ARID ZONE JOURNAL OF ENGINEERING, TECHNOLOGY & ENVIRONMENT
AZOJETE June 2026. Vol.22(2):462-467
Published by the Faculty of Engineering, University of Maiduguri, Maiduguri, Nigeria.
Print ISSN: 1596-2490, Electronic ISSN: 2545-5818
https://doi.org/10.63958/AZOJETE/2026/22/02/023
www.azojete.com.ng

REVIEW ARTICLE

FPGA-BASED DEEP LEARNING PROCESSOR: A REVIEW

S. Hussaini*, C. U. Ngene, P. Y. Dibal, S. J. Bassi

Department of Computer engineering, University of Maiduguri
Borno state, Nigeria.
**Corresponding author's email: salehhussaini05@gmail.com*

ARTICLE INFORMATION

Received: 30th March 2026
Revised: 17th May 2026
Accepted: 19th May 2026

Keywords:
FPGA
Deep learning
Hardware acceleration
Convolutional neural networks
Reconfigurable computing

ABSTRACT

The rapid adoption of deep learning across diverse application areas has intensified the demand for specialized hardware capable of efficiently executing computationally intensive neural networks at scale. Field-Programmable Gate Arrays (FPGAs) have emerged as a prominent solution, offering a unique combination of high parallelism, low latency, and energy efficiency. Despite these advantages, research remains fragmented, with challenges in standardizing architectures, optimizing design flows, and supporting emerging neural network models, including transformers and generative architectures. This review systematically examined recent developments in FPGA-based deep learning processors, focusing on architectural design strategies, optimization methodologies, and deployment across both cloud and edge environments. A structured survey approach was employed, analyzing experimental evaluations and comparative studies of FPGA accelerators for convolutional, recurrent, and next-generation neural networks. The analysis demonstrated that FPGA-based designs consistently achieved superior energy efficiency compared to GPUs and provided scalable solutions for edge inference, though limitations persist in programmability and toolchain maturity. The findings highlighted FPGAs' potential as critical enablers for next-generation intelligent systems while emphasizing the need for higher-level abstractions, automated design-space exploration, and seamless integration with evolving machine learning frameworks.

1.0 Introduction

The rapid evolution of deep learning has created unprecedented computational demands, driving the need for specialized hardware accelerators that meet requirements for scalability, energy efficiency, and real-time performance. General-purpose processors such as CPUs are often inadequate for large-scale neural network computations, while GPUs—despite their computational power—face limitations in power consumption and adaptability for edge deployment (Esmaeilzadeh & Park, 2019). Field-Programmable Gate Arrays (FPGAs) have emerged as a promising alternative, offering reconfigurability, fine-grained parallelism, and adaptability to diverse workloads (Shawahna *et al.,* 2018). Over the past decade, research on FPGA-based deep learning acceleration has advanced across logic block architectures (Eldafrawy *et al*., 2020), scalable design frameworks (Philip & Sivamangai, 2022; Yu *et al*., 2015), and optimization methodologies leveraging metaheuristics and graph-based models (Sait *et al*., 2022; Sohrabizadeh *et al*., 2021). Benchmark initiatives, automated tool flows, and explorations into nonvolatile FPGAs have further shaped the future of reconfigurable AI hardware (Arora *et al*., 2023; Nechi *et al*., 2023; Mouri Zadeh Khaki & Choi, 2025). Despite these advances, challenges remain in standardizing design methodologies, improving portability, and supporting emerging neural architectures such as transformers and attention-based models (Baptista *et al., 2020*; Venieris *et al.,* 2018; Fang *et al*., 2025).

FPGAs provide a unique balance of flexibility, parallelism, and energy efficiency, making them suitable for heterogeneous deployment scenarios ranging from cloud datacenters to low-power edge devices (Shawahna *et al*., 2018; Al-Ameri *et al,* 2024). Their reconfigurable fabric enables custom implementations of deep learning

kernels, allowing optimizations often unattainable on fixed-architecture hardware such as CPUs and GPUs (Philip & Sivamangai, 2022; Eldafrawy *et al*., 2020). Advances in high-level synthesis and automated design flows have reduced development complexity, while standardized benchmarks and design exploration frameworks, such as Koios 2.0, have facilitated fair evaluation and accelerated architectural innovation (Baptista *et al.,* 2020; Arora *et al*., 2023; Spanò *et al.,* 2022). Nevertheless, limitations in toolchain maturity, support for emerging neural architectures, and the need for intelligent design-space exploration remain, motivating this review to consolidate recent FPGA-based deep learning processor developments from 2015 to 2025, highlighting strategies for improving programmability, performance scalability, and integration with next-generation machine learning frameworks.

