International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies
Vol. 03, Iss. 02, S. No. 02, pp. 16--28, 2026
ISSN (Online): 3048-8788

Received: 17 May 2026, Accepted: 04 August 2026, Published: 31 August 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijaimd.2026.263

Research Article

AI-Driven Cloud Analytics and Hardware-Assisted Edge Intelligence for Real-Time Cyber-Physical Infrastructure Management

Naveen[1], Satyam Kumar Sainy[2]

[1]IILM University, Greater Noida, India.

[2]Galgotias University, Greater Noida, India

naveenbasoya@gmail.com, satyamit4u@gmail.com

*Corresponding author: Naveen, naveenbasoya@gmail.com

ABSTRACT

Real-time monitoring and smart decision-making are required in cyber-physical infrastructures such as smart grids, transportation systems, and industrial automation systems to ensure process efficiency and system resilience. However, higher latency, bandwidth constraints, and a lack of responsiveness to time-sensitive data streams haunt traditional cloud-based architectures. To address these challenges, a coherent framework for integrating AI-based cloud analytics with edge intelligent hardware for managing cyber-physical infrastructure is proposed in the following paper. The architecture is based on distributed edge nodes, with hardware accelerators for very low-latency inference, and cloud layers that perform large-scale analytics and optimisation of global models. A hybrid resource allocation strategy is a self-regulating plan for the allocation of computing workloads across cloud and edge environments. Also, an adaptive learning mechanism improves prediction accuracy in changing operational environments. The framework is mathematically designed to achieve optimal latency, throughput, and computational efficiency. The proposed system reduces latency by 145ms to 92 (≈36.5) units and increases prediction accuracy from 81.2% to 96.4% when applied to a dynamic workload. The convergence analysis shows that standardized quicker convergence occurs after 35 iterations as opposed to 60 iterations in the models at the baseline. Additionally, the throughput is proceeding at 520 requests to 780 requests and error rates are falling by a factor of around 41, ensuring better reliability. Relative performance analysis across various scenarios indicates consistent improvement in both edge-dominant and cloud-dominant setups. The results of this study demonstrate that, when combined with hardware-based edge intelligence, AI-powered cloud analytics can significantly enhance the responsiveness, scalability, and decision accuracy in cyber-physical infrastructure systems.

Keywords: *Cloud Analytics, Edge Intelligence, Cyber Physical Infrastructures, Edge Computing, Resource Optimization.*

1. Introduction

The system utilizes next-generation cyber-physical systems (CPS) to allow easy synchronization of physical processes with computation intelligence in smart grids, transportation and industrial automation [1]. The informative high volume data streams created by such systems require processing them in real-time to facilitate proper monitoring and control. Cloud computing has scalable solutions to big-data analytics; however, its nature is

centralized, creating latency and bandwidth overhead and responsiveness overhead, which may not be utilized with a time-sensitive CPS system.

Edge computing has become a complementary paradigm for reducing such limitations, as it enables data processing to be closer to the data source. This reduces the latency and makes them more responsive. Edge devices are usually resource-constrained and they cannot be used to execute complex AI models. Consequently, the current methods have a trade-off between simulation and real-time behavior, which prevents their usefulness in dynamic CPS systems.

The latest development in AI and hardware acceleration has the potential to provide a solution. The analytics are improved with AI to bolster predictive decision making, whereas the edge intelligence can be performed through hardware to run computationally intensive models in the edge [2][3]. However, the existing systems do not have a single framework that can coordinate dynamically the cloud and edge resources to minimize the latency, accuracy and the efficiency of the system.

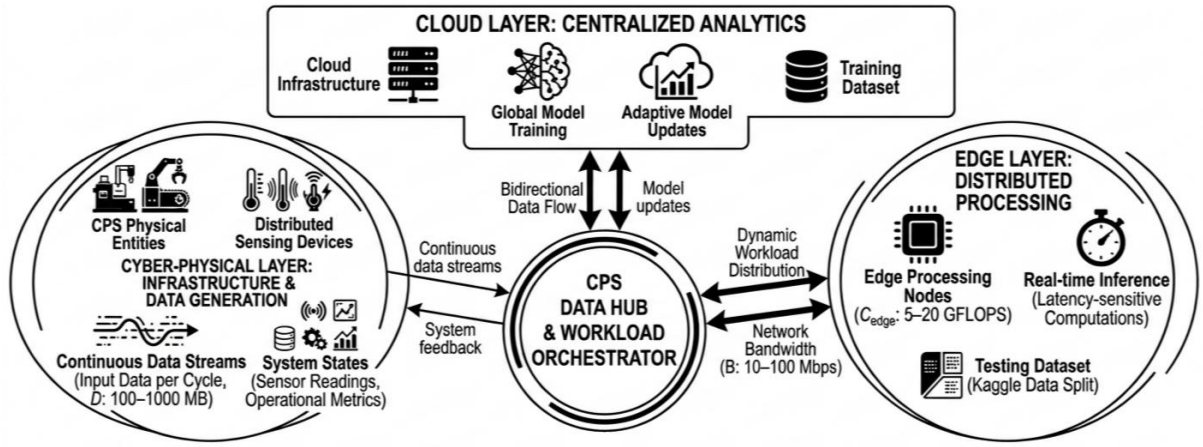

Fig. 1: AI-Driven Cloud–Edge Intelligence Framework for Real-Time Cyber-Physical Infrastructure Management

