Analog artificial intelligence hardware for neural networks: design trends and considerations

Kanchan S. Gorde[1], Sonali M. Sonavane[2], Sonal Hutke[3], Ankush Hutke[4]
[1]Department of Electronics Engineering, Terna Engineering College, University of Mumbai, Navi Mumbai, India
[2]Department of Information Technology, G.H. Raisoni College of Engineering and Management, S.P. Pune University, Pune, India
[3]Department of Electronics & Telecommunication, SIES Graduate School of Technology, University of Mumbai, Navi Mumbai, India
[4]Department of Information Technology, Rajiv Gandhi Institute of Technology, University of Mumbai, Navi Mumbai, India

Article Info

Article history:

Received Jun 12, 2025
Revised Sep 6, 2025
Accepted Sep 27, 2025

Keywords:

Analog very-large-scale integration
Edge artificial intelligence
In-memory computing
Mixed signal circuits
Neural network

ABSTRACT

The increasing deployment of artificial intelligence (AI) in real-time and edge applications intensified the demand for energy-efficient hardware capable of high-throughput processing. Conventional digital processors were constrained by sequential data processing, memory bandwidth limitations, and high-power consumption, making them suboptimal for edge-based AI. This review presented a comprehensive analysis of analog very-large-scale integration (VLSI) design approaches for neural network (NN) implementation focusing on circuit-level architectures including in-memory analog computing, current-mode circuits, switched-capacitor (SC) techniques, and operational transconductance amplifier (OTA)-based designs. Significant hardware design considerations such as process variation, crossbar scalability, precision–linearity trade-offs, and mixed-signal interface challenges were critically examined. Furthermore, training methodologies—spanning offline learning, circuit calibration, and programmability were discussed in the context of analog AI hardware. The review incorporated case studies, recent developments in edge deployment, and a comparative analysis of advanced analog VLSI chips. Key performance evaluation metrics such as accuracy, calibration overhead, noise robustness, and energy per inference, were also addressed. Circuit-level design aspects that impacted the performance, precision, and reliability of analog computing blocks were discussed. The paper concluded by identifying research gaps and future directions for the development of analog AI hardware suitable for real-world edge applications.

Corresponding Author:

Kanchan S. Gorde
Department of Electronics Engineering, Terna Engineering College, University of Mumbai
Sector-22 Nerul, Navi Mumbai, India
Email: kanchangorde@ternaengg.ac.in

1. INTRODUCTION

The rapid developments in artificial intelligence (AI) have enlarged the necessity for hardware accelerators capable of handling the computational demands of neural networks (NN). While digital processors, such as CPUs, GPUs, encounter constraints in scalability, processing speed, and energy consumption [1], [2]. Analog computing presents an alternative due to its potential for lower power consumption and reduced latency. Analog very-large-scale integration (VLSI) circuits utilize basic electrical properties of devices to perform operations like multiplication and accumulation, avoiding the energy overhead associated with digital switching [3]. By operating in the continuous domain with voltages and currents, these circuits enable more compact and energy-efficient computation. Recent progress in device technologies, including memristors, resistive RAM (ReRAM), and floating-gate transistors, has extended the possibilities for analog in-memory computing directly at the storage for data processing [4], [5]. This facilitates the development of rapid, low-power inference engines well suited for edge AI applications.

Despite the significant analog VLSI NN implementations, there is a need for consolidated assessments that highlight the present state of technology, identify challenges, and project future prospects. Even though analog AI hardware is developing rapidly, the field of AI does not have an adequate understanding of how various design choices compare in terms of effectiveness, scalability, and real-world deployment. The present state of research frequently varies among many device kinds, circuit designs, and assessment methodologies, which makes it challenging to draw insightful conclusions or direct future study. In order to overcome this difficulty, this review combines recent research into a cohesive viewpoint.

Although analog systems have limitations such as reduced numerical precision, noise sensitivity, and device variability, advances in circuit design and calibration techniques have improved their viability. Increasing attention in edge intelligence and emerging device technologies spots analog VLSI as a captivating tool for next-generation AI hardware [6]. Essential components such as operational transconductance amplifiers (OTAs), capacitive multipliers, and current mirrors provide fundamental functions like multiply-and-accumulate (MAC) in hardware [7]. Circuit-level simulation of these blocks using tools such as LTspice can aid in analyzing gain nonlinearity, bandwidth limitations, and power-delay trade-offs prior to physical implementation [8].

