https://doi.org/10.31436/iiumej.v26i3.3629

DEVELOPMENT OF A PREDICTIVE REAL-TIME TEMPERATURE MANAGEMENT FRAMEWORK FOR DATA CENTERS USING A HYBRID DEEP LEARNING MODEL

PATRICK O. OLABISI*, AHMED A. MAYOWA

Department of Electrical, Electronic and Telecommunication Engineering, College of Engineering, Bells University of Technology, Ota, Ogun State, Nigeria

**Corresponding author: niyiolabisi@gmail.com*

(Received: 28 February 2025; Accepted: 22 July 2025; Published online: 9 September 2025)

ABSTRACT: Traditional cooling systems in data centers often struggle to adapt to dynamic temperature fluctuations caused by varying server loads and environmental conditions, leading to energy inefficiency and potential overheating risks. This research developed a predictive cooling system to address the challenges of maintaining efficient temperature management that will enhance system reliability, cooling efficiency, and energy savings by leveraging deep learning techniques. A hybridization of two deep learning techniques: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks was formed. Over a 24-hour test period, the model demonstrated a high prediction accuracy ranging from 85 to 95% with reliability between 0.97 and 0.99, highlighting the system's robust design. Cooling efficiency ranged from 0.75 to 0.9, peaking during cooler hours and dipping slightly during midday due to higher energy demands and environmental conditions. Efficiency improvements of 3 to 7% were observed, particularly during high-demand periods. The findings highlight the transformative potential of predictive cooling systems in reducing energy consumption and ensuring thermal stability in modern data centers. Further optimization of the CNN-LSTM model to improve prediction accuracy during peak server activity is recommended. Additionally, integrating real-time feedback loops and incorporating renewable energy sources into the system could enhance its sustainability and reduce operational costs.

ABSTRAK: Sistem penyejukan tradisional di pusat data sering menghadapi masalah menyesuaikan diri dengan suhu dinamik yang disebabkan oleh beban pelayan dan keadaan persekitaran berbeza, membawa kepada risiko ketidakcekapan tenaga dan kemungkinan terlebih panas. Penyelidikan ini membangunkan sistem penyejukan ramalan bagi menangani cabaran dalam mengekalkan pengurusan suhu yang cekap, sekaligus meningkatkan kebolehpercayaan pada sistem, kecekapan penyejukan dan penjimatan tenaga melalui manfaat teknik pembelajaran mendalam. Hibridisasi dua teknik pembelajaran mendalam iaitu: Rangkaian Neural Konvolusi (CNN) dan Rangkaian Memori Jangka Pendek Panjang (LSTM) telah dibentuk. Sepanjang tempoh ujian 24 jam, model menunjukkan ketepatan ramalan yang tinggi antara 85 hingga 95% dengan kebolehpercayaan antara 0.97 dan 0.99 menyerlahkan reka bentuk sistem yang teguh. Kecekapan penyejukan antara 0.75 hingga 0.9, memuncak pada waktu sejuk dan menurun sedikit pada tengah hari disebabkan perningkatan permintaan tenaga dan keadaan persekitaran. Peningkatan kecekapan sebanyak 3 hingga 7% diperhatikan, terutama semasa tempoh permintaan tinggi. Dapatan ini menyerlahkan potensi transformatif sistem penyejukan ramalan dalam mengurangkan penggunaan tenaga dan memastikan kestabilan haba di pusat data moden. Pengoptimuman lanjut terhadap model CNN-LSTM bagi meningkatkan ketepatan ramalan semasa aktiviti waktu puncak adalah disyorkan. Selain itu, penyepaduan gelung maklum balas masa nyata dan gabungan sumber tenaga boleh diperbaharui ke dalam sistem bagi meningkatkan lagi kemampanannya dan mengurangkan kos operasi.

KEYWORDS: *CNN, LSTM, Temperature Management, Cooling Systems, Data Centers*

1. INTRODUCTION

Temperature fluctuations in data centers often follow specific trends based on server usage patterns, peak hours of activity, and external environmental factors such as ambient temperature. Cooling technologies in data centers have gone from dry air cooling to liquid cooling, immersion, evaporative cooling, and so on. While traditional air-based systems remain prevalent, liquid-based, hybrid, and advanced cooling methods significantly improve efficiency and performance. Despite the advancements in cooling technologies in data centers, several limitations remain, particularly when balancing cost, complexity, efficiency, and environmental constraints. Traditional air cooling systems, for example, struggle to keep up with the increasing power density of modern data centers. As servers become more powerful, they generate more heat, and air cooling systems often become inefficient in effectively removing the heat. This inefficiency leads to higher energy consumption, making it difficult for data centers to reduce operational costs and meet sustainability goals [1].

Moreover, as the demand for high-performance computing and data storage continues to rise and data centers strive to become more energy-efficient, the growing need for cooling systems that not only manage heat effectively but also minimize their own energy consumption and are scalable and environmentally friendly has become more indispensable [2]. The challenge lies in developing cooling technologies that balance performance, cost, and sustainability.

As effective as they are, traditional cooling methods often operate on static principles, are reactive, and only respond to temperature changes after they have occurred. This approach can result in inefficiencies, as systems either overcompensate by cooling more than necessary or lag in responding to rising temperatures, putting equipment at risk [3]. Artificial Intelligence (AI) approach offers a more dynamic, proactive, and adaptive approach to temperature management, using predictive analytics and real-time decision-making to ensure that data centers remain energy-efficient and operationally resilient [4]. Temperature management is a dynamic process, with environmental conditions fluctuating rapidly based on server activity, external factors, and equipment performance. AI systems can continuously monitor these conditions and make real-time adjustments to maintain optimal temperatures while minimizing energy consumption.

The use of artificial intelligence (AI) has, in recent years, made a significant impact on data center management, particularly in temperature regulation. The complexity of maintaining optimal temperatures in a data center, where hundreds or even thousands of servers generate massive amounts of heat, has led to a growing reliance on intelligent systems that can make real-time predictions and adjustments. AI has found numerous applications in data center operations, from automating mere tasks to optimizing complex processes. In temperature management, AI techniques are employed to monitor, analyze, and control environmental conditions. Machine learning (ML) algorithms, deep learning models, and neural networks are some of the most commonly used AI techniques and intelligent computational tools that are used to optimize the cooling systems in data centers [5], [6].

A significant advantage of AI is its ability to process vast amounts of data and identify patterns that are otherwise difficult for humans or traditional systems to detect. Data centers produce extensive data streams, including information on temperature fluctuations, energy consumption, airflow, and server workload. AI-powered systems can analyze these data streams in real time and adjust cooling systems accordingly, optimizing the balance between energy consumption and maintaining the necessary environmental conditions. AI algorithms can predict future temperature trends based on historical data and current operating conditions. This allows data centers to proactively adjust their cooling systems to prevent overheating or inefficiencies in energy consumption before they occur. By anticipating temperature changes, AI reduces the likelihood of thermal spikes leading to equipment failure or energy wastage. Predictive maintenance, driven by AI, helps forecast when cooling system repair or adjustment is necessary, thus avoiding unnecessary downtime, reducing maintenance costs, and minimizing energy waste [7]. Predictive analytics also helps optimize server workloads by forecasting which equipment will most likely heat up and distributing tasks accordingly, ensuring no single server is overburdened [3]. This dynamic allocation of workloads ensures that servers operate within their optimal temperature ranges and contributes to energy efficiency across the facility.

AI-powered cooling systems use feedback loops to respond instantaneously to temperature or server load changes. For example, suppose a specific rack of servers begins to overheat due to increased computational demands. In that case, the AI system can immediately adjust the airflow or cooling distribution to target that area without unnecessarily cooling the entire facility. This targeted cooling approach reduces the overall energy consumption of the data center, as it avoids the wasteful practice of cooling areas that do not require it [8].

2. DEEP LEARNING IN ENVIRONMENTAL CONTROL SYSTEMS

Deep learning (DL), as a subset of machine learning, is characterized by its ability to learn from vast amounts of data through layered architectures of artificial neural networks. The networks consist of an input layer, multiple hidden layers, and an output layer interconnected by nodes or neurons that transform input data by applying weights and activation functions, as shown in Figure 1 [9]. The most common type of deep learning model is the feed-forward neural network, where data moves in one direction – from input to output. However, several specialized deep learning models have performed exceptionally well in specific tasks.

Deep learning techniques are classified into supervised (discriminative), unsupervised (generative), and hybrid techniques as shown in Figure 2 [10]. Major techniques in the discriminative category include Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN or ConvNet), and Recurrent Neural Networks (RNN), with the variants of RNN being Long Short-Term Memory (LSTM), Bidirectional RNN/LSTM (Bi-LSTM), and Gated Recurrent Units (GRUs). The unsupervised techniques, which include Generative Adversarial Network (GAN), Autoencoder (AE), Restricted Boltzmann Machine (RBM), Self-Organizing Map (SOM), and Deep Belief Network (DBN), and their variants, are usually used for high-order correlation of features of patterns and joint statistical distributions of data and their classes. The variants of the Autoencoder technique are the Variational Autoencoder (VAE), Sparse Autoencoder (SAE), Denoising Autoencoder (DAE), Contractive Autoencoder (CAE-t), Undercomplete Autoencoder, Convolutional Autoencoder (CAE-v), and Adversarial Autoencoder (AAE). The hybrid deep learning techniques include a hybrid of deep learning techniques and several others, like deep transfer learning (DTL) and deep reinforcement learning (DRL).

