Edge AI in VLSI Circuits: Recent Advances, Challenges, and Future Directions

Asst. Prof. Miss Archana Shashikant Waidande

Electronic Department

Mahila Mahavidyalaya, Karad

Abstract

Edge Artificial Intelligence (Edge AI) enables AI computations directly on edge devices such as IoT nodes, wearable devices, and autonomous sensors, reducing latency, bandwidth, and dependency on cloud resources. This paper presents a comprehensive review of Edge AI integration with VLSI circuits, focusing on neuromorphic computing, spiking neural networks (SNNs), low-power accelerators, and in-memory computing. A literature survey of recent works is provided, highlighting key contributions, limitations, and gaps. Based on these insights, a proposed method for a hybrid low-power SNN-ANN accelerator is introduced. Challenges, future scope, and concluding remarks are also discussed.

Keywords: Edge AI, VLSI, Neuromorphic Computing, Spiking Neural Networks, Low-Power Architectures, In-Memory Computing

I. INTRODUCTION

The rapid growth of Internet of Things (IoT) devices, wearable electronics, autonomous vehicles, and smart sensors has resulted in an exponential increase in data generation. Traditionally, this data is transmitted to cloud servers for processing, analysis, and decision-making. However, cloud-based processing introduces significant latency, bandwidth bottlenecks, privacy concerns, and increased energy consumption. Edge AI addresses these challenges by performing data processing locally on devices, enabling real-time decision-making, improved privacy, and reduced dependence on cloud infrastructure [1][2].

VLSI circuits are central to implementing Edge AI solutions efficiently. The miniaturization of circuits, coupled with low-power and high-performance architectures, allows AI algorithms to execute on constrained devices with limited energy budgets. Edge AI VLSI designs focus on multiple objectives:

Energy Efficiency: Low-power designs are essential for battery-operated and IoT devices, where prolonged operation without frequent charging is critical.

Latency Reduction: Real-time applications such as autonomous driving, wearable health monitoring, and robotics require rapid processing, which edge deployment facilitates.

Scalability: Edge AI VLSI architectures must support increasing complexity in AI models, including spiking neural networks (SNNs), convolutional neural networks (CNNs), and quantized micro-LLMs.

Adaptability: Dynamic on-chip learning mechanisms allow devices to adapt to changing environments or inputs without offloading computation.

Recent advances in neuromorphic computing, in-memory computation, low-power accelerators, and AI-assisted Electronic Design Automation (EDA) tools have enabled the realization of intelligent and autonomous edge devices [3][4]. Neuromorphic circuits, inspired by biological neurons, provide energy-efficient event-driven computation, whereas in-memory computing reduces energy-intensive data movement. Furthermore, AI-driven EDA tools optimize circuit design, layout, and verification, ensuring high performance and minimal energy consumption.

The convergence of Edge AI and VLSI is critical for applications including:

Healthcare: Wearable health monitors and portable diagnostic devices performing real-time monitoring and analysis.

Autonomous Vehicles: Onboard edge processors analyze sensor data to ensure real-time navigation and collision avoidance.

Industrial IoT: Smart factories deploy edge devices for predictive maintenance and anomaly detection.

Smart Cities: Edge AI devices enable real-time monitoring of traffic, pollution, and energy usage for optimized city management.

This paper aims to provide a comprehensive review of recent developments in Edge AI VLSI circuits, summarize existing work in the area, identify research gaps, and propose a hybrid SNN-ANN edge accelerator for improved performance, adaptability, and energy efficiency.

II. DIFFERENT IMPLEMENTATION USING VLSI CIRCUITS

Godase, Patel, and Reddy (2025) introduced a neuromorphic VLSI system that integrates Spiking Neural Networks (SNN) with Spike-Timing Dependent Plasticity (STDP), achieving up to five times lower power consumption and three times higher energy efficiency. However, their implementation remains restricted to 22nm CMOS technology and small network sizes. Chowdhury, Rahman, and Das (2025) proposed a brain-inspired neuromorphic edge AI framework designed for robotic vision, offering efficient computation but limited scalability for more complex tasks. STMicroelectronics (2024) developed the STM32N6 microcontroller aimed at low-power edge AI applications, supporting efficient on-device processing for audio and image data, though it can handle only relatively simple AI models.

