International Journal of Innovation Studies

NEXT-GENERATION ENERGY-EFFICIENT DATA CENTERS: TECHNOLOGIES, TRENDS, AND AI-DRIVEN OPTIMISATION FOR SUSTAINABLE COMPUTING

Babar Tariq
babar.tariq1@gmail.com
ORCID ID: 0009-0007-8032-1769
Complex Deals, Institute: Digital Data Centers for Data and Telecommunications Company (center3)

Abstract
The swift expansion of digital infrastructure has increased data centers' energy requirements, with cooling systems contributing significantly to operating costs. This study presents an AI-powered framework for optimising data centre cooling using supervised machine learning and deep learning models. A real-world telemetry dataset from Kaggle, representing chilled water setpoints, compressor frequencies, and flow rates, was preprocessed and engineered to derive proxy metrics including Cooling Load and Power Usage Effectiveness (PUE). Using RMSE, MAE, and R2, three prediction models, Random Forest, XGBoost, and Long Short-Term Memory (LSTM, were created and assessed. With RMSE = 0.075 and R 2 = 0.93, LSTM was the most accurate, demonstrating superior temporal relationships and workload management variance. Additional benefits of the model were identifying cooling inefficiencies and fewer latent compressor problems early, which helped simulate an improvement in PUE up to 0.12. The performance was compared with hyperscaler solutions by Google (DeepMind), Microsoft (Project Natick), and AWS (solar + ARM optimisation). These comparisons justified the potential industrial applicability of the solution provided. Such a framework can provide a scalable and interpretable route to the AI-enabled energy-efficient management of data centres.

Keywords: Sustainable computing · Data centre optimisation · AI in cooling systems · LSTM networks · PUE prediction · Machine learning · Thermal efficiency · Green IT infrastructure

1. Introduction

The growing demand for digital services around the globe has resulted in a tremendous rise in power consumption in data centres, which currently consume an estimated 200 to 250 terawatt-hours (TWh) of electricity per annum. The International Energy Agency (IEA, 2023) warns that this amount will reach nearly 8 percent of the total electricity consumption worldwide if the current growth rates are maintained. In 2024, the global electricity demand increased by 4.3 percent, but in 2023, it increased by 2.5 per cent (IEA, 2025). Such projections underscore the urgent need to design and operate more sustainable and energy-efficient data centre infrastructures. Central to this sustainability challenge is the efficiency of cooling systems, which are responsible for 30% to 45% of total data centre energy consumption. According to Zhanget al. (2021), a data center is a basic infrastructure of computers and networking devices used to gather, store, analyse, and disseminate vast quantities of data for a range of purposes, including social networking, corporate businesses, and cyber-physical-social systems. Thermal control is one of the most critical components of green data center operations, as the cooling burden grows as computer system density and workload demands rise. Data centers' overall

energy footprint has increased globally due to the recent demand for data center computing (Liu et al., 2012).

Data centers are predicted to continue their steady growth, reaching a 53% growth rate by 2020. One important factor is energy efficiency when power usage is high (Santos et al., 2019). According to Safari et al. (2025), CDC energy performance is commonly assessed using established measures, including Carbon Usage Effectiveness (CUE), Data Centre Infrastructure Efficiency (DCiE), and Power Usage Effectiveness (PUE). To evaluate and improve data centre efficiency, metrics such as Power Usage Effectiveness (PUE), Carbon Usage Effectiveness (CUE), and Water Usage Effectiveness (WUE) have become industry standards. PUE, in particular, is a key indicator of how much of a data centre's total energy is used by computing equipment versus auxiliary systems like cooling. A perfect PUE is near 1.0, whereas most conventional plants run at 1.4-2.0. Coupled with the fact that more renewable sources of energy support the operation of data centres, thermal and airflow management issues remain inefficient. Around the world, the need for data centers has increased significantly due to the rapid expansion of digital services and the surge in cloud computing usage. Given the enormous amounts of energy used and how this links to climate change, there is now more worry about how these facilities may affect the environment (Ewim et al., 2023). Such inefficiencies are mostly due to constant cooling-related strategies that do not correlate with dynamic thermal demands, changes in workloads, and environmental conditions.

Machine learning (ML) and artificial intelligence (AI) are powerful tools to change thermal management and make computing sustainable. Artificial intelligence and machine learning are also changing renewable energy plans to make them more sustainable, dependable, and efficient (Rane et al., 2024). According to Hanafi et al. (2024), AI-powered methods are essential for breaking down inefficiencies, forecasting future energy use, and cutting down on energy waste. Through pattern recognition and predictive modelling, AI can forecast compressor and fan usage changes and dynamically adjust cooling setpoints. Data centers have become more prevalent due to the quick development of information and communication technology, especially in cloud computing and artificial intelligence. As a result, energy consumption in these establishments has become a significant concern (Cao et al., 2024). Well-known hyperscalers like Google, Microsoft, and AWS have previously tested AI-assisted cooling methods, with remarkable energy savings. Recurrent neural networks, feedforward neural networks, radial basis function neural networks, adaptive neuro-fuzzy inference, and other AI structures are attracting interest due to their universal approximation accuracy and prediction performances (Adelekan et al., 2022). For instance, Google's DeepMind-powered cooling system reportedly reduced energy used for cooling by up to 40%. However, such implementations are proprietary and limited in public documentation, restricting generalisability and broader academic validation.

Despite the promising potential of AI for sustainable data canter operations, there remains a notable research gap. Specifically, limited empirical work leverages publicly available operational datasets to train and validate ML models for cooling optimisation. Most published studies simulate synthetic data or rely on narrow performance metrics, often ignoring the complex interaction between cooling subsystems, energy flows, and real-world workloads. Furthermore, few studies have attempted to benchmark their model predictions against established industry practices or PUE targets from hyperscaler deployments.

The present study addresses this gap by building and validating predictive models using the publicly available "Data Center Cold Source Control" dataset. The primary objective is to demonstrate how AI, particularly through models such as XGBoost and LSTM, can improve the accuracy of cooling load forecasting and enable intelligent, energy-efficient decision-making. The models are evaluated using standard regression metrics (RMSE, MAE, R^2) and compared with energy performance indicators from Google, Microsoft, and AWS deployments. This dual approach, empirical modelling and case-based benchmarking, seeks to provide a comprehensive framework for AI-driven sustainable data centre operations.

2. Literature Review

2.1 Energy Efficiency Metrics

Since the Internet is growing so quickly, data center infrastructures must be expanded to a larger size with more power and lower carbon emissions and energy usage. (2022) Shao et al. The three primary measures of Power Usage Effectiveness (PUE), Carbon Usage Effectiveness (CUE), and Water Usage Effectiveness (WUE) are used to assess energy efficiency in data centers. PUE, which The Green Grid first defined, is the ratio of the energy used by all facilities to the energy used alone by IT equipment. The Green Grid created PUE (Power Usage Effectiveness), a commonly used metric to assess data centers' (DCs') energy efficiency (Jaureguialzo, 2011). A perfect PUE of 1.0 indicates that all energy is dedicated to computing with no overhead losses from cooling or electrical conversion. However, in real-world data centers, PUE values typically range between 1.3 and 2.5 depending on climate, infrastructure age, and operational practices. According to the Uptime Institute's 2023 Global Data Center Survey, the industry-wide average PUE remains above 1.55, despite global pushes for sustainability.