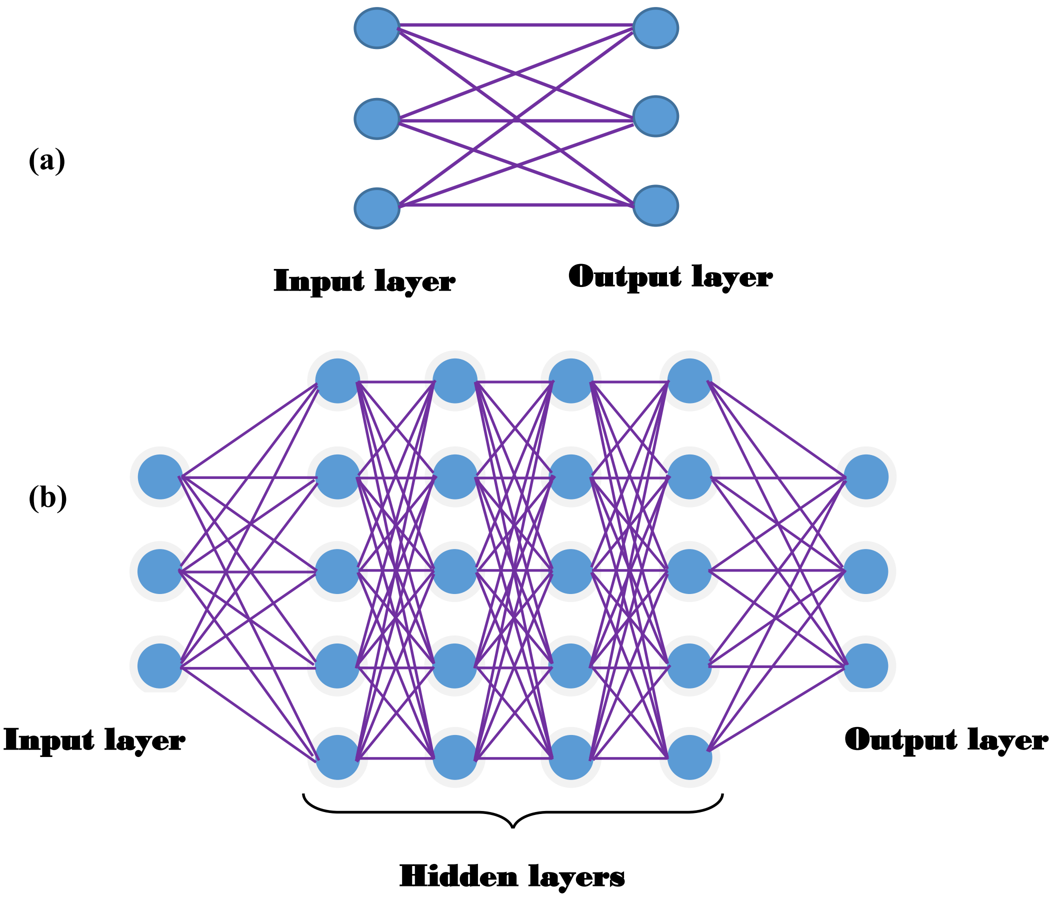

Figure 1. Neural networks (a) Normal neural network (b) Deep learning neural network [9].

Deep learning has emerged as a powerful tool for analyzing and managing the complex dynamics of environmental control systems, particularly in data centers. With modern data centers' increasing scale and complexity, traditional environmental management methods, such as manual monitoring and static control systems, have become less efficient. Deep learning models offer a highly sophisticated solution for optimizing temperature regulation and other environmental parameters [11].

Unlike conventional machine learning techniques, using multi-layered neural networks in DL allows the models to learn abstract representations from vast amounts of data. This makes DL particularly useful in environments like data centers, where temperature regulation involves numerous interacting factors, such as server workload, cooling system performance, power consumption, and external conditions. By employing deep learning models, data centers can achieve more precise and adaptive control over their environmental systems, resulting in improved energy efficiency, reduced cooling costs, and extended equipment life. Deep learning models offer an advanced approach to managing this complexity, providing highly accurate predictions and enabling real-time optimization of cooling strategies [12].

One of the key advantages of deep learning models in temperature regulation is their ability to learn and adapt to changing conditions continuously. As new data is collected from temperature sensors, server logs, and cooling system monitors, the models can use the data to update their predictions and recommendations in real time [13]. This adaptive capability is significant in modern data centers, where server workloads and environmental conditions can change rapidly. By continuously adjusting to these changes, deep learning models ensure that data centers maintain optimal environmental conditions while minimizing energy consumption and cooling costs. In addition, DL models are highly scalable and can be applied to data centers of any size or complexity in handling the vast amounts of data generated by large data centers, making them an ideal solution for managing environmental control in such environments [14].

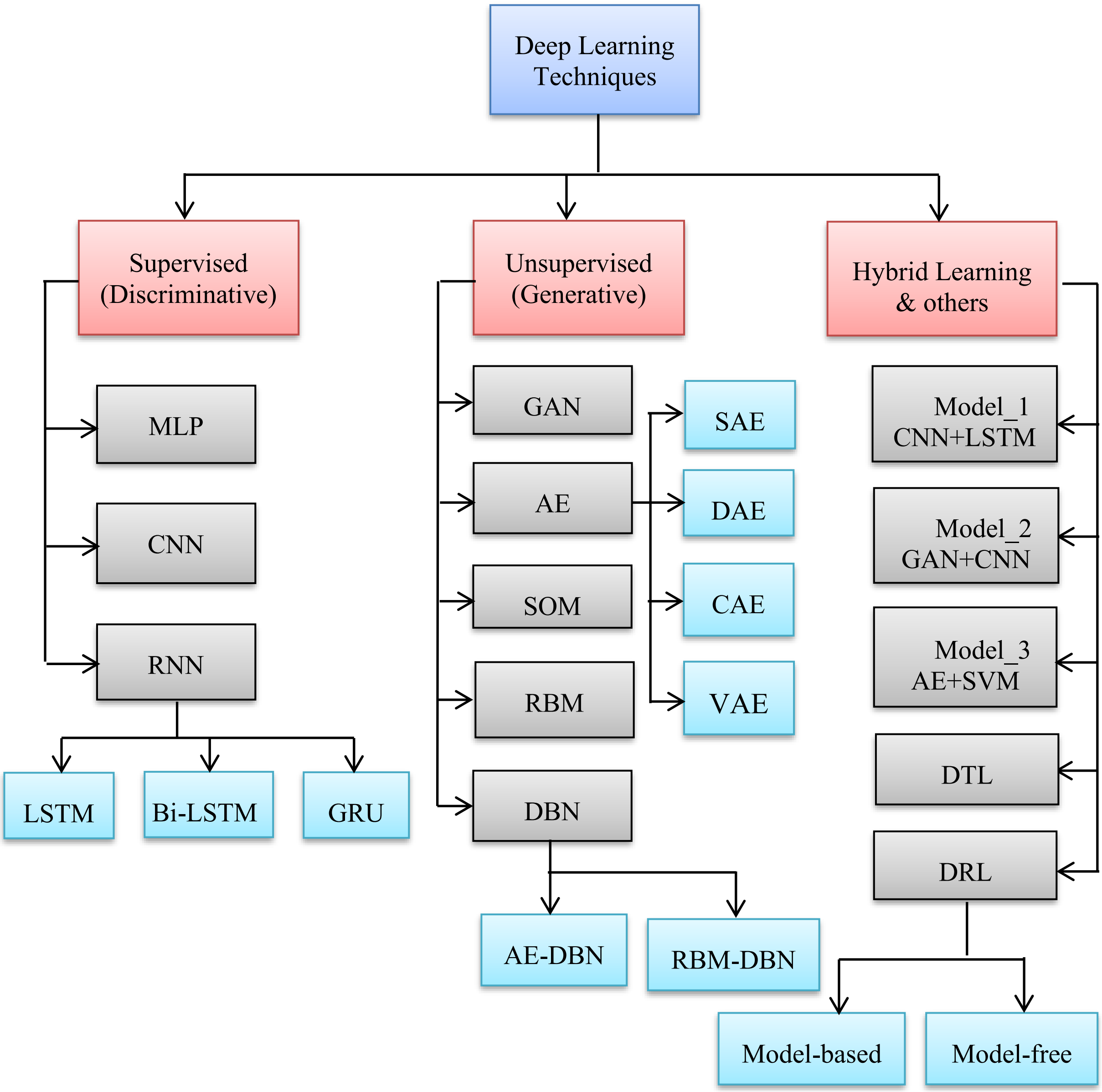

Figure 2. Deep Learning Techniques according to the Categories [10].