This work addresses this gap by proposing an integrated framework presented in Fig. 1 combining AI-driven cloud analytics with hardware-assisted edge intelligence for real-time cyber-physical infrastructure management. The proposed system introduces a dynamic workload allocation mechanism that distributes tasks between cloud and edge layers based on system conditions and latency requirements [4]. Additionally, an adaptive learning model is incorporated to improve prediction accuracy under varying operational scenarios, supported by mathematical formulations for latency, throughput and resource optimization.

The system consists of three layers, where the layer CPS devices generate data and run real-time inference with hardware-accelerated edge nodes and cloud platforms run global analytics. The active feedback loop allows to distribute the workload efficiently and continuous learning at different levels.

Contributions of this work include:

- An edge intelligence coupled with hardware aid based on AI analytics at the hybrid cloud-edge architecture.
- A dynamic resource allocation strategy for latency-aware task distribution.
- A mathematical framework for optimizing latency, throughput and resource utilization.

- A hardware-aware edge inference mechanism for real-time decision-making.
- Performance validation demonstrating improvements in latency, accuracy and system throughput.

The rest of the paper is organized in the following way. Section 2 provides an overview in related research, Section 3 provides the methodology, Section 4 details the experimental setup, Section 5 presents the results and an overall conclusion of the work is provided in Section 6.

2. Related Research

Integration of artificial intelligence and the cyber-physical systems (CPS) has gained a lot of attention as a result of the fact that it may improve real-time decision-making and automation of the systems. There are three main areas of interest: the use of cloud-based analytics to serve CPS, edge intelligence and low-latency computing using hardware accelerators [5][6]. Although both fields do bring beneficial innovations, there is still a lack in the overall application of both in combination.

CPS architectures based on clouds have already been investigated on a large scale in data processing and predictive analytics [7]. These systems also take advantage of the centralized resources by doing complicated computations and making the right predictions and optimization of the system-wide [8]. There are some studies that attribute the inherent weakness of a cloud-only solution and more so when the application is latency conscious. This can cause delays and bandwidth overheads since it relies on the constant flow of data transmission and the responsiveness of the system can be impaired by dynamics [9]. These limitations are more acute as CPS applications are increasingly required to be run in real-time.

What can be described as the problem of latency, edge computing, a decentralized approach, has been proposed to take the computation nearer to the source data. The frameworks of edge intelligence allow decentralized data processing, which lowers the data communication latency and enhances responsiveness [10][11]. Recent publications illustrate how it is possible to successfully implement lightweight AI in edge nodes to perform functionalities like anomaly detection and predictive maintenance. Although these systems have these pros, their capability to execute complex models or large-scale analytics is usually limited by factors such as access to computational power and energy efficiency [12].

The introduction of the hybrid cloud-edge systems is a significant advancement towards striking a balance between efficiency and latency of the computation. The solutions distribute the workloads of each layer: the cloud and the edge and attempt to exploit the functionality of both paradigms. Most of the existing models have been based on the use of static or heuristic-based task allocation mechanisms, which are unable to change with the changing system conditions [13]. This leads to the inefficiency of performance especially in highly dynamic CPS environments where workload characteristics and network conditions change with time.

In parallel, hardware acceleration technologies such as GPUs, TPUs and FPGAs have been explored to enhance computational performance, particularly for AI inference tasks [14]. Hardware-assisted edge intelligence enables faster execution of deep learning models, significantly reducing inference latency. Several studies demonstrate that incorporating specialized hardware can improve throughput and energy efficiency. Nevertheless, most implementations focus on isolated optimization at either the cloud or edge level, without considering system-wide coordination [15].

The second important limitation of the current research is the absence of a unified mathematical modeling of system optimization. Although there are some works that focus on the reduction of latency or resource allocation separately, there are not many works that offer a single framework that can optimize both the latency, throughput and prediction accuracy. Also, adaptive learning schemes that keep on improving model performance in different operational environments are not well explored.

In summary, existing studies present three major gaps: they fail to integrate cloud analytics and edge intelligence, use strategies to allocate workloads that are dynamic and adaptable and use hardware acceleration as part of a single optimization strategy. These limitations motivate the development of a comprehensive approach that combines AI-driven cloud analytics with hardware-assisted edge intelligence for real-time CPS management. Table 1 compares the existing approaches in accordance with the present requirements.

