-charge, and the LSTM-predicted energy demand and cooling load for the next 1 h and 3 h. Including forecast values as state features enables pre-emptive action selection, directly addressing G2. The composite action covers four coupled subsystems: VM placement and live migration decisions, per-server CPU frequency and voltage scaling via DVFS, CRAH supply temperature and cooling-tower fan speed adjustments, and grid-versus-battery power dispatch fractions. Governing IT scheduling and cooling jointly addresses G3 and G4. The scalar reward balances: minimizing total facility energy consumption, maintaining SLA compliance, maximizing renewable energy utilization, and penalizing CPU over-heating above 75°C.

Training

The agent is trained using the Deep Q-Network (DQN) algorithm with experience replay and a periodically updated target network. The network consists of two fully connected hidden layers (256 and 128 units with ReLU activations) and is retrained weekly on data collected from live deployment.

10.4. Pipeline Integration

At each 15-minute interval the pipeline executes three sequential steps.

- **Step 1 — Predict:** The LSTM ingests the latest 60-step observation window and outputs energy and thermal load forecasts for the next 1 h and 3 h ahead.
- **Step 2 — Decide:** The DRL agent observes the combined state vector and selects the resource control action via the trained Q-network.
- **Step 3 — Act and Feedback:** The chosen action is dis-patched to the Kubernetes control plane, DVFS governors, and building management system. Observed outcomes are logged and returned to both models for continuous learning.

10.5. Expected Outcomes and Evaluation Protocol

The framework will be evaluated on Google Cluster Traces augmented with synthetic renewable generation profiles. Performance will be reported across five metrics: MAE (kWh), R2, total energy reduction (%), PUE improvement, and SLA compliance rate (%), directly addressing G6. The framework is anticipated to achieve energy reductions of 20–30% relative to reactive baselines, with PUE improvements toward 1.2–1.3 and SLA compliance above 95% under peak load conditions.

Conclusion

This survey examined ten recent IEEE-published works spanning the core domains of data center energy optimization for green computing: bio-inspired cloud orchestration [1, 5], deep-learning-driven energy forecasting [2, 3], precision cooling control [6, 9, 10], edge data center power management [4], AI-workload scheduling [7], and renewable energy integration [8]. Across these domains, the reviewed literature collectively demonstrates that meaningful energy reductions—from 9.2% in precision air conditioning [6] to 45% in integrated cloud frameworks [5]—are attainable through algorithmic progress applied at multiple infrastructure levels. The highest-impact contributions share a common architectural philosophy: tight coupling of optimization intelligence with orchestration layers, predictive-prescriptive pipeline designs, electrothermal-aware resource management, and co-design with renewable supply chains. The structured gap analysis identified seven cross-cutting deficiencies, clustered into prediction and control gaps (G1, G2), integration gaps (G3, G4), and evaluation and sustain-ability gaps (G5, G6, G7). The most critical are the absence of a closed-loop prediction–control architecture (G1) and the disjoint treatment of cooling and compute optimization (G3). The proposed methodology addresses these gaps through a stacked LSTM coupled with a DQN-based DRL agent whose unified action space simultaneously governs VM scheduling, DVFS setpoints, cooling controls, and renewable dispatch. Closing the identified research gaps will demand sustained interdisciplinary collaboration bridging machine learning, systems engineering, thermal engineering, and environmental sustainability science. Achieving industry sustainability targets requires not a reduction in computing capability but a fundamental elevation in the energy intelligence embedded within computing infrastructure.

References

[1]. B. M. Beena, P. C. Ranga, V. Holimath, S. Sridhar, S. S. Kamble, S. P. Shendre, and M. Y. Priya, “Adaptive energy optimization in cloud computing through containerization,” IEEE Access, vol. 13, pp. 159548–159565, Sep. 2025.

[2]. Z. Zhang, W. Huang, W. Li, Y. Zhang, Z. Shao, Z. Lian, and Y. Yang, “Data center energy consumption prediction and optimization based on attentional-LSTM and cutting plane method,” in Proc. 10th Int. Conf. Cyber Security and Information Engineering (ICCSIE), 2025, pp. 594–598.

[3]. X. Zhang, D. Niu, X. Liu, B. Deng, and Q. Ai, “Deep learning-based data centre energy efficiency analysis and optimisation research,” in Proc. 7th Int. Conf. Energy Systems and Electrical Power (ICESEP), 2025, pp. 1032–1036.

[4]. C. Liu, X. Zhang, W. Gao, W. Peng, and M. Chen, “Energy con-sumption optimization of edge data centers based on load aggregation and computing power migration,” in Proc. 6th Int. Conf. Advanced Electrical and Energy Systems (AEES), 2025, pp. 933–938.

[5]. A. Badhan, R. Kaur, P. Arora, and R. Garg, “Energy efficient cloud computing: Strategies for reducing data center power consumption,” in Proc. 3rd Int. Conf. Augmented Intelligence and Sustainable Systems (ICAISS), 2025, pp. 1156–1162

[6]. C. Jiang, R. Shao, and J. Yang, “Energy-saving optimization of preci-sion air conditioning system in data center based on E-WOA algorithm,” in Proc. 8th Int. Conf. Electrical, Mechanical and Computer Engi-neering (ICEMCE), 2024, pp. 26–31.

[7]. P. C. Melnatami Krishnaram, “Green AI: Optimizing energy efficiency of workloads for sustainable data centers,” in Proc. IEEE Green Technologies Conference (GreenTech), 2025, pp. 148–152.

[8]. H. Ettahri, H. Gualous, J. Sabor, and M. F. Yaden, “Optimization and management of hybrid renewable energy system using PSO algorithm: Data center application,” in Proc. 4th Int. Conf. Innovative Research in Applied Science, Engineering and Technology (IRASET), 2024, pp. 1–6.

[9]. Y. Li and J. Cen, “Research on energy-saving and optimization control of data center cooling systems,” in Proc. 7th Int. Conf. Internet of Things, Automation and Artificial Intelligence (IoTAAI), 2025, pp. 524–527.

[10]. S. Naganandhini, C. Sundar, G. Venkatesan, and R. Sathya, “Towards energy-efficient data centres: A comprehensive analysis of cooling strategies for maximizing efficiency and sustainability,” in Proc. Int. Conf. Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), 2023, pp. 1–7.