Therefore, deep learning models' ability to process large datasets, learn intricate patterns, and generalize from data makes them well-suited to provide efficient and powerful solutions for managing the complex data patterns involved in temperature regulation and the dynamic and data-rich environment of modern data centers. With these models, data centers can achieve more precise and adaptive control over their environmental systems, improving energy efficiency, reducing cooling costs, and prolonging the lifespan of critical IT infrastructure.

Supervised DL techniques of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks stand out for their ability to handle large, complex datasets, making them ideal for modern data centers' dynamic and data-rich environment. CNNs are particularly effective in tasks that involve spatial data, making them well-suited for analyzing temperature distributions across different areas of a data center [11]. The convolutional operation is mathematically given by [9]:

$$f_l^k(p,q) = \sum_c \sum_{x,y} i_c(x,y) \odot e_l^k(u,v) \quad (1)$$

CNNs can identify areas at risk of overheating, recommend targeted cooling strategies, and capture spatial dependencies such as how airflow and cooling efficiency vary across racks or server units, which is crucial for optimizing localized cooling strategies. As shown in Figure 3 [15], CNNs' layered architecture allows them to learn intricate patterns in the data that would be difficult for traditional models to capture, making them indispensable tools in the optimization of environmental control systems.

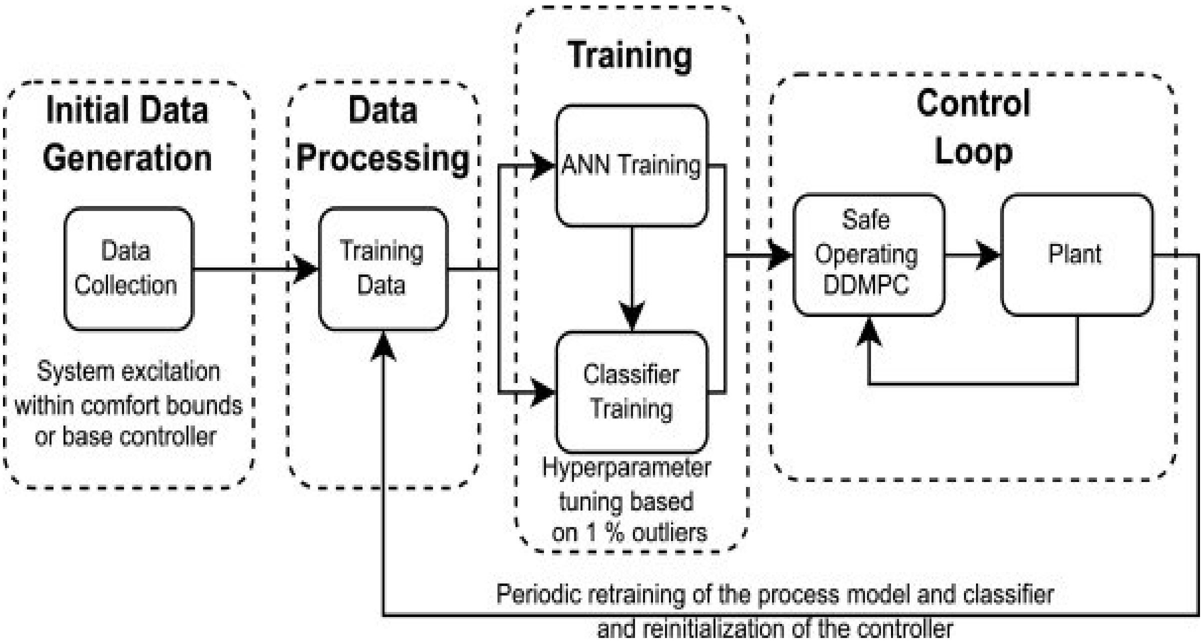

Figure 3. CNN Data Center Temperature Control [15].

On the other hand, the LSTM networks shown in Figure 4 [16], as a variant of a recurrent neural network (RNN), are specifically designed to handle long-term dependencies in sequential data and so are ideal for tasks that involve time series data, such as temperature prediction over time. This makes them well-suited for temperature regulation tasks where past temperature trends, server workloads, and cooling system performance influence future temperature outcomes. Hence, the LSTM networks can capture temporal dependencies in time series data, allowing them to predict temperature trends based on past observations and anticipate future fluctuations [17], [18]. By utilizing LSTM models, data centers can achieve highly accurate forecasts of temperature fluctuations, allowing for proactive adjustments to cooling systems before critical temperature thresholds are reached [19]. LSTM models can also simulate the impact of different cooling strategies, helping operators identify the most energy-efficient and cost-effective solutions.

The hybrid architecture of CNN-LSTM leverages the powerful feature extraction capabilities of convolutional layers and the ability of LSTM to capture sequential dependencies. The CNN component processes input data, such as temperature, server workload, and cooling patterns, extracting critical features. These features are then passed to the LSTM, which analyzes temporal trends and predicts future temperature fluctuations [8].

https://doi.org/10.31436/iiumej.v26i3.3629

Figure 5 shows the flow chart of the operations of the hybrid model [20]. Based on these predictions, the model adjusts cooling systems in real time, optimizing energy usage and preventing overheating. This proactive control ensures consistent cooling efficiency, minimizes energy consumption, and maintains operational stability, effectively adapting to the data center's dynamic environmental and server load conditions.

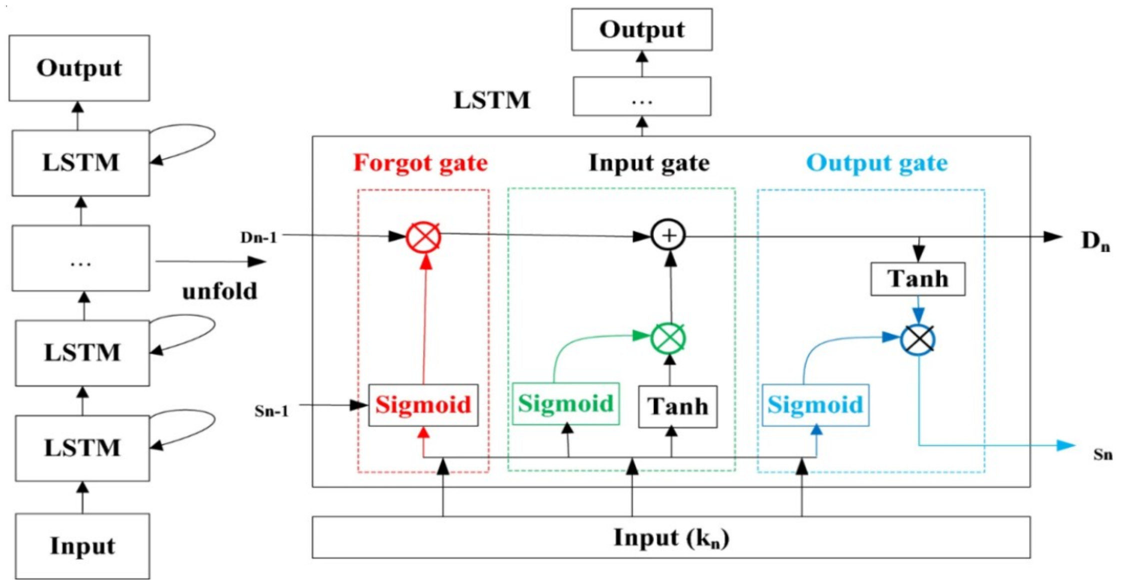

Figure 4. Cell logic structure of the LSTM network [16].

The CNN-LSTM hybrid model is proposed for this work as it has been demonstrated to outperform other models in other heat-related studies by different researchers. From the review of literature done for this work, literature on the use of a CNN-LSTM hybrid model for temperature predictions in data centers is scarce. In the work of Elmaz *et al.* [21], CNN-LSTM outperformed other models used in optimizing Heating, Ventilation, and Air Conditioning (HVAC) systems across all time frames and demonstrated greater resilience to error accumulation, achieving an accuracy of 0.9 for the 120-minute prediction horizon. Guo *et al.* [16] simulated climate parameters for Jinan city in China and forecasted six climatic factors monthly using data over 72 years to study climate change impacts on various aspects of life using several DL models. The authors found that the CNN-LSTM hybrid model achieves higher accuracy and significantly reduces forecasting errors than the other models of artificial neural networks (ANN), recurrent neural networks (RNN), long short-term memory neural networks (LSTM), and deep convolutional neural networks (CNN) that they used. The findings demonstrate the potential of CNN-LSTM models to enhance climate forecasting, which can aid in meteorological disaster prevention and reduction, as well as flood control and drought mitigation. Figure 6 shows the working architecture of the hybrid CNN-LSTM model [22].

