International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

Energy-Efficient Algorithms for Edge AI on lOT Devices

Ms. Bhavyashree M[1], Sreeresh G[2], Sreya Chandran[3], Anand Goutham[4], Sreehari P S[5], AdhinkrishnaUS[6]
[1]Assistant Professor, Bachelor of Computer Application, Yenepoya (Deemed to be University), Bangalore, Karnataka, India.
[2,3,4,5,6]UG, Bachelor of Computer Application, Yenepoya (Deemed to be University), Bangalore, Karnataka, India.
Emails:*bhavyashree.m.blr@yenepoya.edu.in[1],sreereshggirish@gmail.com[2],chandransreya2005@gmail.com[3], anandgoutham11@gmail.com[4], thesreehari.me@gmail.com[5], adhinkrishnaus132@gmail.com[6]*

Abstract
Internet of Things Devices: Data Processing, there exist many Internet of Things devices in our life. So we need to implement data processing to Internet of Things devices. We can use Edge Artificial Intelligence in place of cloud computing all the time. Edge Artificial Intelligence can be used for executing data processing to Internet of Things devices. Then it will be more effective to reduce the delay, network usage and privacy. However, most of the Internet of Things devices have limited battery, processing power and memory. So, the Artificial Intelligence algorithm can consume energy in order to run in Internet of Things devices. Then it will reduce the life time and energy efficiency to Internet of Things devices. This research is about that importance of energy efficient algorithms to Edge Artificial Intelligence in Internet of Things environment. This research is about that conventional artificial intelligence models consume very large power. This research is also about that we need to create the algorithms which can run in limited resource environment. We can do many things to do that, for example: creating power efficient model, compressing model, removing unnecessary part of the model, representing data by numbers, and so on and so forth. All of these can reduce the energy consumption while maintaining the accuracy. This research is about maintaining the accuracy of the model while reducing the energy consumption. This is of critical importance in relation to their real-world feasibility. We have an Energy- algorithms device for the Internet of Things that can predict and anticipate what will happen. It can also respond quickly. It can do this without having to replace or recharge its batteries many times.

Keywords: *Edge Artificial Intelligence (Edge AI), Internet of Things (IoT), Energy-efficient algorithms Model compression and quantization, Sparse neural networks, adaptive voltage and frequency scaling (DVFS).*

1. Introduction
Internet of things is rapidly growing in the number of connected devices billions of them sensing collecting and transmitting data at any given instant. These devices are used for different problems like healthcare smart farming industrial automation and smart transportation. In most cases the data of these devices are uploaded to the cloud for further analysis and processing. But there are some problems associated with this such as high delay high internet usage and privacy. To resolve this problem Edge artificial intelligence Edge AI is becoming more and more popular and it processes the data and makes decisions locally at the edge nodes either at the device or the edge nodes. With local processing and local decision making at the edges, bandwidth usage is decreased and it results in a significant improvement in response time, privacy and reliability. Since most of the IoT are battery-powered devices, having limited memory and computing resources, it is not possible to run AI algorithms on them. In particular, many of the AI algorithms require high computing resources and power.When implementing the AI algorithms on the edge devices the energy efficiency of the algorithm is of prime importance. Because if

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

the AI algorithm is not energy efficient then it will consume the battery very fast hence the life time of the device will be decreased. Also, the energy efficiency is also important for the feasibility of the application. Hence researchers in this field are proposing energy efficient algorithms that give the same/similar accuracy while consuming less computing and power resources. There are a few techniques to solve this problem:

- Model compression.
- Light weight neural network architecture.
- Pruning.
- Quantization.
- Data processing techniques.

