International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424
Volume 3 Issue 3 , March 2026
DOI: https://doi.org/10.71366/ijwos03032669161

Edge computing and sustainable, low-power AI systems

Manoj A[1], Mrs K.Kalaivani[2]
[1] *Pg & Research Department of Computer Science, Sri Ramakrishna College Of Arts & Science*
[2] *Assistant Professor, Pg & Research Department of Computer Science, Sri Ramakrishna College Of Arts & Science.*

Article Info

Article History:

Published:05 March 2026

Publication Issue:
Volume 3, Issue 3
March-2026

Page Number:
62-71

Corresponding Author:
Manoj A

Abstract:

The increasing use of artificial intelligence (AI) in smart environments, industrial automation, healthcare monitoring, and Internet of Things (IoT) networks has increased the computational requirements and power consumption of cloud computing infrastructure. The use of cloud computing for AI processing faces challenges such as high latency, high bandwidth consumption, privacy concerns, and environmental emissions due to the large number of data centers. This work introduces a sustainable and low-power AI model based on the smart edge architecture. It includes energy-efficient machine learning algorithms deployed at the edge for making smart decisions. This model achieves optimal performance using lightweight model deployment, dynamic resource allocation, and hardware-aware optimizations such as model quantization and hardware pruning techniques. A modular architecture is proposed that increases efficiency by focusing on data acquisition, edge processing, AI inference, energy management, and cloud synchronization. Performance metrics include latency, energy consumption, compute efficiency, and inference quality. The latter, in particular, is important for providing efficient AI capabilities on devices with limited resources. Results of experiments show that local processing reduces network and energy use compared to processing in the cloud. The proposed framework enables scalable and efficient deployment of AI while minimizing the environmental impact and maintaining performance to support sustainable computing.

Keywords: Edge Computing, Sustainable Artificial Intelligence, Low-Power AI Systems, Energy Efficient Computing, Edge AI, Green Computing, IoT Intelligence, Lightweight Machine Learning

1. Introduction

AI technologies have been applied to smart cities, health monitoring systems, smart transport systems, and industrial automation. Classical AI applications, however, are typically execution-centric, delegating the processing of massive volumes of data generated by connected devices to centralized cloud computing infrastructures. Many drawbacks exist despite the increased processing power offered by cloud infrastructure, including increased latency, the requirement of a reliable internet connection, privacy issues and the high power consumption.The rapid increase in the number of connected devices within the Internet of Things environment has further accentuated the requirement for a decentralized computational paradigm that can process analytics in real time.

Edge computing has emerged as a solution to these requirements by physically moving the computation and storage capacity of devices closer to the source of data.

However, the use of artificial intelligence models on edge devices also brings forth new challenges that are associated with limited processing power and battery capacity. Conventional deep learning models are computationally intensive and require a lot of memory, and hence, they cannot be used directly on low-power devices without the need for optimization. Thus, sustainable AI development also involves creating intelligent systems that are accurate and efficient.

Some of the current trends in research include the adoption of Green AI principles that aim to ensure that the model is environmentally sustainable during training and deployment. Sustainable AI models are designed to be carbon emission-reducing through the optimization of algorithm efficiency and the reduction of unnecessary data transfer between devices and servers. Edge AI models can be very helpful in the achievement of sustainability goals.

The proposed research work indicates a sustainable low-power AI system with an edge computing architecture that can be applied for efficient real-time intelligence with low power consumption.

The proposed framework indicates the need to add modular processing elements that are supported by energy-optimized strategies.

2. RELATED WORK

Recent breakthroughs in the field of artificial intelligence and distributed computing have inspired the use of edge intelligence to address the challenges posed by the centralized processing architecture. Various researchers have investigated frameworks that combine machine learning algorithms with edge devices to support real-time analytics.

The initial cloud-based AI systems relied heavily on cloud servers for processing, resulting in high bandwidth usage and poor latency response in applications such as autonomous vehicles and healthcare.

Another domain of interest is energy efficiency, which has recently received attention because of the environmental issues associated with the large data centers that are being used for AI-related tasks. Methods such as model compression, which include pruning, distillation, and quantization, have been proposed to allow the execution of deep learning models on embedded systems.

Recent research on edge AI frameworks has been centered on the role of hardware-aware optimization methods.

Neural processing units and low-power GPUs have made it possible to perform AI inference on mobile devices. Adaptive scheduling methods for workload have also been considered for managing task distribution between edge nodes and cloud servers, depending on resource availability.

3. PROBLEM STATEMENT

The increasing adoption of artificial intelligence in distributed smart environments has resulted in a huge dependency on cloud infrastructure for large-scale data processing. Although cloud computing is scalable, there are some working challenges associated with the use of cloud computing in environments where latency is a concern and energy consumption is low. The constant data stream from edge devices to cloud servers results in increased bandwidth consumption and network congestion, which affects applications such as autonomous monitoring systems,healthcare diagnostics, and industrial automation.Another major concern that has been associated with the deployment of AI in a centralized fashion is the high energy consumption. The data centers consume a lot of electrical energy in terms of processing the workload and the cooling systems, which results in increased operational costs and carbon emissions.

