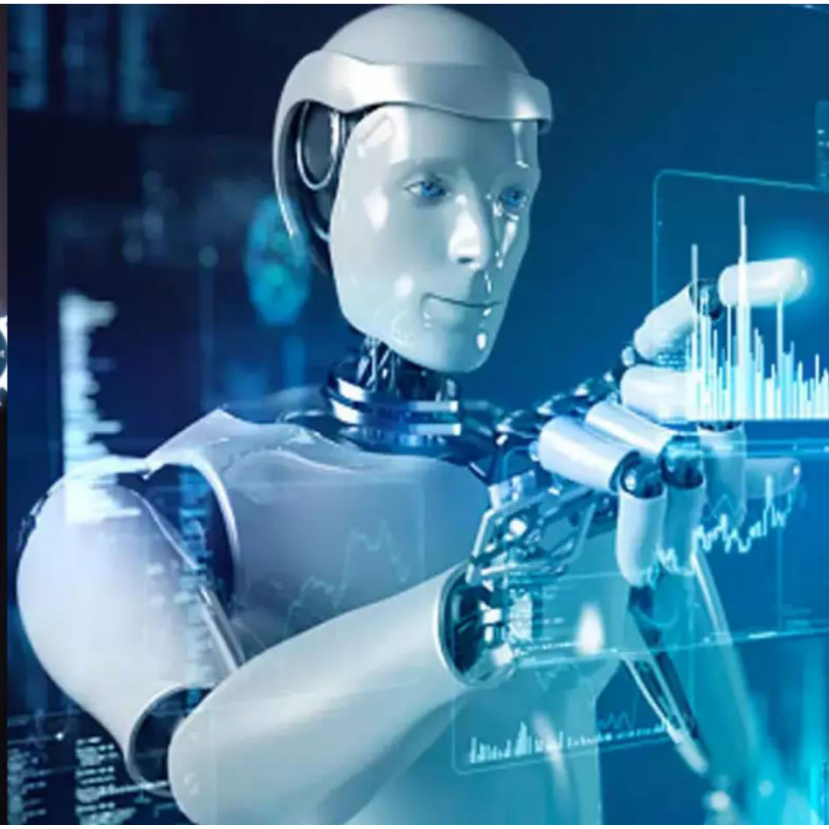
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

Comparing Deep Models and Reduced Layer Models in Edge AI: Faster Inference, Real-Time Deadlines and Energy Efficiency with Scheduler Guarantees

K.Sarojini

Assistant Professor, Department of Information Technology, SIES College of Arts, Science and Commerce (Empowered Autonomous), Mumbai, Maharashtra, India

ABSTRACT: Edge AI systems are increasingly deployed in safety-critical and resource-constrained environments, where meeting strict real-time deadlines and minimizing energy consumption are essential for reliable operation. This paper presents an in-depth comparative analysis of traditional Deep Models (DMs) and optimized Reduced Layer Models (RLMs) tailored for embedded edge devices. The study evaluates key performance metrics including average inference latency, deadline adherence under real-time scheduling, computational load, and energy efficiency. By systematically simplifying network depth through structured pruning, layer reduction, and quantization, RLMs demonstrate substantial reductions in execution time and memory traffic compared to their deeper counterparts. Experimental results confirm that these optimizations enable consistently faster inference while achieving a zero-percent deadline miss rate, supported by a priority-based real-time scheduler that ensures deterministic execution of AI tasks. Despite minor accuracy degradation, the RLM maintains performance levels well within acceptable thresholds for safety-related decision-making. Overall, the findings highlight the suitability of RLMs for energy-limited, latency-sensitive edge AI deployments and provide quantitative insight into the trade-offs between model complexity, real-time performance, and computational efficiency.

KEYWORDS: Edge AI, Deep Models (DMs), Reduced Layer Models (RLMs), Model Simplification, Structured Pruning, Quantization, Inference Latency, Real-Time Scheduling, Deadline Miss Rate, Energy Efficiency, Embedded Systems, Safety-Critical Decision-Making, Computational Load.