Using these techniques, we can construct low complexity models without any loss in performance. Thus, an energy efficient algorithm will enable the IoT devices to carry out complex tasks such as. Edge AI is extracted as a promising technology for IoT systems. The paper reviews the energy- algorithms for Edge AI techniques. The energy algorithm can lower Edge AI energy consumption, improve the performance and satisfy the low-power requirements of IoT systems. In addition, the paper elaborates the trade-off between resource efficiency and accuracy of Edge AI in practical systems

2. Literature Review

Which analyze the use of Artificial Intelligence driven approaches to increase the energy efficient of IoT networks. The paper discuss about the increasing number of IoT devices and the increase in energy consumption. This is a sustainability problem. The researcher sugggets using AI driven models like neural network, decision tree, reinforcement learning to predict and optimise the energy consumption patterns. The paper conclude that AI models can analyse a lot of data collected from the IoT settings to find the patterns of energy consumption and optimise them accordingly. The research results indicate that the application of AI in the network IoT is essential for survival. Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems: Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems presents a fast object detection system which is lightweight and appropriate for edge based IoT system. The research work addressed the problem of high energy consumption and latency in the existing end-to-end deep learning based object detection system. The research work proposed a frame difference method with the help of AI classifiers to make the system energy efficient. The proposed system was tested on various edge devices like AMD Alveo U50, Jetson Orin Nano and Hailo-8 AI Accelerator. The proposed system provides higher accuracy, lower latency and higher energy efficiency compared to the existing system. The research work explicitly indicates that the lightweight object detection system can be suitable for real-time IoT application. [1] In this paper we propose the notion of energy consumption modelling of IoT devices with the help of neural networks. How the neural networks can be used for resources allocation and tasks scheduling so as to minimise wasteful energy consumption. The system can predict the time of high consumption based on the previous usage pattern and react automatically. So we will observe a significant reduction of overall energy consumption and system stability. We also discuss the significance of employing the intelligent prediction models for energy aware IoT. [2] evaluates the frame difference approach as a power efficient alternative to complex optical flow and end-to-end detection approaches. the approach uses pixel difference comparison between frames and only processes when motion is detected. the approach is well suited for edge IoT devices, it can avoid redundant processing, thus saving power. experimental results demonstrate better latency and energy efficiency for fast moving object detection. [1] To emphasize the contribution of edge computing in achieving more energy efficient IoT, as it can mitigate the cloud computing dependence by reducing the energy consumption of data transmission. This is shown in the Reference Paper 1 when the hardware accelerators are utilized to implement efficient object detection algorithms and in the Reference Paper 2 when the AI based energy efficient techniques are implemented at the edge of the IoT network. Both the papers reveal that the combination of AI and edge computing can help in achieving more scalable and sustainable IoT.[1][2]

3. Methodology

Our work takes a methodical, experimental approach

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

to compressing high-performance AI to operate under the power constraints of IoT.

3.1. Method & Goals

Our work takes a methodical, multi-step analysis. Instead of profiling a set of pre-trained models, we construct and analyze a set of optimizations (from structural pruning to precision scaling) to demonstrate how the underlying accuracy vs. energy trade-offs of a combination of methods can be tuned for IoT [3].

3.2. Baseline discussion

We discuss Edge Ai market, and come up with following 3 issue. Latency(delay) on cloud processing, and its privacy. Because user's data is need to send to cloud for processing. “memory wall” for normal neural net on micro controller. Because controller cannot hold large neural net parameter because of its limited memory. There is no benchmark to measure the trade-off between battery life and predictions accuracy.

3.3. Optimization Framework

We discuss 3 way for energy optimization on Edge AI device Structural Optimzation. Prune the neurons.

- **Compression:** compress the size [4-5].
- **Numerical scale. Quantization:** convert to 8bit integer or other representation instead of 32bit float.

Architecture selection. lightweight network, e.g. MobileNet and TinyML-framework. Data optimization. make preprocessing small so cpu use more time for low power mode.

