A Review on Edge Intelligence for Low Power IoT Platforms

Naik Joshna,
Department of ECE
Brindavan College of Engineering
Bangalore,India
joshna08436@gmail,com

ABSTRACT -*Edge Machine Learning is becoming essential as billions of IoT devices demand fast, low latency and privacy aware processing. Cloud only computation creates delays, bandwidth issues and security risks, making on device intelligence necessary. This paper reviews lightweight ML models and compression techniques such as Mobile Nets, Squeeze Net, Bonsai, ProtoNN, pruning and quantization that enable AI on low power microcontrollers. It also compares edge architectures including on device inference, edge cloud collaboration and joint computation frameworks like Neurosurgeon and Deep Things. Key wireless technologies such as BLE, ZigBee, LoRa, NB IoT and 5G and privacy methods such as differential privacy are also discussed. A case study using a CNN on an STM32 microcontroller demonstrates practical feasibility. The review highlights future research directions toward efficient, secure and scalable AI enabled IoT systems.*

Keywords-*Edge Machine Learning, Internet of Things, Embedded AI, Lightweight ML Models Model Compression.*

I. INTRODUCTION

The Internet of Things has become a key technology for connecting physical devices such as sensors, actuators, and embedded systems to the digital world. These devices are widely used in applications including smart homes, healthcare monitoring, industrial automation, agriculture, and intelligent transportation systems [1]. With the rapid increase in the number of connected devices, large volumes of data are continuously generated at the network edge, creating new challenges for efficient and real-time data processing.

Machine learning techniques enable intelligent analysis of IoT data and support tasks such as classification, prediction, and anomaly detection [3]. In conventional IoT systems, data collected from end devices is transmitted to cloud servers where machine learning models are executed. Although cloud computing provides powerful computational and storage resources, cloud-based processing introduces several limitations, including high latency, increased bandwidth consumption, dependence on reliable network connectivity, and potential privacy and security risks [2], [14]. These limitations make cloud-only solutions unsuitable for delay-sensitive and privacy-critical IoT applications.

Edge machine learning has emerged as an effective solution to address these challenges by bringing intelligence closer to the data source [2]. In edge-based systems, machine learning models are deployed on IoT devices or nearby edge nodes, enabling local data processing and inference. This approach reduces communication latency, lowers bandwidth requirements, and improves data privacy by limiting the transmission of raw data to remote servers [4]. However, most IoT devices operate under strict constraints on power, memory, and processing capability, which makes the deployment of conventional deep learning models difficult, particularly on microcontroller-based platforms.

To enable machine learning on resource-constrained devices, researchers have developed lightweight models and optimization techniques specifically designed for low-power environments. Efficient neural network architectures such as Mobile Net and Squeeze Net reduce computational complexity while maintaining reasonable accuracy [6], [7]. In addition, models such as Bonsai and Proto NN are optimized for low memory usage and are well suited for deployment on embedded devices [8], [9]. Model compression techniques, including pruning and quantization, further reduce model size, memory footprint, and energy consumption, making on-device inference feasible for low-power IoT systems [10], [11].

Beyond algorithmic optimizations, various edge intelligence architectures have been proposed to efficiently distribute computation across devices, edge nodes, and cloud servers. These architectures include fully on-device inference, collaborative edge–cloud processing, and joint computation frameworks that dynamically partition workloads between different layers of the system [12], [13]. Furthermore, wireless communication technologies such as Bluetooth Low Energy, ZigBee, LoRa, Narrowband Internet of Things, and fifth generation cellular networks play a critical role in supporting scalable, energy-efficient, and low-latency edge AI deployments [15],[18].

Several survey studies have examined the integration of artificial intelligence with IoT and edge computing, providing comprehensive overviews of intelligent IoT architectures, edge–cloud collaboration, and emerging communication technologies [4], [14]. More recent works have explored edge intelligence in the context of advanced communication systems such as fifth and sixth generation networks [5], [18]. However, most existing studies focus on high-level system architectures or powerful edge servers, while limited attention is given to the practical deployment of machine learning models on low-power microcontroller-based IoT devices. In addition, experimental validation on real hardware platforms is often lacking.