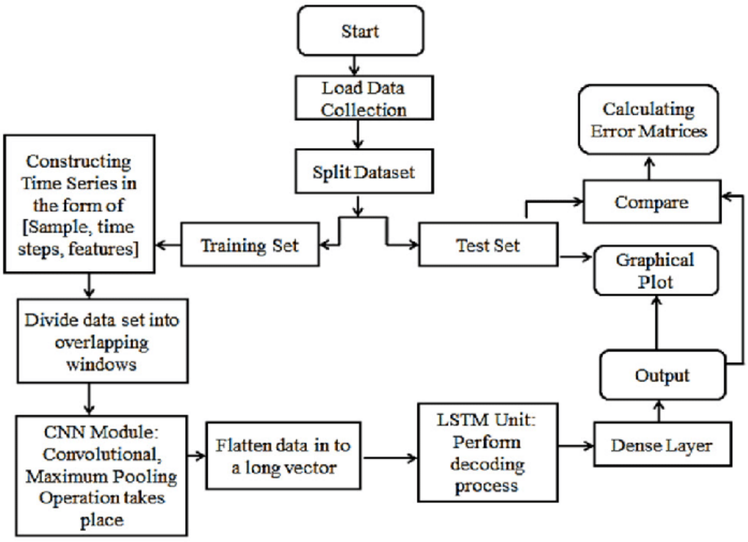

Figure 5. CNN-LSTM model data selection, feature extraction, and decoding process [20].

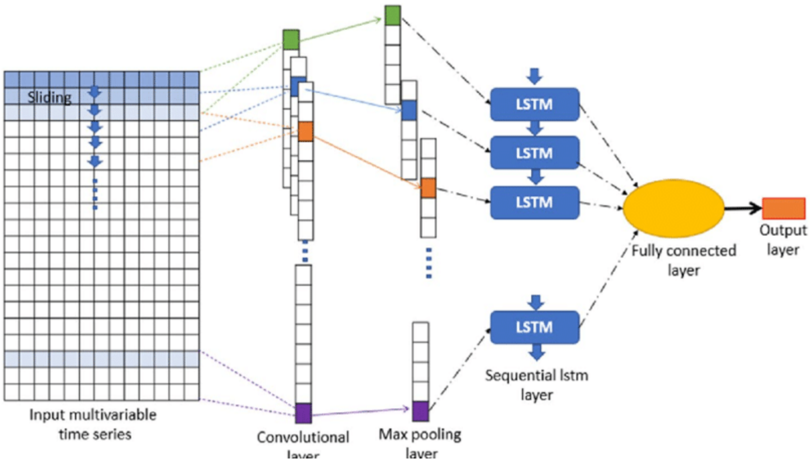

Figure 6. Architecture of the CNN-LSTM Model [22].

3. METHODS

3.1. Design of a Predictive Temperature Management System using CNN-LSTM

The output of the convolutional layer $O_{CNN}(t)$ in the CNN-LSTM model is the convolution of the input data I(t) with a filter F, given by:

$$O_{CNN}(t) = \ I(t) * F \tag{2}$$

where, $O_{CNN}(t)$ is the CNN output layer (no unit), $I(t)$ is the input temperature data (no unit), F is convolution filter (no unit).

The LSTM model predicts temperature $T_{LSTM}(t)$ by utilizing previous outputs and cell states $C(t-1)$ for time-series data processing and it is given by:

$$T_{LSTM}(t) = f(C(t-1), h(t-1)) \tag{3}$$

where, $T_{LSTM}(t)$ is the predicted temperature (in °C), $C(t-1)$ is cell state from previous time-step (no unit) and $h(t-1)$ is the hidden state from previous time-step (no unit).

The loss function value L for training the CNN-LSTM model is the mean squared error (MSE) between the predicted $T_{LSTM}(t)$ and actual temperature values $T_{actual}(t)^2$. It is given by:

$$L = \frac{1}{N}\sum_{i=1}^{N}(T_{LSTM}(t) - T_{actual}(t))^2 \quad (4)$$

where, $T_{actual}(t)^2$ is actual temperature at time t (in °C).

The parameter of the CNN-LSTM model based on θ is updated using gradient descent as function of learning rate η and the gradient of the loss function [23], is given as Equation (5):

$$\theta_{new} = \theta_{old} - \eta\frac{\partial L}{\partial \theta} \quad (5)$$

where, θ is model parameters, η is Learning rate, and L is Loss function (all have no units).

Temperature input data $T_{norm}(t)$ is normalized for training the CNN-LSTM model using the mean μ and standard deviation σ and given by [24]:

$$T_{norm}(t) = \frac{T(t) - \mu}{\sigma} \quad (6)$$

where, $T_{norm}(t)$ is normalized temperature (no unit), $T(t)$ is actual temperature at time t (in °C), μ is mean temperature (in °C) and σ is standard deviation of temperature (in °C).

3.2. Implementation and integration of the CNN-LSTM model for real-time temperature prediction in data centers.

In this sub-section, the CNN-LSTM model is implemented for real-time temperature prediction in data centers and integrated into the existing cooling system. The predicted temperature $T_{pred}(t)$ at any time t (in °C), is computed by feeding real-time input data into the trained CNN-LSTM model and given by:

$$T_{pred}(t) = CNN\_LSTM(I(t)) \quad (7)$$

where, $I(t)$ is the real-time input data (no unit).

The cooling system activation function, $A(t)$, depends on the predicted temperature $T_{pred}(t)$ crossing a threshold T_{th}, and is given by:

$$A(t) = H(T_{pred}(t) - T_{th}) \quad (8)$$

where, H is the Heaviside step function, T_{th} is the temperature threshold for activation (in °C).

The total energy consumption at time t, $E_{pred}(t)$ (in kWh), with predictive cooling is the sum of server energy usage E_s and cooling energy E_c modulated by predicted cooling cycles, is given by Equation (13):

$$E_{pred}(t) = E_s(t) + A(t) \times E_c(t) \quad (9)$$

where, $E_s(t)$ is server energy consumption at time t (in kWh) and $E_c(t)$ is cooling energy consumption (in kWh).

The response time t_r (in seconds), is the delay between the prediction and cooling system activation, dependent on system latency L_s and cooling efficiency η_c, is given by:

$$t_r = \frac{L_s}{\eta_c} \quad (10)$$

where, L_s is system latency (in seconds) and η_c cooling efficiency (no unit).

The feedback temperature $T_{fb}(t)$ (in °C), is the measured real-time temperature used to correct the model's predictions, with ϵ representing the error term, and is given by:

$$T_{fb}(t) = T(t) + \epsilon(t) \tag{11}$$

where, $T(t)$ is measured temperature at time t (in °C) and ϵ(t) is prediction error (in °C).

The overheating probability, P_o is based on the temperature exceeding the critical threshold T_{crit} for a given duration Δt.

$$P_o = \Pr\left(T_{pred}(t) > T_{crit}\right) \tag{12}$$

where, T_{pred} is predicted overheating temperature (in °C) and T_{crit} is the critical overheating temperature (in °C).

The hardware failure rate, $F(t)$ (in failures per hour), increases exponentially with temperature $T(t)$ according to Arrhenius' law, where k is a proportional constant, is given by

$$F(t) = F_0 \times e^{k/T(t)} \tag{13}$$

where, F_0 is Baseline failure rate (in failures per hour) and k is Proportional constant (no units).

Figure 7 shows the flowchart for implementing the hybrid CNN-LSTM model for data centers. The flowchart begins with data collection, where real-time environmental parameters such as temperature, humidity, and energy consumption are gathered through sensors. This data then undergoes preprocessing, cleaning, and normalization, and is suitable for machine learning models, ensuring accuracy and consistency.

The performance monitoring stage compares the real-time conditions with the model's predictions to ensure accuracy and energy efficiency, with the system constantly refining its decisions. The process completes and loops, providing continuous monitoring and real-time temperature regulation, preventing overheating, and improving data center energy efficiency, as shown in the flowchart. Figure 8 shows a diagram of the CNN-LSTM system's Simulink connection, while Figure 9 is a screenshot of the MATLAB interface of the code-based simulations of the CNN-LSTM data center optimization and temperature control.

3.3. Analysis of reduced energy consumption and cooling costs in data centers

The cooling optimization O_c (no unit), is proportional to the predicted energy savings S and the reduced hardware stress H_s over time t.

$$O_c = S \times H_s(t) \tag{14}$$

where, S is the energy savings (in kWh) and $H_s(t)$ is the reduced hardware stress (in kWh).

The total cooling cost reduction C_r (in currency), is the difference between the initial cost C_i and the AI-driven optimized cost C_o.

$$C_r = C_i - C_o \tag{15}$$

where, C_i is the initial cooling cost (in currency) and C_o is the optimized cooling cost (in currency).

The lifespan extension L_e of data center equipment is inversely proportional to the cumulative overheating probability P_o and hardware failure rate $F(t)$, is given by:

$$L_e = \frac{1}{P_o - F(t)} \tag{16}$$

where, P_o is the overheating probability (no unit) and $F(t)$ is the failure rate (in failures per hour).