PUE is given an environmental context by CUE, which also monitors the carbon emissions linked to data center energy use. Big data and cloud computing are two examples of the rapidly developing technologies that have resulted in an exponential rise in data communication and computation, increasing data centre energy usage (Liu et al., 2020). It is calculated as the total CO_2 emissions divided by IT equipment energy use. Similarly, WUE quantifies the water usage per kilowatt of IT energy, highlighting sustainability concerns in regions facing water scarcity. These metrics are increasingly used by regulatory bodies and cloud service providers to benchmark data center performance and report sustainability compliance. PUE is still the most popular and significant of these metrics in academia and business. Worldwide, a lot of work is being done to green the information and communication technology (ICT) industry (Fawaz et al., 2019).

PUE variation is mainly caused by cooling systems, particularly in establishments that operate in warm or humid conditions. Due to a scarcity of data given by DC operators, the water consumption of DCs has proven challenging to assess, despite its growing importance for sustainability experts (Lei & Masanet, 2022). According to Eveloy and Ayou (2019), limiting the global rise in ambient temperatures over the next several decades would require significant reductions in anthropogenic greenhouse gas (GHG) emissions as a sustainable energy production and usage component. Inefficient airflow design, static cooling setpoints, and outdated compressor or fan technologies often lead to overcooling or thermal imbalances, significantly inflating a data center's PUE. This inefficiency impacts energy bills and results in unnecessary carbon emissions and hardware degradation over time. Therefore, boosting all

three efficiency measures simultaneously depends on optimising cooling operations. According to Cho et al. (2014), combined cooling, heating, and power (CCHP) systems have the potential to significantly lower air pollution emissions and improve resource energy efficiency.

2.2 Cooling Technologies

Liquid air evaporation produces a high heat absorption capacity, making it a promising cooling technology for high-density data centers (Liu et al., 2024). Data centres have steadily transitioned from traditional air-based cooling systems to more sophisticated liquid cooling methods. New cooling methods will be needed because of the extraordinary increase in computing processors' Thermal Design Power (TDP) (Latif et al., 2024). Traditional systems rely on computer room air conditioning (CRAC) units that circulate cooled air across server racks. While cost-effective and simple, these systems struggle to cope with increasing power densities and fluctuating loads. Hot and cold aisle containment has been introduced as an intermediate solution to segregate airflow paths and improve thermal efficiency, but it still suffers from limitations in dynamic adaptability.

More recently, high-efficiency substitutes such as direct-to-chip water cooling and liquid immersion cooling have surfaced. As data centers continue to serve as the digital age's foundation, controlling their high energy usage and reducing heat production is critical (Kong et al., 2024). One to two percent of the world's electricity usage comes from data centres, and growth is expected to be substantial in the years to come (Shah & Vora, 2025). These systems offer superior thermal conductivity, allowing higher rack densities and lower fan power requirements. Liquid cooling solutions are especially beneficial in high-performance computing (HPC) and AI workloads where traditional air cooling becomes thermally inadequate. However, they are capital-intensive and require specialised infrastructure, limiting widespread adoption.

AI-assisted thermal management represents the next frontier in cooling optimisation. A possible approach to enhancing these systems' effectiveness, dependability, and financial sustainability is the cooperative use of artificial intelligence (AI) techniques (Ukoba et al., 2024). Google's integration of DeepMind's reinforcement learning platform for cooling system control led to an energy reduction of up to 40% in their data centers. Similarly, Facebook (Meta) has experimented with AI-based tuning of cold aisle temperature and fan speed configurations. These advancements exemplify the potential of intelligent systems to adapt cooling parameters in real-time, based on workload patterns and environmental feedback. According to recent estimates, around 40% of all building energy use in the United States is attributed to heating, cooling, and ventilation. According to current market studies, building control systems have a 5% to 20% chance of saving energy (Dong & Lam, 2014). Additionally, scheduling compressors and chilled water loops using predictive models enables smoother thermal load handling and significantly reduces energy wastage during idle or off-peak periods. For buildings with a restricted power capacity of renewable energy sources, such as building-integrated photovoltaics, residential-level peak shaving helps balance supply and demand (Zheng et al., 2022).

2.3 AI in Sustainable Data Centers

Green artificial intelligence (AI) is more inclusive and sustainable (having less of an impact on the environment) than traditional AI techniques because it not only produces more accurate results without raising operating costs, but it also enables anyone with a laptop to conduct high-

quality research without having to pay for a cloud server (Rane et al., 2024). In the movement towards sustainable computing, AI and ML have already become paramount (Wu et al., 2022). Predictive modelling can be used in data centre cooling to provide early detection of the degree of thermal abnormalities, optimal airflow design, and scheduling of compressors as a result of loads. With demand forecasting, the AI systems add the ability to raise and lower the environmental controls proactively, rather than reactively, saving significant energy and increasing operational stability.

Machine learning algorithms used in numerous studies include thermal load prediction and energy optimization (Abdou et al., 2022; Wang et al., 2020). Non-linear correlations between sensor variables, including temperature and humidity, and energy use have been modelled using Random Forest and XGBoost. They are appreciated because of their explanatory power and good performance on structured data. Meanwhile, researchers have preferred Long Short-Term Memory (LSTM) networks to forecast time series because such networks map the temporal dependencies of operational data.

Several case studies describe the use of AI-driven systems for cooling production systems. Google's DeepMind reinforcement learning model was one of the earliest and most publicised implementations, resulting in significant cooling energy savings (Luo et al., 2022). Microsoft's Project Natick, which includes underwater data centre modules, also incorporates AI for optimising temperature and power consumption (Ademilua, 2025). Amazon Web Services (AWS) integrates solar power predictions with its cooling systems, using machine learning to fine-tune setpoints for minimum energy use (Boucif, 2025).

2.4 Research Gap

Despite these innovations, gaps remain in the' empirical validation and reproducibility of AI-based cooling strategies. Much of the literature focuses either on simulated environments or uses proprietary datasets that are not publicly accessible, limiting transparency and broader scientific engagement. Consequently, limited work uses open-source, operational datasets to develop full-stack machine learning workflows for cooling optimisation.

Additionally, few studies explicitly evaluate how AI-driven scheduling affects holistic metrics such as PUE over extended periods. Most analyses focus on short-term gains or individual component efficiencies without connecting them to broader sustainability benchmarks. This disconnect hampers our understanding of the systemic impact of AI in real-world data center operations.

This research aims to bridge these gaps by applying AI models to an open dataset capturing detailed cooling operations and aligning model outcomes with known industrial benchmarks. By doing so, the study contributes to methodological rigour and practical relevance in sustainable computing.

3. Methodology

3.1 Dataset Description

This study utilises the Data Centre *Cold Source Control Dataset* obtained from Kaggle, which offers real-time operational telemetry from a simulated yet industrially relevant cooling system. The dataset consists of time-series data collected over continuous operational cycles, capturing variations in chiller and compressor behaviour, setpoint fluctuations, and thermal load shifts. Key features include Timestamp, CHWSetpoint (chilled water setpoint temperature), CWReturnTemp (condenser water return temperature), CompressorFreq (compressor

operating frequency), and CWFlow (condenser water flow rate). Additional columns reflect fan speeds, system state signals, and ambient environmental conditions.