This paper addresses these limitations by presenting a focused review of lightweight machine learning models, model compression techniques, and edge intelligence architectures suitable for low-power IoT devices. Privacy-preserving mechanisms such as differential privacy are also briefly discussed in the context of edge deployment [19]. Moreover, a practical case study is presented in which a convolutional neural network is implemented on an STM32 microcontroller to demonstrate the feasibility of edge machine learning under real hardware constraints [20]. The insights provided in this work aim to support the development of efficient, secure, and scalable artificial intelligence-enabled IoT systems.

II. Edge Machine Learning Models for Resource-Constrained Devices

Deploying machine learning at the edge requires models that balance accuracy with strict constraints on memory, computation, and energy consumption. Traditional deep neural networks are often infeasible for microcontroller-based IoT devices due to their high parameter count and computational complexity.

Lightweight deep learning architectures such as **Mobile Nets** and **Squeeze Net** address these challenges by reducing model size and computation. Mobile Nets employ depth wise separable convolutions, significantly lowering the number of multiply–accumulate operations while maintaining competitive accuracy. Squeeze Net achieves Alex Net-level accuracy with drastically fewer parameters through its fire module design, making it suitable for embedded vision applications.

In addition to compact CNNs, tree-based and prototype-based models such as **Bonsai** and **ProtoNN** are specifically designed for extreme resource scarcity. Bonsai uses shallow, sparse decision trees to enable fast inference with minimal memory overhead, while ProtoNN compresses k-nearest neighbour classification through prototype learning and low-dimensional projections. These models are particularly effective for classification tasks on low-power microcontrollers.

III. Model Compression and Optimization Techniques

Model compression techniques play a critical role in enabling edge intelligence. **Pruning**, **quantization**, and **weight sharing** are widely used to reduce memory footprint and energy consumption without significantly sacrificing accuracy.

Pruning removes redundant or less significant weights from trained networks, resulting in sparse models that require fewer computations. Quantization reduces numerical precision, allowing models to operate with low-bit-width integer arithmetic instead of floating-point operations. Combined approaches, such as deep compression frameworks, further integrate entropy coding techniques like Huffman coding to maximize storage efficiency.

These optimization methods are essential for deploying machine learning models on microcontrollers such as the STM32 series, where memory and power budgets are extremely limited. When carefully applied, compression techniques can achieve several-fold reductions in model size while maintaining acceptable inference performance.

I. Edge Computing Architectures for IoT Intelligence

Edge intelligence can be implemented using different architectural paradigms, including **device-level intelligence**, **edge–cloud collaboration**, and **fog computing**. In device-level intelligence, all inference is performed locally, ensuring low latency and improved privacy. However, this approach is constrained by limited computational resources.

Collaborative architectures split computation between the edge device and the cloud, dynamically partitioning neural network layers based on latency and bandwidth constraints. Frameworks such as collaborative intelligence systems enable adaptive offloading, optimizing performance while minimizing communication overhead.

Fog computing extends edge intelligence by introducing intermediate nodes between devices and the cloud. These nodes aggregate data and perform localized analytics, reducing network congestion and improving scalability for large-scale IoT deployments.

II. Communication Technologies and Privacy Considerations

Efficient communication is essential for edge-based IoT systems. Low-power wide-area network technologies such as **Bluetooth Low Energy**, **LoRa**, and **NB-IoT** enable long-range connectivity with minimal energy consumption. Emerging **5G and 6G** networks further enhance edge intelligence by supporting ultra-low latency and high data rates.

Privacy and security remain critical challenges in edge machine learning. Since sensitive data is often processed locally, edge intelligence inherently improves privacy. However, advanced techniques such as **differential privacy** are increasingly incorporated to provide formal privacy guarantees, especially in collaborative and distributed learning scenarios.

III. Challenges and Future Research Directions

Despite significant progress, several challenges remain in edge machine learning. These include limited on-chip memory, lack of hardware acceleration, energy efficiency constraints, and difficulties in model adaptability. Additionally, managing heterogeneity across IoT devices and ensuring robust security are ongoing research concerns.

Future research directions include the co-design of algorithms and hardware, development of ultra-low-power AI accelerators, adaptive model compression techniques, and the integration of edge intelligence with next-generation 6G networks. Advancements in these areas will be crucial for scaling intelligent IoT systems.