With the rapid adoption of AI, the sustainability of cloud computing is being questioned.Edge devices like sensors, smart cameras, wearable technology, and IoT gateways produce enormous amounts of data in real-time. But edge devices also work under very tight constraints, such as a limited battery life, less memory availability, and less processing power. Conventional deep learning models are computationally complex and cannot be directly deployed on edge devices without any modifications.

Also, there are privacy and security concerns when personal user data is constantly being sent over public networks for processing. Health monitoring or smart personal space applications need intelligence to be developed at the edge to maintain data privacy.However, the existing edge AI systems are only partially addressing the latency issue while neglecting the energy monitoring and sustainability analysis components.

4. OBJECTIVES OF THE STUDY

The main objective of this research work is to design a sustainable and energy-efficient artificial intelligence framework that can support intelligent decision-making at the edge of the network. The designed framework will be able to support real-time analytics with less computational overhead and environmental impact.

The objectives of this research study are as follows:

- To design an edge computing framework that can support data processing at the edge to reduce latency and bandwidth usage.
- To design energy-efficient artificial intelligence models that can be executed on low-power devices
- using model compression, quantization, and pruning techniques.
- To integrate energy monitoring modules that can assess the power consumption of devices during AI

inference tasks.

- To improve user privacy by reducing data transmission outside the system and implementing localized decision-making.

5.PROPOSED SYSTEM ARCHITECTURE

The proposed system brings forth a modular edge computing architecture that is intended to facilitate sustainable and low-power artificial intelligence processing. The proposed system architecture is focused on localized data intelligence, energy optimization, and adaptive workload management for efficient system operation.

The framework comprises three large functional modules:

A. Data Acquisition Module

The data acquisition module acquires real-time data inputs from sensors, cameras, wearable devices, or IoT-enabled devices. The data could comprise environmental data, images, audio signals, or system status depending on the domain of application.Noise removal, normalization, and feature extraction, which are pre-processing tasks, are carried out at this level.

B. Edge Processing Module

This is as opposed to sending the data to servers for processing. The data processing at the edge involves data aggregation, temporary storage, anomaly detection, and workload scheduling.

The module is also responsible for deciding whether the computation should be done at the edge or whether cloud help is required.

C. Low-Power AI Inference Module

This module symbolizes the intelligence of the proposed system. Low-power machine learning and deep learning models are used for inference.

Optimization techniques used are:

- Model quantization, which reduces the precision of numbers.
- These techniques greatly reduce the memory and computational requirements of the model while still having a reasonable prediction accuracy.

6. SYSTEM METHODOLOGY

A. Requirement Analysis

The initial phase of the system development methodology is requirement analysis. The functional requirements of the system comprise real-time data acquisition, localized AI inference, adaptive workload management, and cloud synchronization. The non-functional requirements of the system include low latency, low power consumption, high prediction accuracy, and data privacy preservation.Taking into consideration the limitations of edge devices, which include memory constraints, lower processing power, and battery capacity, the system must be able to work within tight computational budgets while ensuring acceptable AI performance

B. Model Selection and Optimization

Conventional deep learning models are computationally intensive and cannot be run on low-power devices without optimization. As a result, efficient models are selected and optimized using the following methods:

- Model Quantization

The bit precision of the model is decreased from 32-bit floating-point to 8-bit or lower, resulting in a substantial decrease in memory and computational complexity.

- Model Pruning

The unnecessary connections in the neural network are removed to optimize the tasks.

- Knowledge Distillation

Knowledge is transferred from a pre-trained large model (teacher model) to an efficient model (student model).

C. Edge Deployment Strategy

The deployment includes:

1. Runtime monitoring services
2. Lightweight inference engines
3. Resource-awarescheduling modules

   The deployment strategy is designed in such a way that the inference tasks are carried out within a specified energy budget.

7. ENERGY-AWARE EDGE AI PROCESSING ALGORITHM

The algorithm is designed to promote efficient AI processing while consuming less energy.

Algorithm 1: Energy-Aware Edge AI Inference

Algorithm 2: Input: Sensor data stream D

Output: Optimized inference result R

Initialize system parameters and energy threshold values.

Algorithm 3: Perform preprocessing (noise removal, normalization).

Algorthim 4: Else if E ≤ threshold:

Switch to low-complexity inference mode.

Else if E ≤ threshold:

a. Transition to low-complexity inference mode

b. Lower model precision or inference rate.

c. Produce approximate result R.

Algorithm 5: Log energy consumption and performance data.

Algorithm 6: If long-term storage needed:

- Send summarized result to cloud.

Algorithm 9: Repeat for next data cycle.

