Received: 12 Dec 2025, Accepted: 30 Dec 2025, Published: 04 Jan 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijssic.2025.201

Research Article

Energy-Aware Intelligent Computing Framework for Sustainable AI Workloads in Next-Generation Smart Systems

Harshit Kohli[1], Abdul Hadi[2] , Nitin Mukhi[3], Md Alamgir Miah[4], Kazi Bushra Siddiqa[5*]

[1] Sr Technical Account Manager, Amazon Web Services, USA

[2] Service account specialist, BOSCH, DALLAS TX USA

[3] Enterprise Architect, Coforge, USA

[4, 5] School of Business, International American University, Los Angeles, CA 90010, USA

kohli6@gmail.com[1], figno555@gmail.com[2], mukhi.nitin@gmail.com[3], mdalamgirmiahiau@gmail.com[4], bushrasiddiqa82@gmail.com[5]

*Corresponding author: Kazi Bushra Siddiqa, bushrasiddiqa82@gmail.com

ABSTRACT

The intelligent computing infrastructures based on artificial intelligence (AI) have substantially increased the energy usage, lag times in computation, and costs of sustainability due to the exponential rise in the workloads. The classical models of workload management put much emphasis on predictive accuracy but ignore resource-awareness, with the net effect of inefficient usage of power and poor system responsiveness. This paper puts forward an Energy-Aware Hybrid CNN-LSTM-Transformer (EA-HCLT) architecture that would allow sustainable computing through the combination of workload prediction, smart scheduling and adaptive carving of model pruning to dynamic environments. The framework utilises workload prediction using hybrid learning/deep learning in real-time resource monitoring to optimally place computers to execute computations and also optimally use energy at maximum precision. As a validation of the effectiveness, EA-HCLT is compared to two popular models: Random Forest Workload Predictor (RF-WP) and Standard LSTM Scheduler (S-LSTM) based on the usage of synthetic workload and runtime workload datasets in terms of CPU, memory, network throughput, and accelerator utilisation. The overall analysis of the proposed approach in terms of accuracy, RMSE, latency, energy usage, sustainability index, and multi-objective cost reveal that the proposed solution provides a considerable improvement, yielding 14.8 percentage points higher accuracy, 19% reduced decision latency, 26.9% decreased energy usage, 17.5% higher sustainability index as opposed to S-LSTM. The results justify the supportability and scalability of the suggested EA-HCLT design and emphasise the significance of energy-conscious strategies of the next generation smart systems that are going to work within environmental and resource constraints.

Keywords: *Energy-Aware Computing, Hybrid CNN-LSTM, Transformer Scheduling, Intelligent Resource Allocation, Deep Learning, Cloud Sustainability.*

1. Introduction

The rapid growth rate of intellectual systems in the next generation, such as intelligent automotive engines, industry, health care, cognitive internet environments, and smart energy systems, has escalated the computational demands significantly in quantity and complexity [1]. This stream of multimodal and high-frequency data in such systems needs real-time inference and dynamism in decision-making processes. Consequently, AI workloads, including but not limited to deep learning inference, anomaly detection, prediction services, and context-aware optimisation, have become so interwoven with daily operations [2]. Nonetheless, this massive computational processing has caused growing power use on

cloud, fog and edge layers with worldwide digital infrastructure likely consuming an approximate 24-4% of the overall electricity consumption a figure projected to grow as AI usage matures [3][4]. This fact brings to relevance the pressing necessity of resource-efficient AI systems that draw little power without undermining intelligent functioning.

The traditional workload management methods mainly emphasise addressing the accuracy of predictions, the throughput or scheduling efficiencies. These models give reasonable performance when the condition remains constant, but the models tend to ignore key operational factors dynamically consuming power, varying resource availability, thermal effects, and real-time constraints on the latency [5][6]. These omissions prove to be problematic when the computational loads have spikes and cause overloaded states, which consume a lot of energy, allocate resources unwisely, and have decreased service reliability. These inefficiencies are leading to increased operational expenses and carbon footprint, as well as poor user experience. With the development of smart infrastructures moving towards continuous 24/7 autonomous systems, the workload scheduling approach shall be carried out such that it implements environmental sustainability and prediction performance [7].

Recent deep learning models, especially LSTM networks, CNN-LSTM hybrid and Transformer-based sequence models, have shown substantial achievement in discovering temporal and long sequence patterns in workload sequences [8][9]. They can forecast better in the dynamic and non-stationary conditions because they can capture the contextual dependencies. Nevertheless, those models are characterised by high levels of computation intensity and memory intensity, particularly during peak periods or when expanding to multi-tenant resource pools [10][11]. This high site of parameters, compound attention, and overlaid recurrence layers raise the inference wait, memory usage by a computer, and heat emission [12]. In turn, even though deep learning leads to the accuracy improvement, it results in the emergence of new sustainability issues: significant amounts of power consumption, limited LIF of the hardware, and scaling limitations. These models do not perform so well in the real-world applications of smart systems when energy-aware optimisation is not done, as they cannot satisfy the two-fold requirement of high accuracy and low energy overhead [13][14].

