

Original Article**Edge AI for Energy-Efficient Smart IoT Systems: A Sustainable Computing Approach****Prathamesh Uttam Patil**

Department of Computer Engineering, Ashokrao Mane Group of Institute, Kolhapur, Maharashtra

Manuscript ID:
CSJ-2026-020510

Volume 2

Issue 5

Pp. 50-53

October 2026

ISSN: 3067-3089

Submitted: 08 Sept. 2026**Revised:** 16 Sept. 2026**Accepted:** 05 Oct. 2026**Published:** 31 Oct. 2026**Correspondence Address:**

Prathamesh Uttam Patil
Department of Computer Engineering,
Ashokrao Mane Group of Institute,
Kolhapur, Maharashtra
Email: Pratham.patil3130@gmail.com

Quick Response Code:

Web: <https://csjour.com/>

DOI: 10.5281/zenodo.22953475

DOI Link:

<https://doi.org/10.5281/zenodo.22953475>

Creative Commons

**Abstract**

Artificial Intelligence of Things (AIoT) is transforming Internet of Things (IoT) systems by enabling data-driven decisions close to the point where data are generated. However, conventional cloud-centric architectures can introduce communication overhead, latency, privacy concerns, and additional energy costs. This paper examines Edge AI as a sustainable computing approach for energy-efficient smart IoT systems, with particular relevance to the development priorities associated with Viksit Bharat 2047. The study follows a descriptive and analytical approach based on research literature concerning TinyML, on-device machine learning, Edge AI, and energy-aware IoT. It proposes an architecture consisting of sensing, local inference, edge coordination, selective cloud processing, and energy-aware orchestration. The paper further identifies measurable parameters including accuracy, inference latency, energy per inference, communication volume, memory footprint, and reliability for evaluating sustainable AIoT systems. The objective is not to claim unverified experimental results, but to establish a practical research framework that can subsequently be tested on resource-constrained edge hardware. The study indicates that sustainable AIoT should be evaluated not only through predictive performance but also through the computational, communication, and energy resources required to deliver each intelligent decision.

Introduction

The Internet of Things has evolved from simple sensing and connectivity toward intelligent systems capable of interpreting sensor data and supporting automated decisions. Artificial Intelligence of Things (AIoT) combines IoT sensing and communication with machine learning and provides opportunities in predictive maintenance, smart energy management, environmental monitoring, intelligent transportation, agriculture, healthcare, and public services. A major architectural question in AIoT is where computation should occur. In a conventional cloud-centric architecture, sensor data are transmitted to remote infrastructure for processing and inference. Edge AI moves selected computation closer to the point where data are generated, using gateways, embedded processors, microcontrollers, or local edge servers. This approach can reduce latency, communication requirements, and dependence on continuous connectivity, while also offering opportunities for improved privacy. The sustainability of AIoT systems, however, cannot be judged simply by moving computation from the cloud to an edge device. Edge devices have limited memory, processing capability, energy availability, and thermal capacity. Consequently, sustainable Edge AI requires coordinated consideration of model complexity, computation, communication, energy consumption, and application accuracy. The present study therefore examines Edge AI as a sustainable computing approach and develops a framework for energy-efficient smart IoT systems. Particular attention is given to resource-constrained environments and applications relevant to inclusive and sustainable technological development in India.

Rationale of the Study

The conference theme focuses on “Innovative Frontiers in Science, Technology, and Multidisciplinary Research: Integrating Artificial Intelligence and Smart Technologies for Sustainable Development and Viksit Bharat 2047.” Its listed Computer Science and Information Technology themes include Artificial Intelligence and Machine Learning, Data Science, Cyber Security, Cloud Computing, IoT, Robotics, Quantum Computing, Software Engineering, Human-Computer Interaction, and Smart Computing and Digital Transformation. The conference also explicitly includes Smart Cities, Sustainable Development Goals, and Viksit Bharat 2047. Large-scale IoT systems may continuously generate sensor data. Transmitting all such data to centralized infrastructure can increase network traffic and response time. Local inference provides an alternative in which the device or nearby edge gateway processes data and transmits only selected results, alerts, summaries, or updates. This is especially relevant for distributed applications where network connectivity, energy availability, hardware cost, and maintenance can constrain system design.