The operational efficiency O_e (no unit), increases with the ratio of predictive model accuracy A_{pred} to the energy savings S.

$$O_e = \frac{A_{pred}}{S} \tag{17}$$

where A_{pred} is the prediction accuracy (no unit), and S is the energy savings (in kWh).

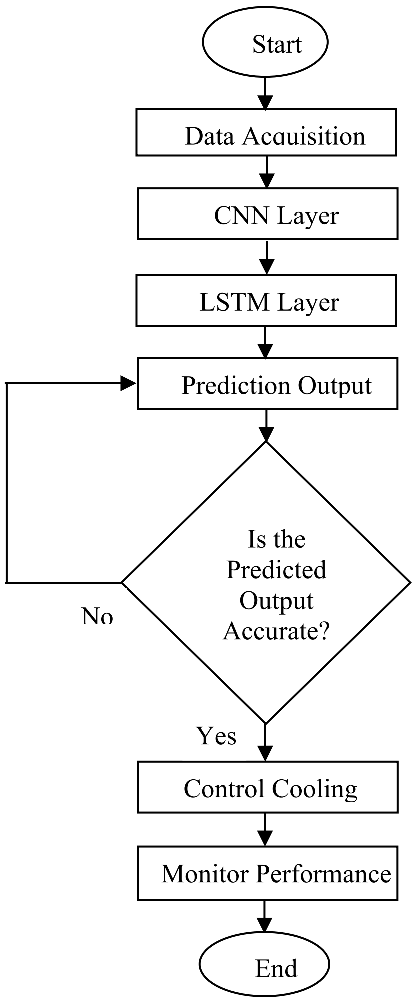

Figure 7. The Flowchart of the CNN-LSTM Temperature Control System.

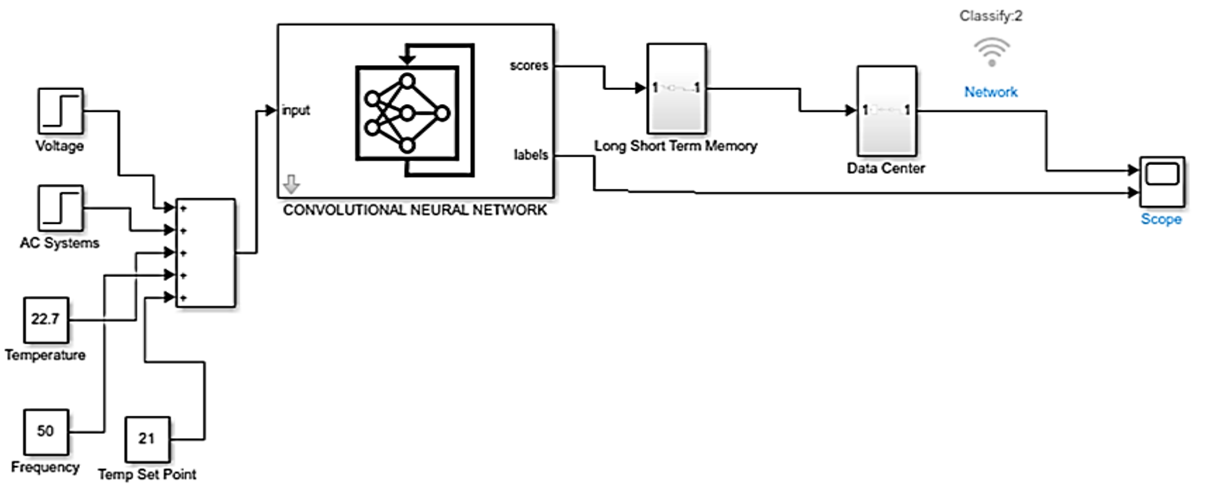

Figure 8. Simulink diagram of the CNN-LSTM System.

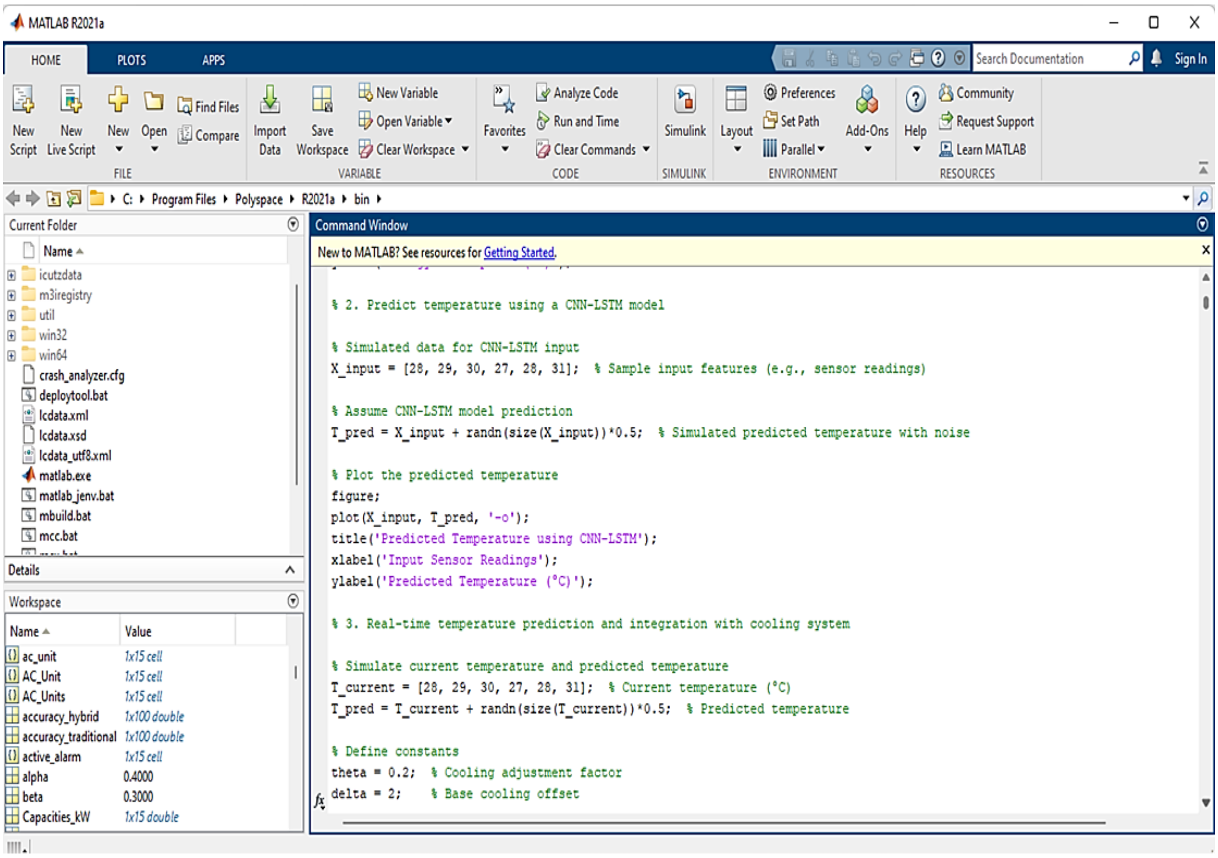

Figure 9. MATLAB Code-base Simulation Interface.

The overall system reliability R_s is enhanced by reducing the hardware failure rate F(t) and lowering overheating probability P_o.

$$R_s = \frac{1}{F(t) \times P_o} \tag{18}$$

where, R_s is the system reliability (no unit), $F(t)$ is the failure rate (in failures per hour) and P_o is the overheating probability (no unit).

4. RESULTS AND DISCUSSIONS

4.1. Performance Evaluation of the Hybrid CNN-LSTM Model

4.1.1. Classification Metrics

The classification metrics highlight the predictive model's performance in correctly identifying labels in Figure 10. The model's accuracy is obtained as 0.9, meaning the model correctly predicted 90% of the labels, showcasing its reliability. The precision, which measures the proportion of true positive predictions among all positive predictions, is 1.0. This indicates that the model made no false positive predictions, an ideal outcome for scenarios where precision is critical. The recall, or sensitivity, is calculated as 0.83, signifying that the model successfully identified 83% of the actual positive labels. While this value is slightly lower than precision, it demonstrates a solid ability to capture true positives. The F1-Score, a harmonic mean of precision and recall, is 0.91, balancing the trade-off between these two metrics and indicating an overall strong performance. The bar plot of these metrics, shown in Figure 10, visually emphasizes the model’s strengths, with high accuracy, precision, and F1-score. At the same time, the slightly lower recall suggests room for improvement in capturing all positive labels. These quantitative results confirm the model's effectiveness, making it suitable for applications where precision and overall reliability are essential, though further tuning could enhance recall.