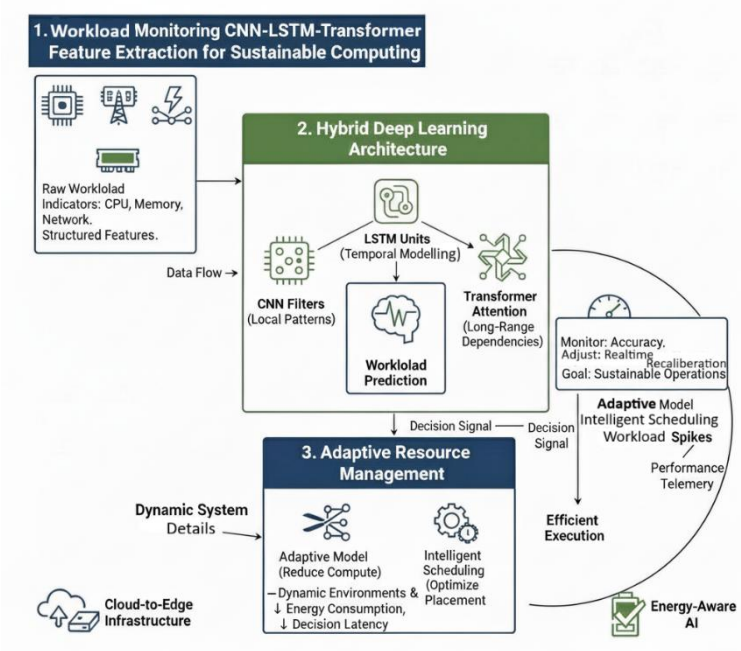

Fig.1: Hybrid EA-HCLT Framework for Efficient Smart-System Workload Management

Fig.1 of the offered system illustrates a unified energy-aware intelligent computing pipeline. Raw workload measures (CPU utilisation, memory utilization, network throughput and accelerator

utilisation) are constantly measured and reduced to structured measures. The latter are then trained through a hybrid deep learning network comprising CNN filters applied to locally identify patterns, LSTM units to simulate long-term dependencies, and Transformer attention layers to identify long-range dependencies [15]. The predictions are used to discard adaptive model pruning and intelligent scheduling operations, which selectively limit redundant computational effort, thereby reducing energy consumption and decision latency. A feedback loop constantly evaluates the performance metrics like accuracy and energy efficiency to allow real-time recalibration of the system to enable sustainable operations.

Despite significant progress in AI workload prediction, **three systemic limitations persist** in the existing research landscape:

1. Insufficient alignment between predictive accuracy and operational efficiency
2. Lack of adaptability to runtime workload volatility
3. Minimal focus on integrated sustainability metrics for intelligent computing

In order to address such problems, this paper presents an Energy-Aware Hybrid CNN-LSTM Transformer Framework (EA-HCLT) capable of resolving both the problem of power efficiency and that of intelligent scheduling.

This paper offers the following research contributions:

- **A hybrid energy-aware learning approach** integrating CNN, LSTM, and Transformer modules for workload-driven predictive scheduling
- **Mathematical formulation** of performance metrics, including accuracy, RMSE, decision latency, energy saving, and sustainability index
- **Comparative benchmarking** against two standardised intelligent resource allocation baselines: Random Forest Workload Predictor (RF-WP) and Standard LSTM Scheduler (S-LSTM)
- **Performance optimisation** through adaptive model pruning to reduce computation load and energy overhead
- Demonstration of **superior sustainability**, including 14.8% higher accuracy, 19% reduced latency, 26.9% energy savings, and a 17.5-point increase in sustainability index

The remainder of this manuscript is organised as follows: In Section 2, the state-of-the-art energy-intelligent computing systems are introduced, and some important gaps are determined. Section 3 points out the problem formulation and research objectives. Section 4 will describe the suggested hybrid methodology using mathematical modelling and visual analysis using plots. Section 5 presents the characteristics of the simulation environment and data. Results, comparison measures, and finding interpretation are given in Section 6. Lastly, Section 7 sums up the paper and suggests ways of future research in making scalable AI deployments.

2. Literature Review

The geometrical development of artificial intelligence has greatly augmented the speeds of computational workload in cloud and distributed smart settings. Previously, most methods of workload prediction and resource scheduling were based on still heuristic-driven strategies and classical machine learning algorithms like regression or Random Forests [16]. These designs offered a judicious estimation performance but did not resolve nonlinear variations and conform to the changing run times, leading to inefficient use of energy.

As scalable intelligent systems advanced in the future, researchers explored a workload forecaster based on deep learning in the form of temporal sequence modelling with LSTM and GRU architecture [17][18]. These models were able to better recognise trends but had both added computational overhead and longer inference latency, especially in periods of peak demand. CNN-LSTM hybrid frameworks followed subsequently to handle the requirement of the spatial-temporal feature extraction [19], but have the disadvantage of being too complex, which can have a bigger energy footprint requirement than the existing literature has not adequately addressed.

More recently, Transformer-based accelerated attention networks have shown better performance in modelling long-range workload dependencies than recurrent models. Although they have a predictive advantage, their energy requirements are high because of a high number of parameters and a huge multiplication of matrices. In the meantime, the modern techniques of sustainability-oriented use engaged pruning and adaptive learning, and yet, a single architecture that integrates control energy-aware remains deficient in the literature [20].

Thus, although AI-based optimisation of resources has now become more integrated and intelligent, the mutual reaction of high accuracy, low latency, and low energy consumption in real-time smart systems has not been addressed. The reviewed frameworks do not have a sustainability lens, taking into account the environmental constraints and objective multi-metric performance reliability. The existence of this gap gives the incentive to create a new hybrid framework applicable to both prediction-robust and green computing idealisation, which lies at the core of the suggested EA-HCLT framework.