3.4. Experimental Architecture

We adopt the real life deployment 3 tier architecture:

- **Perception Layer:** all the platform raw sensor data
- **Edge Execution Layer:** the platform where the distilled model runs and decisions are made in realtime
- **Cloud Storage:** it is only for long term storage and/or retraining and the edge device is the key player [6]

3.5. Deployment and System Assessments

We train and construct our models in Tensorflow Lite and Pytorch Mobile as they can run on many edge devices. Then we “distill” the model with our optimality training method. After that we evaluate and compare with a set of hard metrics:

Power Consumption: the (mW) of usage in inference

- **Resource Utilization:** the max RAM and Flash.
- **Accuracy-Energy Trade-off:** relative accuracy loss for 1% of energy saved Our work tak [7].

3.6. Comparative Synthesis & Synthesis

A cross-functional analysis of the findings is the last stage. The study determines which particular pruning and quantization combinations provide the best scalability for real-time IoT applications by mapping the trade-offs between accuracy and power consumption.

4. Result and Analysis

This system uses a video frame analysis technique in which consecutive video frames are analyzed for motion using a frame differencing technique [8] Shown in Figure 1.

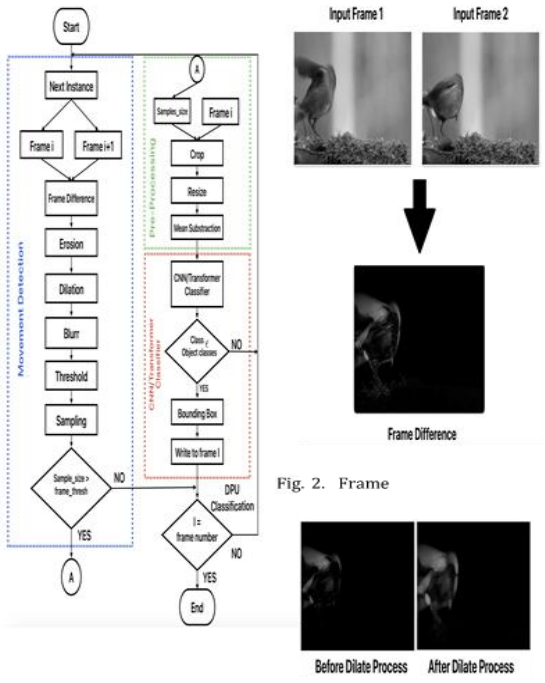

Fig. 2. Frame

Figure 1 Flowchart of Motion Detection and Object Classification Using Frame Differencing and CNN-Based Processing

Image preprocessing techniques such as erosion, dilation, blurring, and thresholding are carried out in order to remove noise from images and enhance the appearance of objects. After the preprocessing stage,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

the region of interest is cropped and resized for further processing [9]. The image features extracted from the preprocessing stage are used as input for a Convolutional Neural Network (CNN) model. The CNN model has an input layer, hidden layer, and output layer that learns the pattern and classifies the identified object. The system generates a bounding box around the identified object and shows the classification result on the video display. The results indicate that the image processing and neural network technique enhances the accuracy of object identification and classification Shown in Figure 2.

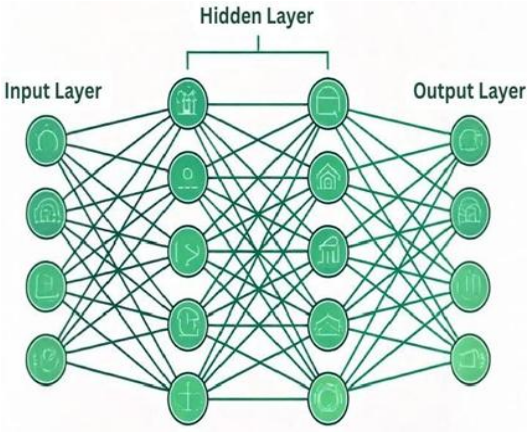

Figure 2 Architecture of an Artificial Neural Network (ANN) With Input, Hidden, And Output Layers

