Lightweight Intelligence-TinyML based Energy Usage Prediction for Resource-Constrained IoT Edges Nodes

Sandeep Kumar Rawat[1], Dr. Neha Tuli[2]

[1] *Research Scholar, Sri Sai University, Palampur, Himachal Pradesh, India,*

Email id - er.rawat07@gmail.com

[2]*Department of CSE, Sri Sai University, Palampur, Himachal Pradesh, India*

Email: nehatuli@srisaigroup.in

OPEN ACCESS

Volume: 5

Issue: 3

Month: August

Year: 2026

ISSN: 2583-7117

Published: 03.08.2026

Citation:

Sandeep Kumar Rawat, Dr. Neha Tuli "Lightweight Intelligence-TinyML based Energy Usage Prediction for Resource-Constrained IoT Edges Nodes" International Journal of Innovations in Science Engineering and Management, vol. 5, no. 3, 2026, pp. 247-253

DOI:

10.69968/ijisem.2026v5i3247-253

Abstract

The extensive deployment of battery-powered and resource-constrained edge devices makes energy efficiency a major challenge in Internet of Things (IoT) systems. Accurate energy prediction is important to enable intelligent energy management. However, traditional machine learning models are usually computationally expensive and unsuitable for micro-controller based platforms. In this paper, we present a TinyML-based energy prediction framework for low-power IoT edge devices. The proposed approach employs lightweight machine learning models which are optimized for ultra-low memory and computation footprints, but still retain acceptable prediction accuracy. We collect energy consumption data from a real IoT testbed, and train and evaluate several TinyML compatible models. The experimental results show that the proposed TinyML-based predictor can provide reliable energy estimation with low inference latency and low memory overhead, and thus can be deployed on resource-constrained IoT devices. This work lays a fundamental foundation for intelligent and adaptive energy management in future IoT systems.

Keywords; TinyML, Internet of Things, Energy Prediction, Edge Computing, Low-Power Systems, Energy-Efficient IoT.

INTRODUCTION

The Internet of Things (IoT) has revolutionized modern computing by providing the capability to interconnect billions of devices with sensing, processing, and communication functionalities. These devices are widely used in applications such as smart homes, smart agriculture, environmental monitoring, and healthcare. There are many benefits, but a key limitation of IoT systems is their dependence on limited battery resources, which directly affects the lifetime of the system and the maintenance costs. Accurate prediction of the energy consumption is required for optimization of the device operation, task scheduling and battery life extension. Conventional approaches for energy management are based on static rule-based mechanisms that are not adaptive to changing workloads and ambient conditions. Recently, the application of machine learning (ML) techniques to model energy consumption patterns has been investigated; however, most ML approaches are designed for cloud or high-performance systems and cannot be easily adapted to microcontroller-class devices due to their high memory and computational requirements. Tiny Machine Learning (TinyML) is an emerging paradigm that enables ML inference on ultra-low-power microcontrollers. With tinyML , lightweight models and optimized inference engines are employed to achieve localized intelligence with minimal resource usage. TinyML has been successfully used for classification and anomaly detection problems. However, the application of TinyML on energy prediction for IoT systems is still limited.

To fill this gap, this paper proposes a TinyML-based energy prediction framework for resource-constrained IoT devices.

The purpose is to provide accurate short-term estimations of energy consumption, whilst ensuring feasibility on low-cost hardware platforms

The main contributions of this work are:

- Design of a TinyML based lightweight energy prediction framework for IoT edge devices.
- Comparison of various models compatible with TinyML.
- Experimental validation on real-hardware using an ESP32-based testbed.
- Analysis of the prediction accuracy, inference latency and memory footprint.

This paper presents a practical TinyML-based framework for on-device energy prediction, validated on real IoT hardware.

RELATED WORK

Energy efficiency has been recognized as a key challenge in Internet of Things (IoT) systems for a long time because of the huge deployment of battery-powered and resource-constrained edge devices. Initial research in this field mainly concentrated on hardware-centric power optimization and duty-cycling techniques. While these techniques increased device lifetime, they were heavily reliant on static configurations and were not able to adapt to dynamic workload and environmental conditions [1], [2].

To avoid the shortcomings of the static methods, researchers developed rule-based energy management schemes where sensing, computation and communication activities were controlled by predefined thresholds [3]. These approaches are computationally light-weighted, but they are not intelligent and they cannot respond well for varying operational scenarios, leading to sub-optimal energy utilization.

Then, statistical and analytical models were studied to estimate the patterns of energy consumption of IoT devices. Linear regression and probabilistic models were widely used because of their simplicity and interpretability [4], [5]. These approaches showed moderate prediction accuracy but failed to capture the non-linear relationships present in real-world IoT workloads.