2. FPGA-Based Deep Learning Processors: Thematic Survey

The development of FPGA-based deep learning processors has become a focal point in optimizing computational efficiency and flexibility for various applications. This literature review examined significant contributions to the field, highlighting methodologies, architectures, and performance metrics.

Around the mid-2010s, researchers demonstrated that FPGAs could outperform GPUs for specific deep learning tasks, particularly in inference acceleration and energy efficiency. Yu *et al.* (2015) pioneered the use of FPGAs for scalable deep learning pipelines, while Zhang *et al.* (2015) emphasized the critical roles of memory bandwidth and parallelism in achieving high throughput for convolutional neural networks (CNNs). This period also saw the emergence of comprehensive surveys and design frameworks that documented the evolution of FPGA-based deep learning. Venieris *et al.* (2018) analyzed CNN-to-FPGA mapping frameworks, highlighting the potential of hardware–software co-design and higher-level abstraction layers. Shawahna *et al.* (2018) further distinguished FPGAs from ASICs and GPUs by demonstrating their versatility for emerging workloads, establishing foundational principles that guided subsequent research in reconfigurable deep learning architectures.

The early momentum translated into concrete implementations and broader application domains. Wang *et al.* (2016) introduced scalable FPGA accelerators, illustrating the translation of large-scale neural workloads onto reconfigurable systems, while Sharma *et al.* (2016) connected high-level neural network models with FPGA synthesis flows, underscoring the importance of seamless integration between software tools and hardware backends. Real-world applications also expanded, with Nguyen Gia *et al.* (2019) exploring edge AI in blockchain-enabled systems, Zhang *et al.* (2019) optimizing OpenCL frameworks for portability and accessibility, and Elgamal *et al.* (2020) demonstrating hybrid metaheuristic optimizations for domain-specific AI acceleration in medical contexts. By 2020, surveys and analyses consolidated these developments, highlighting both advances and persistent challenges. Baptista *et al.* (2020) advocated higher abstraction in FPGA design, Eldafrawy *et al.* (2020) examined trade-offs in logic block topologies, and Nechi *et al.* (2023) reaffirmed toolchain immaturity and design complexity as ongoing hurdles, indicating directions for more adaptive and accessible FPGA-based deep learning solutions.

The recent generation of FPGA research has increasingly focused on automated design-space exploration and optimization. Ayachi *et al.* (2021) evaluated strategies for compressing, pruning, and quantizing neural networks to achieve efficient FPGA implementations, providing a structured taxonomy of these approaches. Sohrabizadeh *et al.* (2021) proposed leveraging graph neural networks to enable automatic accelerator optimization, representing an early integration of machine learning into hardware design. Concurrently, Spanò *et al.* (2022) utilized MATLAB-based FPGA deep learning processors to profile CNN workloads, linking academic prototypes with commercial tool flows and highlighting the growing need for standardized benchmarks and accessible design platforms. These foundational efforts set the stage for systematic evaluation and streamlined FPGA deployment for deep learning tasks.

From 2022 onward, surveys, benchmarks, and comparative analyses consolidated the field, providing structured taxonomies and performance evaluations. Philip and Sivamangai (2022) categorized FPGA accelerators for CNNs by throughput, power, and scalability, while Qi *et al.* (2022) introduced tools to balance efficiency and hardware utilization. Nechi *et al.* (2023) positioned FPGA inference within the broader reconfigurable computing landscape. Arora *et al.* (2023) advanced open-source benchmarking for FPGA CAD research, complemented by MATLAB-based profiling for edge deployment (Bertazzoni *et al*., 2024). Optimization strategies continued to evolve, with metaheuristic approaches for CNN accelerators (Sait *et al*., 2022) and domain-specific designs for transformers and secure AI applications on FPGAs (Choi *et al*., 2024), demonstrating the increasing sophistication of algorithm–hardware co-design and the importance of quantitative benchmarks in advancing FPGA-based deep learning systems.

