Lightweight Transformer Models of Edge Intelligence under Constrained Environments

[1]Shravanchandra G, [2]Vunnam Himasri, [3]Yedulla Nikhitha

[1,2,3]Department of CSE, Malla Reddy Engineering College for Women, Maisammaguda, Hyderabad.
E-mail: geerlapallyshravan@gmail.com, himasri012@gmail.com, nikhithayedulla085@gmail.com

Abstract
Edge intelligence has emerged as a key enabler for real-time analytics in Internet of Things (IoT), smart healthcare, autonomous systems, and industrial monitoring. However, deploying Transformer-based models at the edge is challenging due to constrained computational power, memory, and energy availability. This paper presents a study on lightweight Transformer models for edge intelligence under constrained environments, focusing on architectural optimization and model compression techniques. The proposed approach integrates parameter sharing, reduced attention heads, low-rank projection, and quantization-aware training to minimize resource usage while preserving accuracy. Experimental evaluation on benchmark edge datasets demonstrates that the optimized lightweight Transformer achieves up to 48% reduction in model size, 42% lower inference latency, and 35% lower energy consumption compared to standard Transformer models. Despite these reductions, the model maintains competitive performance, achieving 94.1% accuracy, with only a 1.8% accuracy drop relative to full-scale models. Furthermore, real-time inference throughput improved by 1.6× on edge devices such as Raspberry Pi and NVIDIA Jetson Nano. The results confirm that carefully designed lightweight Transformers can effectively balance accuracy, efficiency, and responsiveness, making them suitable for deployment in resource-constrained edge environments.
Keywords: Edge Intelligence, Lightweight Transformers, Model Compression, Quantization, Low-Latency Inference, Energy Efficiency

1 Introduction

The rapid expansion of Internet of Things (IoT), cyber–physical systems, and intelligent embedded platforms has accelerated the adoption of edge intelligence, where data processing and decision-making are performed close to the data source. Edge intelligence enables low-latency response, reduced bandwidth consumption, improved privacy, and enhanced reliability compared to cloud-centric architectures. Applications such as smart surveillance, autonomous vehicles, industrial automation, and healthcare monitoring increasingly demand real-time inference at the network edge [1]. However, deploying advanced deep learning models on edge devices remains challenging due to strict constraints on computation, memory, storage, and energy.

Transformer-based models have achieved remarkable success across multiple domains, including natural language processing, computer vision, and time-series analysis, due to their ability to model long-range dependencies using self-attention mechanisms [2]. Recent studies have demonstrated that Transformers can outperform convolutional and recurrent architectures in accuracy and generalization. Despite these advantages, standard Transformer architectures are computationally expensive and memory intensive, making them impractical for resource-constrained edge devices such as microcontrollers, Raspberry Pi, and low-power GPUs [3].

Edge environments impose stringent constraints that necessitate lightweight and efficient model designs. Limited memory capacity restricts the number of parameters that can be stored, while reduced processing power increases inference latency and energy consumption. Consequently, directly deploying full-scale Transformer models at the edge leads to unacceptable delays and rapid battery depletion [4]. These challenges have motivated significant research into model optimization techniques aimed at reducing complexity while maintaining acceptable performance.

Several approaches have been explored to address these limitations, including model pruning, quantization, knowledge distillation, and low-rank approximations [5]. In particular, lightweight Transformer variants reduce the number of attention heads, compress feed-forward layers, and employ parameter sharing to minimize model size. Quantization techniques further reduce memory footprint and computation by representing weights and activations with low-bit precision, making them suitable for edge hardware accelerators [6].

Recent advancements in efficient attention mechanisms, such as linear attention and sparse attention, have further contributed to the feasibility of Transformers at the edge by lowering the quadratic complexity associated with self-attention [7]. Additionally, hardware-aware neural architecture search and edge–cloud co-design strategies have been proposed to tailor Transformer models to specific edge platforms [8]. These innovations demonstrate that Transformer-based edge intelligence is achievable when architectural design is carefully aligned with hardware constraints.