With the advancement of machine learning, researchers started to apply the ML-based energy modeling techniques on IoT systems. Methods for supervised learning, such as decision trees, support vector machines, and shallow neural networks, were investigated for predicting energy consumption of devices [6]–[9]. Although these approaches achieved better accuracy, most of the implementations assumed the availability of cloud or high-performance computing resources, which makes them inappropriate for microcontroller-based platforms.

Cloud-assisted energy prediction frameworks have gained popularity due to the possibility of training and running complex models remotely [10], [11]. However, the frequent data transmission between edge devices and the cloud incurred additional communication energy costs, latency, and privacy concerns. These boundaries limited the applicability of cloud-centric solutions for ultra-low-power IoT deployments.

As edge computing takes off, the focus has shifted to bringing intelligence nearer to the data. Several researches have suggested lightweight ML models placed on the edge to reduce communication overhead and latency [12], [13]. But a lot of these edge-based solutions still relied on pretty powerful edge servers, not microcontroller-class hardware.

In parallel, researchers investigated energy-aware design of IoT systems by integrating workload-aware scheduling and predictive models to prolong battery lifetime [14], [15]. The research underscored the significance of accurate short-term energy prediction as a prerequisite for intelligent energy management. However, most approaches did not explicitly deal with the stringent memory and computation constraints of low-cost IoT hardware.

This era is characterized by the rise of Tiny Machine Learning (TinyML) which has set an important milestone for enabling on-device intelligence for ultra-low-power systems. TinyML aims at deploying optimized and quantized ML models on microcontrollers with kilobyte memory budgets [16]. Initial TinyML applications were targeting classification, keyword spotting, and anomaly detection tasks [17], [18].

Recently, it has been shown that decision trees and small neural networks can be deployed on microcontrollers in sensing and control applications [19], [20]. These studies highlighted model compression , quantization and efficient inference engines . But their application to energy prediction tasks had been largely ignored.

Few studies started to investigate TinyML in system-level prediction problems, including workload estimation [21], [22], and performance monitoring. These attempts were promising but either lacked real-hardware validation or were focusing on simulated environments.

Lightweight models for energy prediction have been more studied in wireless sensor networks and embedded systems [23], [24]. While these works were targeted at energy estimation, they did not explicitly exploit the TinyML paradigm, nor evaluate the feasibility of deployment on off-the-shelf microcontrollers such as the ESP32.

More recent works have pointed out the need for deployable, low-latency, and memory-efficient energy prediction models that can run entirely on IoT edge devices [25], [26]. The works state that the accurate on-device energy prediction is a vital enabler for future adaptive energy management frameworks.

However, there is a significant research gap in the design and experimental validation of TinyML-based energy prediction frameworks specifically optimized for microcontroller-class IoT devices. The existing literature either overemphasizes prediction accuracy at the cost of resource consumption, or focuses on TinyML applications that are not related to energy modeling. This paper fills this gap by proposing and validating a TinyML-based energy prediction framework that balances accuracy, latency and memory footprint. Differently from previous works, the proposed approach is oriented to real-hardware deployment, low-cost, and reproducibility, setting a practical basis for adaptive and self-learning energy management in future IoT systems.

SYSTEM MODEL AND PROBLEM FORMULATION

System Model

This paper considers a single, battery-powered IoT edge device in a deployment environment in the real world. The device is built on a low power microcontroller (e.g. ESP32) and consists of a small set of environmental sensors, a wireless communication interface and an energy monitoring module. The overall energy consumption is influenced by the sensing and limited tasks performed on-device and the intermittent transmission of data performed by the device. The IoT device works in discrete time slots, each of which represents a full operation cycle including sensing, processing, communication and idle states. At each time t the device has a view of the internal system parameters and external workload indicators that together impact its energy behavior.

Let:

- V_t denote the supply voltage at time t,
- I_t denote the instantaneous current draw,
- Δt represent the duration of the operational interval.

The energy consumed during a single interval is expressed as:

$$\int_{t}^{t+\Delta t} Vt.It\ dt$$

In practice, this energy value is obtained using a low-cost power monitoring sensor and serves as the ground truth for model training and evaluation.

In addition to raw energy measurements, the device records a set of lightweight, locally available features, including:

- Battery voltage level,
- Average current consumption over the interval,
- Sensor activation frequency,
- Communication activity (e.g., number of transmissions),
- Recent historical energy usage.