2. BACKGROUND: CLASSIFICATION OF ANALOG VLSI ARCHITECTURE FOR NEURAL NETWORKS

NN implementations on analog VLSI exhibit design methodologies depending on the physical realization of neural functions. These categories are summarized in Figure 1.

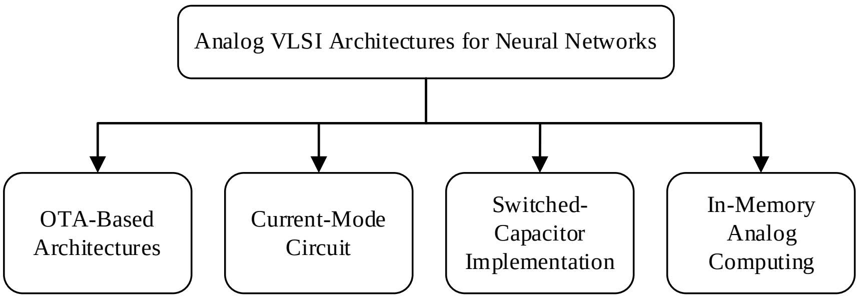
Analog VLSI Architectures for Neural Networks
OTA-Based Architectures
Current-Mode Circuit
Switched-Capacitor Implementation
In-Memory Analog Computing

Figure 1. Classification of analog VLSI architectures for NN

2.1. Operational transconductance amplifier-based architecture

OTA-based designs are preferred for their ability to convert voltage inputs into linear current outputs, enabling efficient matrix-vector multiplications (MVM). These architectures are advantageous in low-power applications and the effective implementation of dense layers in NNs [1], [7]. Recent advances integrate adaptive biasing and gain control to enhance the dynamic range and linearity of OTAs for improving computational accuracy [8]. Figure 2 depicts an OTA-based neuron implementation used in analog NNs illustrating differential pair configuration, biasing, and current-mode output.

2.2. Current-mode circuits

Current-mode circuits enable compact and power-efficient analog multiplication and summation by processing signals. This approach reduces parasitic effects and supports high-speed operations [2]. These circuits leverage current mirrors, translinear loops, and current conveyors to perform neural operations such as weighted summation and activation. Stability and precision are improved through active feedback mechanisms and current reflectors integrated within these circuits [9]. Their inherent suitability for low-voltage operation and high-bandwidth signal processing makes them attractive for edge AI applications. Figure 3 illustrates a current-mode neuron circuit commonly used in analog NNs. The design uses input and output currents to function those benefits in deep submicron technologies such as low voltage operation, high speed response, and improved scalability. The elements like biasing branches, translinear loops, and current mirrors are integrated into the circuit to carry out analog activation and summing.

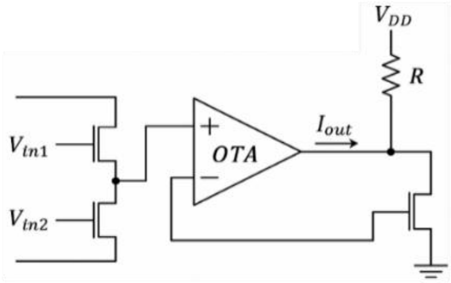

Figure 2. An OTA-based artificial neuron

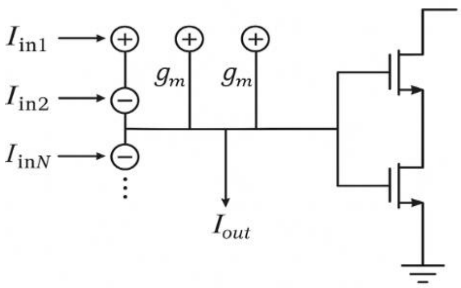

Figure 3. Current mode neuron circuit

2.3. Switched-capacitor implementations

Switched-capacitor (SC) networks perform time-domain signal processing to control over weights through charge transfer between capacitors. A typical SC circuit, as illustrated in Figure 4, uses complementary switching phases to alternately sample and transfer charge emulating resistance. SC circuits are compatible with mixed-signal systems and can interface with digital components [3]. Recent implementations integrate non-volatile memory elements for adaptive tuning and reconfiguration, enhancing system flexibility and programmability [10].