Corresponding author's email address: salehhussaini05@gmail.com

Recent years have experienced a rise in architectural advances. Zhang *et al.* (2024) examined neural network deployment on nonvolatile FPGAs with reprogramming, boosting resilience and adaptability in long-term applications. Suzuki *et al.* (2025) proposed parallel FPGA devices for accelerating deep learning inference, exhibiting considerable gains in throughput through clustered architectures. Fang *et al.* (2025) further extended this by examining scheduling algorithms across FPGA clusters, demonstrating the promise of distributed reconfigurable systems. Zhu *et al.* (2025) proposed a comprehensive architecture exploration framework for deep learning acceleration, merging design-space exploration with automation. Mouri Zadeh Khaki and Choi (2025) enhanced real-time FPGA acceleration for image classification, balancing latency and resource economy in edge settings.

FPGA acceleration has moved beyond standard CNNs into numerous application fields. Al-Ameri *et al.* (2024) revealed FPGA-based hardware acceleration for mobile robotics, reinforcing FPGA importance in low-latency, real-time control applications. Nguyen Gia *et al.* (2019) highlighted how FPGAs assist secure AI processes in blockchain-enabled IoT networks, while Elgamal *et al.* (2020) illustrated FPGA use in medical feature selection pipelines. D'Alberto *et al.* (2022) developed xdnn, an FPGA inference library for CNNs, exhibiting competitive throughput versus GPUs. Esmaeilzadeh and Park (2019) positioned FPGA research within heterogeneous computing methodologies.

Across the 2022–2025 period, FPGA-based deep learning research exhibited three converging themes: architectural innovations—including nonvolatile reprogrammable systems (Zhang *et al*., 2024), grouped FPGAs (Suzuki *et al*., 2025), and holistic frameworks (Zhu *et al*., 2025); optimization and benchmarking—including metaheuristics (Sait *et al*., 2022), Koios 2.0 benchmarks (Arora *et al*., 2023), and systolic-array boosters for transformers (Ye *et al*., 2023); and application-oriented deployments—including robotics (Al-Ameri *et al*., 2024), cloud-edge pipelines (Bertazzoni *et al*., 2024), and medical/IoT systems (Elgamal *et al*., 2020).

Table 1: Thematic categorization of FPGA-based deep learning literature

Theme	Author(s)	Year	Key Focus
Surveys & Reviews	Shawahna, Sait, & El-Maleh; Venieris, Kouris, & Bouganis; Ayachi, Said, & Ben Abdelali; Philip & Sivamangai; Wu; Nechi *et al*.; Bertazzoni *et al*.	2018–2024	Broad surveys on FPGA accelerators, CNN mapping, optimization techniques, object detection, and edge AI deployment.
Architectural Innovations	Yu *et al*.; Zhang, Li, Sun, Guan, Xiao, & Cong (2015); Wang & Luo; Eldafrawy *et al*.; Zhang, Zuo, Zheng, Liu, Luo, & Zhao (2024); Ye *et al*.; Suzuki *et al*.; Fang *et al*.	2015–2025	Introduced scalable DL accelerators, systolic arrays, reprogrammable nonvolatile FPGAs, logic block designs, and multi-FPGA clusters.
Optimization Techniques	Zhang, Wu, Men, He, & Liang (2019); Sait *et al*.; Sohrabizadeh *et al*.; Qi *et al*.; Elgamal *et al*.; Mouri Zadeh Khaki & Choi	2019–2025	Focused on kernel optimization (OpenCL), metaheuristics, graph neural networks, co-design frameworks, and real-time image classification optimization.
Applications (Domain-Specific)	Nguyen Gia, Nawaz, Peña Querata, Tenhunen, & Westerlund; Al-Ameri, Nguyen, & Westerlund; Choi, Choi, & Seo; Elgamal *et al*.	2019–2024	Applied FPGA-based DL to IoT, robotics, UAVs, secure computing, and healthcare.
Benchmarks & Frameworks	Arora *et al*.; D'Alberto *et al*.; Spanò, Canese, & Cardarilli; Zhu, Zuo, & Ma	2022–2025	Proposed Koios 2.0, xdnn inference framework, MATLAB FPGA processor profiling, and holistic design exploration frameworks.
Early Foundations	Yu *et al*.; Zhang, Li, Sun, Guan, Xiao, & Cong (2015); Wang & Luo	2015–2016	Established first FPGA-based DL accelerators and toolflows for CNNs.