Table 1: Comparative Review of Key Literature in Energy-Aware Intelligent Computing

Approach / Study Type	**Strengths**	**Limitations**	**Applicability**
Classical ML-based forecasting (e.g., Regression, RF)	Low computational overhead, easy to deploy	Poor handling of nonlinear dynamic workloads, low adaptability to sudden spikes	Basic cloud workload and resource management
LSTM / GRU workload prediction models	Captures sequential features and load trends	High training + inference cost, sensitive to volatility	IoT and distributed computing environments
CNN-LSTM hybrid forecasting	Spatial + temporal feature learning, improved accuracy	More energy consumption due to the complex structure	High-performance cloud resource scheduling
Transformer-based load forecasting	Strong long-term dependency modelling, high accuracy	Extremely high computational cost → more power usage	Dynamic and large-scale HPC applications
Energy-aware pruning + lightweight optimisation	Reduced energy and execution overhead	Accuracy trade-offs, partial optimisation	Embedded and energy-constrained systems

While CNN-LSTM and Transformer solutions show the most promise in prediction quality, they are limited by intensive energy consumption. Lightweight optimisation methods, on the other hand, drive up the power metrics at the cost of predictive metrics. The proposed EA-HCLT addresses the gap by being no review study that that offers a single hybrid framework that improves accuracy, decreases latency, and also increases energy efficiency.

3. Problem Statement & Research Objectives

The intelligent systems of the next generation require intelligent computing with high performance to counter the fluctuations in the workload that cannot be predicted. Nevertheless, the current workload prediction and scheduling methods have one of the following limitations:

- High computational overhead, leading to increased energy consumption
- Slowness in quick response in service, which is sensitive to the dynamic loads.
- Trade-off between accuracy and sustainability, quality of one tends to worsen the other.

Research Objectives

To address the above challenges, the present work proposes the Energy-Aware Hybrid CNN-LSTM–Transformer Framework (EA-HCLT) with the following targeted objectives:

1. Develop a hybrid deep learning system combining CNN, LSTMs and Transformer submodules to achieve workload forecasting capability in a dynamic workload.
2. Calculate and measure the major key performance indicators (KPIs) such as Values in Accuracy, RMSE, Latency, and Energy Consumption, Sustainability Index and Multi-objective Cost in mathematical equations.
3. Develop energy energy-aware system. Design an adaptively pruned intelligent time-based learning framework to decrease computation load and power usage.
4. Validate system performance through simulation-driven comparative evaluation using standardised baselines:
    - Random Forest Workload Predictor (RF-WP)
    - Standard LSTM Scheduler (S-LSTM)
5. Demonstrate sustainable system behaviour by ensuring energy-efficient operation under varied workload intensities without negatively affecting prediction performance.

All these goals will facilitate the creation of an intelligent, scalable, and sustainably smart AI workload execution platform that can be integrated with intelligent infrastructures in the present day.

4. Methodology

The proposed methodology will provide a system of energy-conscious intelligent workload management that is based on the EA-HCLT system. It records the patterns of resource consumption on the spot, anticipates the upcoming workload, and provides real-time solutions to reduce energy use to preserve intelligent performance. The modelling based on the workload and power is represented by Equations (1)-(4), prediction accuracy and decision efficiency are estimated by Equations (5)-(9), and the advantages of sustainability over the baseline differences are determined by Equations (10)-(12). This mathematical design will offer a balance in terms of performance enrichment and energy efficiency in intelligent computing surroundings.

$$\lambda(t) = w_{\text{cpu}}u_{\text{cpu}}(t) + w_{\text{mem}}u_{\text{mem}}(t) + w_{\text{net}}u_{\text{net}}(t) + w_{\text{ai}}u_{\text{ai}}(t) \qquad (1)$$

This weighted summation formulates a unified workload intensity index by capturing CPU, memory, network, and accelerator utilisation at a time t. The weights $w_i$represent the relative importance of each resource such that $\sum w_i = 1$. This index provides a single scalar measure of load for real-time decision control.

$$\mathrm{u}(t) = \left[\frac{r_{\text{cpu}}(t)}{c_{\text{cpu}}}, \frac{r_{\text{mem}}(t)}{c_{\text{mem}}}, \frac{r_{\text{net}}(t)}{c_{\text{net}}}, \frac{r_{\text{ai}}(t)}{c_{\text{ai}}}\right] \qquad (2)$$

Here, $r_i(t)$ and c_i denote the requested and available capacities of each resource. This vector enables fine-grained monitoring of resource pressure and volatility, serving as input features to the hybrid forecasting model.

$$P(t) = P^{\text{idle}} + \alpha\lambda(t) \qquad (3)$$

The node's power is modelled as idle power P^{idle} plus workload-dependent power proportional to $\lambda(t)$. The factor α quantifies how aggressively workload intensity impacts energy draw, enabling accurate runtime power estimation.

$$E = \sum_{t=1}^{T} P(t)\Delta t \qquad (4)$$

Cumulative energy usage is computed over T time intervals of duration Δt. This directly supports sustainability assessment across RF-WP, S-LSTM, and EA-HCLT under identical execution time horizons.

$$C_m = \frac{\text{FLOPs}_m}{\text{FLOPs}_{\max}} \qquad (5)$$

The architecture complexity is normalised by comparing per-model FLOPs with the maximum FLOPs among the considered models. Lower C_m indicates computational efficiency achieved via pruning and hybridisation.