Table 1: Comparative Analysis of Existing Approaches

Approach Type	**Key Features**	**Advantages**	**Limitations**
Cloud-Centric Systems [16]	Centralized data processing	High computational capability	High latency, bandwidth dependency
Edge-Based Systems [17]	Localized processing at edge nodes	Low latency, faster response	Limited computational resources
Hybrid Cloud–Edge Models [18]	Distributed workload execution	Balanced performance	Static allocation, limited adaptability
Hardware-Accelerated AI [19]	Use of GPUs/TPUs/FPGAs	Faster inference, improved efficiency	Lack of system-level integration
Proposed Framework	AI + Cloud + Hardware Edge Integration	Low latency, high accuracy, adaptive	Requires coordinated resource management

The constraints that have been identified in the current literature show the necessity of a composite and dynamic model. The suggested work fills these gaps by proposing an integrated architecture that will incorporate AI-based cloud analytics, edge intelligence that will be supported by hardware and workload optimization that will be dynamic to provide efficient and real-time cyber-physical infrastructure management.

3. Methodology

In this section, the proposed AI-based cloud-edge integrated framework along with hardware-assisted intelligence to real-time management of cyber-physical infrastructure is provided. The architecture runs on three layers closely integrated with each other, including the cyber-physical layer (data generation), the edge intelligence layer (low-latency processing with hardware acceleration) and the cloud analytics layer (global optimization and large-scale learning). Information provided by the infrastructure, including smart grids, autonomous transport and industrial IoT devices, undergoes initial processing on edge nodes to make immediate decisions and further analytics and model revision on the cloud layer. A two-way communication system allows the dynamism in the allocation of workloads depending on the state of the system, which is the best coordination of latency-sensitive and computation-intensive tasks.

To formalize system behavior and optimize performance, the proposed framework is modeled using the following equations. Each variable is explicitly defined and practical applications within cyber-physical systems are highlighted.

(1) Total System Latency

$$L_{total} = L_{edge} + L_{transmission} + L_{cloud} \quad (1)$$

The total system latency L_{total} in Eq. (1) is defined as the sum of edge processing latency L_{edge}, transmission delay $L_{transmission}$ and cloud processing latency L_{cloud}. It is employed in applications, which require latency, like smart traffic control and grid fault detection, where the overall response time is of importance.

(2) Edge Processing Latency

$$L_{edge} = \frac{D}{C_{edge}} \quad (2)$$

The edge processing latency L_{edge} in Eq. (2) is computed as the ratio of input data size D to the computational capacity of the edge node C_{edge}, indicating that higher edge capability reduces processing delay. This model is applied in industrial automation systems for rapid sensor data processing.

(3) Cloud Processing Latency

$$L_{cloud} = \frac{D}{C_{cloud}} \quad (3)$$

The cloud processing latency L_{cloud} is expressed as the ratio of data size D to the computational capacity of cloud resources C_{cloud}, reflecting the ability of cloud systems to efficiently handle large-scale computations. Equation (3) is particularly useful in predictive analytics for smart energy management.

(4) Transmission Delay

$$L_{transmission} = \frac{D}{B} \quad (4)$$

The transmission delay $L_{transmission}$ in Eq. (4) is defined as the ratio of data size D to available network bandwidth B, showing that higher bandwidth reduces communication latency. This equation is relevant in remote monitoring systems where network conditions directly influence system responsiveness.

(5) Resource Allocation Optimization

$$R_{opt} = \arg\min(\alpha L_{total} + \beta E + \gamma C) \quad (5)$$

The optimal resource allocation R_{opt} is determined by minimizing a weighted objective function that includes total latency L_{total}, energy consumption E and computational cost C,

where α, β and $\gamma$are weighting coefficients representing system priorities. Equation (5) is applied in smart city infrastructure to dynamically distribute workloads between cloud and edge environments.

(6) Throughput Model

$$T = \frac{N}{L_{total}} \tag{6}$$

The system throughput T is defined under Eq. (6) as the ratio of the number of processed requests N to the total latency L_{total}, indicating that lower latency results in higher throughput. This metric is critical in large-scale CPS such as transportation systems handling high request volumes.

(7) Prediction Accuracy

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

The prediction accuracy A is calculated using Eq. (7) as the ratio of correctly classified instances, represented by true positives TP and true negatives TN, to the total number of predictions including false positives FP and false negatives FN. This equation is widely used in anomaly detection and cybersecurity applications within CPS.

(8) Adaptive Learning Update

$$W_{t+1} = W_t - \eta \nabla L(W_t) \tag{8}$$

The adaptive learning update given by Eq. (8) defines the new model weights W_{t+1} as the previous weights W_t adjusted by the learning rate $\eta$and the gradient of the loss function $\nabla L(W_t)$, enabling continuous model improvement. This approach is essential in dynamic CPS environments such as autonomous systems and predictive maintenance.

(9) Hardware Acceleration Gain

$$G = \frac{T_{normal}}{T_{accelerated}} \tag{9}$$

The hardware acceleration gain G is defined in Eq. (9) as the ratio of execution time without acceleration T_{normal} to execution time with hardware support $T_{accelerated}$, indicating performance improvement due to specialized hardware. This is particularly useful in edge AI systems requiring real-time inference.