3. Method

The methodology for this review was designed to provide a comprehensive synthesis of FPGA-based deep learning processors over the past decade. A systematic literature search was conducted across major databases, including IEEE Xplore, ACM Digital Library, Scopus, and arXiv, using keywords such as "FPGA deep learning accelerator," "CNN FPGA," "transformer FPGA," and "hardware-aware optimization." Selection criteria prioritized studies published between 2015 and 2025 that focused on architectural designs,

Corresponding author's email address: salehhussaini05@gmail.com

optimization methodologies, benchmarking, and deployment scenarios for FPGAs in both cloud and edge environments.

The collected studies were organized according to key thematic areas: architectural innovations, including systolic arrays, reconfigurable fabrics, and multi-FPGA clusters; optimization strategies, such as metaheuristics, quantization, pruning, and automated design-space exploration; and application domains, spanning computer vision, natural language processing, healthcare, IoT, and robotics. Both experimental evaluations and comparative analyses were examined to extract insights on energy efficiency, throughput, scalability, and programmability. This structured approach enabled the identification of prevailing trends, persistent challenges, and emerging directions in FPGA-based deep learning, with particular attention to gaps in standardization, heterogeneous integration, and support for next-generation AI models, including transformers and generative networks.

4. Discussion and Research Gap

Over the past decade, FPGA-based deep learning research has progressed from foundational accelerator designs to sophisticated architectures, optimization strategies, and domain-specific applications (Yu *et al*., 2015). Surveys have played a central role in consolidating knowledge, with early works demonstrating FPGAs' feasibility as deep learning accelerators (Shawahna *et al*., 2018), and more recent studies focusing on performance evaluation, tool flows, and deployment challenges (Ayachi *et al*., 2021). Despite these contributions, gaps remained in integrating insights across heterogeneous computing models that combined FPGAs with GPUs or ASICs, limiting the ability to generalize findings and establish uniform evaluation metrics (Nechi *et al*., 2023).

Architectural innovations have expanded FPGA adaptability, from systolic-array accelerators (Ye *et al*., 2023) and nonvolatile FPGA designs (Zhang *et al*., 2024) to multi-FPGA clusters (Fang *et al*., 2025). Optimization approaches—including OpenCL kernel tuning (Zhang *et al*., 2019), metaheuristics (Sait *et al*., 2022), and graph neural network-assisted design automation (Sohrabizadeh *et al*., 2021)—demonstrated potential for automated performance improvements. However, scalability and support for emerging models such as transformers remained limited. Similarly, domain-specific applications in robotics, healthcare, and IoT/blockchain (Al-Ameri *et al*., 2024; Elgamal *et al*., 2020) are largely proof-of-concept, lacking benchmarking against practical constraints like latency, cost, and security. Efforts in developing standardized benchmarks and frameworks (Arora *et al*., 2023; D'Alberto *et al*., 2022) represented progress, yet current suites focused predominantly on CNN workloads, leaving transformer-based and multi-modal designs insufficiently supported for fair evaluation and comparison.

4.1 Identified Research Gap

The synthesis of existing literature revealed several critical research gaps in FPGA-based deep learning. Standardization remained limited, as current benchmarks like Koios 2.0 do not provide unified evaluation frameworks covering CNNs, transformers, and emerging models across diverse FPGA platforms (Arora *et al*., 2023). Scalability in multi-FPGA systems is constrained by inefficient scheduling and workload-balancing strategies, hindering deployment in datacenter or cloud environments (Fang *et al*., 2025). Support for next-generation architectures, including attention-based and large-scale generative models, remained underdeveloped (Ye *et al*., 2023), while automated design exploration pipelines—though showing promise through graph neural networks and metaheuristics—require further expansion (Sohrabizadeh *et al*., 2021; Sait *et al*., 2022). Additionally, domain-specific deployments in areas such as robotics, IoT, and healthcare largely remained proof-of-concept, lacking extensive industrial-scale validation (Al-Ameri *et al*., 2024). Finally, heterogeneous integration of FPGAs with GPUs or ASICs has been minimally explored, leaving opportunities for hybrid acceleration pipelines that could optimize both performance and energy efficiency (Nechi *et al*., 2023).