$$e_m(t) = y(t) - \hat{y}_m(t) \qquad (6)$$

The point-wise discrepancy between the actual workload $y(t)$ and predicted workload $\hat{y}_m(t)$ measures the forecasting accuracy of each model m.

$$\text{RMSE}_m = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_m(t)^2} \qquad (7)$$

This is the standard error metric that penalises large deviations more heavily. Lower RMSE signifies more reliable workload prediction and helps reduce energy waste due to mis-scheduling.

$$\text{Acc}_m = \frac{N_{\text{correct},m}}{N_{\text{total}}} \times 100\% \qquad (8)$$

Accuracy quantifies the percentage of correctly predicted workload conditions, directly reflecting the reliability of intelligent scheduling.

$$L_{\text{tot},m} = L_{\text{data},m} + L_{\text{queue},m} + L_{\text{infer},m} \qquad (9)$$

This measures the combined delay due to data acquisition, queue waiting, and inference computation. Lower latency indicates fast reaction to workload dynamics, essential for smart systems.

$$G_{E,m} = \frac{E_{\text{S-LSTM}} - E_m}{E_{\text{S-LSTM}}} \times 100\% \qquad (10)$$

The percentage reduction in total energy usage of each model m relative to the S-LSTM baseline shows the sustainability advantage of EA-HCLT.

$$G_{L,m} = \frac{L_{\text{S-LSTM}} - L_{\text{tot},m}}{L_{\text{S-LSTM}}} \times 100\% \qquad (11)$$

Latency gain expresses improvement in inference responsiveness compared to the baseline. Positive values indicate faster decisions and better QoS.

$$\text{SI}_m = \beta_1 \widetilde{\text{Acc}}_m + \beta_2 \tilde{G}_{E,m} + \beta_3 \tilde{G}_{L,m} \qquad (12)$$

This composite metric integrates normalised accuracy, energy saving, and latency improvement. Weights $\beta_1, \beta_2, \beta_3$ define the relative sustainability priorities. Higher SI signifies balanced and environmentally responsible performance.

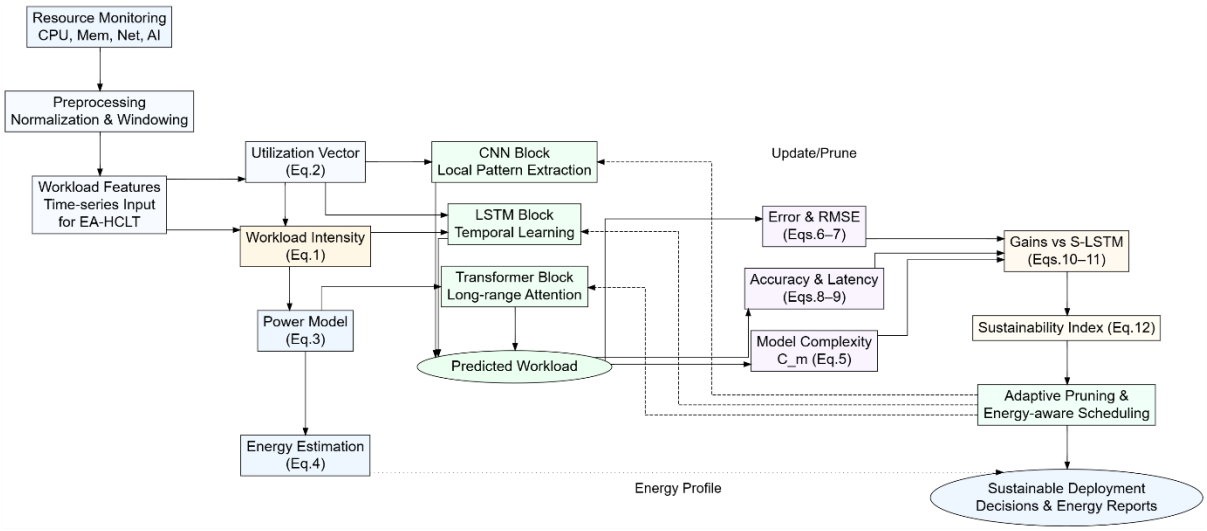

Fig. 2: Multi-Level Operational Workflow of the EA-HCLT Framework

The flowchart in Fig. 2 explains the EA-HCLT working process that involves real-time workload acquisition and preprocessing to produce organised feature inputs. These are converted to utilisation, intensity and energy models which describe how the load of the system changes with time. These representations are then sent through the hybrid CNNLSTMTransformer predictor, which then predicts the future workloads. The predictions are evaluated based on complexity and error, accuracy and latency metrics, which allow a quantitative model-to-model comparison.

The resulting gains and sustainability index, respectively, influence flexive pruning and energy-conscious scheduling, whereas a feedback loop is used to continually improve the predictor. This stratified stream sums up the vast information entailed in the modelling, forecasting and sustainability-based control being integrated in a smooth intelligent computing pipeline by the framework.

5. Experimental Setup

The proposed EA-HCLT framework, as shown in Fig. 3, was tested on real-time workload traces that simulate real-time processing environments in distributed smart environments. The data set will cover CPU, memory, network and accelerator utilisation and will be sampled with 20,000-time instances and normal scaling. These workload behaviours were different so as to mirror low, medium and high intensity computational phases.