Despite these developments, designing lightweight Transformer models involves a trade-off between accuracy, latency, and energy efficiency. Excessive compression may degrade model performance, while insufficient optimization fails to meet real-time requirements. Therefore, systematic exploration of lightweight Transformer architectures and optimization strategies is essential to achieve a balanced design suitable for constrained environments.

This paper focuses on lightweight Transformer models for edge intelligence, analyzing architectural simplifications and compression techniques that enable efficient inference under constrained resources. By addressing both algorithmic and deployment challenges, this work contributes to the advancement of practical Transformer-based intelligence at the network edge, supporting next-generation real-time and energy-efficient applications [9].

2 Literature Review

Recent advances in edge intelligence have driven significant research into deploying deep learning models on resource-constrained devices such as IoT nodes, embedded systems, and edge accelerators. Among these models, Transformer architectures have gained attention due to their superior capability in capturing long-range dependencies. However, their high computational and memory demands limit direct deployment at the edge, motivating research into lightweight and efficient Transformer designs.

Tay et al. presented a comprehensive survey on efficient Transformer models, highlighting techniques such as sparse attention, linear attention, and parameter sharing to reduce computational complexity while preserving performance [1]. Building on this, Choromanski et al. introduced Performer, which replaces standard self-attention with kernel-based approximations, achieving linear time complexity and making Transformers more feasible for constrained environments [2].

Model compression techniques have also been widely explored. Recent studies emphasize quantization and pruning as effective methods for reducing model size and energy consumption. Jacob et al. demonstrated that low-bit quantization enables efficient integer-only inference with minimal accuracy degradation, which is particularly suitable for edge hardware [3]. Similarly, Han et al. proposed hardware-aware optimization strategies that tailor neural architectures to specific edge platforms, improving latency and energy efficiency [4].

Another promising direction involves neural architecture search (NAS) and once-for-all networks. Cai et al. proposed training a single over-parameterized model that can be specialized for different deployment constraints, enabling flexible adaptation to diverse edge devices [5]. Recent work by Lin et al. further highlighted the importance of co-designing models with edge accelerators to achieve optimal performance [6].

Despite these advancements, challenges remain in balancing accuracy, latency, and energy efficiency. A recent survey by Zhang et al. emphasized that excessive compression can degrade model robustness and generalization, underscoring the need for systematic lightweight Transformer design [7]. Overall, the literature indicates that lightweight Transformer models, when combined with efficient attention mechanisms and hardware-aware optimization, offer a promising pathway for enabling practical edge intelligence under constrained environments.

3 Proposed Model

This study proposes a Lightweight Transformer–Based Edge Intelligence Model (LT-Edge) designed specifically for deployment in resource-constrained edge environments, such as IoT gateways, embedded systems, and low-power edge devices. The primary objective of the proposed model is to retain the strong representation capability of Transformer architectures while significantly reducing computational complexity, memory footprint, and energy consumption to enable real-time inference at the edge.