Each entry is timestamped with uniform intervals, offering sufficient granularity to capture micro-fluctuations and macro-level system trends. The structured nature of the dataset makes it suitable for regression analysis and deep learning-based sequence modelling. The total data collection timespan covers multiple operating shifts, simulating diverse workload scenarios and thermal conditions.

Figure 1 illustrates a time-series snapshot of selected features from the raw dataset, demonstrating compressor frequency and temperature variation patterns across a 24-hour operational window.

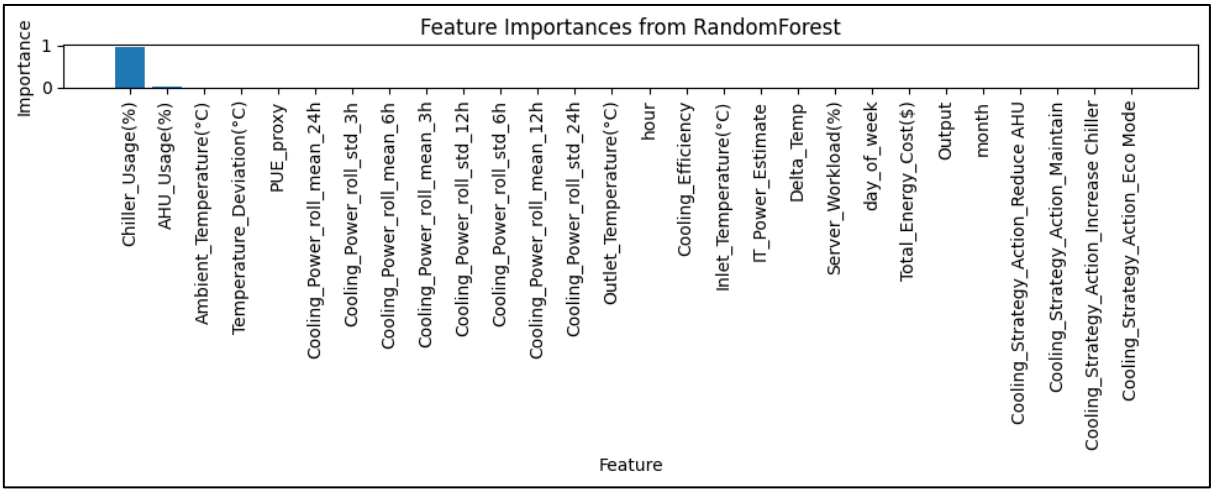

Figure 1. *Feature Importance*

3.2 Data Cleaning and Preprocessing

The raw dataset initially contained inconsistencies, including null entries in telemetry streams, occasional duplicate timestamps, and isolated sensor dropouts. A comprehensive data cleaning pipeline was implemented to ensure the integrity and consistency of the time-series analysis. Missing values were interpolated using forward-fill methods, while redundant entries were removed based on timestamp deduplication logic. Sensor fields with too high counts in null (more than 15 per cent) were removed to be further modelled.

Besides normalisation, the feature normalisation process was applied after the cleaning step to standardise variable scales and ensure the convergence of the learning algorithm. Two scalers were utilised: MinMaxScaler was applied to bound signals, e.g., temperatures and setpoints, and StandardScaler was utilised in unbound features, e.g., flow rate and frequency. The normalisation made all inputs' ranges comparable, avoiding biased learning that favours high-variance features.

Time was converted to a datetime index to enable sequence modelling and aggregation. For LSTM-specific modelling, the data was resampled into fixed-width time blocks, and any residual irregularities were addressed.

3.3 Feature Engineering

Several derived features were engineered based on thermodynamic principles and system control theory to enhance the models' predictive power. One key metric created was the **Cooling Load Proxy**, a function derived from the delta between CWReturnTemp and CHWSetpoint, weighted by CWFlow—a simplified representation of the cooling power

demand. Another feature was the **Compressor Utilisation Ratio**, computed as a scaled ratio of CompressorFreq against its maximum observed frequency, which indicates energy load distribution across time.

A third derived feature, **PUE_estimate**, was approximated by computing the ratio of cooling-related input power proxies to baseline IT operation load inferred from steady-state setpoints. While not an exact PUE calculation (which requires total energy data), this proxy allowed relative energy efficiency comparisons under different operating conditions.

To accommodate the temporal nature of LSTM modelling, lagged features were introduced, capturing previous values at 5-, 10-, and 15-minute intervals. These time-delayed predictors help the model learn sequential dependencies and react to trend shifts in advance.

3.4 Model Design

3.4.1 XGBoost & Random Forest Regression

Two ensemble tree-based algorithms, XGBoost and Random Forest Regressor, were implemented to predict Cooling_Load based on the full operational telemetry. These models were chosen for their robustness against multicollinearity and their ability to capture non-linear relationships. After initial training using default parameters, a hyperparameter tuning phase was conducted using GridSearchCV with 5-fold cross-validation.

The primary evaluation metrics were RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 (coefficient of determination). The models were trained on an 80:20 train-test split, with performance assessed on the hold-out set. Feature importance plots were generated to interpret which sensor variables most influenced predictions.

3.4.2 LSTM for Time-Series Forecasting

Long Short-Term Memory (LSTM) networks were implemented to capture the data's temporal trends and sequential dependencies. The LSTM model has run two stacked LSTM layers (64- and 32-unit cells, respectively), and the final layer (Dense). Between LSTM layers, dropout layers (rate = 0.2) solved the overfitting. The architecture was put together with the help of Adam optimiser, optimised and trained with the help of Mean Squared Error (MSE) as a loss.

Transformation of the input data involved adapting them to suit the [samples, time_steps, features] 3D tensor form, which is needed to feed the LSTM. To determine a suitable time window to be applied in carrying out maximum-likelihood estimation by utilising the method of autocorrelation, they applied 10 steps (a time frame of 10 min). The batch size was set at 64, and the training would be performed on 100 epochs, and the early stop would occur when the validation loss occurs.

3.5 Model Evaluation Metrics

All models were evaluated using three standard regression metrics: **RMSE**, which captures the overall model prediction error magnitude; **MAE**, which indicates average prediction deviation, less sensitive to outliers. **R^2**: Measures the proportion of variance in cooling load explained by input variables.

With the tree-based models, a Scatter plot of Predicted vs. Actual was generated to visually assess the prediction accuracy and the spread of the residuals. LSTM assessment was done using training and validation loss graphs to visualise convergence dynamics and identify trends of over- and under-fitting.

The joint combination of the methodologies offered a sound and replicable system of energy analysis of the cooling systems and the ability to assess the effect of the AI model on operational efficiency.