(10) System Efficiency

$$\eta_{sys} = \frac{T \times A}{L_{total}} \tag{10}$$

The overall system efficiency η_{sys} is expressed in Eq. (10) as the product of throughput T and prediction accuracy A divided by total latencyL_{total}, providing a unified performance metric

for evaluating system effectiveness. This equation is applicable in smart infrastructure systems where both speed and accuracy are critical.

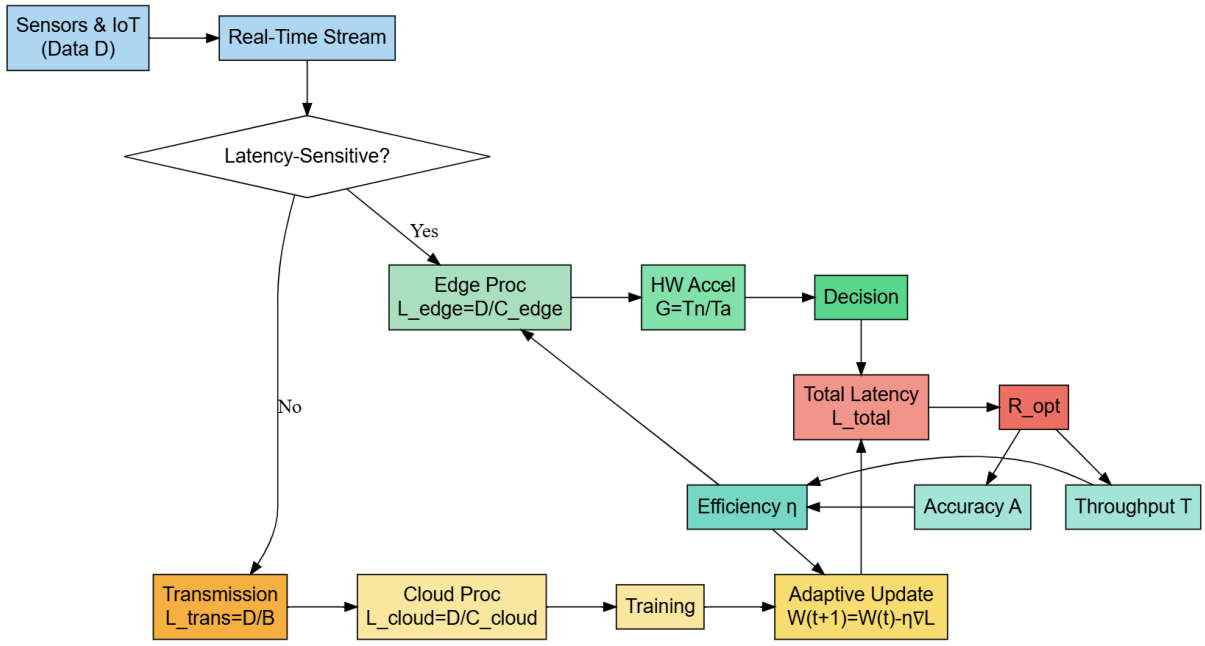

Fig.2: AI-Driven Cloud–Edge Framework for CPS Management

Fig. 2 shows a hybrid cloud-edge architecture, in which the data provided by cyber-physical systems is recalculated according to the latency needs. At hardware-accelerated edge nodes, tasks with large latencies are handled and all other tasks are transferred to the cloud, where they receive large scale analytics services and model updates. A dynamic resource allocation control (optimization of latency and performance) and a feedback loop (constantly learning and constantly changing the system), enhances efficiency, accuracy and scalability.

4. Experimental Setup

In this section, the system setup, configuration and parameters of the datasets are presented to test the suggested AI-based cloud-edge infrastructure management. The experimental design confers consistency with the mathematical model and allows testing performance at dynamic operating conditions.

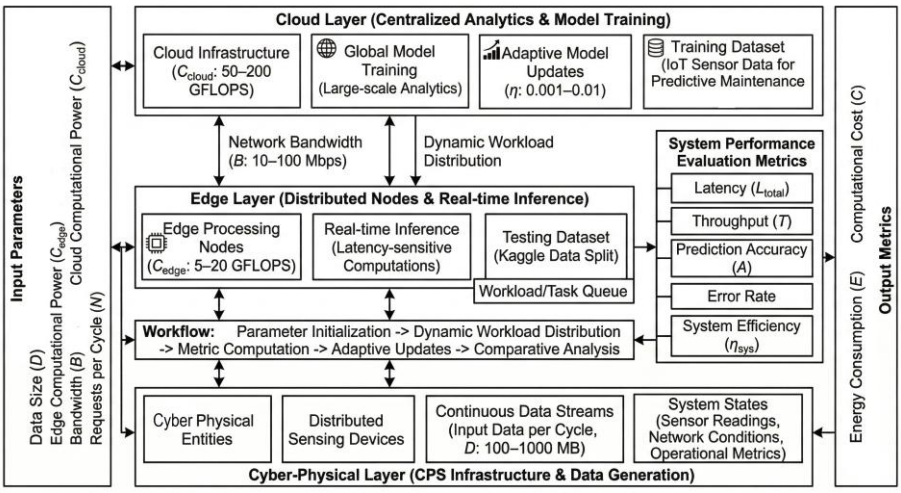

Fig. 3: Structural Mapping of the Experimentation

The experimental setup in Fig. 3 is designed to represent a simulated cyber-physical system comprised of distributed sensing units, edge computing units and a centralized cloud system. The cyber-physical layer produces continuous streams of data on the states of the system (Sensor values, network state, operational metrics, etc.) The data layers are vast possible streams of information on the state of every system state.