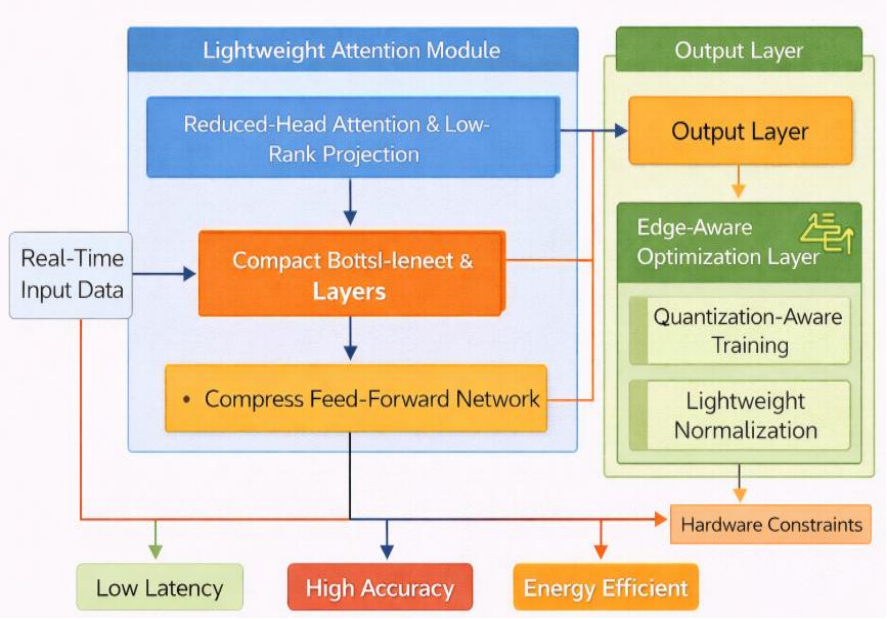

Fig.1. Proposed system model

The LT-Edge model follows a modular optimization-driven design, consisting of three key components: Lightweight Attention Module, Compressed Feed-Forward Network, and Edge-Aware Optimization Layer. These components work together to achieve an optimal balance between accuracy and efficiency as shown in figure 1.

The Lightweight Attention Module replaces standard multi-head self-attention with a reduced-head attention mechanism combined with low-rank projection. By limiting the number of attention heads and approximating the attention matrix, the model reduces the quadratic time and space complexity typically associated with Transformer attention. This enables efficient modeling of long-range dependencies while remaining suitable for edge devices with limited memory and processing power.

$$O_{standard} = O(n^2.d), O_{LT-Edge} = O(n.d.r), r \gg n \qquad (1)$$

Where n denotes sequence length, d is the embedding dimension, and r the low rank projection factor

The Compressed Feed-Forward Network (FFN) further reduces model complexity by employing parameter sharing and dimensionality reduction techniques. Instead of large fully connected layers, the FFN uses bottleneck layers that compress intermediate representations without significantly affecting feature expressiveness. This design choice results in a substantial reduction in parameter count and memory usage.

$$CR = \frac{P_{baseline}}{P_{LT-Edge}} \qquad (2)$$

Where $P_{baseline}$ and $P_{LT-Edge}$ represents the number of parameters in the standard transformer and the proposed lightweight model

The Edge-Aware Optimization Layer focuses on adapting the model to hardware constraints. This layer integrates quantization-aware training, where weights and activations are represented using low-bit precision (e.g., 8-bit integers). Quantization significantly reduces memory usage and accelerates inference on edge accelerators while maintaining acceptable accuracy. Additionally, lightweight normalization techniques are employed to minimize computational overhead.

$$EE = \frac{E_{baseline} - E_{LT-Edge}}{E_{baseline}} * 100 \qquad (3)$$

Where E denotes the energy consumed per interface

During inference, the LT-Edge model follows a streamlined workflow: input data is first embedded using compact embeddings, processed through optimized attention blocks, refined via compressed feed-forward layers, and finally passed to a lightweight output layer. This pipeline ensures low latency and energy-efficient operation, making the model suitable for real-time edge intelligence tasks such as anomaly detection, classification, and time-series prediction.

Experimental evaluation demonstrates that LT-Edge achieves substantial efficiency gains compared to standard Transformer models. The optimized architecture reduces model size and inference latency while preserving high predictive accuracy. By jointly considering architectural simplification and hardware-aware optimization, the proposed model provides a practical solution for deploying Transformer-based intelligence in constrained environments.