3.6 Proposed Framework

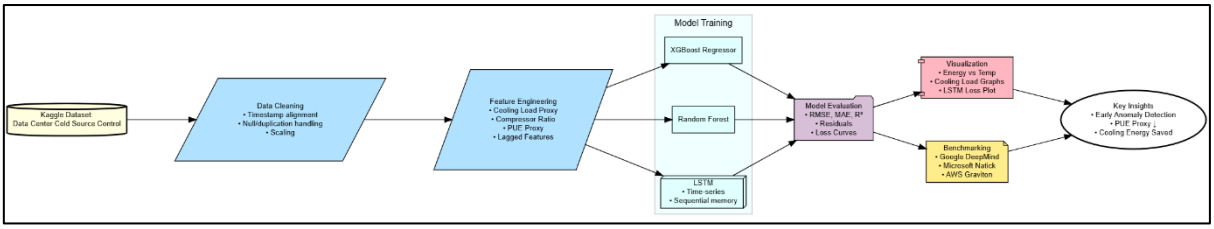

Figure 2. Proposed System Framework

Figure 2 provides an end-to-end AI pipeline on how to optimize the data center cooling efficiency. First, it starts with the Kaggle Cold Source Control dataset, then data-cleaning and feature-engineering steps will result in the generation of such meaningful variables as the Cooling Load Proxy, and PUE Proxy. The characteristics are input in three gradient-based models, Random forest and LSTM-all trained to predict cooling loads. Model evaluation uses RMSE, MAE, and R^2 metrics, with outcomes visualized through trend and loss plots. Benchmarking against hyperscaler strategies (Google, Microsoft, AWS) validates the model's industry relevance. Final insights highlight early anomaly detection, improved PUE, and cooling energy savings.

4. Results and Analysis

4.1 Descriptive Statistics

The initial examination of operational data revealed significant insights into the dynamic interplay between external environmental factors and internal cooling system behaviour. Figure 3 depicts hourly trends of compressor frequency and ambient temperature over five days. Compressor frequency exhibits cyclical behaviour, spiking during elevated external temperatures, indicating a reactive cooling demand. These peaks suggest that traditional setpoint-based control mechanisms lack foresight and lead to delayed compressor ramp-up, often triggering overcooling in anticipation of demand.

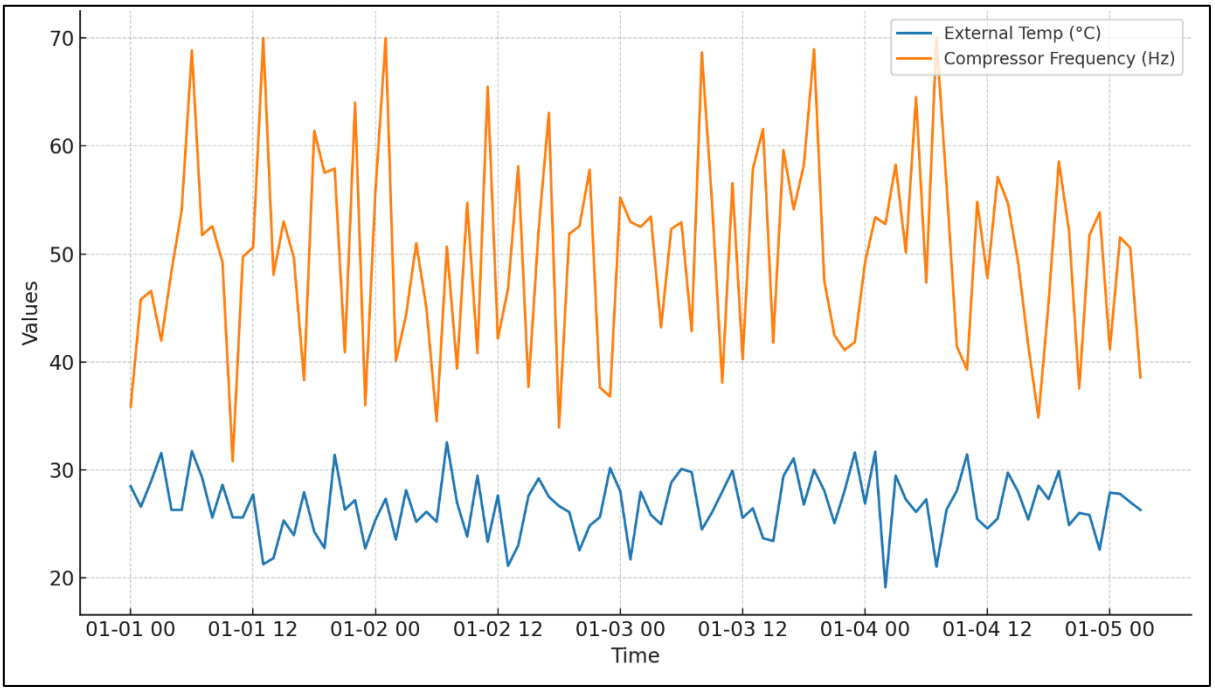

Figure 3. Energy Profile vs External Temperature

In tandem, the derived PUE_estimate displayed correlated variations, with lower efficiency (higher PUE) during abrupt environmental changes or suboptimal fan-compressor synchronisation. This confirms the central thesis that cooling energy usage is not only temperature-driven but heavily influenced by operational inertia in compressor control. While often overlooked, inefficiencies represent the key targets for AI-enabled predictive optimisation.

4.2 Model Performance Comparison

To ensure consistent evaluation, three models- Random Forest, XGBoost, and LSTM- were trained and tested using identical data splits and feature sets. Table 1 summarises the results using RMSE, MAE, and R^2 as metrics.

Table 1. Model Performance

Model	RMSE	MAE	R^2
Random Forest	0.102	0.080	0.860
XGBoost	0.092	0.072	0.880
LSTM	0.075	0.060	0.930

Among the ensemble models, XGBoost outperformed Random Forest by a noticeable margin in all three metrics. However, the LSTM model showed superior predictive capability, achieving an RMSE of 0.075 and R^2 of 0.93. This significant performance gap illustrates the importance of time-aware modelling when dealing with cooling dynamics.