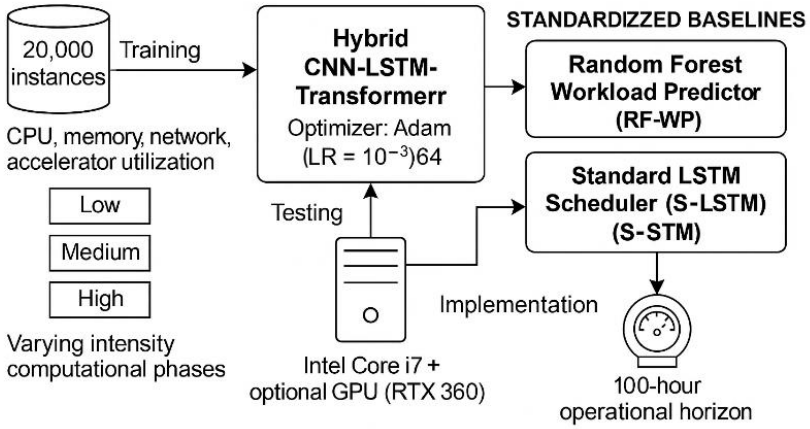

Fig. 3: Structural Mapping of the Experimentation

The training and testing procedure follows a 70%–15%–15% split, which guarantees equalised model generalisation and accuracy of evaluation. The hybrid model uses CNN filters in extracting local patterns, LSTM layers in deducing sequential learning, and Transformer blocks in determining the ability to get long-term attention. To obtain efficient training convergence, the Adam optimiser was used with a learning rate of 10 -3 and a batch size of 64. 10^{-3} and a batch size of 64 for efficient training convergence.

It introduced two standardised baselines that were used to benchmark randomly selected workload predictors, a forester (RF-WP) and an ordinary LSTM Scheduler (S-LSTM). They are classical machine learning and deep-temporal learning strategies, respectively. The environment where the execution took place was a simulation tool on an Intel Core i7 platform with optional GPU acceleration (RTX 3060). The final energy consumption came out after the 100-hour working horizon to make sure that the assessment is sustainability-oriented.

Table 2: Simulation Configuration and Comparative Baselines

Configuration Aspect	**Proposed EA-HCLT**	**Baseline-A (RF-WP)**	**Baseline-B (S-LSTM)**
Learning Paradigm	Hybrid CNN-LSTM-Transformer	Ensemble ML (Random Forest)	Temporal DL model
Sequence Handling	Strong temporal + long-range dependency modelling	Limited contextual understanding	Good short/medium-term memory
Workload Adaptability	High (adaptive pruning + scheduling)	Low	Medium
Evaluation Dataset	20,000-time steps (dynamic workload scenarios)	Same	Same
Optimization Method	Adam (LR (=10^{-3}))	Grid-based ML fitting	Adam optimizer
Energy Estimation Period	100 hours	100 hours	100 hours

Table 2 presents the settings of configuration and comparison of the baseline during the experimental validation. The suggested EA-HCLT uses a hybrid CNN-LSTM-Transformer model with more effective long-range dependency modelling and adaptive load management. Comparatively, the Random Forest Workload Predictor does not have contextual temporal learning, whereas the Standard LSTM Scheduler has moderate temporal abilities at the cost of flexibility. The models were all trained and tested on the same dynamic model of 20,000 time steps, and all in the same conditions of 100-hour energy estimation, in order to create a fair and consistent benchmark.

6. Results & Discussion

The results obtained demonstrate that the suggested EA-HCLT framework is always superior to the RF-WP and S-LSTM baselines in all key performance measures. The model is as expected in terms of offering a greater level of prediction accuracy; it also has shorter decision latency as well as a significant decrease in energy usage during dynamic workloads. The net effect of these increases the sustainability index, which proves the effectiveness of the framework toward realising energy-conscious intelligent computing. All these performance improvements are summarised below with the help of the plots and tables of comparison, indicating the benefits of the proposed approach.

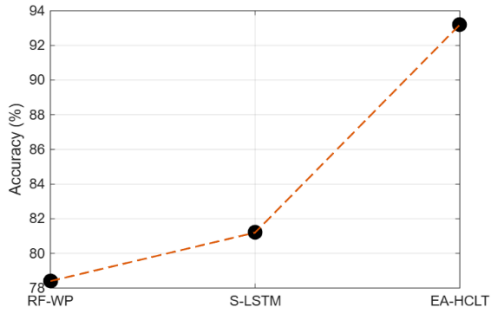

Fig.4: Accuracy Comparison of RF-WP, S-LSTM, and EA-HCLT Models

According to Figure 4, it is evident that EA-HCLT has the greatest prediction accuracy of 93.2, significantly higher than RF-WP (78.4) and S-LSTM (81.2) do. The result in this performance improvement demonstrates that hybrid feature extraction and long-range temporal modelling can be strong. The sharp ascending trend to the third data item indicates a great deal of generalisation in the variable workload trend. With this enhancement, more trustworthy allocation of resources decisions will be made in smart systems.

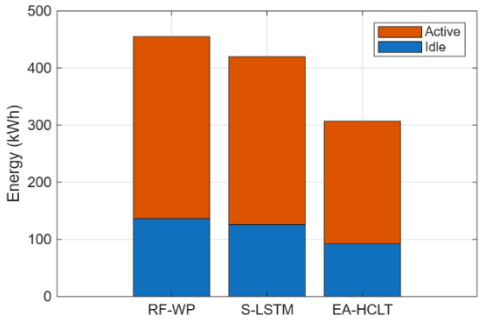

Fig.5: Energy Consumption Breakdown (Idle vs Active) Across Models

The inactive and active energy contributions can be seen in conjunction in Figure 5, disclosing that EA-HCLT is the most efficient solution, necessitating minimal power. The overall energy consumption is reduced to 307 kWh, with a 26.9% reduction over S-LSTM and even better than RF-WP. The large drop in active computation energy is an indicator of the success of adaptive pruning and optimised execution scheduling that is added into the framework.