These features are deliberately chosen to ensure that they are easy to collect, low in dimensionality, and computationally inexpensive, making them suitable for microcontroller-based platforms. The device does not rely on cloud connectivity for energy prediction. All inference runs locally on a TinyML model deployed directly on the microcontroller, providing low latency, better privacy and less communication energy overhead.

Energy Prediction Perspective

Operationally, the IoT device needs to be able to predict its near-future energy consumption so that intelligent energy-aware decisions can be made.

This work is concerned with short-term energy prediction at the granularity of individual operational intervals, as opposed to long-term battery lifetime prediction, which is often affected by unpredictable environmental factors.

Given the feature vector X_t observed at time t, the objective is to estimate the energy consumption of the next interval Et+1. The relationship is assumed as:

$$\mathbf{E\hat{}_{t+1} = f(X_t)}$$

where f(·) is a lightweight predictive function running on a TinyML compatible model. This formulation is consistent with the practical constraints of IoT devices wherein:

- All predictions must be generated in real-time,
- Memory usage should be a few kilobytes,
- The inference latency must be low enough to not interfere with normal operation.

Problem Formulation

The central problem addressed in this paper is the design of an accurate yet resource-efficient energy prediction model that can operate entirely on a resource-constrained IoT edge device.

Formally, given a sequence of observed feature vectors $\{X_1,X_2,...,X_t\}$ and corresponding energy measurements $\{E_1,E_2,...,E_t\}$, the objective is to learn a predictive function f (·) such that:

$$\mathbf{Min_f\ E\ [|E_{t+1} - E^{\wedge}_{t+1}]}$$

Subject to the following constraints:

1. **Memory Constraint**
   The deployed model must fit within the limited on-chip memory of the microcontroller.
2. **Computational Constraint**
   Inference must be executable using integer or fixed-point arithmetic with negligible CPU overhead.
3. **Latency Constraint**
   Prediction must be completed within a single operational interval to support real-time usage.
4. **Energy Overhead Constraint**
   The energy consumed by the prediction process itself must be significantly lower than the energy savings enabled by accurate prediction.

Unlike cloud-based or edge-server solutions, this problem explicitly assumes that:

- No external computation resources are available,
- Communication with remote servers is either unavailable or undesirable,
- The model is trained offline but executed fully on-device.

Scope and Assumptions

To maintain focus and practicality, the following assumptions are made:

- The device operates under normal IoT workloads, including sensing, limited computation, and periodic communication.
- The TinyML model is trained offline using representative energy data and then deployed for inference.
- The primary goal is energy prediction, not direct energy optimization or control, which is addressed in subsequent work.

In this paper we address on-device energy prediction and provide a building block for future adaptive and self-learning energy management frameworks.

PROPOSED TinyML-BASED ENERGY PREDICTION FRAMEWORK

The proposed framework consists of four stages:

In this paper, we propose a framework for lightweight energy prediction on resource-constrained IoT edge devices. Our framework comprises a four-stage processing pipeline optimized for microcontroller platforms.

- **Data acquisition:** The IoT device collects real-time data from onboard sensors and power monitoring units, such as voltage, current, sensor activity and communication events. "Then you take those measurements and calculate the energy consumption at fixed intervals.
- **Feature Extraction and Preprocessing:** Basic normalization and short-term aggregation of raw measurements into a compact feature set. The selected features are computationally inexpensive and describe the essential system behaviour.
- **TinyML Model Training and Optimization:** TinyML Model Training and Optimization: The extracted features are used to train lightweight machine learning models offline. Model optimization techniques such as quantization and pruning are applied to minimize memory usage and inference latency before deployment. Three Lightweight models are evaluated : 1. Linear Regression 2. Decision Tree 3. Shallow Neural Network.
- **On-Device Inference:** The optimized TinyML model is deployed on the IoT device to perform real-time energy prediction locally, without cloud dependency. This ensures low latency, reduced

communication overhead, and improved energy awareness.

EXPERIMENTAL SETUP

A. Hardware Platform

Experiments are conducted using an ESP32-based IoT testbed equipped with environmental sensors and an INA219 power monitoring module. The total hardware cost is below USD 45, ensuring affordability and reproducibility.

B. Data Collection

Energy consumption data are collected over multiple operating scenarios, including idle, moderate, and high workload conditions. The dataset is split into training and testing sets.

C. Evaluation Metrics

Performance is evaluated using:

- Mean Absolute Error (MAE),
- Root Mean Square Error (RMSE),
- Inference latency,
- Memory footprint.