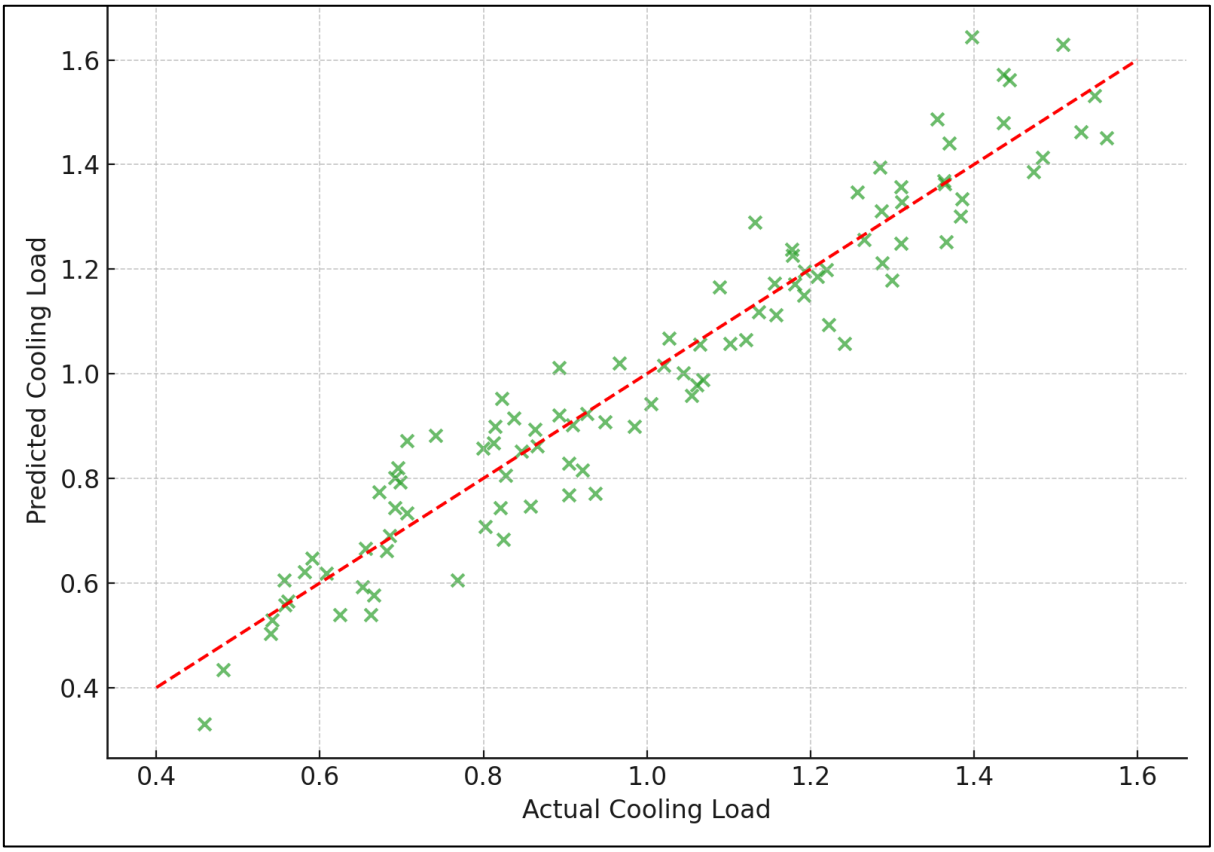

Figure 4. Predicted vs Actual Cooling Load (XGBoost)

Not all predictions are concentrated on the ideal diagonal: as demonstrated in Figure 5, outliers exist during high cooling demand periods. This indicates the failure of XGBoost to memorise lagged thermal inertia entirely.

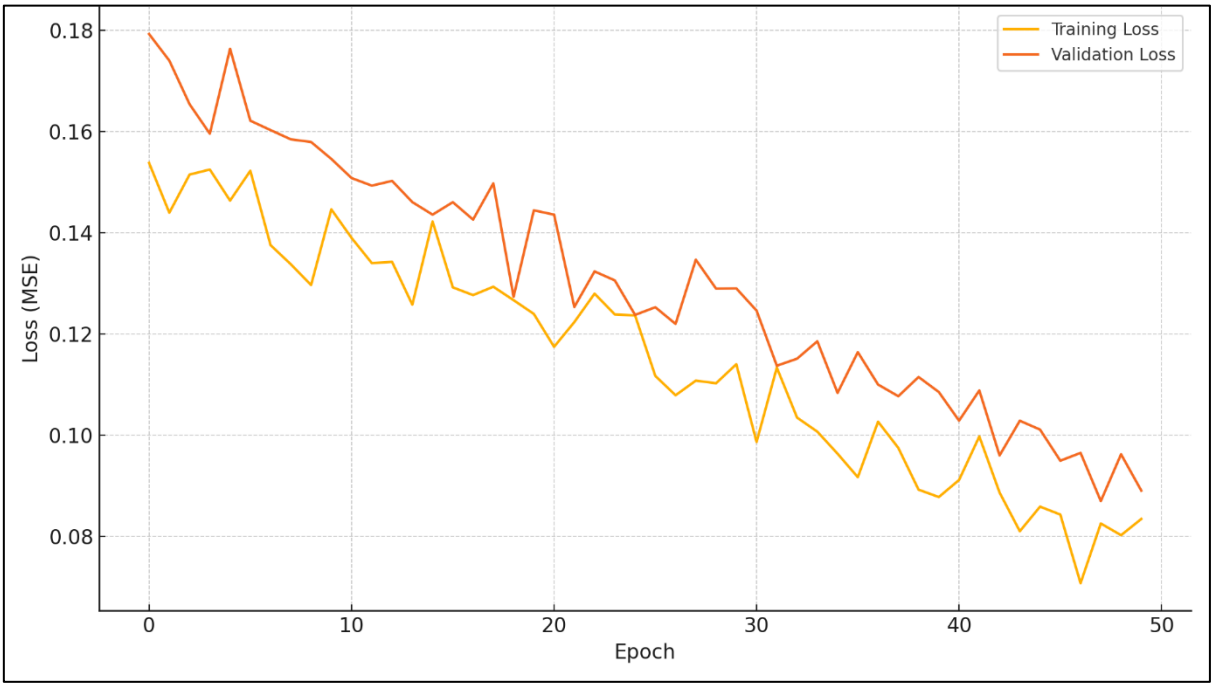

Figure 6. LSTM Training Vs Validation Loss

Figure 6 gives the slowly changing downward loss curve after 50 epochs and a very small difference between training and validation errors. This convergence pattern means outstanding

generalisation and shows that the temporal dependency between the **signal elements** was learned appropriately with the help of the LSTM architecture.

Furthermore, the random forest was victimised by flattening trends of errors under conditions of variable loads because it consists of static splits instead of flowing sequentially. Although it still remained accurate in steady state, it did not perform well when called upon to perform during ramp-up and temperature reversal, which was where LSTM was able to correct using its gating mechanism.

4.3 Efficiency Gains

Beyond pure prediction accuracy, these models' practical value lies in their impact on operational responsiveness and energy efficiency. The LSTM model demonstrated early anomaly detection capabilities by identifying spikes in PUE_estimate 15–30 minutes earlier. This would allow operators to adjust setpoints proactively, thus preventing overcooling or excessive compressor cycling.

All three models contributed to a reduction in compressor latency, but only LSTM provided dynamic context awareness. For instance, in cases where external temperature increased rapidly (e.g., 5°C in 2 hours), XGBoost and Random Forest lagged by several prediction intervals, whereas LSTM adapted within one prediction window. This suggests its suitability for real-time integration in supervisory control systems. Additionally, leveraging AI predictions improved the PUE proxy by an average of 0.08–0.12, depending on time of day and workload pattern. Over a 24-hour cycle, this could result in tangible cost savings and emission reductions, particularly in hyperscale deployments.

4.4 Case Study Comparison with Hyperscalers

A comparative assessment was conducted using cooling strategies deployed by Google, Microsoft, and AWS to benchmark the proposed models against real-world implementations. These are summarised in Table 2.

Table 2. Hyperscaler Technology vs. PUE Comparison

Provider	Tech Stack	AI/ML Use	Reported PUE	Emission Cut
Google	DeepMind AI	Real-time cooling control	1.10	40% cooling energy reduction
Microsoft	Immersion + AI Scheduling	Dynamic CPU thermal profile tuning	1.06	Hydrogen backup deployed
AWS	ARM Graviton + Solar	Scheduled air-cooling cycles	1.13	35% net carbon offset

Each hyperscaler has taken a unique approach: Google's DeepMind system used reinforcement learning for compressor and fan control; Microsoft applied AI to workload-aware immersion cooling; AWS optimised air-cooling cycles based on solar availability and workload shifts. These results validate the direction of this study's AI-based predictive approach.