2.4. In-memory analog computing

In-memory analog computing (AIMC) architectures perform MAC operations by encoding weights of memory cells. Using memristive crossbar arrays and Kirchhoff's current law, these systems achieve parallel current summation directly within the memory hardware [4], [5]. Recent research emphasizes on improving device endurance, retention, and reducing device-to-device variability to enhance reliability in analog memory arrays [6], [11]. Figure 5 illustrates a memristor-based crossbar array used for in-memory analog computing. Each cross-point in the array consists of a programmable memristive device that stores the weight value. Input voltages are applied to the word lines (rows), and the resulting output currents on the bit lines (columns) represent the analog dot-product operation, enabling efficient matrix–vector multiplication directly in memory. This architecture minimizes data movement and supports parallel computation, making it highly suitable for energy-efficient NN inference at the edge.

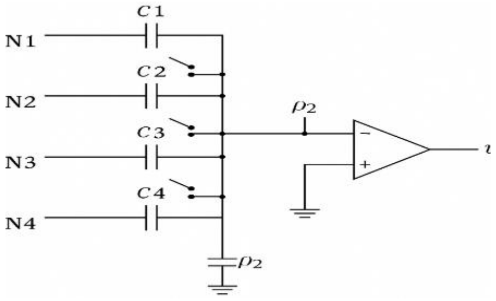

Figure 4. SC neuron circuit

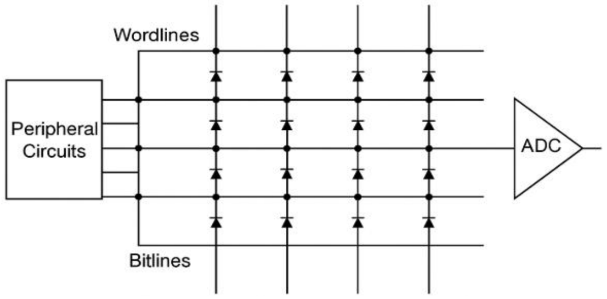

Figure 5. Memristor based crossbar array

3. METHODS: ANALOG HARDWARE PRIMITIVES FOR NEURAL NETWORKS

NN implementations using analog hardware depend on basic circuit elements to perform accumulation, multiplication, and nonlinear activation. These hardware primitives design influences performance parameters such as silicon area, computational throughput and energy efficiency. Neural computation in analog VLSI depends on MVM is realized through capacitive charge-sharing techniques or current-mode multipliers. An example is the high-throughput multiply-accumulate unit based on low-voltage rapid single-flux quantum circuits, offers superior energy efficiency and is suitable for high-speed signal processing applications [12]. Current-mode approaches require careful layout and design of current summing nodes to avoid nonlinearity and signal loss over long interconnects.

Memory devices such as ferroelectric field-effect transistors (FeFETs) and ReRAM are integrated into in-memory computing architectures for analog MVM. By improving the memory-compute bottleneck, these devices enable local processing of neuronal weights [13]. Recent analog AI chips combine dense memory arrays with digital post-processing to enhance computational accuracy [14]. Analog activation functions implemented using translinear elements or piecewise-linear circuits provide hardware-efficient and cost-effective nonlinear functions like sigmoid or ReLU without relying on digital lookup tables [15]. To maintain signal integrity across the neural processing pipeline, analog accumulators, integrators, and buffers are essential. Design techniques such as offset cancellation, gain adjustment, and adaptive biasing are used to offset thermal drift and device mismatch [16]. On-chip learning engines benefit from dynamic bias management and floating-gate storage for improved reliability [17]. Table 1 compares the primary analog hardware primitives used in NN implementations.

Table 1. Comparison of analog hardware primitives for NN implementation

Hardware primitive	Key function	Common implementation	Strengths	Limitation
Matrix-vector multiplier	Core MAC operation	Capacitive DACs, current-mode, and crossbars	High throughputs, energy efficient	Limited precision, susceptible to noise
Activation function	Applies non-linearity to neuron outputs	Translinear circuits, differential pairs	Low power, fast response	Limited flexibility for complex functions
Analog mentor weights	Store and process neural weights	ReRAM, FeFETs, and floating-gate transistors	Enables in-memory computing	Endurance, drift over time
Accumulator integrator	Accumulates intermediate outputs	Charge integration circuits	Area efficient	Leakage offset errors
Signal conditioning	Maintains signal integrity	Bias tuning, gain amplifiers	Improves analog reliability	Additional circuit complexity

3.1. Design considerations for analog primitives

Recent analog architectures highlight improved signal purity, inclusive training support and continuous integration with digital systems. Important design aspects and techniques utilized in analog VLSI neural hardware are discussed below:

3.1.1. Mixed-signal interfaces

Analog neural cores require analog-to-digital converter (ADC) and digital-to-analog converters (DAC) to interface with digital control units, memory blocks, or input/output modules. The resolution, sampling rate, and power efficiency of these converters affect overall system performance. For energy-constrained edge devices, high-speed but low-resolution ADCs, such as successive-approximation register ADCs, are adopted [18].