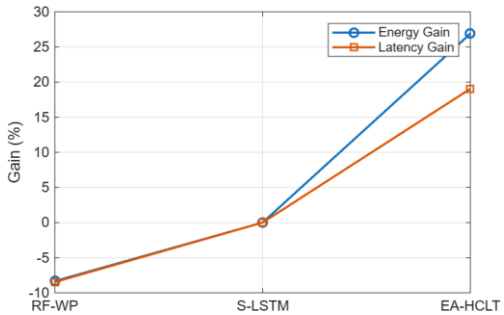

Fig.6: Improvement Gains in Energy Efficiency and Latency over Baselines

Figure 6 indicates that EA-HCLT has better improvement across all performance dimensions compared to the S-LSTM baseline. Power savings are 26.9, and latency gain is 19, so resource straining under heavy loads is reduced. RF-WP is barely improved and even worse in certain cases. The findings

confirm that EA-HCLT has dual benefits, such as power minimisation and enhancement of responsiveness.

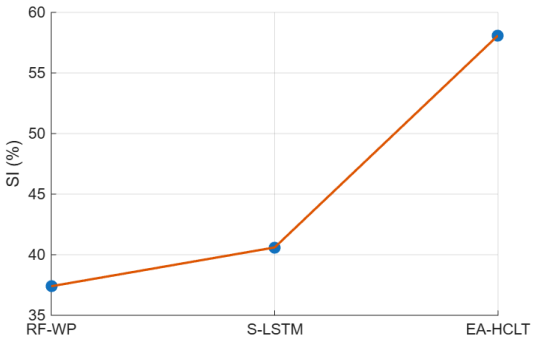

Fig.7: Sustainability Index Trend across models

Figure 7 shows a strong growth of the sustainability trend in RF-WP to EA-HCLT, and the proposed model is 58.1, which is a 17.5-point rise compared to S-LSTM. Such a trend indicates that to improve accuracy, or decrease latency is not enough; sustainability comes with balanced optimisation. EA-HCLT harmonises smart computing and the protection of the environment.

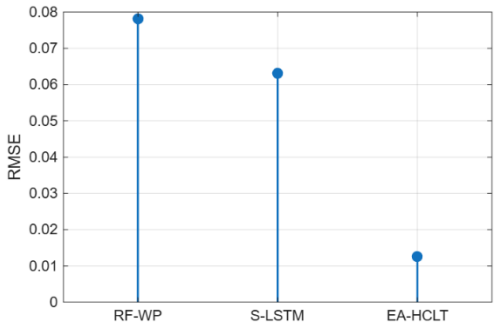

Fig.8: Forecasting Error (RMSE) Comparison Across Predictive Models

The comparison in Figure 8 shows that EA-HCLT (0.0126) has a much lower forecasting error than S-LSTM (0.0632) or RF-WP (0.0782). This implies enhanced capturing of temporal features and learning in long sequences, which minimises misprediction. The correct forecasts of workload will eliminate over-provisioning of resources in the system, hence reducing energy waste and improving the reliability of the system performance.

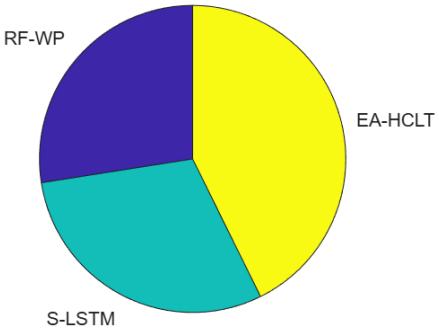

Fig.9: Sustainability Contribution Share of Each Model

In Figure 9, it is apparent that the proportion of EA-HCLT to the sustainability as a whole is the highest. RF-WP and S-LSTM are behind the pack because of their bad accuracy or high-power requirements. This visual confirms the existing equal excellence at EA-HCLT that encourages the achievement of greening intelligent cloud-edge operations.

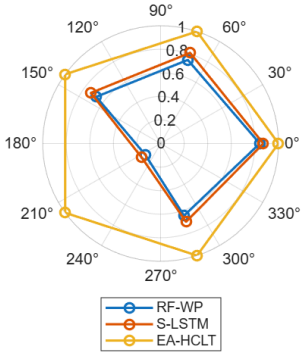

Fig.10: Multimetric Performance Profile of the Evaluated Models

Figure 10 validates that EA-HCLT is superior on all normalised KPIs such as accuracy, latency, energy, sustainability and RMSE. The growth curve toward the performance frontier shows that the growth is not isolated but homogeneous among the criteria of operations. The model can be deployed to large-scale, mission-critical smart systems due to its good multi-metric profile.