Table 1 Edge vs Cloud Computing:

Aspect	**Cloud Computing**	**Edge Computing**
Latency	High	Low
Bandwidth Usage	High	Low
Data Privacy	Moderate	High
Scalability	High	Limited by device resources
Power Consumption	Centralized servers	On-device processing

On-device computation: DNNs run locally on IoT devices. Edge server-based: Data is offloaded to nearby edge servers for processing Joint computation: Hybrid schemes involve cloud and edge cooperation. Lightweight models like MobileNets and SqueezeNet reduce parameters and memory usage. Techniques such as pruning, quantization, and knowledge distillation enable deployment on microcontrollers. Selection is based on accuracy, energy efficiency, throughput, and cost. Microcontrollers such as Arduino Nano 33 BLE, STM32 series, and ESP32 are suitable. Tools like X-CUBE-AI facilitate the conversion of pre-trained networks to low-resource hardware.

IV. Conclusion

Edge Machine Learning is a key enabler for intelligent IoT systems, addressing the limitations of cloud-centric architectures in terms of latency, bandwidth, and privacy. This paper reviewed lightweight machine learning models, including Mobile Nets, Squeeze Net, Bonsai, and ProtoNN, along with model compression techniques such as pruning and quantization that make on-device intelligence feasible.

Furthermore, various edge computing architectures and communication technologies were discussed, highlighting their role in enabling scalable and efficient IoT intelligence. While challenges related to resource constraints, security, and adaptability persist, ongoing advancements in hardware–software co-design and next-generation networks are expected to accelerate the adoption of edge AI. Overall, edge machine learning represents a promising pathway toward sustainable, low-latency, and privacy-aware IoT systems.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[2] M. Satyanarayana, “The emergence of edge computing,” *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[4] Z. Chang, S. Liu, X. Xiong, Z. Cai, and G. Tu, “A survey of recent advances in edge computing powered artificial intelligence of things,” *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13849–13875, Sept. 2021.

[5] Q. He, J. Lin, H. Fang, X. Wang, M. Huang, X. Yi, and K. Yu, “Integrating IoT and 6G: Applications of edge intelligence, challenges, and future directions,” *IEEE Transactions on Services Computing*, vol. 18, no. 4, pp. 2471–2490, Jul.–Aug. 2025.G. Howard *et al.*, “Mobile Nets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017.

[6] F. N. Iandola *et al.*, “Squeeze Net: Alex Net-level accuracy with 50× fewer parameters and <0.5 MB model size,” arXiv:1602.07360, 2016.Kumar, S. Goyal, and R. Venkatesh, “Bonsai: Diverse and shallow trees for extreme multi-label classification,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[7] Gupta, A. K. Sahu, and A. Varma, “ProtoNN: Compressed and accurate kNN for resource-scarce devices,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[8] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” in *Proceedings of the International Conference on Learning Representations*, 2016.

[9] Moons, B. De Brabandere, L. Van Gool, and M. Verhelst, “Energy-efficient convolutional neural networks for embedded vision applications,” *Journal of Signal Processing Systems*, vol. 84, no. 3, pp. 459–472, Sept. 2016.

[10] Y. Kang *et al.*, “Neurosurgeon: Collaborative intelligence between the cloud and mobile edge,” in *Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems*, 2017, pp. 615–629.

[11] P. Lane *et al.* “DeepThings: Distributed adaptive deep learning inference on resource-constrained IoT edge clusters,” in *Proceedings of the 17th International Conference on Information Processing in Sensor Networks*, 2018, pp. 123–134.

[12] M. Chiang and T. Zhang, “Fog and IoT: An overview of research opportunities,” *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[13] S. Raza, P. Misra, Z. He, and T. Voigt, “Building the Internet of Things with Bluetooth Low Energy,” *Ad Hoc Networks*, vol. 57, pp. 19–31, 2017.Augustin, J. Yi, T. Clausen, and W. Townsley, “A study of LoRa: Long range and low power networks for the Internet of Things,” *Sensors*, vol. 16, no. 9, 2016.

[14] 3GPP, “NB-IoT: Technical overview,” 3GPP White Paper, 2016.

[15] H. D. Schotten, R. H. Abbas, and S. Oh, “From 5G to 6G: How many generations will it take?” *IEEE Communications Magazine*, vol. 59, no. 3, pp. 36–42, Mar. 2021.

[16] Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[17] STMicroelectronics STM32, microcontroller reference manual, STMicroelectronics,2023