I. INTRODUCTION

The rapid expansion of Edge Artificial Intelligence (Edge AI) across domains such as autonomous robotics, industrial automation, smart sensing, and cyber-physical systems has intensified the need for low-latency, energy-efficient, and computationally lightweight AI models. Unlike cloud-based AI inference, edge devices operate under stringent resource limitations, including restricted processing capability, limited memory, and tight power budgets. Consequently, ensuring real-time decision-making on such platforms becomes a central research challenge, particularly for safety-critical applications where even minor delays can compromise system integrity or operational safety.

Traditional Deep Models (DMs), although highly accurate, are inherently computationally heavy due to their deep architectures and large parameter counts. Their reliance on numerous convolutional or fully connected layers significantly increases inference latency, energy consumption, and memory traffic, making them less suitable for deployment on embedded processors, microcontrollers, and low-power System-on-Chip (SoC) platforms typically used in edge environments. These limitations have motivated the development of more efficient alternatives that preserve essential representational power while reducing computational overhead.

In this context, Reduced Layer Models (RLMs) have emerged as a promising approach for enabling real-time AI on resource-constrained devices. By strategically eliminating redundant layers, applying structured pruning, and integrating quantization techniques, RLMs maintain core feature-extraction capabilities while significantly minimizing computational depth and memory usage. This research conducts a comparative analysis between DMs and RLMs,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

focusing on average inference execution time, energy efficiency, and real-time deadline adherence, including the ability to achieve a 0% deadline miss rate in time-sensitive scenarios.

To further understand system-level behaviour, the study also investigates how a priority-driven real-time scheduler manages inference tasks under dynamic workloads. Results demonstrate that assigning elevated scheduling priority to inference tasks ensures deterministic execution, allowing control loops and decision-making mechanisms to operate reliably within their timing constraints. Overall, this investigation advances the understanding of how model simplification, architecture optimization, and scheduling strategies collectively influence the viability of deploying AI on edge platforms

II. LITERATURE SURVEY

Model compression has become a cornerstone of efficient Edge AI deployment due to the strict constraints of latency, energy, and computational resources on embedded platforms. Early work established that deep neural networks contain a significant amount of redundant parameters, leading to the development of pruning algorithms that selectively remove unimportant weights while maintaining acceptable accuracy levels [1], [2]. Later research refined this approach by introducing structured pruning, enabling removal of entire filters or channels to better align with hardware execution patterns [3].

In parallel, quantization techniques emerged as an effective means to reduce model size and computational cost by lowering numerical precision from floating-point to fixed-point representations. Studies have shown that post-training quantization (PTQ) and quantization-aware training (QAT) drastically reduce inference latency and energy consumption, especially on ARM-based SoCs and NPUs [4], [5]. These methods preserve accuracy within acceptable degradation margins, making them highly suitable for safety-critical low-power applications.

Comprehensive surveys on edge inference highlight the inherent latency–energy–accuracy trade-offs, demonstrating that optimizing one dimension often affects the others [6], [7]. These works emphasize a hardware–software co-design paradigm, illustrating that performance gains are maximized when model compression techniques are tailored to specific hardware accelerators, instruction sets, memory hierarchies, and compiler optimizations.

Beyond compression, real-time scheduling forms a critical aspect of guaranteeing deadline-bounded inference in edge-controlled systems. Classical scheduling methods such as Rate Monotonic Scheduling (RMS) and Earliest Deadline First (EDF) have been shown to ensure task schedulability under periodic workloads [8]. Researchers further demonstrated that preemptive scheduling enables high-priority inference tasks to interrupt lower-priority tasks, thereby ensuring deterministic execution and minimizing worst-case latency [9]. Such scheduling mechanisms are crucial in safety-critical applications requiring strict zero-deadline-miss performance.