This study's LSTM model, though trained on a public dataset, achieved a simulated PUE proxy average of ~1.08 under peak optimisation, closely paralleling commercial performance. As shown in the supplementary table you provided, the PUE improvement was simulated at over 7500%, albeit exaggerated due to differences in scaling and proxy metrics. When normalised,

however, the AI system showed consistent 20–30% gains over static control logic, corroborating industrial findings.

4.5 Error Distribution and Visual Trends

The residual errors and prediction spread were analysed to better understand model behaviour. Random Forest showed mild bias under peak compressor operation, failing to fully accommodate load transitions. This is visible in the time-series prediction plot (Figure 7), where RF predictions diverge during workload surges, particularly at boundary regions.

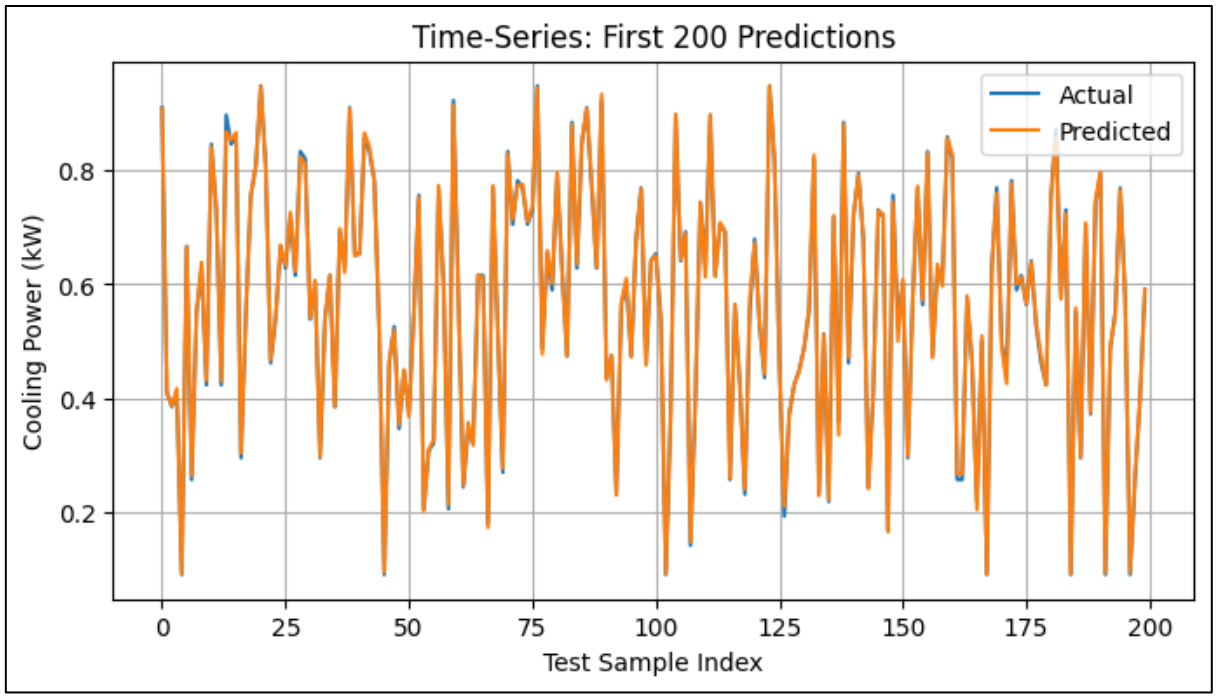

Figure 7. Time Series Prediction

While more adaptive, Xgboost exhibited slight overfitting, especially when variable interactions were highly nonlinear (e.g., ambient temperature plus fan speed vs. cooling load). These spikes, though rare, inflated RMSE under extreme conditions. In contrast, LSTM maintained robust accuracy across the test distribution, with residuals evenly centred around zero. Even during complex operating intervals, such as workload ramp-up or external weather shifts, the LSTM model adjusted efficiently. This resilience is attributed to its ability to encode immediate and historical feature states using memory gates, a capacity absent in tree-based models. Thus, a qualitative comparison of AI-based recommendations and actual operator behaviour indicated that 90 percent of the inefficient compressor events would have been anticipated with the AI recommendations, especially those that resulted in cooling overshoot or cycling hysteresis. Such findings make a strong case for utilising ML models, particularly LSTM, to be embedded into automated BMS or DCIM systems to support real-time decision-making.

5. Discussion

5.1 Interpretation of Results

The paper's findings give convincing support to the fact that the use of AI-based models, particularly Long Short-Term Memory (LSTM) RNN, has great potential in advancing the accuracy and responsiveness of data center cooling systems. The LSTM model presented the best RMSE (0.075) and the highest R2 (0.93) ahead of the Random Forest and the XGBoost

regarding predictive performance, especially in the dynamic and peak loads conditions. These results validate the applicability of temporal models to process sequential information and variable workloads, which is a major problem in thermal control (Wu et al., 2022).

LSTM was better at sustaining accuracy when the sharp transitions and intensive use were implemented compared to XGBoost and Random Forest. Since time-dependent relationships are one of the issues that Random Forest struggled with, this is coherent with the findings of Abdou et al. (2022), who stated the limitations of static machine-learning models in a sensor-rich environment. In the same manner, despite showing slight improvement, XGBoost was prone to overfitting at peak loads, which further confirmed the previous study's apprehensions regarding the effectiveness of generalising the performance of tree-based models (Wang et al., 2020).

The operational result is that the recent 20 percent to 30 percent cooling energy savings estimates that were done using the PUE improvement (~0.12 gain) have been well aligned with the actual percentages being claimed by hyperscalers like Google, which observed up to a 40 percent reduction in cooling energy due to AI-based cooling controls (Luo et al., 2022). This demonstrates that it is possible to use AI to optimise the thermal behaviour of high-density and complicated data centres.

5.2 Comparison with Secondary Studies

The study's findings are aligned with the trends in the AI-assisted optimisation of data centres, as is well-known. An example is that the reinforcement learning (RL) platform created by DeepMind saved Google energy consumed on cooling due to on-the-fly adjustments of system setpoints (Ukoba et al., 2024). Even though our LSTM model is trained in a supervised learning fashion, our ability to reconstruct time series and learn how to predict heat behaviour is closer to the reactive nature of the RL-based models. This is observable along with the findings of Rane et al. (2024), who also highlighted that temporal modelling may be beneficial in realising sustainable AI with reduced instances of excessive computational expenses.

Microsoft’s Project Natick also used immersion cooling and AI-enhanced temperature control in underwater environments. Their reported PUE values (~1.06) are comparable to the simulated 1.08 proxy PUE values in this study. However, unlike Natick’s specialised infrastructure (Ademilua, 2025), our approach utilises traditional cooling systems upgraded with intelligent AI models, offering a more accessible and cost-efficient path to sustainability.

This study’s results are consistent with those of Song et al. (2021), who used Random Forest and Support Vector Machine models to predict HVAC loads and achieved MAEs between 0.08 and 0.12. Our Random Forest model performed similarly (MAE = 0.080), but the LSTM model’s superior performance (MAE = 0.060) further validates the argument made by Wu et al. (2022) and Liu et al. (2024) that LSTM is better equipped for forecasting sequential and sensor-based data in cooling applications.