3.1.2. Precision and linearity trade-offs

Analog primitives have limitations due to device nonlinearity, thermal noise, and mismatch, which degrade inference accuracy and signal fidelity compared to digital counterparts. To alleviate these effects, digital correction, calibration circuits, and mixed-signal compensation techniques are mostly employed, balancing circuit complexity against performance gains [19].

3.1.3. Scalability issues in crossbar arrays

Large-scale analog crossbar arrays used in in-memory MVM suffer from voltage drops, sneak-path currents, and size-dependent performance degradation. Addressing these requires peripheral circuit compensation, hierarchical array partitioning, and advanced materials engineering [20].

3.1.4. Process variation and reliability

Analog circuits exhibit higher sensitivity to fabrication process variations and environmental changes. Reliability is enhanced trimming, feedback calibration, and on-chip learning mechanisms that adapt dynamically to changing conditions over time and temperature. Recent work demonstrated improved resilience of in-memory training hardware under asymmetry and variability through on-chip adaptation mechanisms [21].

3.1.5. Circuit-level design aspects

At the transistor level, the implementation of analog primitives such as multipliers and integrators require proper biasing and matching. OTA-based MAC units, current mirrors and capacitive integrators used in analog computation must be optimized for linearity, offset, and temperature stability. A schematic of analog MAC unit is shown in Figure 6. These units form the core computational engines in analog NN hardware and influence accuracy, linearity, and power consumption. Layout techniques such as common-centroid placement and guard rings are employed to mitigate mismatch and noise coupling in analog cores [22].

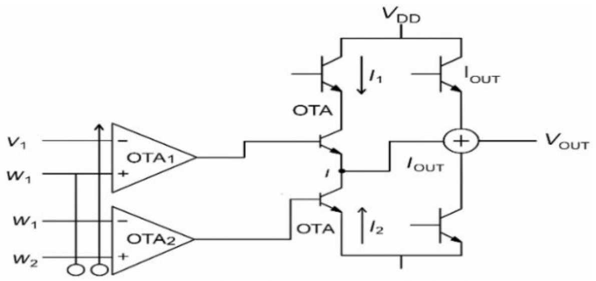

Figure 6. Analog MAC block utilizing OTAs and current mirrors

For ultra-low power edge devices, OTA-based circuits work best, while in-memory computation provides parallelism and scalability. Although current-mode circuits are excellent at high-speed operations, they are sensitive to noise. Circuits using SC provide greater accuracy at the expense of area efficiency. Table 2 provides a comparative summary of major analog VLSI architectures in terms of power efficiency, computational speed, silicon area, and scalability.

Table 2. Comparative summary of analog VLSI architectures

Architecture	Power	Speed	Area	Scalability	Key strength
OTA-based	Very low	Moderate	Small	Moderate	Energy efficiency
In-memory computing	Low	High	High density	High	Parallelism, density
Current-mode circuits	Low	Very high	Moderate	Limited by noise	High-speed ops
SC	Moderate	Moderate	Area-heavy	Limited	Precision control

4. METHODS: TRAINING CHALLENGES AND DEPLOYMENT STRATEGIES

4.1. Training techniques and circuit calibration

Training NNs on analog hardware presents several challenges as non-idealities, unpredictability, and restricted programmability of the device. Analog circuits use voltage levels, currents, or the physical states of memory units to encode parameters, in contrast to digital systems that use floating-point arithmetic to update weights exactly. This complicates convergence during training by introducing drift, restricted update granularity, and stochastic behaviour.

Offline training is a popular method in which the final weights are transmitted to the analog hardware after the network has been trained in a high-precision digital simulation. This reduces adaptability but avoids the complexity of in-situ training. This has been addressed by hybrid training loops that enable partial learning in hardware by fusing digital back propagation with analog forward passes [22], [23]. In addition, circuit-level calibration methods including bias tuning, redundancy, and closed-loop correction have been incorporated into analog designs to preserve accuracy when temperature and temporal drift are present [24]. These techniques are important in fluctuating edge situations for long-term deployment. Common deployment issues are compiled in the Table 3, along with mitigation techniques.