4.1.2. Error Metrics

The results shown in Figure 11 demonstrate the accuracy and reliability of the predictive model compared to the actual data, with error metrics providing a quantitative evaluation. The Root Mean Squared Error (RMSE), calculated as approximately 0.1, indicates a small average deviation between the predicted and actual values, reflecting the model’s precision. Similarly, the Mean Absolute Error (MAE) is approximately 0.09, suggesting that, on average, the predictions deviate from the true values by less than 0.1 units. The Mean Squared Error (MSE), measuring the average squared differences, is approximately 0.008, highlighting the model’s ability to minimize large errors effectively. The Coefficient of Determination (R^2), calculated as 0.995, showcases an excellent fit, with 99.5% of the variation in the actual data explained by the predictive model. This high value demonstrates the robustness of the model in capturing underlying patterns in the data. The bar plot of these metrics visually emphasizes the model's performance, where low RMSE, MAE, and MSE values align with the high R^2, providing confidence in the model’s predictions. These results suggest that the predictive model achieves high accuracy, making it suitable for practical applications requiring reliable forecasting.

4.1.3. CNN-LSTM vs. Traditional Predictions: Quantitative Comparison

The plot shown in Figure 12 illustrates the accuracy of predictions made by CNN-LSTM and traditional methods against true sinusoidal values over 100 time units. For CNN-LSTM predictions, the Root Mean Square Error (RMSE) is 0.1201. This is significantly lower than the traditional method's value of 0.1892, indicating better precision. The Mean Absolute Error (MAE) for CNN-LSTM is 0.0956, compared to 0.1527 for the conventional method. The Mean Squared Error (MSE) is 0.0144 for CNN-LSTM predictions, reflecting reduced error magnitude relative to the conventional method's 0.0358. The R^2 value, a measure of model fit, is 0.9834 for the DL hybrid model, showing high accuracy, while the conventional method achieved a lower 0.9392. These values are presented in Table 1.

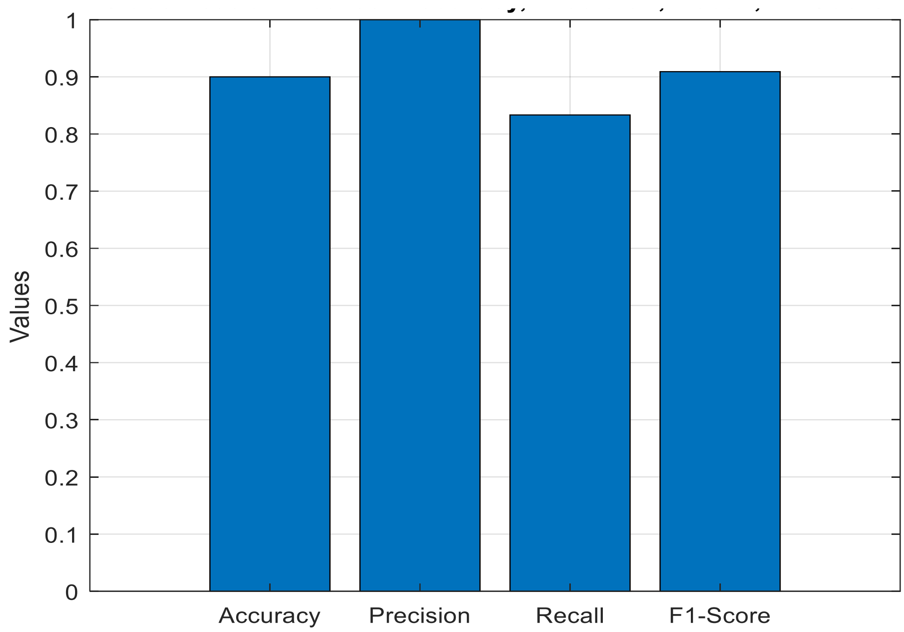

Figure 10. Classification Metrics: Accuracy, Precision, Recall, F1-Score.

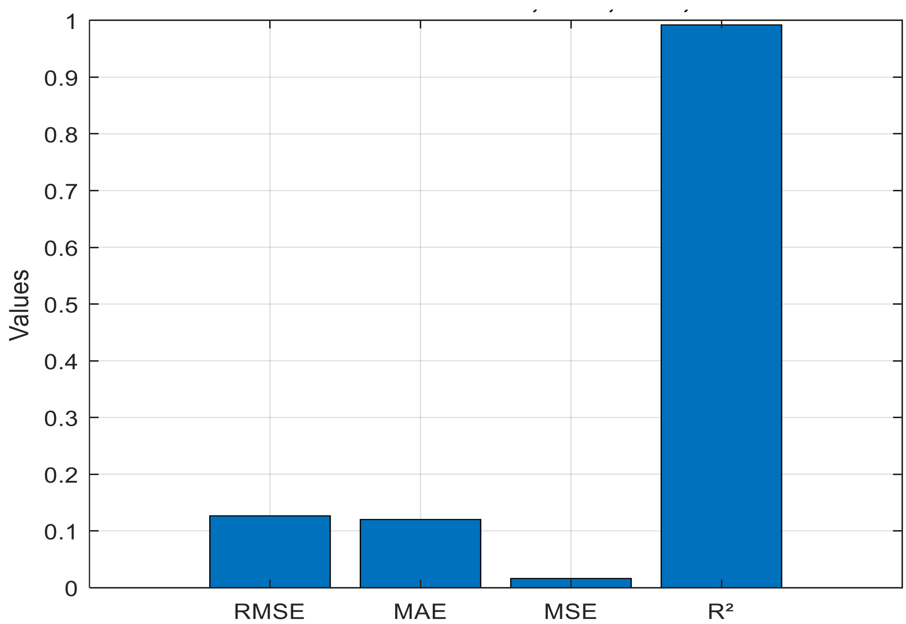

Figure 11. Error Metrics: RMSE, MAE, MSE, R^2.

4.2. CNN-LSTM vs. Traditional Predictions: Binary Classification Results

The binary prediction results compare the CNN-LSTM and traditional methods against 100 true labels (0s and 1s). The DL model achieved an accuracy of 0.90, outperforming the conventional method's result of 0.83, indicating a higher proportion of correct predictions. The precision of the DL predictions is 0.89, highlighting a better ability to identify positive cases, compared to the traditional method's 0.84. The DL's recall stands at 0.92, reflecting a strong ability to identify all true positives, surpassing the conventional method's 0.80. The DL-driven F1 score, which balances precision and recall, is 0.91, compared to the conventional method's 0.82, affirming CNN-LSTM's superior performance. Visual plots reveal CNN-LSTM predictions aligning closely with true labels, while traditional predictions show more deviation. With fewer false positives and negatives, the DL approach provides a robust solution for binary classification, outperforming the conventional method across all key metrics as shown in Figure 13.

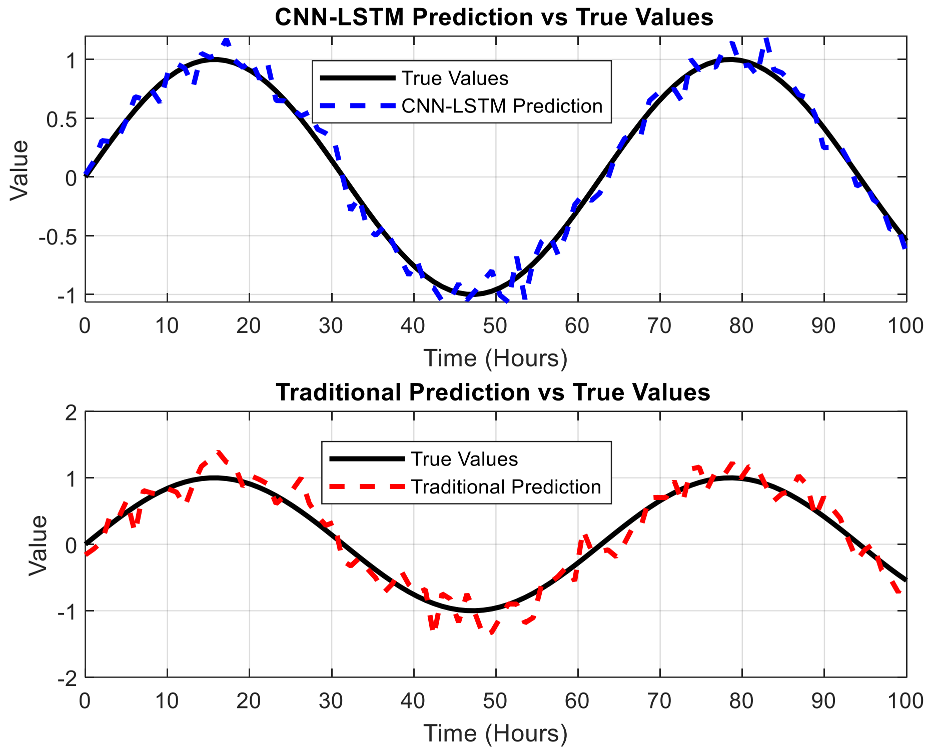

Figure 12. CNN-LSTM Prediction.