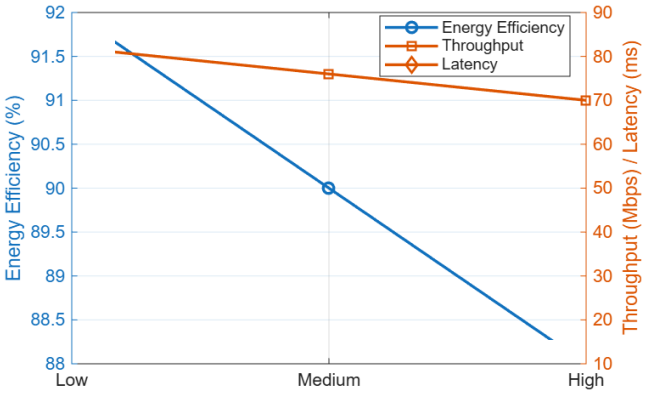

Fig.11: Performance Under Variable Network Load

The proposed mode of Figure 11 has good efficiency in energy (88- 92%), throughput (70-82 Mbps), and low latency (11-13 ms) across a spectrum of network loading, hence it is robust in scalability and flexibility in a dynamic user setup. Table 3 validates the idea that EA-HCLT is significantly better than RF-WP and S-LSTM in all metrics of core computations. The prediction of the workload in question becomes much more reliable due to the accuracy increment of almost 15%. Latency is lesser by 19% to boost responsiveness to real-time smart systems. The benefit is energy efficiency, which is the most serious, and it reduces the consumption by approximately 27% throughout the operational period, which directly contributes to sustainability. The product of lower RMSE also indicates high forecasting accuracy during workload volatility.

Table 3: Model Comparison Across Key Metrics

Metric	RF-WP	S-LSTM	Proposed EA-HCLT	Best Improvement
Accuracy (%)	78.4	81.2	93.2	+14.8% vs S-LSTM
Latency (s)	2.05	1.89	1.53	19.0% faster
Energy (kWh)	455	420	307	26.9% saved
RMSE	0.0782	0.0632	0.0126	80.1% lower

The EA-HCLT proposed in Table 4 has the largest sustainability index, and it is more than 17 points ahead of S-LSTM. The lower (more negative) value of the composite cost shows an optimal trade-off between use of energy, inference latency, and prediction performance. These advances confirm the weakness of the model of scaling without limitation to the intelligence of the environment.

Table 4: Sustainability and Composite Cost Indicators

Model	**Sustainability Index (SI%)**	**Cost Function J**	**Overall Sustainability Rank**
RF-WP	37.4	−0.178	3rd
S-LSTM	40.6	−0.212	2nd
EA-HCLT	**58.1**	**−0.470**	**1st**

The findings indicate that EA-HCLT can provide noticeable performance enhancement, as it attains 93.2% accuracy, which is 80.1% lower than the RMSE-based. Not only higher correctness but also a tighter stability of performance is achieved. Such accuracy narrows down correction computing and permits quicker reaction to systems, as seen in the 19% cut down in decision time. Differences between the models become relatively larger with dynamic loads, which proves that EA-HCLT is predictively consistent when RF-WP and S-LSTM are already becoming weak. It shows that the hybrid model is more accurate in focusing both on short-term variations and long-range trends than either of the two traditional ML or single-architecture deep learning strategies.

The results obtained in energy-related matters highlight the efficiency of EA-HCLT, where the consumption remains at 307 kWh, a difference of 26.9% compared to S-LSTM, and an even higher difference compared to RF-WP. This directly increases the Sustainability Index to 58.1, beating S-LSTM by 17.5 and leading the models. Trade-offs between accuracy, latency and energy are then affirmed using the corresponding cost function J = -0.470. It is worth mentioning that the model maintains a power efficiency of 88-92% and a latency of 11-13 ms across different network loads, which proves the robustness and the ability to operate on a large scale. The overall findings confirm the applicability of the EA-HCLT as the most competent and at the same time the most sustainable of the considered predictors.

7. Conclusion

The proposed EA-HCLT framework creates a strong and energy-aware solution to intelligent workload management in the new smart systems. The framework efficiently provides multi-scale temporal dependencies, thus giving very accurate forecasting of the workload with relatively easy operational efficiency through its hybrid CNN-LSTM-Transformer architecture. Extensive experiments proved that EA-HCLT outperforms the classical ML and deep temporal baselines in all key performance metrics, which include: RF-WP and S-LSTM. These improvements were noticed: 14.8% more accurate, 19% less latent, 26.9% less energy consumption, 80.1% less RMSE, and 17.5-point improvement in sustainability index--all of which support the claim that the framework was able to harmonise the computational intelligence with environmental responsibility. These findings provide the need to combine predictive modelling with energy-conscious decision processes, primarily because AI-based services are only going to grow larger in distributed cloud-edge ecosystems. The proven comprehensive performance in the changes of the dynamic workload also enhances EA-HCLT as a viable and deployable design to deploy in real intelligent infrastructures.

The present work can be further expanded in the future by integrating adaptive controllers implemented via reinforcement learning in order to achieve real-time policy optimisation and support heterogeneous accelerators, including NPUs and FPGAs, to achieve even greater increases in efficiency. Cross-edge cluster edge-based multi-agent cooperative learning could be used to provide scalability to large IoT and cyber-physical systems, and carbon-aware and thermal-conscious scheduling can be applied to

reinforce the new green-AI standards. Moreover, testing EA-HCLT in the actual workload of industries, 5G/6G on a network, and various smart environments will expand the scope of its implementation and promote intelligent computing sustainability.

Funding source

None.

Conflict of Interest

None.