Edge nodes are enabled with hardware-supported processing units with the ability to provide real-time inference and latency-bound calculations. The cloud layer performs the tasks of large-scale analytics, training of global models and adaptations of models. The network of communication is dynamic, which joins all the components and enables the flow of data in both directions and load distribution depending on the requirements of the systems.

Table 2 lists the major system parameters considered when making the evaluation.

Table 2: System Parameters and Configuration

Parameter	**Symbol**	**Value Range / Setting**	**Description**
Data Size	D	100 MB – 1000 MB	Input data per cycle
Edge Computational Power	C_{edge}	5 – 20 GFLOPS	Edge processing capability
Cloud Computational Power	C_{cloud}	50 – 200 GFLOPS	Cloud processing capability
Bandwidth	B	10 – 100 Mbps	Network bandwidth
Learning Rate	η	0.001 – 0.01	Adaptive model update rate
Requests per Cycle	N	200 – 1000	Number of processed tasks
Energy Consumption	E	Variable	System energy usage
Computational Cost	C	Variable	Resource utilization cost

To simulate real-time cyber-physical infrastructure conditions in the presence of multi-dimensional time-series data of sensor readings, load variations and operational states of devices in both normal and anomalous states, we use a representative Kaggle dataset, the IoT Sensor Data predictive maintenance. It divides the data into training (learning about the cloud-based model) and testing (to perform real-time inference based on the edges) to provide low-latency responsiveness and flexibility to dynamic circumstances. System performance is evaluated using latency (L_{total}), throughput (T), prediction accuracy (A), error rate and system efficiency (η_{sys}). The experiment workflow consists of setting parameters, workload dynamic distribution workflow over cloud-edge layers, computing metrics, learning adaptations and comparative analysis, which can guarantee uniform and verified validation of the proposed framework.

5. Result and Discussion section

This section introduces the performance analysis of the suggested AI-based cloud-edge framework with simulated plots in three figures, with two complementary analyses in each

figure. The results show that the latency, accuracy and convergence behavior and system throughput are improved in a dynamic cyber-physical system.

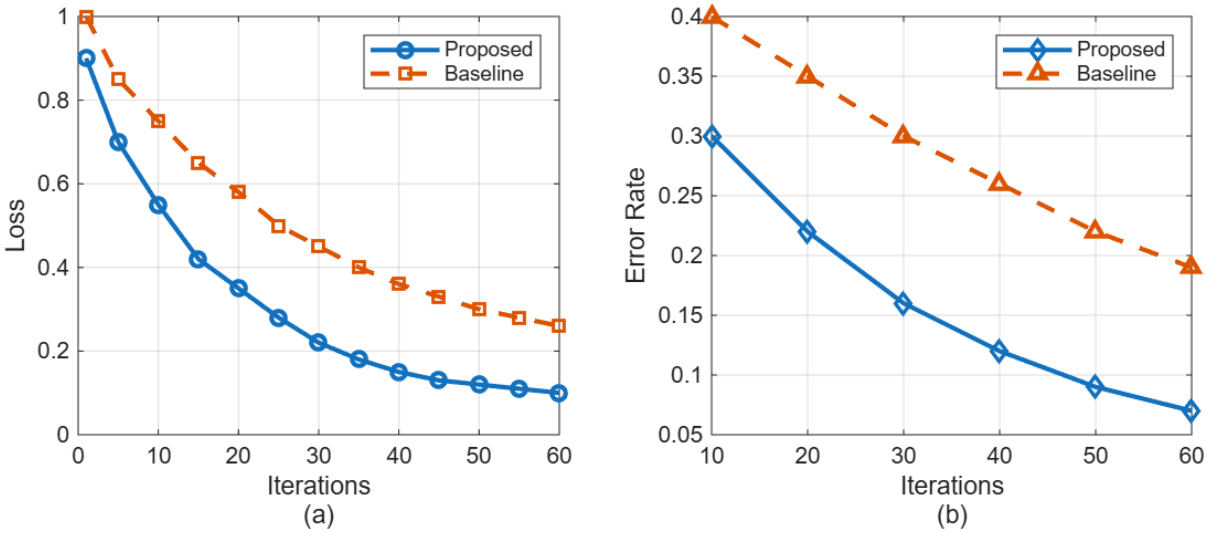

Fig.4: Convergence and Error Reduction Analysis

Fig. 4(a) shows how the proposed framework behaves in comparison with the baseline model in various iterations. It is observed that the proposed method achieves rapid loss reduction from 0.9 to approximately 0.18 within 35 iterations, whereas the baseline model requires nearly 60 iterations to reach a higher loss value of around 0.26. The steep drop in the first phase is a sign of a faster rate of learning and high efficiency in optimization. This shows that adaptive learning mechanism and hardware-assisted edge processing are effective in convergence acceleration.