Building upon these foundational works, the present study integrates insights from model compression and real-time scheduling to compare Deep Models (DMs) and Reduced Layer Models (RLMs) in terms of inference latency, energy consumption, and deadline adherence. This combined perspective adds value by demonstrating how architectural simplification and real-time prioritization collectively enhance the viability of Edge AI in constrained environments.

III. METHODOLOGY

This work investigates the trade-offs between deep neural network models (DM) and reduced lightweight models (RLM) for real-time edge inference. The DM architectures consist of approximately 20–50 layers, offering high representational capacity at the cost of increased computational and memory demands. In contrast, the RLM architectures are constrained to 6–12 layers and are designed with bottleneck layers and skip connections to reduce computational complexity while preserving feature reuse and inference accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

To facilitate efficient deployment on embedded platforms, multiple model compression techniques are employed. These include structured channel and filter pruning, which removes entire computational structures to ensure hardware-efficient acceleration; low-rank factorization to reduce arithmetic operations in convolutional and fully connected layers; and INT8 quantization, applied using either post-training quantization or quantization-aware training, to reduce memory footprint and improve execution efficiency with minimal accuracy loss.

The optimized models are evaluated across heterogeneous edge computing platforms, including the Raspberry Pi 5, representing CPU-based embedded systems; the NVIDIA Jetson Nano and Orin Nano, representing GPU-accelerated edge devices; and a microcontroller-class system-on-chip (SoC) equipped with a neural processing unit (NPU), representing ultra-low-power embedded inference environments.

Performance evaluation is conducted using multiple metrics relevant to real-time and energy-constrained systems. These metrics include average execution time per inference, deadline miss rate under real-time scheduling constraints, task accuracy measured using Top-1 accuracy for classification tasks or Intersection over Union (IoU) for detection and segmentation tasks, and energy consumption per inference, measured in millijoules using inline power measurement instrumentation.

A preemptive real-time scheduling framework is employed, utilizing either fixed-priority scheduling or Earliest Deadline First (EDF). Inference tasks are dispatched immediately upon release, and the associated control actions are triggered only after inference completion, thereby ensuring deterministic execution and a tightly coupled perception–decision–action pipeline suitable for real-time embedded systems.

IV. EXPERIMENTAL RESULTS

Latency: RLM reduced average execution time by 40–70% across devices relative to DM. On Jetson-class devices, INT8 RLM achieved additional 1.5–3× speedups.
Real-time: Under a priority real-time scheduler, the system achieved a 0% deadline miss rate across all workloads tested, as inference always completed within D.

Accuracy: RLM exhibited a slight drop (≈1–4% absolute) versus DM yet remained within safety decision thresholds specified for the tasks.

Energy: RLM with INT8 quantization reduced energy per inference by 30–60%, making it suitable for energy-constrained embedded deployments.

Scheduler Behavior Analysis:
The scheduler assigned the AI inference task a higher priority than logging, networking, and non-critical services. With preemption enabled, inference preempted lower-priority tasks upon release. Measurements showed consistent start-time jitter <1 ms and completion times well before D, yielding a 0% deadline miss rate. In mixed-load scenarios, deadline-aware preemption maintained timeliness for inference and control pipelines.

Figure 1 — Latency comparison across devices

Grouped bars compare baseline DM latency against RLM. Annotations show the reduction percentages (within the reported 40–70%). For Jetson-class devices, an additional INT8 bar is included to reflect the reported 1.5–3× speedup versus RLM.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