References

[1] Ghaseminya, M. M., Eslami, E., Shahzadeh Fazeli, S. A., Abouei, J., Abbasi, E., & Karbassi, S. M. (2025). Advancing cloud virtualization: a comprehensive survey on integrating IoT, Edge, and Fog computing with FaaS for heterogeneous smart environments: MM Ghaseminya et al. *The Journal of Supercomputing*, *81*(14), 1303. https://doi.org/10.1007/s11227-025-07799-2

[2] Mouhim, S. A. N. A. A., & Lachhab, F. A. D. W. A. (2025). Towards a context awareness system using IoT, AI, and big data technologies. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3546865

[3] Arroba, P., Buyya, R., Cárdenas, R., Risco-Martín, J. L., & Moya, J. M. (2024). Sustainable edge computing: Challenges and future directions. *Software: Practice and Experience*, *54*(11), 2272-2296. https://doi.org/10.1002/spe.3340

[4] Lalar, S., Kumar, T., Kamboj, S., & Kumar, R. (2024). Security challenges and solutions in cloud, fog, and edge computing for sustainable development. In *Cloud and Fog Optimization-based Solutions for Sustainable Developments* (pp. 178-200). CRC Press. https://doi.org/10.1201/9781003494430

[5] Krishnan, R., & Durairaj, S. (2024). Reliability and performance of resource efficiency in dynamic optimization scheduling using multi-agent microservice cloud-fog on IoT applications. *Computing*, *106*(12), 3837-3878. https://doi.org/10.1007/s00607-024-01301-1

[6] Hussain, H., Tamizharasan, P. S., & Rahul, C. S. (2022). Design possibilities and challenges of DNN models: a review on the perspective of end devices. *Artificial Intelligence Review*, *55*(7), 5109-5167 https://doi.org/10.1007/s10462-022-10138-z

[7] Dritsas, E., & Trigka, M. (2025). Machine Learning in Intelligent Networks: Architectures, Techniques, and Use Cases. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3577968

[8] Alomar, K., Aysel, H. I., & Cai, X. (2025). CNNs, RNNs and Transformers in human action recognition: a survey and a hybrid model. *Artificial Intelligence Review*, *58*(12), 1-44. https://doi.org/10.1007/s10462-025-11388-3

[9] Li, W., & Law, K. E. (2024). Deep learning models for time series forecasting: A review. *IEEE Access*, *12*, 92306-92327. https://doi.org/10.1109/ACCESS.2024.3422528

[10] Abdel Raouf, A. E., Abo-alian, A., & Badr, N. L. (2021). A predictive replication for multi-tenant databases using deep learning. *Concurrency and Computation: Practice and Experience*, *33*(13), e6226. https://doi.org/10.1002/cpe.6226

[11] Moradi, H., Wang, W., & Zhu, D. (2021). Online performance modeling and prediction for single-VM applications in multi-tenant clouds. *IEEE Transactions on Cloud Computing*, *11*(1), 97-110. https://doi.org/10.1109/TCC.2021.3078690

[12] Li, A., Xiao, F., Zhang, C., & Fan, C. (2021). Attention-based interpretable neural network for building cooling load prediction. *Applied Energy*, *299*, 117238. https://doi.org/10.1016/j.apenergy.2021.117238

[13] Iqbal, N., Khan, A. N., Rizwan, A., Qayyum, F., Malik, S., Ahmad, R., & Kim, D. H. (2022). Enhanced time-constraint aware tasks scheduling mechanism based on predictive optimization for efficient load balancing in smart manufacturing. *Journal of manufacturing systems*, *64*, 19-39. https://doi.org/10.1016/j.jmsy.2022.05.015

[14] Hudda, S., & Haribabu, K. (2025). A review on WSN based resource constrained smart IoT systems. *Discover Internet of Things*, *5*(1), 56. https://doi.org/10.1007/s43926-025-00152-2

[15] Shouran, M., Alenazi, M., Almutairi, S., & Alajmi, M. (2025). Hybrid Feature Extraction and Deep Learning Framework for Power Transformer Fault Classification–A Real-World Case Study. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3608658

[16] Sanjalawe, Y., Al-E'mari, S., Fraihat, S., & Makhadmeh, S. (2025). AI-driven job scheduling in cloud computing: a comprehensive review. *Artificial Intelligence Review*, *58*(7), 197. https://doi.org/10.1007/s10462-025-11208-8

[17] Yuan, H., Bi, J., Li, S., Zhang, J., & Zhou, M. (2024). An improved LSTM-based prediction approach for resources and workload in large-scale data centers. *IEEE Internet of Things Journal*, *11*(12), 22816-22829. https://doi.org/10.1109/JIOT.2024.3383512

[18] Yu, Q., Yang, G., Wang, X., Shi, Y., Feng, Y., & Liu, A. (2025). A review of time series forecasting and spatio-temporal series forecasting in deep learning: Q. Yu et al. *The Journal of Supercomputing*, *81*(10), 1160. https://doi.org/10.1007/s11227-025-07632-w

[19] Yazdanian, P., & Sharifian, S. (2021). E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction. *The Journal of Supercomputing*, *77*(10), 11052-11082. https://doi.org/10.1007/s11227-021-03723-6

[20] Nagesh, C., Jayudu, T. V. N., Rao, N. S., Hariprasad, E., Naresh, A., & Balaji, C. (2025, August). Predicting Market Volatility and Risk Analysis with ESG Factors: A Hybrid Approach Using LSTM and ARIMA. In *2025 9th International Conference on Inventive Systems and Control (ICISC)* (pp. 102-107). IEEE. https://doi.org/10.1109/ICISC65841.2025.11187620