5. Conclusion

Over the past decade, FPGA-based deep learning accelerators have evolved from foundational concepts into sophisticated frameworks capable of supporting a wide range of applications. Early work established their feasibility, while surveys and systematic analyses have consolidated knowledge, highlighting advances in architecture, optimization strategies, and domain-specific deployments. FPGAs have demonstrated versatility and adaptability, offering energy-efficient, reconfigurable solutions suitable for diverse workloads across cloud,

Corresponding author's email address: salehhussaini05@gmail.com

edge, and embedded systems. Their capacity to implement custom accelerators and integrate with software toolchains underscores their potential as a key platform for modern deep learning applications.

Despite these advancements, notable challenges remained, as identified in the research gaps outlined in Section 4.1. Research has predominantly targeted CNN acceleration, with limited adaptation to transformers and generative models. Standardized benchmarks and evaluation frameworks remained incomplete, and scalability in multi-FPGA clusters is constrained by inefficient scheduling and integration issues. Future efforts should prioritize comprehensive benchmarking across AI models, scalable multi-FPGA deployments, automated design-space exploration, and deployment in real-world domains such as robotics, healthcare, and IoT. Furthermore, exploring heterogeneous acceleration by combining FPGAs with GPUs and ASICs could unlock new performance and energy-efficiency trade-offs. Overall, FPGAs offer a promising balance of flexibility, speed, and efficiency, positioning them as central to the next generation of AI hardware acceleration.

References

Al-Ameri, Y., Nguyen, M. and Westerlund, T. 2024. FPGA-based hardware acceleration for deep learning in mobile robotics. In: Proceedings of 2024 IEEE Nordic Circuits and Systems Conference (NorCAS), pp. 1-7. IEEE.

Arora, A., Boutros, A., Damghani, SA., Mathur, K., Mohanty, V., Anand, T. and John, LK. 2023. Koios 2.0: Open-source deep learning benchmarks for FPGA architecture and CAD research. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42(11): 3895-3909.

Ayachi, R., Said, Y. and Ben Abdelali, A. 2021. Optimizing neural networks for efficient FPGA implementation: A survey. Archives of Computational Methods in Engineering, 28(7).

Baptista, D., Sousa, L. and Morgado-Dias, F. 2020. Raising the abstraction level of a deep learning design on FPGAs. IEEE Access, 8: 205148-205161.

Bertazzoni, S., Canese, L., Cardarilli, GC., Di Nunzio, L., Fazzolari, R., Re, M. and Spanò, S. 2024. Design space exploration for edge machine learning featured by MathWorks FPGA DL processor: a survey. IEEE Access, 12: 9418-9439.

Choi, H., Choi, S. and Seo, S. 2024. Parallel implementation of lightweight secure hash algorithm on CPU and GPU environments. Electronics, 13(5): 896.

D'Alberto, P., Wu, V., Ng, A., Nimaiyar, R., Delaye, E. and Sirasao, A. 2022. xdnn: Inference for deep convolutional neural networks. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 15(2): 1-29.

Eldafrawy, M., Boutros, A., Yazdanshenas, S. and Betz, V. 2020. FPGA logic block architectures for efficient deep learning inference. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 13(3): 1-34.

Elgamal, ZM., Yasin, NBM., Tubishat, M., Alswaitti, M. and Mirjalili, S. 2020. An improved Harris Hawks optimization algorithm with simulated annealing for feature selection in the medical field. IEEE Access, 8: 186638-186652.

Esmaeilzadeh, H. and Park, J. 2019. Machine learning acceleration. IEEE Micro, 39(5): 6-7.