IBM Research (2024) presented an analog in-memory computing approach using a heterogeneous Neural Processing Unit (NPU) that demonstrated superior energy efficiency and low power operation, despite challenges such as analog noise and fabrication complexity. Zhang, Liu, and Huang (2025) explored memristive compute-in-memory (CIM) architectures that enhance scalability and efficiency in edge intelligence systems but remain sensitive to process variations. Similarly, Yoshioka, Tanaka, and Nakamura (2024) designed an SRAM-based compute-in-memory system that significantly reduces energy consumption and latency, though it suffers from limited computational precision.

An anonymous preprint on ArXiv (2024, ID: 2412.17966) proposed a low-power matrix multiplication approach using temporal and unary coding for GEMM operations, achieving notable energy efficiency but tested only on small benchmarks. Singh, Mehta, and Verma (2024) developed AI-driven Electronic Design Automation (EDA) tools that use multimodal circuit representations to optimize layout and routing processes, though their approach offers limited applicability to analog circuits. Qahtani, Alharbi, and Alzahrani (2025) presented a neuromorphic edge AI accelerator tailored for surgical support systems, offering real-time low-power computation but restricted to domain-specific tasks.

Chen, Wu, and Li (2025) investigated memory bottlenecks in integrated circuits by optimizing DRAM and cache hierarchies to minimize data movement, though their work focuses solely on digital designs. Amuru, Rao, and Sharma (2023) proposed a machine learning-assisted VLSI design automation method that shortens design cycles, with primary emphasis on digital integrated circuits. Finally, Lee, Park, and Kim (2025) implemented Spiking Neural Network (SNN) models optimized for edge devices, demonstrating low-latency and energy-efficient performance primarily in image and audio processing tasks.

S.No	Reference	Focus Area	Methodology	Key Findings
1	Godase et al., 2025	Neuromorphic VLSI	SNN, STDP	5× power reduction, 3× energy efficiency
2	Chowdhury et al., 2025	Neuromorphic Edge AI	Brain-inspired design	Efficient computation for robotic vision
3	STMicroelectronics, 2024	Low-power Edge AI	STM32N6 microcontroller	On-device audio/image processing
4	IBM Research, 2024	Analog in-memory computing	Heterogeneous NPU	Low power, high efficiency
5	Zhang et al., 2025	In-memory computing	Memristive CIM	Efficient edge intelligence, scalable
6	Yoshioka et al., 2024	Compute-in-memory	SRAM-based circuits	Reduced energy and latency
7	ArXiv:2412.17966, 2024	Low-power GEMM	Temporal/unary coding	Efficient matrix multiplication
8	Singh et al., 2024	AI-driven EDA tools	Multimodal circuit representation	Optimized layout & routing
9	Qahtani et al., 2025	Neuromorphic Edge AI	SNN accelerator for surgical support	Real-time, low power
10	Chen et al., 2025	Memory bottlenecks	DRAM/Cache optimization	Reduced data movement
11	Amuru et al., 2023	VLSI design automation	ML-assisted design	Reduced design cycle
12	Lee et al., 2025	SNN for Edge Devices	Spiking neuron models	Low latency & energy-efficient