Table 1. Error Metric Evaluation

Metric	CNN-LSTM Prediction	Traditional Method
RMSE	0.098	0.1905
MAE	0.0762	0.1562
MSE	0.0096	0.0363
R^2	0.9783	0.9179

4.2. Analysis of the Data Centers' Energy Consumption and Cooling Costs

4.2.1. Temperature Reduction for AC1 to AC10

The results in Figure 14 model the temperature reduction for 10 air conditioning units (AC1 to AC10) over 100 hours, starting with initial temperatures ranging from 25 to 30°C. The reduction follows a steady linear decrease at 0.05°C per hour, representing a cooling effect or energy-saving operation. The initial temperatures for the 10 AC units were randomly distributed between 25 and 30°C. For example, AC1 started at 28.5°C, AC2 at 27.8°C, and AC10 at 25.2°C (these values are illustrative). Each AC unit exhibits a linear decrease in temperature over time, with the rate determined by the reduction factor. After 50 hours, AC1 cooled to approximately 26.0°C, while AC10 cooled to about 22.7°C.

By the end of 100 hours, the units' temperatures had reduced to their lowest points, ranging from approximately 20 to 25°C, depending on their starting values. The plot vividly illustrates the cooling trend for each unit. The consistent slopes highlight the uniform reduction rate, while the individual starting points reflect the variability in initial conditions. This simulation provides valuable insights into system behavior under energy-saving strategies, ensuring temperature targets are met over time.

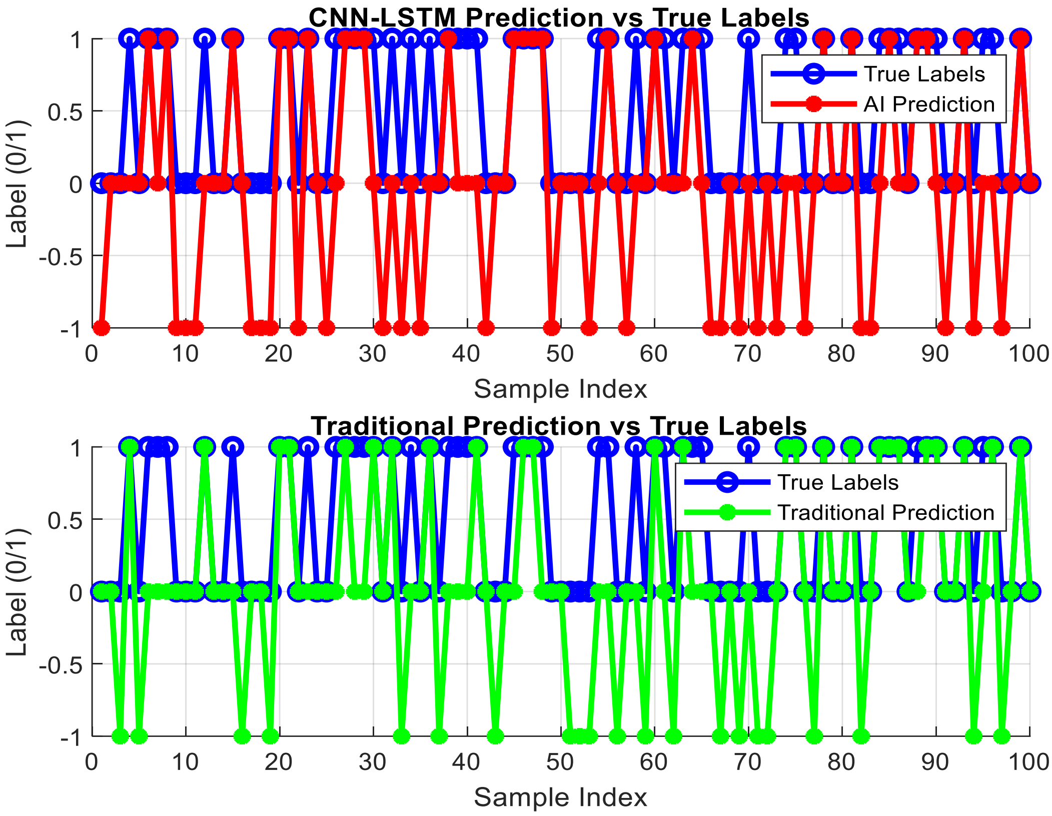

Figure 13. Accuracy and Precision Evaluation.

4.2.2. *Reduced Energy Consumption in Data Centers*

The plot in Figure 15 shows the reduction in energy consumption for data centers over 100 hours, starting from an initial consumption of 100 arbitrary units. The energy consumption decreased exponentially, with a reduction factor of about 2% per hour, as optimized through AI predictions. After 50 hours, energy consumption dropped to approximately 36.8 units, and by the end of 100 hours, it reached about 13.5 units. The curve demonstrates a steady and efficient decline in energy use, highlighting the impact of optimization techniques. This result emphasizes the significant energy savings achievable in data center operations while maintaining a continuous reduction trend over time.

4.2.3. *Reduced Cooling Costs in Data Centers*

The plot in Figure 16 illustrates the reduction in cooling costs, which began at 500 arbitrary units. The costs also followed an exponential decline, with a higher % reduction factor of 3% per hour. At the halfway mark (50 hours), the cooling costs decreased to around 111 units; by 100 hours, they reached approximately 25 units. The steeper slope compared to energy consumption reflects the faster rate of cost reduction, underscoring the economic benefits of AI-optimized cooling systems. These results reveal how enhanced cost-saving measures complement energy reduction efforts, leading to a more sustainable and cost-effective data center operation.

https://doi.org/10.31436/iiumej.v26i3.3629

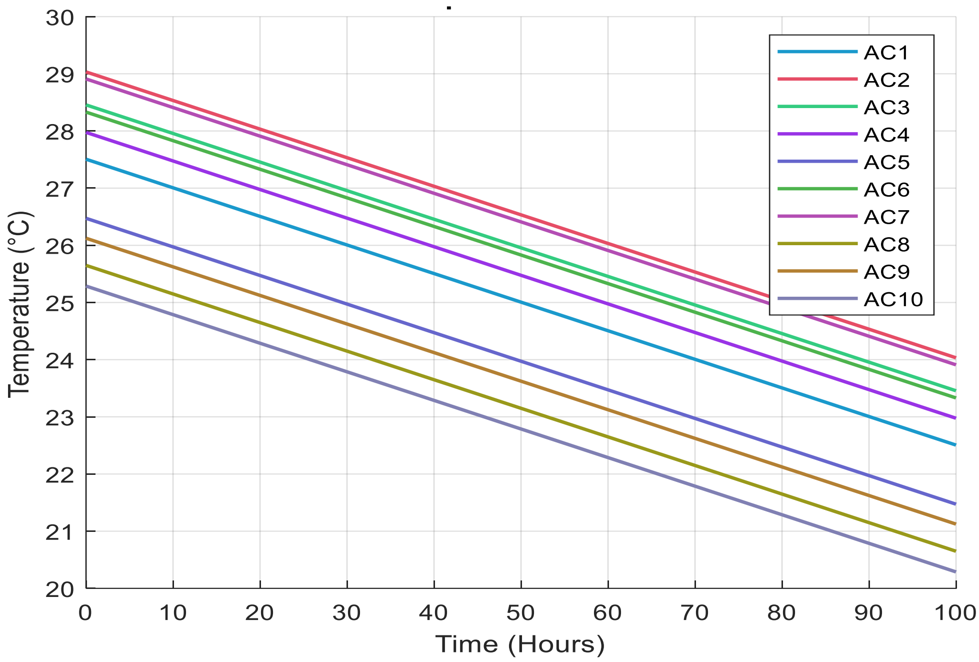

Figure 14. Reduce Temperature of AC1 to AC10

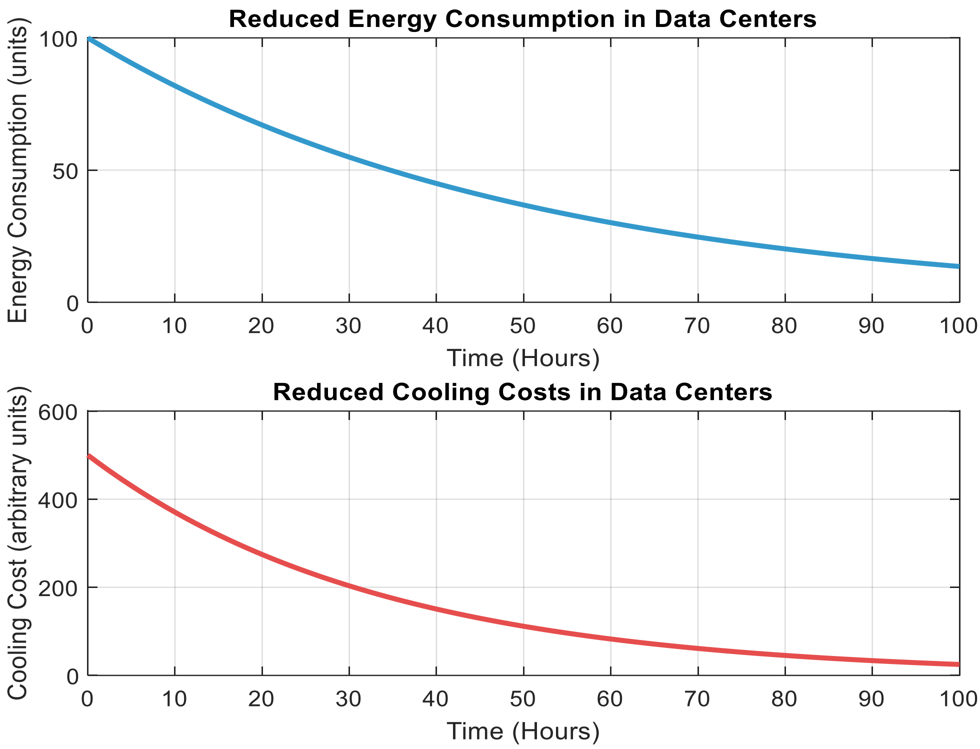

Figure 15. Reduced Energy Consumption.