Algorithm Properties

- Adaptive computation control
- Unnecessary processing reduction
- Selective cloud communication
- Energy-limited optimization

The algorithm guarantees the system runs continuously even when power is limited and still provides acceptable accuracy.

8. IMPLEMENTATION DETAILS

The proposed sustainable low-power artificial intelligence framework was implemented with a modular software architecture to enable scalability and efficient deployment in heterogeneous edge environments.

The implementation is centered on lightweight execution, resource monitoring, and adaptive intelligence processing.

A. Hardware Environment

The system is intended for deployment on resource-constrained edge platforms such as embedded single-board computers and IoT gateways. The typical deployment setup would involve low-power processors with integrated GPUs or neural acceleration.

The hardware setup enables:

- Multi-core CPU processing
- Optimization of limited RAM usage
- Battery-friendly operation
- Wireless communication interfaces.

Sensors with wired or wireless interfaces are continuously generating real-time environmental or system data.

B. Software Environment

The software implementation employs a lightweight runtime environment that is capable of executing machine learning inference.

The software components are:

- Python-based processing framework.
- Lightweight machine learning inference libraries.
- Containerized deployment support for modular updates.
- Edge runtime monitoring services.

Data preprocessing modules include normalization and feature extraction prior to model inference. The system supports real-time streaming data processing pipelines.

C. AI Model Deployment

Machine learning models were optimized and transformed into lightweight models that can be deployed on edge inference.

The steps involved in deployment are:

- Training models using large datasets in a centralized setup.

D. Energy Monitoring Integration

An energy monitoring feature is always collecting system information such as:

- CPU usage
- Memory usage
- Battery usage
- Device temperature.

9. TESTING AND EXPERIMENTAL RESULTS

A. Unit Testing

The system modules were tested individually,the accuracy of data acquisition was assessed by testing the continuous sensor data streams.The accuracy of AI inference was assessed by testing the benchmark datasets.The energy monitoring capability was tested by simulating workload variations.The modules were able to accomplish their respective tasks without system failure.

B. Integration Testing

Integration testing was performed to ensure smooth interaction between the modules.

The system functioned well in terms of data transfer from acquisition to decision output without any loss of packets or processing delay.

C. System Testing

System testing was done to test actual scenarios with continuous data streams.

The results include:

- Robust performance after prolonged usage.
- Lower latency than cloud processing.

D. Experimental Observations

Experimental assessment showed that local AI inference resulted in a substantial reduction in network dependency,the model ensured a consistent level of prediction accuracy with a reduced number of communication occurrences and lower power consumption.

10. SUSTAINABILITY AND CARBON FOOTPRINT ANALYSIS

Artificial intelligence computing workloads are major energy consumers in the global computing ecosystem. There is a need for sustainable computing practices to reduce the carbon footprint.

The proposed framework supports sustainability in the following ways:

- Less reliance on cloud computing.
- Dynamic control of computations.

11.CONCLUSION AND FUTURE SCOPE

The study has proposed a sustainable edge computing framework to facilitate low-power artificial intelligence deployment in a distributed setup. The framework combines lightweight machine learning models with adaptive energy monitoring to facilitate real-time intelligence with low computational complexity.

The experimental study has shown the effectiveness of the proposed framework in achieving lower latency, enhanced privacy preservation, and lower energy consumption compared to traditional cloud computing-based AI processing systems. The framework is scalable and maintainable with the ability to facilitate localized intelligence.

The proposed work has demonstrated a sustainable edge computing framework for low-power artificial intelligence to be deployed in a distributed manner.

The framework integrates lightweight machine learning models with adaptive energy monitoring for real-time intelligence with low computational complexity.

References

1. X. Ran et al., “DeepDecision: A Mobile Deep Learning Framework for Edge Video Analytics,” IEEE INFOCOM, 2018.
2. Y. Mao, C. You, J. Zhang, K. Huang, and K. Letaief, “Mobile Edge Computing: Survey and Research Outlook,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, 2017.
3. A. Abadi et al., “TensorFlow: Large-Scale Machine Learning Systems,” USENIX Symposium, 2016.
4. H. Esmaeilzadeh et al., “Dark Silicon and Energy Efficiency Challenges,” IEEE Micro, vol. 32, no. 3, pp. 122–134, 2012.

5. R. Schwartz et al., “Green AI,” Communications of the ACM, vol. 63, no. 12, pp. 54–63, 2020.
6. J. Redmon and A. Farhadi, “YOLOv3: Real-Time Object Detection,” arXiv, 2018.
7. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks,” arXiv, 2014.
8. Y. Bengio et al., “Representation Learning: A Review,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.
9. M. Satyanarayanan, “The Emergence of Edge Computing,” Computer, vol. 50, no. 1, pp. 30–39, 2017.
10. S. Deng, H. Zhao, W. Fang, J. Yin, and A. Dustdar, “Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence,” IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7457–7469, 2020.