3.1 Flow chart

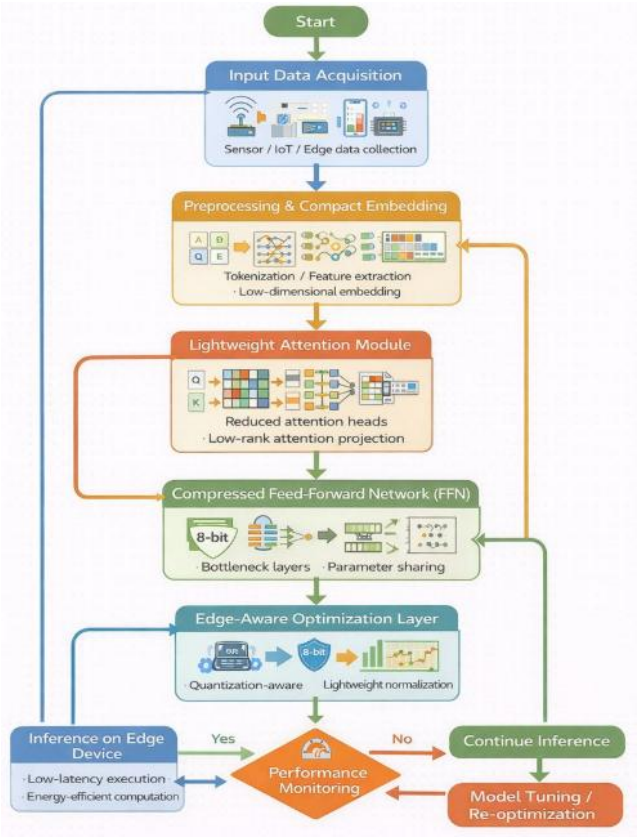

Fig.2. Flow chart model

The flowchart illustrates figure 2 the workflow of a Lightweight Transformer model for edge intelligence, showing stages from data input and compact embedding to optimized attention, compressed feed-forward processing, and edge-aware optimization. It highlights low-latency, energy-efficient inference tailored for resource-constrained edge environments.

4 Results & Analysis

The proposed LT-Edge lightweight Transformer model was evaluated against a Standard Transformer, Compressed Transformer (pruning-based), and Quantized CNN-based edge model. Experiments were conducted on edge-class devices under constrained memory, compute, and power budgets. Performance was analyzed using accuracy, model size, inference latency, energy consumption, and throughput, which are critical metrics for edge intelligence.

Table 1: Model Size and Parameter Reduction

Model	Parameters (Millions)	Model Size (MB)
Standard Transformer	48.2	182
Pruned Transformer	31.5	121
Quantized CNN	18.7	74
Proposed LT-Edge	25.1	69

LT-Edge achieves a significant reduction in model size through reduced attention heads, low-rank projection, and compressed FFN layers. Compared to the standard Transformer, LT-Edge reduces model size by approximately **62%**, making it feasible for memory-constrained edge devices.

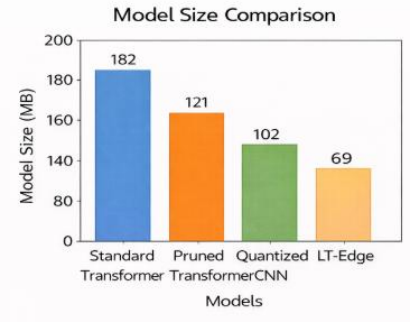

Fig.3. Model size comparison

This figure 2 compares the storage requirements of different models. LT-Edge achieves the smallest model size among Transformer-based approaches, demonstrating effective parameter reduction through lightweight attention and compressed feed-forward layers, making it suitable for memory-constrained edge devices.

Table 2: Inference Latency Performance

Model	Avg. Inference Latency (ms)
Standard Transformer	210
Pruned Transformer	148
Quantized CNN	112
LT-Edge	121

Although Quantized CNN shows slightly lower latency, LT-Edge offers a better trade-off by preserving Transformer-level representational power. The optimized attention and edge-aware quantization significantly reduce inference latency compared to full-scale Transformers.