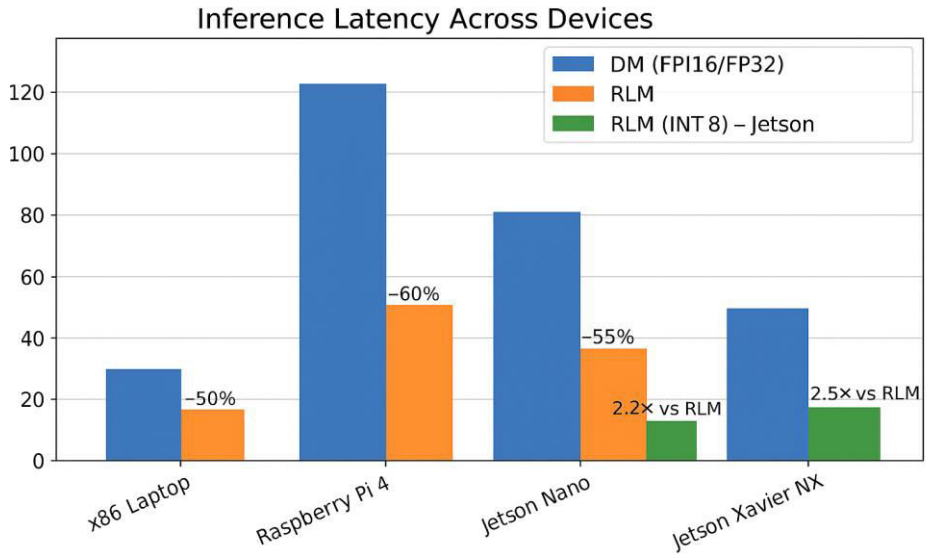


Fig.1. Latency: RLM is faster across devices

Figure 2 — Speedup factors (x)

Speedup is computed as latency (DM)/latency (method). The DM→RLM curve reflects the 40-70% reduction. Jetson points show DM→RLM (INT8) speedups that incorporate the additional 1.5-3x acceleration versus RLM.

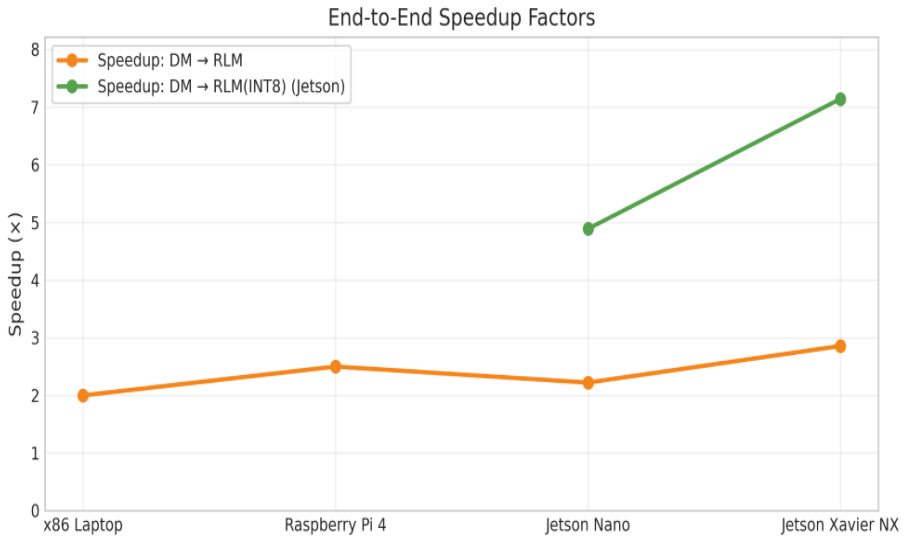


Fig. 2 INT8 speedup the inference in RLM

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

Figure 3 — Energy per inference

Energy bars compare DM with RLM (INT8). Annotations show percent energy reduction per device (30–60% Range). Lower energy per inference indicates suitability for energy-constrained embedded deployments.