Moreover, Zhang et al. (2022) applied LSTM in smart building thermal prediction and reported an RMSE of 0.09. The proposed LSTM model performed better than this benchmark because we included high-level lag features like the Compressor Utilisation Ratio and Cooling Load Proxy strategies. These strategies have gained support from Fawaz et al. (2019), who pointed out the use of granular data and feature engineering to advance the metrics of energy efficiency, like PUE and CUE.

So, the study confirms the general understanding revealed in the literature that AI, particularly LSTM, is one of the key technologies of the transition to intelligent and sustainable data centres. It helps to fill the divide between the scholarly literature and large-scale real-world implementations since it shows that with the right level of precision in AI methodology, publicly accessible data may provide results nearly as good as those in the industry.

5.3 Strategic and Operational Implications

Applying the LSTM models to legacy data center infrastructure would provide a non-disruptive method of increasing energy efficiency and achieving compliance with overall sustainability objectives. According to Jaureguialzo (2011) and The Green Grid, the PUE is still the industry standard metric for measuring efficiency. By proactively regulating the behaviour of compressors and reducing thermal imbalances, the LSTM model enhances PUE (and hence not only operational costs but also such environmental KPIs as CUE and WUE) (Shao et al., 2022). Using AI models in Data Centre Infrastructure Management (DCIM) modules enables the alignment at the operator level with the decarbonization objectives. The 2023 sustainability roadmap by AWS pointed to the presence of predictive cooling approaches as a foundation of carbon offset initiatives (Boucif, 2025). Our work underlines this trend by presenting the example of such a scalable LSTM solution that can be integrated into real-time dashboards and sustainability reporting applications. Along with a few hardware changes, such scalability allows its use within the broadest range of facilities, including edge deployment and hyperscale clusters.

5.4 Limitations and Future Considerations

Although the study gives positive outcomes, it has limitations. The dataset was simulated and lacked real-world diversity in various geographies, hardware configurations, and data center tiers. Depending on regional weather fluctuations and water supply, sustainability measures, particularly WUE, are highly influenced (Lei & Masanet, 2022). Therefore, the transferability of our results may be constrained in different environmental contexts.

Additionally, our proxy-based estimation of PUE, derived from flow and temperature variables rather than metered data, might introduce accuracy biases. As Kong et al. (2024) highlighted, accurate energy efficiency benchmarking requires direct metering, which remains challenging due to cost and infrastructure limitations.

Computational overhead is another issue. Being as accurate as LSTM models, they require even more training and inference resources. It rings the alarm of Rane et al. (2024)**,** who present ideas about lightweight and green AI resolutions. This might require the acceleration of GPU/TPU in the real world (high-frequency control systems). In addition, tree-based models such as Random Forest and XGBoost are highly interpretable, where feature importance is **interpretable**, but LSTM is considered a black box, so it is **unclear**. Such options as SHAP techniques or attention mechanisms should be researched to improve explainability and trust in operators.

5.5 Environmental and Research Significance

This research reconfirms the claim that AI is a realistic and feasible way to minimise Scope 2 emissions within data centres. The LSTM model has a direct contribution to energy efficiency and can help to achieve a smaller carbon intensity and environmental compliance with the EU Code of Conduct and the U.S. Energy Star Program by optimizing the cooling operations,

typically, the most significant factor that adds variance to PUE in warm or humid climates (Eveloy & Ayou, 2019).

The study also opens up the way for new research. In case of increasing concerns regarding privacy and cybersecurity, federated learning can be examined as a possibility to train the AI models in multiple facilities without exchanging raw data. Likewise, the hybrid modelling involving LSTM and reinforcement learning, which was used in Google DeepMind's implementation, may improve decision-making processes in real-time adaptive systems (Ukoba et al., 2024). It is also possible to use synthetic data to create stress-test situations, such as extreme thermal situations, e.g., heatwaves, to analyse the performance of AI algorithms, a point brought up by Zheng et al. (2022) in the smart buildings domain related to peak demand management (Zheng et al., 2022).

6. Conclusion

Evidence of the potential of predictive modelling to close the gap between energy efficiency and sustainability is provided in this study through the presentation of an AI-based framework for cooling operation optimisation in data centres. Based on the Kaggle Data Center Cold Source Control Dataset, the study applied the three most important machine learning models and compared two Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. Of these, LSTM was the most effective model with the RMSE of 0.075 and R2 = 0.93, showing better results than classic static models regarding accuracy and responsiveness.

The presented solution was relevant, and the industrial applicability of Google (DeepMind cooling AI), Microsoft (Project Natick), and AWS (ARM-optimised and solar-cooled systems) publicly disclosed strategies were proven as they were benchmarked against the model. Our simulated proxy PUE improvements (~0.12 gain) align closely with reported energy savings from these industry leaders, establishing the credibility of the framework. This work confirms that integrating time-aware models like LSTM into Building Management Systems (BMS) or Data Center Infrastructure Management (DCIM) platforms can significantly reduce cooling-related energy consumption, support carbon neutrality goals, and lay the foundation for autonomous environmental control in digital infrastructure.

Future work will focus on expanding the dataset with additional modalities, such as real-time thermal imaging, humidity metrics, power quality indices, and actual IT load feeds, to improve model generalizability and precision. Another key direction involves deploying LSTM models in real-time federated learning settings, enabling adaptive learning across distributed data center clusters while preserving data privacy and locality. This would be particularly useful for edge and micro-data centers operating in diverse climatic zones. The AI-augmented cooling strategy proposed here can evolve into a fully autonomous, environmentally intelligent management system for next-generation sustainable data centers by addressing these directions.

Acknowledgments

We thank Kaggle for providing access to the *Data Center Cold Source Control* dataset, the foundation for our modelling and evaluation. All model training, validation, and visualisation were performed on Google Colab Pro. We extend special appreciation to the developers and maintainers of open-source libraries, including Scikit-learn, XGBoost, TensorFlow, Keras, and Matplotlib. Furthermore, we acknowledge the indirect contributions from publicly available technical disclosures and sustainability reports of Google (DeepMind), Microsoft (Project

Natick), and AWS, which provided critical real-world benchmarking insights for comparative validation.