Table 3. Deployment challenges and strategies in analog VLSI NN for edge AI

Challenges	Mitigation strategy
Device non-linearity	Use of linearizing circuits, compensation algorithms
Process variation	Statistical calibration, adaptive tuning
Limited bit precision	Quantization-aware training, redundancy
Temperature drift	On-chip thermal sensors and dynamic recalibration
Weight retention in NVMs	Periodic refresh, write-verification loops

4.2. Programmability in analog hardware

One of the fundamental constraints for analog VLSI is programmability. Post-deployment modifications are challenging because the majority of analog accelerators are designed for fixed network topologies and weight distributions. Programmable non-volatile memories, such as resistive-RAM and electrochemical-RAM, are used as weight stores in some designs; nevertheless, frequent programming causes endurance problems and the accumulation of analog noise [25]. Modular cores and reconfigurable interconnects have been investigated recently, allowing for some flexibility in data flow and model structure [26]. However, the degree of programmability is still significantly lower than that of digital accelerators, necessitating early development co-design with the model architecture. An explicit workflow spanning from training to inference is required for the implementation of analog VLSI devices for NN. Training in digital systems starts offline, and then the trained weights are converted into analog representation for storage in RRAM or memristive devices. Analog circuits carry out the required calculations during inference, while DACs are used to transform the outputs back to digital format. This procedure guarantees great accuracy and energy-efficient calculation. An analog AI inference workflow combines digital training with analog inference to enhance energy efficiency, throughput, and memory utilization, as outlined in the following key stages:

- Offline training: model is trained using high-precision floating-point arithmetic in digital systems (GPUs).
- Digital-to-analog weight conversion: trained weights are encoded into analog form for storage in non-volatile memory (memristors or RRAM).
- Analog inference: inference computations are performed using low-power analog circuits.
- Digital output conversion: DACs convert analog results to digital outputs for post-processing.
- Optional feedback loop: used for calibration and performance refinement.

4.3. Deployment considerations for edge artificial intelligence

Analog NN implementation at the edge has system-level limitations. Because of its low energy consumption per operation, analog computing is appealing because power efficiency and silicon area are the top concerns. However, issues including resistance to environmental noise, real-time inference delay, and integrating with digital sensors need to be resolved. Additionally, as analog systems expand beyond small network sizes, scalability problems arise. In order to minimize latency and energy overhead, co-integration with ADC/DAC interfaces and hierarchical memory access patterns is necessary [27]. Several analog accelerators have proven to be capable of processing images and signals in real time while adhering to stringent edge-power budgets [28]. A comparison of the analog and digital techniques to edge AI deployment is presented in the Table 4.

Table 4. Comparison of analog and digital approaches in edge AI

Metrics	Analog VLSI	Digital processors
Power consumption	Ultra-low (nW–µW range)	Higher (mW–W range)
Inference latency	Extremely low (~ns–µs)	Moderate (~µs–ms)
Precision	Low to moderate (4–8 bits)	High (8–32 bits)
Area efficiency	High due to in-memory computation	Lower due to separate memory and logic
Scalability	Limited by crossbar size and noise	Easier with standard process nodes
Programmability	Moderate (via NVM or floating gate)	High (via software updates)

4.4. Comparison of recent analog very-large-scale integration chips

Table 5 presents a comparison of selected recent analog VLSI chips targeting NN acceleration. These chips reflect diverse design approaches for advancing analog NN, with an emphasis on low power consumption and in-memory computing.

Table 5. Comparison of recent analog VLSI chips for edge AI applications

Chip/architecture	Technology (mm)	Architecture type	Application	Notable features
IBM analog AI chip	14	Memristive crossbar	DNN inference	High accuracy, in-situ training ability
Cerebras WSE	16	Mixed-signal array	Large-scale DNNs	High throughput, energy efficient
Brain chip Akida	28	Event-driven analog	Edge AI	Low power spiking NN accelerator
Mythic analog matrix processor	65	Analog matrix processor	Edge inference	High-density analog matrix, low latency
ISAAC accelerator	45	RRAM-based crossbar	CNN inference	In-memory computing, scalable arrays

4.5. Case studies

Recent research explores deploying analog VLSI chips in image recognition, biomedical sensing, and real-time classification tasks. For instance, the Akida chip has been used in wearable EEG classification, and the Mythic processor supports object detection at the edge. Emerging trends in deployment strategies include:

- In-memory computing: advances in ReRAM and phase-change memory (PCM) enable direct computation within memory arrays, minimizing memory access latency and energy use [27].
- Hybrid analog-digital systems: combining analog inference units with digital control logic improves system adaptability and enables real-world applications [22], [25].
- Crossbar arrays for ai acceleration: analog crossbars are essential for fast, efficient MVM, as seen in accelerators like Mythic's AMP [29] and ISAAC [30].