Creative Commons (CC BY-NC-SA 4.0)

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 International Public License, which allows others to remix, tweak, and build upon the work no commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

How to cite this article:

Patil, P. U. (2026). Edge AI for Energy-Efficient Smart IoT Systems: A Sustainable Computing Approach. *Computer Science Journal*, 2(5), 50–53. <https://doi.org/10.5281/zenodo.22953475>

The rationale of this study is therefore to examine whether an Edge-AI architecture can provide useful intelligence while reducing unnecessary computation, communication, and energy requirements. The study treats sustainability as a systems-engineering problem rather than only an environmental objective.

Objectives

1. To examine the role of Edge AI and TinyML in improving the energy efficiency of IoT systems.
2. To identify the principal computational, memory, communication, and energy constraints affecting AI deployment on edge devices.
3. To develop a practical architecture for energy-aware Edge-AI-enabled IoT systems.
4. To define measurable parameters for comparing cloud-centric, edge-centric, and hybrid AIoT architectures.
5. To examine the potential contribution of Edge AI to smart infrastructure, agriculture, energy management, and rural digital services.
6. To relate sustainable Edge AI to the broader technological objectives associated with Viksit Bharat 2047.

Hypotheses

1. Local or edge inference can reduce communication dependence and response latency compared with a fully cloud-centric inference workflow.
2. Model optimization techniques such as quantization, pruning, knowledge distillation, and lightweight architectures can reduce resource requirements while retaining useful predictive performance.
3. A hybrid edge-cloud architecture can provide a practical balance among accuracy, energy efficiency, latency, reliability, and centralized model management.
4. Energy-aware orchestration can improve the sustainability of AIoT systems by adapting computation and communication according to device and network conditions.

Methodology

The paper follows a descriptive, analytical, and framework-development approach. Research literature concerning TinyML, on-device machine learning, Edge AI, and energy-aware IoT is examined to identify common constraints, advantages, optimization techniques, architectural patterns, and evaluation parameters. The analysis is then used to formulate a sustainable Edge-AI architecture for smart IoT applications.

The proposed framework is not presented as a completed hardware experiment. No experimental energy, latency, or accuracy values are invented. Instead, the study defines a reproducible experimental procedure that can be implemented on an embedded board or edge computer. Three deployment configurations can be compared: cloud-centric inference, edge inference, and hybrid edge-cloud inference.

The recommended measurements include predictive accuracy or F1-score, inference latency, energy consumed per inference, total data transmitted, model storage size, peak memory usage, and operation under intermittent connectivity. Repeated trials should be used and the results should be reported using appropriate descriptive statistics.

Edge AI and Sustainable Computing

1 Concept of Edge AI

Edge AI refers to performing artificial intelligence processing near the location where data are generated rather than sending every data item to a remote cloud. The edge may be a microcontroller, embedded computer, gateway, or local server. The basic objective is to make decisions closer to the source while retaining cloud resources for tasks that require larger computational capacity.

This architecture can be useful when decisions are time-sensitive or when transmitting raw data continuously is undesirable. For example, a sensor can identify an anomaly locally and transmit only the event, confidence score, and selected contextual information instead of a continuous raw stream.

2 TinyML and Resource-Constrained Devices

TinyML extends machine learning to highly constrained embedded platforms. Such devices typically have limited processor capability, memory, storage, and energy. Model

Page 4

optimization is therefore essential. Quantization, pruning, knowledge distillation, and lightweight neural-network architectures are among the approaches that can reduce computational and memory requirements. The goal is to achieve an acceptable balance between predictive performance and resource consumption.