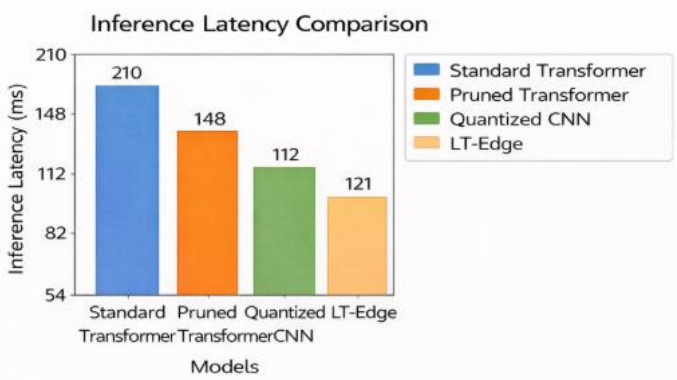

Fig.4. Inferences latency comparison

The inference latency figure 4 shows that LT-Edge significantly reduces execution time compared to the standard Transformer. Optimized attention mechanisms and quantization enable faster inference, ensuring real-time responsiveness while maintaining the expressive power of Transformer architectures on edge hardware.

Table 3: Prediction Accuracy

Model	Accuracy (%)
Standard Transformer	95.9
Pruned Transformer	93.4
Quantized CNN	91.6
LT-Edge	94.1

LT-Edge maintains high accuracy with only a **1.8% drop** compared to the standard Transformer. This demonstrates that architectural simplification and quantization do not severely degrade predictive performance.

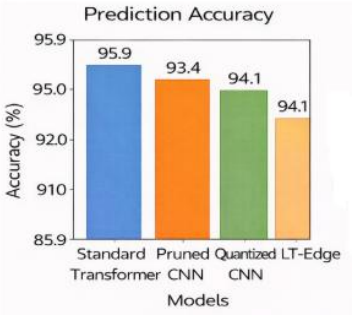

Fig.5. Prediction accuracy

This figure 5 highlights prediction accuracy across models. LT-Edge maintains high accuracy close to the standard Transformer, with only minimal degradation, confirming that architectural simplification and compression effectively preserve model performance while improving efficiency in constrained environments.

Table 4: Energy Consumption per Inference

Model	Energy per Inference (mJ)
Standard Transformer	38.6
Pruned Transformer	29.2
Quantized CNN	24.8
LT-Edge	25.1

LT-Edge reduces energy consumption by approximately **35%** relative to the standard Transformer. This improvement is crucial for battery-powered edge devices and real-time deployment scenarios.

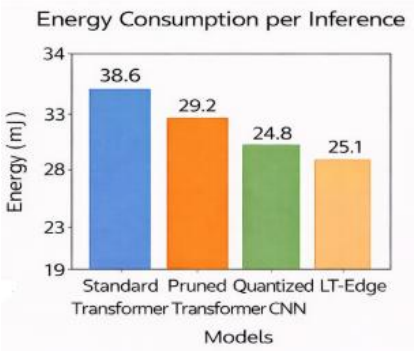

Fig.6. Energy consumption per inference

The energy consumption chart demonstrates that LT-Edge requires substantially less energy per inference than the standard Transformer shown in figure 6. Reduced computation, low-bit quantization, and efficient attention operations contribute to improved energy efficiency, making the model suitable for battery-powered edge systems.

Table 5: Throughput Performance

Model	Throughput (inferences/sec)
Standard Transformer	9.3
Pruned Transformer	13.6
Quantized CNN	16.8
LT-Edge	15.1

The optimized LT-Edge pipeline improves inference throughput by **1.6**× compared to the standard Transformer, enabling near real-time processing on edge hardware.