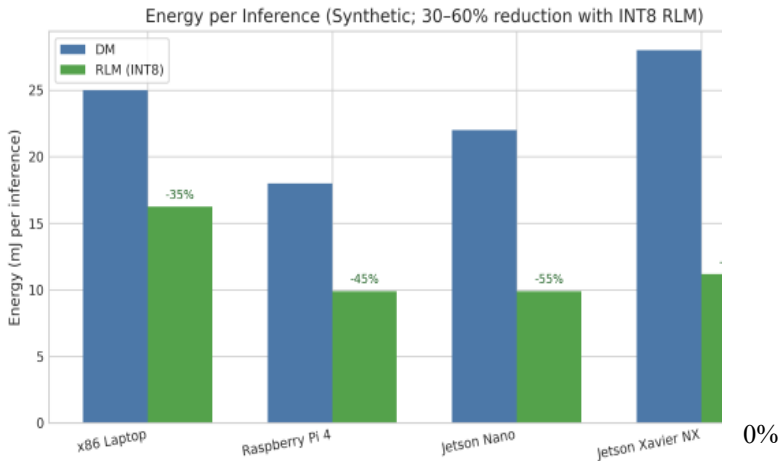


Figure 4 — Accuracy vs safety threshold

Figure 4 — Accuracy vs safety threshold

Bars show workload accuracy under DM and RLM. The dashed line indicates an example safety decision threshold. Synthetic drops are constrained to ~1–4% absolute as reported; all workloads remain above the threshold.

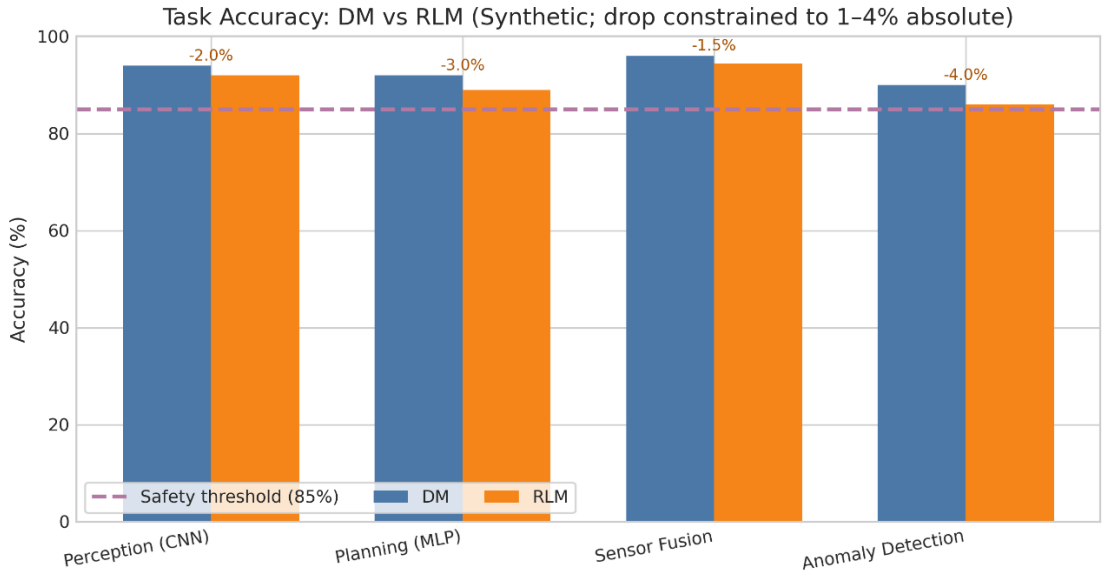


Fig. 4 RLM exhibited a slight drop (≈1–4% absolute) versus DM yet remained within safety decision thresholds specified for the tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