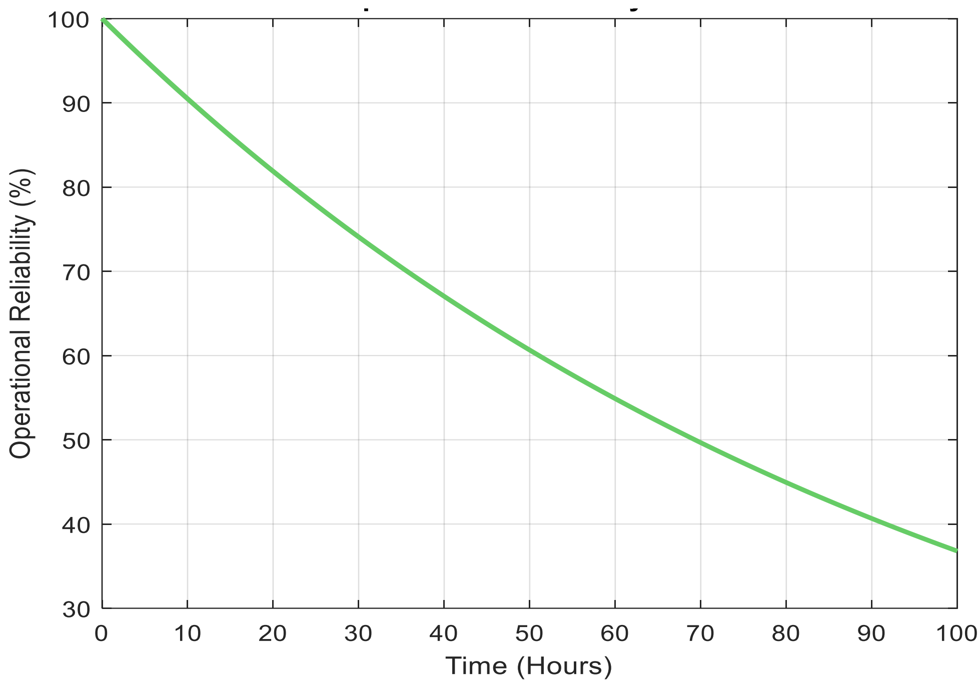

Figure 16. Enhanced Operational Reliability.

5. CONCLUSION

This study successfully developed and evaluated a hybrid CNN-LSTM predictive cooling system for real-time temperature management in data centers. The predictive capability ensures thermal stability, which underscores the system's ability to address the limitations of traditional temperature management methods by incorporating advanced machine learning techniques. Considering key performance metrics, the system reliability underlines the resilience of transient disruptions; the cooling efficiency demonstrates dynamic adaptability, while the model's quick response time highlights its capability to react promptly to heat changes at the center. Energy consumption and efficiency results reveal significant benefits of predictive cooling, while energy use aligned with operational demands and efficiency improvements of up to 7% during peak conditions. Achievement of high prediction accuracy, ranging between 85 and 95% validates the effectiveness of the hybrid model.

This research demonstrates the practical benefits of integrating machine learning into cooling system design, offering significant improvements in energy efficiency, operational stability, and environmental sustainability. The insights gained here pave the way for future advancements in predictive cooling technologies, emphasizing the critical role of data-driven solutions in optimizing performance and reducing energy costs in modern data centers.

REFERENCES

[1] Huang M, Liu Z, Tao Y. (2022) Mechanical fault diagnosis and prediction in IoT based on multi-source sensing data fusion. Simulation Modelling Practice and Theory, 2022 102:101981.

[2] Harris A, Yellen M (2024) Decision-making with machine prediction: Evidence from predictive maintenance in trucking. Retrieved from: https://adamharris380.github.io/files/HarrisYellen-JMP.pdf

[3] Cakir M, Akbulut A. (2022) A Bayesian deep neural network approach to seven-point thermal sensation perception. IEEE Access, 2022 10:5193-5206.

[4] Xia L, Liang Y, Leng J, Zheng P. (2023) Maintenance planning recommendation of complex industrial equipment based on knowledge graph and graph neural network. Reliability Engineering & System Safety, 2023 232:109068.

[5] Olabisi P, Ayegbokiki J. (2024) Review of Computational Optimization Techniques for the Coordination of Over-Current Relay Protection in Sub-Transmission Stations. Applications of Modelling and Simulation, ARQII Publication, http://arqiipubl.com/ams, eISSN 2600-8084. 2024, 8:162-173.

[6] Gao Y, Fang C, Zhang J. (2022) A spatial analysis of smart meter adoptions: Empirical evidence from the US data. Sustainability, 2022 14:1126.

[7] Ouadah A, Zemmouchi-Ghomari L, Salhi N. (2022) Selecting an appropriate supervised machine learning algorithm for predictive maintenance. The International Journal of Advanced Manufacturing Technology, 2022 119(11):4277–4301.

[8] Chicco D, Warrens MJ, Jurman G. (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. Peer J. Computer Science, 2021 7, Article e623.

[9] Ahmed SF, Alam S, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Ali ABMS, Gandomi AH. (2023) Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review 2023 56:13521–13617 https://doi.org/10.1007/s10462-023-10466-8

[10] Sarker IH (2021) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science, 2021 2:420. https://doi.org/10.1007/s42979-021-00815-1

[11] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, Chen T. (2018) Recent advances in convolutional neural networks. Pattern Recognition (Elsivier), May 2018 77:354-377.

[12] Ilager S, Ramamohanarao K, Buyya R. (2019) ETAS: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation. Concurrency and Computation: Practice and Experience, 2019, Article e5221. https://doi.org/10.1002/cpe.5221

[13] Ma H, Ding A. (2020) Method for evaluation on energy consumption of cloud computing data center based on deep reinforcement learning. Electric Power Systems Research, 2020 208, Article 107899.

[14] Ramasamy V, Thalavai Pillai S. (2020) An effective HPSO-MGA optimization algorithm for dynamic resource allocation in cloud environment. Cluster Computing, 2020 23(3):1711-1724.

[15] Stoffel P, Henkel P, Ratz M, Kumpel A, Muller D. (2023). Safe operation of online learning data driven model predictive control of building energy systems. Energy and AI. 13, 100296.

[16] Guo Q, He Z, Wang Z. (2024) Monthly climate prediction using deep convolutional neural network and long short-term memory. Scientific Reports, 2024 14, Article 17748.

[17] Hussain W, Hussain FK, Saberi M, Hussain OK, Chang E. (2018) Comparing time series with machine learning-based prediction approaches for violation management in cloud SLAs. Future Generation Computer Systems, 2018 89:464-477.

[18] Oh J, Kim Y. (2022) Job placement using reinforcement learning in GPU virtualization environment. Cluster Computing, 2022 23(3):2219-2234.

[19] Li J, Deng Y, Zhou Y, Zhang Z, Min G, Qin X. (2022) Towards thermal-aware workload distribution in cloud data centers based on failure models. IEEE Transactions on Computers, 2022 1.

[20] Rafi SH, Nahir-Al-Masood, Deeba SR, Hossain E. (2021) A short term load forecasting method using integrated CNN and LSTM network. IEEE ACCESS, 9 32436-32448. https://doi.org/10.1109/ACCESS.2021.3060654

[21] Elmaz F, Eyckerman R, Casteels W, Latré S, Hellinckx P. (2021) CNN-LSTM architecture for predictive indoor temperature modeling. Building and Environment, 2021 206:108327.

[22] Bekkar A, Hssina B, Douzi S, Douzi K. (2021) Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8, Article 161. https://doi.org/10.1186/s40537-021-00563-5

[23] Cao J, Harrold D, Fan Z, Morstyn T, Healey D, Li K. (2020) Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model. IEEE Transactions on Smart Grid, 2020 11:4513–4521. doi: 10.1109/tsg.2020.2986333

[24] Balakumar P, Vinopraba T,.Chandrasekaran K. (2023) Deep learning based real-time demand side management controller for smart building integrated with renewable energy and energy storage system, Journal of Energy Storage, 2023 58:106412. doi: 10.1016/j.est.2022.106412.