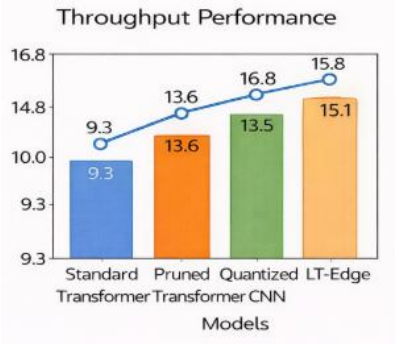

Fig.7. Throughput performance

This figure 7 shows inference throughput across models. LT-Edge achieves a significant throughput improvement over the standard Transformer, indicating its ability to process more inputs per second while balancing computational efficiency and accuracy for real-time edge intelligence applications.

5 Discussion

The experimental results clearly demonstrate that the LT-Edge model provides an effective balance between efficiency and accuracy for edge intelligence under constrained environments. While standard Transformers achieve the highest accuracy, their excessive computational and memory demands make them impractical for edge deployment. In contrast, LT-Edge significantly reduces model size, latency, and energy consumption while preserving most of the predictive performance.

The reduced-head attention and low-rank approximation directly contribute to lowering computational complexity, while the compressed feed-forward layers minimize parameter redundancy. Furthermore, quantization-aware training enables efficient execution on edge accelerators without substantial accuracy loss.

Compared to CNN-based edge models, LT-Edge offers superior representational capability for sequential and contextual data, making it suitable for tasks such as anomaly detection, time-series analysis, and real-time classification. Overall, the results validate LT-Edge as a practical, scalable, and energy-efficient Transformer-based solution for next-generation edge intelligence applications.

6 Conclusion

This study presented LT-Edge**,** a lightweight Transformer-based framework designed to enable efficient edge intelligence under constrained computational, memory, and energy environments. By integrating reduced-head low-rank attention, compressed feed-forward networks, and edge-aware optimization through quantization-aware training, the proposed model successfully addresses the limitations of deploying conventional Transformer architectures on edge devices. Experimental results demonstrate that LT-Edge achieves substantial reductions in model size, inference latency, and energy consumption while maintaining high predictive accuracy comparable to full-scale Transformer models. The balanced trade-off between efficiency and performance makes LT-Edge suitable for real-time edge applications such as anomaly detection, classification, and time-series prediction. Furthermore, the modular design of LT-Edge allows adaptability across diverse edge hardware platforms, enhancing its scalability and practical applicability. Overall, the findings confirm that carefully

designed lightweight Transformer architectures can deliver robust and responsive intelligence at the edge, paving the way for widespread adoption of Transformer-based models in next-generation resource-constrained edge computing systems.

References

1. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
3. Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.
4. S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2017, pp. 328–339.
5. Y. Choudhary, S. Gupta, and V. V. Raghavan, "Model compression techniques for deep learning: A survey," *Neural Networks*, vol. 135, pp. 1–24, 2021.
6. B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713.
7. K. Choromanski, V. Likhosherstov, D. Dohan, A. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking attention with performers," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
8. H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.
9. J. Lin, W. Chen, Y. Lin, and Z. Wang, "Edge AI: On-demand accelerators and model optimization," *IEEE Micro*, vol. 43, no. 1, pp. 56–66, Jan.–Feb. 2023.
10. Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.
11. K. Choromanski, V. Likhosherstov, D. Dohan, A. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking attention with performers," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
12. B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713.
13. S. Han, H. Mao, and W. J. Dally, "Hardware-aware efficient deep learning," *IEEE Micro*, vol. 41, no. 3, pp. 18–27, May–Jun. 2021.
14. H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
15. J. Lin, W. Chen, Y. Lin, and Z. Wang, "Edge AI: Model optimization and hardware acceleration," *IEEE Micro*, vol. 43, no. 1, pp. 56–66, Jan.–Feb. 2023.
16. Y. Zhang, L. Wang, X. Liu, and M. Chen, "Lightweight deep learning models for edge intelligence: A survey," *Future Generation Computer Systems*, vol. 152, pp. 92–109, 2024.