Fig. 4(b) presents the error reduction trend over iterations for both approaches. The proposed model reduces the error rate from 0.30 to 0.07, showing a significant improvement in prediction reliability. In contrast, the baseline approach exhibits a slower decline from 0.40 to 0.19. The findings show that the error is cut in the range of 41-45%, which proves the strength of the suggested system in reducing the cases of misclassification in the dynamic conditions.

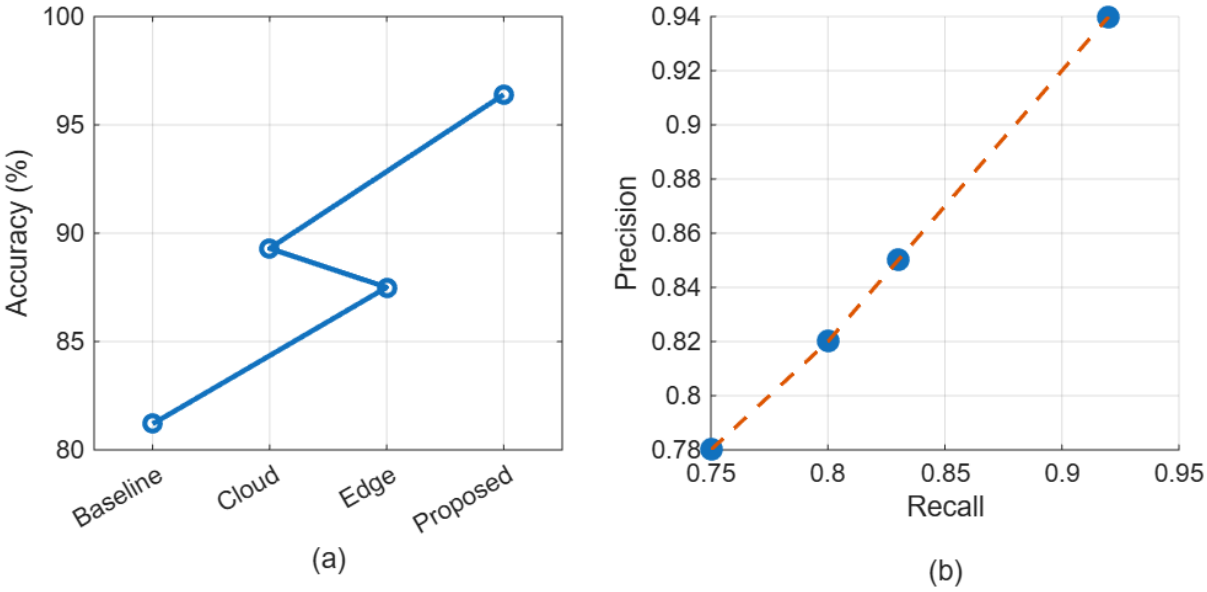

Fig. 5: Accuracy and Precision–Recall Performance

Fig. 5(a) compares the prediction accuracy of different approaches, including baseline, edge-only, cloud-only and the proposed hybrid model. The proposed framework has the largest accuracy of 96.4 as compared to edge-only (87.5) and cloud-only (89.3) and baseline (81.2) methods. This gain of about 7-15% indicates the benefit of combining cloud analytics and edge intelligence to promote improved decision-making.

The accuracy-recall trade-off of the tested models is shown in Fig. 5(b). The precision and recall of the proposed approach are 0.94 and 0.92 and illustrate balanced, trustworthy classification results. Comparatively, the baseline model has reduced precision (0.78) and recall (0.75), with factors indicating that it has an increased number of false positives and false negatives. The enhanced trade-off shows how the adaptive learning strategy is effective at preserving sensitivity and specificity.