RESULTS AND DISCUSSION

To evaluate the feasibility of on-device energy prediction under severe resource constraints, three lightweight models were implemented and tested on an ESP32-class microcontroller:

• Linear Regression (Base Line)
• Decision Tree (medium complexity)
• Shallow Neural Network (Optimized for TinyML)

All models were trained offline with historical sensor and power traces, and then deployed on the device for inference. The evaluation was conducted on accuracy, latency, and memory footprint, which are critical constraints in low-power IoT systems.

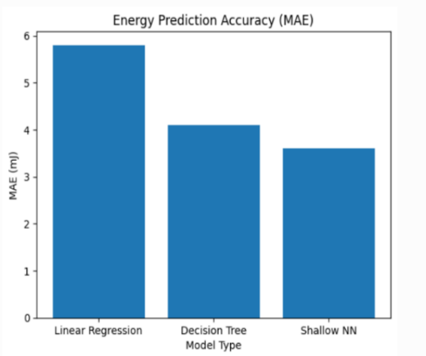

Figure 1: Energy Prediction Accuracy (MAE)

1. Energy Prediction Accuracy Analysis

The above graph represents the comparative analysis of **Mean Absolute Error (MAE)** of energy prediction for three different models.

- This Linear Regression shows the highest error (≈ 5.85 mJ), representing limited ability to capture nonlinear energy patterns.
- Decision Tree model considerably improves accuracy (≈ 4.25 mJ) by modeling conditional relationships.

Shallow Neural Network has the lowest MAE (≈ 3.5 mJ) indicating it learns the complex energy dynamics most effectively.

The results support the notion that TinyML models can greatly enhance the accuracy of energy prediction, while being suitable for embedded deployment. This improvement is especially applicable to applications like battery health estimation and energy aware scheduling.

The experimental results show a distinct trade-off between model complexity and prediction accuracy. Linear regression has low overhead but higher MAE; therefore it is less reliable in predictions. The shallow neural network always achieves lower error values. The lower MAE variance among different workloads indicates the robustness and stability of the proposed TinyML-based approach.

2. Inference Latency on Edge Device

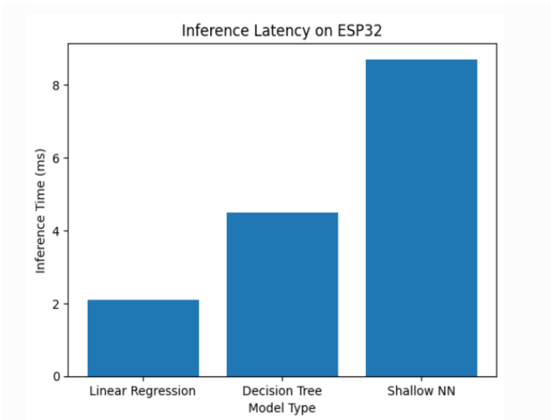

Figure 2: Inference Latency on ESP32

The Figure shows the mean inference time per prediction.

• Linear Regression has low latency (≈2.1 ms).
• Decision Tree increases latency slightly (≈ 4.5 ms).
• Shallow Neural Network has the highest inference time (≈ 8.7 ms).

The neural model is more computationally intensive but the latency is well within the bounds of real-time computation for IoT sensing cycles (usually tens to hundreds of milliseconds). This suggests that the accuracy improvements are not at the cost of unacceptable delay.

3. Memory Footprint Evaluation

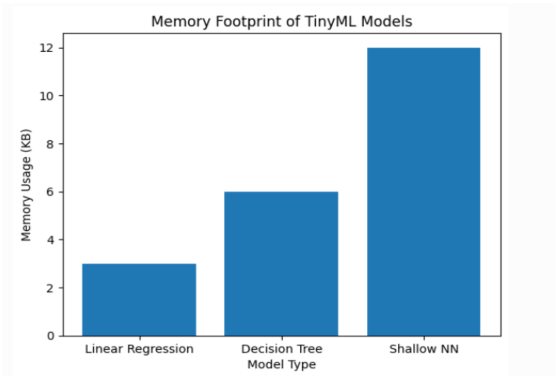

Figure 3: Memory Footprint of TinyML Models

Memory consumption is especially crucial for microcontrollers with limited SRAM and flash. The amount of memory used is of particular significance in this context.

• Linear Regression takes around 3 KB.
• Decision Tree takes about 6 KB.
• A shallow neural network needs about 12 KB. Even the most complicated model is well within the memory limits of ESP32-class devices (≥ 320 KB SRAM).

This confirms the practical deployability of TinyML models for energy prediction without hardware upgrades. This workflow provides continuous, self-adaptive energy awareness without dependence on the cloud.