5. Discussion

The shift of artificial intelligence from servers to small devices like internet of things devices represents a big shift in how we do computing. We run into problems when we try to put complex models in these small devices since they lack power, memory and cooling capability. Studies have shown that to be successful with artificial intelligence on these devices we have to consider how the hardware and software interact and we have to ensure the system works well as a whole. The Balance of Model Compression. To run large artificial intelligence models on internet of things devices we need to compress the models. We can do this using techniques like pruning, quantization, and knowledge distillation. Pruning can help by removing unnecessary parts of the model to make it smaller and easier to run, but removing parts of the model may not make the model run faster. So we have to be careful how we prune the model. Another way to compress the model is to use quantization to change how the model stores the numbers. This can help reduce memory and energy usage. But we have to be careful because this may reduce how well the model works. Knowledge distillation is a method to make a model that can do about the same as a larger model. We do this by training the model to copy the larger model. Then the smaller model can do the same things as the larger model but it uses less memory and energy. The Need for Hardware and Algorithm to Work Together. We cannot just think about the algorithm we also have to think about the hardware it is running on. some devices, like the raspberry pi are not very good at running models. If we use special devices, like the google coral or NVidia jetson nano we can make the models run much faster and use less energy. We also need to make sure the algorithm is working with the device by changing things like the batch size and input resolution. Neuromorphic Computing: A New Way of Thinking While making models smaller is helpful there is another way of doing things called computing. This is a way of making models that work like the brain by using spikes to communicate. This can make the models use less energy. some devices, like the intel loihi 2 or IBM True North are specially made to run these kinds of models. When we use these devices we can make the models run faster and use much less energy. Making The Whole System Work Well Another question we also need to ask is not just on how the model works, but how the whole system works. We can also apply artificial intelligence to make the whole system consume less energy. We can use machine learning to predict how much energy the whole system will consume, and then the whole system will make changes to consume less. We can also apply the artificia intelligence work on our device of sending the data to the cloud, which makes the whole system consume less energy. What We Still Need to Work On even though we have done a lot, we still have more to do. one problem is there are so many different kinds of devices, so it is hard to make one model that will work well with all the devices. we also need to make it easier to train the models and also to make the models well integrated

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

with the devices. in the future we have to also make sure we want to work on the models and devices integrated together so we can apply artificial intelligence works better with the small devices [10 19].

Conclusion

Deploying artificial intelligence in IOT devices is a key shift from cloud-based processing to edge computing. Although edge ai can solve some of the systemic problems with latency, bandwidth, and data privacy, IOT devices have limited resources, and energy efficient algorithms are essential for Edge AI deployment.This work demonstrates that a equilibrium between model accuracy and energy consumption is necessary for Edge AI, and the equilibrium is achieved by three primary methods:

- **Model Compression:** Prune, quantize and distill complex AI models to millions-fold less computational power on edge.
- **Hardware-Algorithm Co-Design:** Take a holistic approach to solving the problem; Pair advanced algorithms with custom edge hardware (e.g. Google Coral, NVIDIA Jetson Nano) for greatest computational efficiency.
- **New Computing Paradigm:** Leverages neuromorphic-like computing, which uses brain-inspired spike based communication for power efficiency.

While the AI community has done a lot of work in scaling down and optimizing complex AI models for the edge, there is still a long way to go in solving the fragmentation and deployment complexity problems of most of the industry. We need to move closer to more software and hardware co-design. By further improving these energy efficient frameworks, we will unleash the full potential of the connected world and IoT networks and enable real-time, autonomous and power efficient decision at the source. in the end, although a lot of progress has been achieved in terms of slimming down and optimizing AI for the edge, the community still needs to deal with the issues of hardware fragmentation and deployment complexity. The future of Edge AI hinges upon further co-designing software and hardware. By further improving these energy-efficient frameworks, we can unleash the full potential of IoT networks and enable power-efficient and real-time, autonomous, and on-the-fly decision-making at the source.