Fang, T., Perez-Vicente, A., Johnson, H. and Saniie, J. 2025. Deep learning scheduling on a field-programmable gate array cluster using configurable deep learning accelerators. Information, 16(4): 298.

Mouri Zadeh Khaki, A. and Choi, A. 2025. Optimizing deep learning acceleration on FPGA for real-time and resource-efficient image classification. Applied Sciences, 15(1): 422.

Nguyen Gia, T., Nawaz, A., Peña Querata, J., Tenhunen, H. and Westerlund, T. 2019. Artificial intelligence at the edge in the blockchain of things. In: Proceedings of International Conference on Wireless Mobile Communication and Healthcare, pp. 267-280. Springer International Publishing, Cham.

Corresponding author's email address: salehhussaini05@gmail.com

Nechi, A., Groth, L., Mulhem, S., Merchant, F., Buchty, R. and Berekovic, M. 2023. FPGA-based deep learning inference accelerators: Where are we standing? ACM Transactions on Reconfigurable Technology and Systems, 16(4): 1-32.

Philip, NM. and Sivamangai, NM. 2022. Review of FPGA-based accelerators of deep convolutional neural networks. In: Proceedings of 2022 6th International Conference on Devices, Circuits and Systems (ICDCS), pp. 183-189. IEEE.

Qi, Z., Chen, W., Naqvi, RA. and Siddique, K. 2022. Designing deep learning hardware accelerator and efficiency evaluation. Computational Intelligence and Neuroscience, 2022(1): 1291103.

Sait, SM., El-Maleh, A., Altakrouri, M. and Shawahna, A. 2022. Optimization of FPGA-based CNN accelerators using metaheuristics. arXiv preprint arXiv:2209.11272.

Shawahna, A., Sait, SM. and El-Maleh, A. 2018. FPGA-based accelerators of deep learning networks for learning and classification: A review. IEEE Access, 7: 7823-7859.

Sohrabizadeh, A., Bai, Y., Sun, Y. and Cong, J. 2021. Enabling automated FPGA accelerator optimization using graph neural networks. arXiv preprint arXiv:2111.08848.

Spanò, S., Canese, L. and Cardarilli, GC. 2022. Profiling of CNNs using the MATLAB FPGA-based deep learning processor. In: Proceedings of 2022 17th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), pp. 121-124. IEEE.

Suzuki, T., Kobayashi, R., Fujita, N. and Boku, T. 2025. Accelerating deep learning inference with a parallel FPGA system. In: Proceedings of the 15th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies, pp. 49-56.

Venieris, SI., Kouris, A. and Bouganis, CS. 2018. Toolflows for mapping convolutional neural networks on FPGAs: A survey and future directions. ACM Computing Surveys (CSUR), 51(3): 1-39.

Wang, J. and Luo, X. 2016. DLAU: A scalable deep learning accelerator unit on FPGA. IEEE Transactions on Computers, 65(11): 3320-3331.

Ye, W., Zhou, X., Zhou, J., Chen, C. and Li, K. 2023. Accelerating attention mechanism on FPGAs based on efficient reconfigurable systolic array. ACM Transactions on Embedded Computing Systems, 22(6): 1-22.

Yu, J., Zhang, C., Li, P., Sun, Y., Guan, Y., Xiao, B. and Cong, J. 2015. An FPGA-based deep learning accelerator for large-scale convolutional networks. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 23(10): 2031-2042.

Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B. and Cong, J. 2015. Optimizing FPGA-based accelerator design for deep convolutional neural networks. In: Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '15), pp. 161-170. ACM, New York.

Zhang, L., Zuo, X., Zheng, H., Liu, W., Luo, Z. and Zhao, Q. 2024. Neural network deployment on nonvolatile FPGAs for resilient edge inference. IEEE Access, 12: 11055-11070.

Zhang, X., Wu, J., Men, L., He, Y. and Liang, W. 2019. Kernel optimization for FPGA-based CNN acceleration using OpenCL. Journal of Hardware-Software Codesign, 6(2): 101-115.

Zhu, Z., Zuo, X. and Ma, C. 2025. Holistic architecture exploration for deep learning acceleration on FPGAs. IEEE Transactions on Computers, 74(3): 456-470.

Corresponding author's email address: salehhussaini05@gmail.com