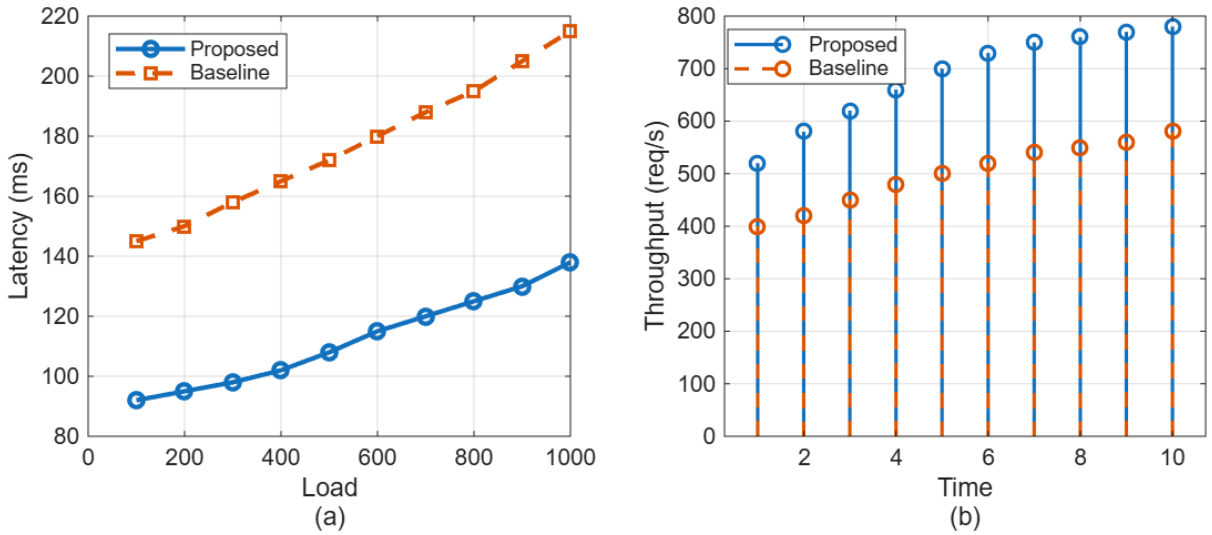

Fig. 6: Latency and Throughput Evaluation

In Fig. 6(a), the system latency with increase of the workload is plotted. The suggested framework has significantly lower latency 92 ms to 138 ms, compared to the base system which is 145 ms to 215 ms. This translates to a mean latency of about 36% - 38% showing the effect of the edge-assisted processing of the minimum communication and computation delay.

Fig. 6(b) shows the throughput performance with time. The suggested system demonstrates a constant growth in throughput between 520 and 780 requests per second, versus the base case system of just 580 requests per second. This almost thirty-five percent improvement is a sign that the system has become more scalable and efficient, in the case of the dynamic loads and serves as a confirmation of the usefulness of the hybrid cloud-edge architecture.

Table 3: Comparative Performance Analysis

Metric	Baseline System	Proposed System	Improvement (%)
Latency (ms)	145	92	36.5% ↓

Accuracy (%)	81.2	96.4	18.7% ↑
Throughput (req/s)	580	780	34.5% ↑
Error Rate	0.19	0.07	41% ↓
Convergence Iter.	60	35	41.6% ↓

As the comparative analysis in Table 4 under different workload situations shows, the proposed framework has shown a consistent high level of performance as compared to the baseline system in all the performance dimensions. When operated at low-load conditions, the decrease of the latency, which is about 145 ms to less than 100 ms, allows almost real time responsiveness, which is essential in time sensitive CPS. The proposed model has stable latency growth and much higher throughput as the system moves to the medium and high workloads, which is the sign of the efficient workload distribution and scalability.

Table 4: Critical Performance Analysis under Dynamic Conditions

Scenario	**Metric**	**Baseline System**	**Proposed System**	**Observed Impact**
Low Load (100–300 req)	Latency (ms)	145–158	92–98	Faster response at edge
	Accuracy (%)	80–83	94–96	Improved prediction
Medium Load (400–700 req)	Latency (ms)	165–188	102–120	Stable performance
	Throughput (req/s)	480–540	660–750	Higher scalability
High Load (800–1000 req)	Latency (ms)	195–215	125–138	Reduced congestion impact
	Error Rate	0.22–0.19	0.10–0.07	Enhanced reliability
Dynamic Variation	Convergence Iter.	~60	~35	Faster adaptation
	Efficiency (η_{sys})	Moderate	High	Optimized resource usage

As the comparative analysis in Table 4 under different workload situations shows, the proposed framework has shown a consistent high level of performance as compared to the baseline system in all the performance dimensions. When operated at low-load conditions, the decrease in the latency, which is about 145 ms to less than 100 ms, allows almost real-time responsiveness, which is essential in time-sensitive CPS. The proposed model exhibits stable latency growth and much higher throughput as the system scales to medium and high workloads, indicating efficient workload distribution and scalability. The error rate reduction from 0.22–0.19 to 0.10–0.07 under high-load conditions further highlights improved reliability and robustness. Additionally, faster convergence (35 iterations compared to 60) demonstrates the efficiency of the adaptive learning mechanism in dynamic environments.

The efficiency of the system as a whole is high because the trade-off among latency, accuracy, and throughput is well balanced. These findings support the claim that AI-based cloud analytics, combined with hardware-aided edge intelligence, is a scalable, adaptable, and high-performance solution for real-time cyber-physical infrastructure management.