References

[1]. Mas Nurul Achmadiah, Afaroj Ahamad, Chi-Chia Sun Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems IEEE INTERNET OF THINGS JOURNAL, VOL.12, NO.3, 1234–1245.
[2]. Owais Raza Energy-Efficient IoT Networks Using AI Driven Smart, Internet, Things. Vol. 1 No. 3 (2024) 203-212.
[3]. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, May 2015.
[4]. T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, “Machine learning meets computation and communication
[5]. D. Wen, X. Li, Q. Zeng, J. Ren, and K. Huang, “An overview of data-importance aware radio resource management for edge machine learning,” Journal of Communications and Information Networks, vol. 4, no. 4, pp. 1–14, Dec. 2019.
[6]. M. M. Amiri and D. Gndz, “Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air,” IEEE Trans. Signal Process., vol. 68, pp. 2155–2169, 2020.
[7]. A. K. Singh, K. R. Basireddy, A. Prakash, G. V. Merrett, and B. M. Al-Hashimi, “Collaborative adaptation for energy-efficient heterogeneous mobile SoCs,” IEEE Trans. on Compute., vol. 69, no. 2, pp. 185–197, 2020.
[8]. X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in Proc. IEEE/CVF Conf. Compute. Vision & Pattern Recognit. (CVPR), Salt Lake City, USA, Jun 18-23, 2018.
[9]. S. Kook, W.-Y. Shin, S.-L. Kim, and S.-W. Ko, “Joint data deepeningand-prefetching for energy-efficient edge learning,” presented at the IEEE International Conf. Commun. (ICC), Rome, Italy, 2023.
[10]. G. E. Hinton and R. R. Salakhutdinov,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0179
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1191 - 1196

"Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[11]. Siva Satya Sreedhar P, I. Carol, Meenakshi, Helina Rajini Suresh, M. Thillai Rani, J. Rejina Parvin Optimizing Energy-Efficient Task Offloading in Edge Computing: A Hybrid AIBased Approach International Journal of Computational and Experimental Science and Engineering Vol. 11-No.2 (2025) pp.1794-1802.

[12]. M. Kiran Myee, & M. Humera Khanam. (2025). GAN and Res Net Fusion a Novel Approach to Ophthalmic Image Analysis for Glaucoma. International Journal of Computational and Experimental Science and Engineering,11(1).https://doi.org/10.22399/ijcesen.683

[13]. Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact in The USA. International Journal of Applied Sciences and RadiationResearch,2(1).https://doi.org/10.22399/ijasrar.19

[14]. Yu F, Zhou Q (2022) Adaptive client selection in resource-constrained federated learning environments. J Cloud Compute Edge Intel 18(4):522–533

[15]. Johnson T, Lee M (2022) Quantization techniques for energy-efficient training in distributed IoT networks. IEEE Trans Neural Netw Learning Syst 33(8):2045–2058

[16]. A. Burrello et al., "Predicting Hard Disk Failures in Data Centers Using Temporal Convolutional Neural Networks," in Euro-Par 2020: Parallel Processing Workshops, ser. Lecture Notes in Computer Science, B. Balis et al., Eds. Cham: Springer International Publishing, 2021, pp. 277– 289

[17]. J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, and B. Martinez, "Edgevits: Competing light-weight cnns on mobile devices with vision transformers," in European Conference on Computer Vision, pp. 294–311, Springer, 2022.

[18]. H. V. Pham, T. G. Tran, C. D. Le, A. D. Le, and H. B. Vo, "Benchmarking jetson edge devices with an end-to-end video-based anomaly detection system," in Future of Information and Communication Conference, pp. 358–374, Springer, 2024.

[19]. S. Bouguezzi, H. B. Fredj, T. Belabed, C. Valderrama, H. Faiedh, and C. Souani, "An efficient fpga-based convolutional neural network for classification: Ad-mobilenet," Electronics, vol. 10, no. 18, p. 2272, 2021.