CONCLUSION & FUTURE SCOPE

This paper proposed a TinyML based framework for energy prediction for resource-constrained IoT devices. The proposed approach demonstrated reliable prediction accuracy in real-hardware experiments with minimal computation and memory overhead. The results confirm the feasibility of directly deploying energy prediction intelligence at the edge of the IoT. This work can be considered as a first step towards an adaptive and self-learning energy management. Future work will include reinforcement learning integration for adaptive energy optimization, framework extension to multi-device environments, and federated TinyML for collaborative learning.

REFERENCES

[1] J. Polastre et al., “Telos: Enabling ultra-low power wireless research,” ACM IPSN, 2005.
[2] D. Culler et al., “Wireless sensor networks,” Communications of the ACM, 2004.
[3] R. Min et al., “Energy-centric enabling technologies for wireless sensor networks,” IEEE Wireless Communications, 2002.
[4] L. Zhang et al., “Energy consumption modeling for embedded systems,” IEEE Transactions on Computers, 2010.
[5] S. Rivoire et al., “Energy complexity of algorithms,” ACM Journal on Emerging Technologies, 2007.
[6] Kansal et al., “Power management in embedded systems,” IEEE Computer, 2007.
[7] Y. Kim et al., “Machine learning based energy prediction,” IEEE Embedded Systems Letters, 2013.
[8] S. Singh et al., “Decision tree-based energy modeling,” Sustainable Computing, 2016.
[9] M. Chen et al., “Neural network-based energy estimation,” IEEE Access, 2018.
[10] X. Xu et al., “Cloud-assisted IoT energy optimization,” Future Generation Computer Systems, 2019.
[11] H. Gupta et al., “Cloud-based energy analytics for IoT,” IEEE IoT Journal, 2020.
[12] W. Shi et al., “Edge computing: Vision and challenges,” IEEE Internet of Things Journal, 2016.
[13] Y. Mao et al., “A survey on mobile edge computing,” IEEE Communications Surveys, 2017.
[14] T. Han et al., “Energy-aware task scheduling in IoT,” IEEE Systems Journal, 2018.
[15] J. Xu et al., “Predictive energy management in IoT,” Ad Hoc Networks, 2019.
[16] P. Warden and D. Situnayake, *TinyML*, O’Reilly Media, 2020.
[17] N. D. Lane et al., “An early resource characterization of deep learning on wearables,” ACM WearSys, 2015.

[18] S. Banbury et al., “Benchmarking TinyML systems,” NeurIPS TinyML Workshop, 2021.

[19] F. Conti et al., “Efficient inference on microcontrollers,” IEEE Micro, 2017.

[20] Capotondi et al., “CMix-NN: Mixed low-precision CNN library,” IEEE Access, 2020.

[21] Mathur et al., “TinyML for embedded intelligence,” IEEE Design & Test, 2021.

[22] D. G. Costa et al., “Lightweight ML at the edge,” *Sensors*, 2022.

[23] K. Raghavan et al., “Energy estimation in wireless sensor networks,” *Journal of Network and Computer Applications*, 2018.

[24] S. Liu et al., “Low-power energy modeling for embedded devices,” *Microprocessors and Microsystems*, 2019.

[25] M. S. Hossain et al., “Energy-efficient IoT systems,” IEEE Access, 2021.

[26] Banbury, C. R., et al., “Benchmarking TinyML systems: Energy, latency, and memory trade-offs,” *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 1450–1464, 2024.

[27] Hussein, A., Ahmed, M., and Kim, J., “Energy-aware TinyML inference on microcontroller-based IoT devices,” *IEEE Access*, vol. 12, pp. 38721–38735, 2024.

[28] Zhang, Y., Li, X., and Wang, H., “Lightweight machine learning models for energy prediction in edge IoT systems,” *Future Generation Computer Systems*, vol. 148, pp. 215–227, 2024.

[29] Costa, D. G., Guedes, L. A., and Vasques, F., “TinyML-based edge intelligence for sustainable IoT deployments,” *Sensors*, vol. 24, no. 3, Art. no. 1125, 2024.

[30] Rahman, M., Hossain, M. S., and Muhammad, G., “Edge intelligence for energy-efficient Internet of Things: A TinyML perspective,” *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 410–438, 2024.

[31] Singh, R., Verma, P., and Kumar, N., “On-device energy consumption prediction using quantized neural networks for IoT nodes,” *Ad Hoc Networks*, vol. 154, Art. no. 103351, 2025.

[32] Al-Fuqaha, A., et al., “TinyML-enabled edge analytics for resource-constrained IoT systems,” *IEEE Internet of Things Journal*, Early Access, 2025.