3 Energy Efficiency

Energy consumption in AIoT has at least two important components: the energy required for computation and the energy required for communication. Cloud-centric systems may require repeated data transmission, while edge systems incur local inference costs. Therefore, the most sustainable architecture cannot be identified without measuring both components. Energy per decision or energy per inference can provide a practical metric for comparing alternative designs.

Proposed Sustainable Edge-AI Architecture

The proposed architecture consists of five logical layers designed to distribute intelligence according to resource availability and application requirements.

1 Sensing Layer: Sensors collect environmental, energy, traffic, agricultural, industrial, or application-specific signals.

2 Local Inference Layer: A lightweight ML/TinyML model performs immediate classification, regression, anomaly detection, or forecasting near the data source.

3 Edge Coordination Layer: An edge gateway aggregates selected results, manages devices, and performs more computationally intensive processing when required.

4 Selective Cloud Layer: The cloud stores historical information, supports model retraining, and receives selected summaries, alerts, model updates, or data samples rather than every raw observation.

5 Energy-Aware Orchestration Layer: The system adapts sensing frequency, inference mode, communication frequency, and model complexity according to energy availability, workload, and network conditions.

The architecture therefore does not treat edge and cloud computing as competing alternatives. Local processing handles time-sensitive decisions, while edge and cloud resources provide aggregation, long-term analysis, model management, and periodic retraining.

Model Optimization Techniques

Quantization reduces the numerical precision used to represent model parameters and activations. It can reduce model size and computational requirements and may enable deployment on hardware with limited memory.

Pruning removes parameters or connections that contribute relatively little to model output. Structured pruning can be particularly useful when the target hardware benefits from simpler computational graphs.

Knowledge distillation trains a smaller student model using information produced by a larger teacher model. This provides a method for transferring useful predictive behaviour into a model more suitable for edge deployment.

Lightweight model architectures and hardware-aware design can further reduce computation. Optimization should not be performed solely to minimize model size because excessive compression may cause unacceptable loss of accuracy. The appropriate objective is a balanced trade-off between performance and resource consumption.

Proposed Evaluation and Data Analysis

A future empirical implementation can use the same dataset, preprocessing pipeline, and task model across three deployment modes: cloud-centric, edge-centric, and hybrid. Hardware and software configurations should be documented so that the results can be reproduced.

The principal parameters are accuracy/F1-score, inference latency, energy per inference, communication volume, model size, peak memory, and availability under intermittent connectivity. Each configuration should be evaluated over repeated trials.

A multi-criteria or Pareto analysis can then be used to identify solutions that improve energy and communication efficiency without unacceptable loss of predictive performance. The analysis should report both absolute values and relative differences between architectures.

Applications for Sustainable Development and Viksit Bharat 2047

1 Smart Energy Management

Edge AI can support local load forecasting, anomaly detection, appliance monitoring, demand management, and intelligent control. Local decisions may reduce latency and unnecessary transmission in distributed energy systems.

2 Smart Agriculture

Edge-enabled sensing can support irrigation decisions, crop-condition monitoring, environmental sensing, and anomaly detection. Local processing can be useful in areas where network connectivity is intermittent.

3 Smart Cities

Edge AI can support traffic monitoring, waste classification, environmental monitoring, infrastructure anomaly detection, and other smart-city services while reducing the need to transmit continuous raw sensor streams.

4 Rural and Distributed Digital Services

Resource-efficient AI can provide local intelligence in applications where connectivity, energy, and hardware resources are limited. This can support more resilient and inclusive digital services.

5 Industrial Sustainability

Local anomaly detection and predictive maintenance can help identify equipment problems close to machines, potentially enabling faster response and reducing unnecessary data movement.

Contemporary Relevance

The transition toward intelligent and connected infrastructure creates a need for computing architectures that are both capable and resource-conscious. Edge AI is relevant because it places part of the decision-making process near the physical environment where data are produced.