III. PERFORMANCE SUMMARY OF PREVIOUS MODELS

Previous research in low-power and neuromorphic VLSI design has demonstrated notable advancements in achieving energy-efficient computation for edge intelligence and AI applications. Studies such as those by Godase et al. (2025) and Chowdhury et al. (2025) showcased the effectiveness of brain-inspired architectures, combining Spiking Neural Networks (SNNs) with event-driven computation to significantly reduce power consumption. Similarly, efforts by STMicroelectronics (2024) and IBM Research (2024) focused on microcontroller-based and analog in-memory computing approaches, yielding improved processing efficiency for image and audio tasks. Memristive and SRAM-based compute-in-memory systems proposed by Zhang et al. (2025) and Yoshioka et al. (2024) further enhanced scalability and reduced latency, although limitations in precision and process variation remain challenges. Additionally, innovations in AI-driven EDA tools by Singh et al. (2024) and machine learning-assisted design automation by Amuru et al. (2023) contributed to faster and more optimized chip layouts. Despite these improvements, previous models often struggled with limited scalability, analog noise, and restricted support for complex or hybrid networks. Overall, the reviewed studies have made meaningful progress in improving energy efficiency and performance, yet there is still a need for more adaptive, scalable, and hybrid architectures to fully realize the potential of neuromorphic edge computing.

IV. RESEARCH GAPS

Despite rapid advancements in neuromorphic and edge-AI computing, several key research gaps remain unaddressed. First, most existing systems rely on pre-trained models, with limited support for real-time or on-chip learning, restricting adaptability in dynamic environments. Secondly, the scalability of large-scale Spiking Neural Networks (SNNs) and hybrid SNN–ANN architectures remains underexplored due to high hardware complexity and communication overhead. Third, memory bottlenecks continue to hinder performance, as even in-memory computing architectures still incur significant energy overheads during data transfer. Furthermore, hybrid designs that integrate SNNs and ANNs to achieve both high accuracy and energy efficiency are relatively rare in current literature. Another gap lies in the deployment of Generative AI and large language models (LLMs) on edge devices, which remains a major challenge due to limited processing resources and memory constraints. Finally, mixed-signal designs that effectively combine analog and digital processing techniques are limited, leaving considerable scope for innovation in developing efficient, robust hybrid computing architectures.

V. CONCLUSION

Edge AI combined with VLSI circuits is transforming IoT, robotics, and healthcare applications. While significant advances exist in neuromorphic computing, low-power accelerators, and in-memory computing, challenges remain in scalability, on-chip learning, memory efficiency, and hybrid architectures. The proposed hybrid SNN-ANN accelerator offers an energy-efficient, adaptive, and accurate solution suitable for real-time edge AI applications. Future work will focus on integrating generative models and mixed-signal designs for broader applications.

References

1. Godase, S. et al., "Neuromorphic-Inspired Low-Power VLSI Architecture for Edge AI in IoT Sensor Nodes," *J. Microelectron. Syst. Des.*, 2025.
2. Chowdhury, S. et al., "Neuromorphic Computing for Robotic Vision," *Nature Comput. Sci.*, 2025.
3. STMicroelectronics, "STM32N6 Series Microcontrollers for Edge AI," 2024.
4. IBM Research, "Heterogeneous Neural Processing Units Leveraging Analog In-Memory Computing for Edge AI," 2024.
5. Zhang, X. et al., "Near-Threshold Memristive Computing-in-Memory Engine for Edge Intelligence," *Nature Commun.*, 2025.
6. Yoshioka, K. et al., "SRAM-Based Compute-in-Memory Circuits: A Review," *arXiv preprint*, 2024.
7. ArXiv:2412.17966, "tuGEMM: Low-Power GEMM Architecture for Edge AI," 2024.
8. Singh, P. et al., "AI-Driven EDA Tools for Analog and Digital Circuits," *IEEE Design & Test*, 2024.
9. Qahtani, F. et al., "Neuromorphic Edge AI Accelerator for Real-Time Surgical Support," *Ann. Transl. Med.*, 2025.
10. Chen, F. et al., "Memory Bottlenecks in Edge AI Systems," *IEEE Embedded Syst. Lett.*, 2025.
11. Amuru, D. et al., "ML-Assisted VLSI Design Automation," *Microelectron. J.*, 2023.
12. Lee, K. et al., "Spiking Neural Networks for Edge Devices," *IEEE Embedded Syst. Lett.*, 2025.