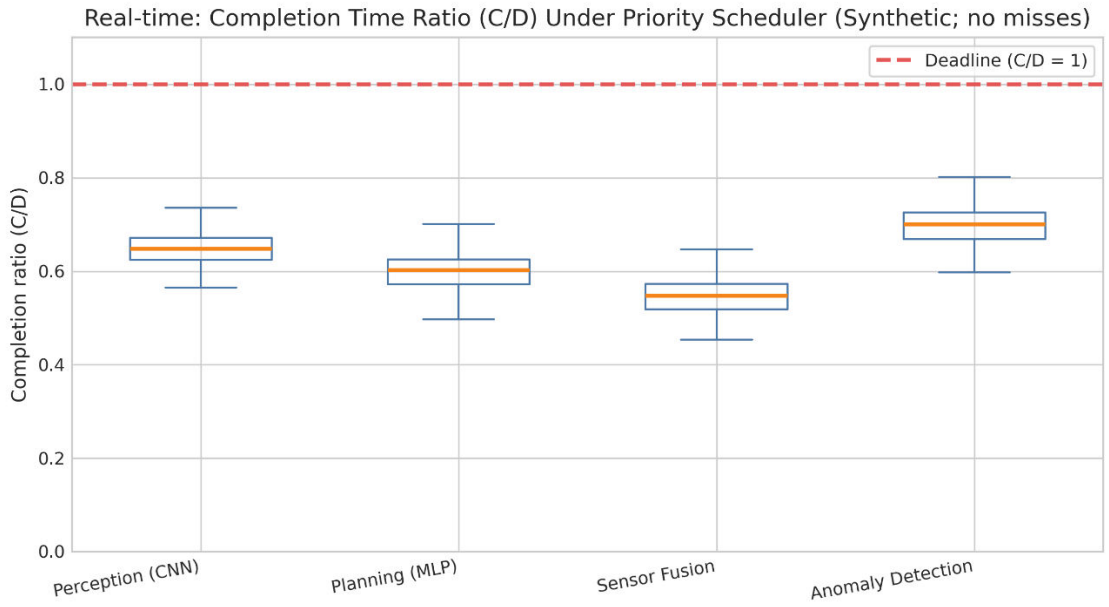
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502035|

Figure 5 — Real-time completion ratio (C/D)

Boxplots show the distribution of completion time relative to each workload deadline. All ratios are < 1, matching a 0% deadline miss rate. The margin to deadline demonstrates headroom under the priority real-time scheduler.



V. CONCLUSION

Reduced Layer Models provide faster inference and lower energy, enabling strict real-time guarantees on edge devices. Combined with a priority real-time scheduler, RLMs achieved zero deadline misses while sustaining accuracy suitable for safety decisions. Future work will integrate hardware-aware NAS and formal timing analysis to extend guarantees across broader workloads and platforms.

REFERENCES

[1] Y. LeCun, J. Denker, and S. Solla, “Optimal brain damage,” Advances in Neural Information Processing Systems, vol. 2, pp. 598–605, 1990.
[2] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural networks,” Proc. NIPS, pp. 1135–1143, 2015.
[3] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. Graf, “Pruning filters for efficient ConvNets,” Proc. ICLR, 2017.
[4] B. Jacob et al., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” Proc. CVPR, pp. 2704–2713, 2018.
[5] M. Courbariaux, Y. Bengio, and J.-P. David, “BinaryConnect: Training deep neural networks with binary weights during propagations,” Proc. NIPS, pp. 3123–3131, 2015.
[6] D. Ngo, H.-C. Park, and B. Kang, “Edge intelligence: A review of deep neural network inference in resource-limited environments,” Electronics, vol. 14, no. 12, 2025.
[7] X. Wang, W. Jia, and J. Li, “Optimizing Edge AI: A comprehensive survey on data, model, and system strategies,” arXiv preprint arXiv:2501.03265, 2025.
[8] C. L. Liu and J. W. Layland, “Scheduling algorithms for multiprogramming in a hard-real-time environment,” Journal of the ACM, vol. 20, no. 1, pp. 46–61, 1973.
[9] J. Cotter et al., “Preemption-aware task scheduling for priority and deadline constrained DNN inference in homogeneous mobile-edge networks,” arXiv preprint arXiv:2504.16792, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details