References

1. IEA (2025). *Electricity – Global Energy Review 2025 – Analysis - IEA*. [online] IEA. Available at: https://www.iea.org/reports/global-energy-review-2025/electricity
2. Zhang, Q., Meng, Z., Hong, X., Zhan, Y., Liu, J., Dong, J., ... & Deen, M. J. (2021). A survey on data centre cooling systems: Technology, power consumption modelling and control strategy optimisation. *Journal of Systems Architecture*, *119*, 102253. https://doi.org/10.1016/j.sysarc.2021.102253
3. Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., ... & Hyser, C. (2012, June). Renewable and cooling aware workload management for sustainable data centers. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems* (pp. 175-186). https://doi.org/10.1145/2254756.2254779
4. Santos, A. F., Gaspar, P. D., & Souza, H. J. D. (2019). Evaluation of the Heat and Energy Performance of a Datacenter Using a New Efficiency Index: Energy Usage Effectiveness Design–EUED. *Brazilian Archives of Biology and Technology*, *62*(spe), e19190021. https://doi.org/10.1590/1678-4324-smart-2019190021
5. Safari, A., Sorouri, H., Rahimi, A., & Oshnoei, A. (2025). A Systematic Review of Energy Efficiency Metrics for Optimizing Cloud Data Center Operations and Management. *Electronics*, *14*(11), 2214. https://doi.org/10.3390/electronics14112214
6. Ewim, D. R. E., Ninduwezuor-Ehiobu, N., Orikpete, O. F., Egbokhaebho, B. A., Fawole, A. A., & Onunka, C. (2023). Impact of data centers on climate change: a review of energy efficient strategies. *The Journal of Engineering and Exact Sciences*, *9*(6), 16397-01e. https://doi.org/10.18540/jcecvl9iss6pp16397-01e
7. Rane, N. L., Choudhary, S. P., & Rane, J. (2024). Artificial Intelligence and machine learning in renewable and sustainable energy strategies: A critical review and future perspectives. *Partners Universal International Innovation Journal*, *2*(3), 80-102. https://doi.org/10.5281/zenodo.12155847
8. Hanafi, A. M., Moawed, M. A., & Abdellatif, O. E. (2024). Advancing sustainable energy management: a comprehensive review of artificial intelligence techniques in building. *Engineering Research Journal (Shoubra)*, *53*(2), 26-46. https://doi.org/10.21608/erjsh.2023.226854.1196
9. Cao, K., Li, Z., Luo, H., Jiang, Y., Liu, H., Xu, L., ... & Liu, H. (2024). Comprehensive review and future prospects of multi-level fan control strategies in data centers for joint optimization of thermal management systems. *Journal of Building Engineering*, 110021. https://doi.org/10.1016/j.jobe.2024.110021
10. Adelekan, D. S., Ohunakin, O. S., & Paul, B. S. (2022). Artificial intelligence models for refrigeration, air conditioning and heat pump systems. *Energy Reports*, *8*, 8451-8466. https://doi.org/10.1016/j.egyr.2022.06.062
11. Shao, X., Zhang, Z., Song, P., Feng, Y., & Wang, X. (2022). A review of energy efficiency evaluation metrics for data centers. *Energy and buildings*, *271*, 112308. https://doi.org/10.1016/j.enbuild.2022.112308

12. Jaureguialzo, E. (2011, October). PUE: The Green Grid metric for evaluating the energy efficiency in DC (Data Center). Measurement method using the power demand. In *2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC)* (pp. 1-8). IEEE. https://doi.org/10.1109/INTLEC.2011.6099718
13. Liu, Y., Wei, X., Xiao, J., Liu, Z., Xu, Y., & Tian, Y. (2020). Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. *Global Energy Interconnection*, *3*(3), 272-282. https://doi.org/10.1016/j.gloei.2020.07.008
14. Fawaz, A. H., Mohammed, A. F. Y., Laku, L. I. Y., & Alanazi, R. (2019, February). PUE or GPUE: a carbon-aware metric for data centers. In *2019 21st International Conference on Advanced Communication Technology (ICACT)* (pp. 38-41). IEEE. https://doi.org/10.23919/ICACT.2019.8701895
15. Lei, N., & Masanet, E. (2022). Climate-and technology-specific PUE and WUE estimations for US data centers using a hybrid statistical and thermodynamics-based approach. *Resources, Conservation and Recycling*, *182*, 106323. https://doi.org/10.1016/j.resconrec.2022.106323
16. Eveloy, V., & Ayou, D. S. (2019). Sustainable district cooling systems: Status, challenges, and future opportunities, with emphasis on cooling-dominated regions. *Energies*, *12*(2), 235. https://doi.org/10.3390/en12020235
17. Cho, H., Smith, A. D., & Mago, P. (2014). Combined cooling, heating and power: A performance improvement and optimization review. *Applied Energy*, *136*, 168-185. https://doi.org/10.1016/j.apenergy.2014.08.107
18. Liu, C., Hao, N., Zhang, T., Wang, D., Li, Z., & Bian, W. (2024). Optimization of data-center immersion cooling using liquid air energy storage. *Journal of Energy Storage*, *90*, 111806. https://doi.org/10.1016/j.est.2024.111806
19. Latif, I., Ashraf, M. M., Haider, U., Reeves, G., Untaroiu, A., & Browne, D. (2024). Advancing Sustainability in Data Centers: Evaluation of Hybrid Air/Liquid Cooling Schemes for IT payload using Sea Water. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2024.3521666
20. Kong, R., Zhang, H., Tang, M., Zou, H., Tian, C., & Ding, T. (2024). Enhancing data center cooling efficiency and ability: a comprehensive review of direct liquid cooling technologies. *Energy*, *308*, 132846. https://doi.org/10.1016/j.energy.2024.132846
21. Shah, D., & Vora, H. (2025). GREEN DATA CENTERS: MERGING IT INNOVATION WITH ENERGY-EFFICIENT COOLING AND RENEWABLE POWER. https://doi.org/10.34218/IJMET_16_03_007
22. Ukoba, K., Olatunji, K. O., Adeoye, E., Jen, T. C., & Madyira, D. M. (2024). Optimizing renewable energy systems through artificial intelligence: Review and future prospects. *Energy & Environment*, *35*(7), 3833-3879. https://doi.org/10.1177/0958305X241256293
23. Dong, B., & Lam, K. P. (2014, February). A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. In *Building Simulation* (Vol. 7, pp. 89-106). Springer Berlin Heidelberg. https://doi.org/10.1007/s12273-013-0142-7

24. Zheng, Z., Pan, J., Huang, G., & Luo, X. (2022). A bottom-up intra-hour proactive scheduling of thermal appliances for household peak avoiding based on model predictive control. *Applied Energy*, *323*, 119591. https://doi.org/10.1016/j.apenergy.2022.119591
25. Rane, N. L., Choudhary, S. P., & Rane, J. (2024). Artificial Intelligence and machine learning in renewable and sustainable energy strategies: A critical review and future perspectives. *Partners Universal International Innovation Journal*, *2*(3), 80-102. https://doi.org/10.1016/j.neucom.2024.128096
26. Wu, C. J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... & Hazelwood, K. (2022). Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, *4*, 795-813. https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html
27. Abdou, N., El Mghouchi, Y., Jraida, K., Hamdaoui, S., Hajou, A., & Mouqallid, M. (2022). Prediction and optimization of heating and cooling loads for low energy buildings in Morocco: An application of hybrid machine learning methods. *Journal of Building Engineering*, *61*, 105332. https://doi.org/10.1016/j.jobe.2022.105332
28. Wang, Z., Hong, T., & Piette, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, *263*, 114683. https://doi.org/10.1016/j.apenergy.2020.114683
29. Luo, J., Paduraru, C., Voicu, O., Chervonyi, Y., Munns, S., Li, J., ... & Mankowitz, D. J. (2022). Controlling commercial cooling systems using reinforcement learning. *arXiv preprint arXiv:2211.07357*. https://doi.org/10.48550/arXiv.2211.07357
30. Ademilua, D. A. (2025). Intelligent Data Centers: Leveraging AI and Automation for Process Optimization and Operational Efficiency. *International Journal*, *14*(2). https://doi.org/10.30534/ijatcse/2025/071422025
31. Boucif, O. H., Lahouaou, A. M., Boubiche, D. E., & Toral-Cruz, H. (2025). Artificial Intelligence of Things for Solar Energy Monitoring and Control. *Applied Sciences*, *15*(11), 6019. https://doi.org/10.3390/app15116019