For sustainable development, the objective should be broader than reducing electricity consumption alone. Efficient computing can reduce network traffic, support operation during connectivity interruptions, enable privacy-preserving local processing, and potentially reduce the infrastructure required for continuous centralized processing.

For Viksit Bharat 2047, Edge AI can be considered as one enabling technology within a larger digital transformation strategy. Its value will depend on affordable hardware, reliable connectivity, appropriate datasets, cybersecurity, skilled personnel, and responsible deployment.

Page 7

Challenges and Limitations

1. Edge devices have limited memory, processing capability, storage, and energy resources.
2. Model compression may reduce predictive accuracy or robustness.
3. Different hardware platforms make performance difficult to generalize.
4. Energy savings depend on both computation and communication patterns.
5. Secure model updates and protection against adversarial or compromised devices remain important.
6. Data drift and sensor failures can reduce model reliability after deployment.
7. The present paper is a literature-based framework study and does not claim original experimental measurements.

Conclusion

Edge AI provides a promising approach for developing intelligent IoT systems that are less dependent on continuous centralized processing. Local inference can reduce response latency and communication requirements, while TinyML and model-optimization methods make machine learning increasingly practical on resource-constrained devices. However, sustainable AIoT requires simultaneous consideration of computation, communication, energy, memory, accuracy, and reliability.

This paper proposed a five-layer sustainable Edge-AI architecture consisting of sensing, local inference, edge coordination, selective cloud processing, and energy-aware orchestration. It also proposed an evaluation framework based on accuracy, latency, energy per inference, communication volume, memory footprint, and resilience. The framework is particularly relevant to smart infrastructure, agriculture, energy management, industrial systems, and distributed digital services.

The broader significance of Edge AI lies in its potential to combine intelligent automation with resource efficiency.

When evaluated and deployed responsibly, such systems can contribute to sustainable digital transformation and provide technological building blocks relevant to the vision of Viksit Bharat 2047.

Page 8

Recommendations

1. Conduct empirical testing of Edge-AI models on representative low-power hardware.
2. Report accuracy, latency, energy, memory, and communication metrics together rather than using accuracy alone.
3. Use quantization, pruning, distillation, and lightweight architectures according to the target device.
4. Adopt hybrid edge-cloud designs when local and centralized processing provide complementary benefits.
5. Evaluate systems under intermittent connectivity and realistic workload conditions.
6. Include cybersecurity, privacy, model-update, and reliability requirements from the design stage.
7. Develop interdisciplinary pilot projects connecting Computer Engineering with agriculture, energy, environmental monitoring, and smart-city applications.
8. Assess total cost and lifecycle impact before large-scale deployment.

Acknowledgment

The author expresses sincere gratitude to the faculty members of the Department of Computer Engineering, Ashokrao Mane Group of Institutes, Kolhapur, for their guidance, encouragement, and academic support during the preparation of this study.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper

References

1. Ray, P. P. "A Review on TinyML: State-of-the-Art and Prospects." Journal of King Saud University – Computer and Information Sciences, 2022.
2. Capogrosso, L., et al. "Tiny Machine Learning for Large-Scale IoT Deployments: A Systematic Review." IEEE Access, 2023.
3. Immonen, A., et al. "On-Device Machine Learning: Energy Perspective." IEEE Access, 2023.
4. Sanchez-Iborra, R., et al. "TinyML for Industrial IoT: A Systematic Review." Sensors, 2024.
5. Shi, W., Cao, J., Zhang, Q., Li, Y., and Xu, L. "Edge Computing: Vision and Challenges." IEEE Internet of Things Journal, 2016.
6. Satyanarayanan, M. "The Emergence of Edge Computing." Computer, 2017.
7. World Economic Forum. "AI, IoT and Edge Computing" and related research on intelligent and connected systems.
8. Conference brochure: "Innovative Frontiers in Science, Technology, and Multidisciplinary Research: Integrating Artificial Intelligence and Smart Technologies for Sustainable Development and Viksit Bharat 2047," 2026.