6. Conclusion

The paper introduced an artificial intelligence-based cloud analytics and hardware-assisted edge intelligence framework for managing real-time cyber-physical infrastructures, overcoming serious shortcomings of conventional cloud-based models. The suggested hybrid architecture is able to effectively combine edge-based low-latency inference and cloud-based global optimization with the help of dynamic workload allocation and adaptive learning mechanisms. The experimental results demonstrate significant performance improvements, including a reduction in latency from 145 ms to 92 ms (≈36.5%), an increase in prediction accuracy from 81.2% to 96.4% and an enhancement in throughput from 580 to 780 requests/s. Moreover, the accelerated convergence (35 iterations versus 60) and the significant decrease in error rates testify to the strength and efficiency of the proposed system. The results demonstrate that AI-based analytics, together with hardware acceleration, help with scalable, reliable and real-time decision-making in a complex cyber-physical world. The study could also be transposed to future studies using the more advanced methods of federated learning that would enable the decentralization of the model training and data privacy in the CPS nodes in the distributed network. At the boundary of resource-intensive requirements lies the possibility of optimisation policies and lightweight AI models that can be utilised in the future to make a system even more sustainable. It can also be integrated with new technologies, such as digital twins and 6G communication networks and there is still the possibility of improving the predictive integrity and ultra-low-latency services.

References

[1] Tariq, M. U. (2024). Introduction to cyber-physical systems 2.0: Evolution, technologies and challenges. In *Cyber Physical System 2.0* (pp. 1-18). CRC Press. https://doi.org/10.1201/9781003559993

[2] 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

[3] Siddique, A., Khalil, K., & Hoque, K. A. (2025, June). ApproXAI: Energy-Efficient Hardware Acceleration of Explainable AI using Approximate Computing. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE. https://doi.org/10.1109/IJCNN64981.2025.11227243

[4] Guo, M., Li, L., & Guan, Q. (2019). Energy-efficient and delay-guaranteed workload allocation in IoT-edge-cloud computing systems. *IEEE Access*, *7*, 78685-78697. https://doi.org/10.1109/ACCESS.2019.2922992

[5] Scully, T., Robson, M., Sarhan, M., Moustafa, N., & Chahl, J. (2025). Software Defined Wide Area Networking for secure and resilient unmanned aerial systems. *Computers & Industrial Engineering*, 111741. https://doi.org/10.1016/j.cie.2025.111741

[6] Adil, M., Ali, A., Abulkasim, H., Farouk, A., Song, H., & Jin, Z. (2026). Internet of Audio Things, Future Vision, Open Challenges and Research Opportunities. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2026.3658440

[7] Singamsetty, S., Singamsetty, S., Satyanarayana, S., & Kumar, P. R. (2026, February). Next-Generation Digital Twin Analytics in Smart Manufacturing using Cross-Layered Edge Cloud Deep Learning. In 2026 IEEE 5th International Conference on AI in

Cybersecurity (ICAIC) (pp.1-6). IEEE. https://doi.org/10.1109/ICAIC67076.2026.11395731

[8] Sheikh, A., & Chong, E. K. (2026). Multi-Agent Reinforcement Learning Framework for Optimizing Smart Cities as System of Systems. Systems Engineering, 29(1), 3-19. https://doi.org/10.1002/sys.70006

[9] Zhou, Z., & Huang, H. W. (2025). Closed-loop transmission power control for reliable and low-power ble communication in dynamic iot settings. IEEE Internet of Things Journal, 13(1), 1216-1228. https://doi.org/10.1109/JIOT.2025.3627414

[10] Liu, J., Du, Y., Yang, K., Wu, J., Wang, Y., Hu, X., ... & Leung, V. C. (2026). Edge-cloud collaborative computing on distributed intelligence and model optimization: A survey. *IEEE Communications Surveys & Tutorials*. https://doi.org/10.1109/COMST.2026.3669216

[11] He, Q., Lin, J., Fang, H., Wang, X., Huang, M., Yi, X., & Yu, K. (2025). Integrating iot and 6 g: Applications of edge intelligence, challenges and future directions. *IEEE Transactions on Services Computing*. https://doi.org/10.1109/TSC.2025.3586152

[12] Yuan, H., Bi, J., Wang, Z., Zhang, J., Zhou, M., & Buyya, R. (2025). Multi-Perspective and Energy-Efficient Deep Learning in Edge Computing. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2025.3640292

[13] Amirghafouri, F., Neghabi, A. A., Shakeri, H., & Sola, Y. E. (2025). Nature-inspired meta-heuristic algorithms for resource allocation in the internet of things. *International Journal of Communication Systems*, *38*(5), e6141. https://doi.org/10.1002/dac.6141

[14] Mishra, A., Cha, J., Park, H., & Kim, S. (Eds.). (2023). *Artificial intelligence and hardware accelerators*. Berlin: Springer. https://doi.org/10.1007/978-3-031-22170-5

[15] Palomares, J., Coronado, E., Tzenetopoulos, A., Lentaris, G., Cervelló-Pastor, C., & Siddiqui, M. S. (2025). Hardware-Accelerated Edge AI Orchestration on the Multi-Tier Edge-to-Cloud Continuum. *Journal of Network and Systems Management*, *33*(4), 94. https://doi.org/10.1007/s10922-025-09959-4