Targeting smart sensors and internet of things (IoT) nodes, co-packaging of analog AI modules with microcontrollers is becoming more popular. These hybrid systems preserve digital programmability while leveraging the ultra-efficient inference capabilities of analog circuits. Analog AI modules are increasingly being used in conjunction with microcontrollers to combine the efficiency of analog circuits with the flexibility of digital control for application in smart sensors and IoT devices. Similarly, companies like Qualcomm and HP Labs are evolving this trend by incorporating memristive technologies, which offer non-volatile memory and low-latency AI processing, into edge AI platforms [31]. These developments emphasize the transition of analog VLSI from a research area to a mainstream solution for scalable and efficient edge intelligence.

5. RESULTS: HARDWARE INNOVATIONS AND INTEGRATION

5.1. Device-level technologies and memory integration

The performance of analog NN accelerators depends on the underlying device technologies used for computation and memory storage. Emerging non-volatile memory devices such as RRAM, PCM, FeFETs, and floating-gate transistors have gained significant attention due to their suitability for analog programmability and non-volatile weight storage.

Among these, RRAM-based crossbar arrays are especially promising due to their low power consumption, high integration density, and compatibility with CMOS processes. However, challenges including non-ideal switching characteristics, resistance drift, device-to-device variability, and degradation with cycling remain significant barriers to widespread adoption [32], [33]. To address these, techniques such as differential pair configurations to cancel common mode variations, write-verify tuning schemes, and digital-assisted analog tuning have been employed and demonstrated to improve accuracy and reliability [34], [35].

Practical implementations like the Mythic Analog Matrix Processor and Syntiant's NDP200 illustrate how analog computing blocks can be integrated with on-chip flash or RRAM to enable inference at microwatt-level power [36], [37]. Furthermore, temperature-aware programming algorithms and selector integration are increasingly employed to enhance analog weight stability and endurance in edge-AI environments [38].

5.2. Integration challenges in analog-centric systems

Analog-centric systems face significant integration hurdles due to the inherent differences in signal domains, noise sensitivity, and design methodologies when interfacing with digital and mixed-signal blocks. Power supply fluctuations, substrate coupling, and capacitive crosstalk can impact analog signal fidelity, especially in dense edge devices. Designers employ shielded interconnect routing, layout symmetry, and deep n-well isolation to minimize analog signal degradation [39]. Moreover, calibration circuitry is often embedded on-chip to dynamically adjust offset and gain errors resulting from temperature or process variations. Unlike digital blocks that scale predictably with technology nodes, analog arrays require meticulous layout tuning to preserve linearity and matching. In recent designs, such as the IBM analog AI Core and SambaNova Reconfigurable Dataflow Unit, careful partitioning of analog tiles, along with hierarchical interconnects and programmable digital overlays, are employed to balance flexibility with performance [40], [41].

To ensure robust deployment, some analog architectures utilize hardware-in-the-loop calibration, where digital units assist analog computations during runtime. This hybrid analog–digital approach has shown success in commercial chips like Brain Chip Akida and Aspinity AML100, offering ultra-low-power operation with enhanced reliability in noisy or dynamic environments [42], [43]. Circuit-level techniques such as differential signaling, deep n-well isolation, and bias tuning circuits are integrated to reduce substrate coupling and supply noise. Low-voltage analog AI cores tend to be susceptible to distortion and drift, which are influenced by the design of the analog devices. To reduce mismatch and nonlinearity across computational arrays, precise transistor sizing, layout symmetry, and matching are crucial in highly scaled analog circuits [44].

5.3. Layout-aware design and mixed signal integration

Analog neural accelerators are particularly susceptible to layout-level problems including parasitic, device mismatch, and noise coupling from digital logic. Techniques like common-centroid placement, guard rings, and shielded routing are crucial for maintaining signal integrity in OTAs and current mirrors. In order to minimize interference in mixed-signal system-on-chip (SoC) integration, an analog core should be placed near memory and separated from noisy digital blocks [45]. The development of automated, layout-aware analog design tools has considerably simplified this process. For instance, floor planning enabled by reinforcement learning greatly enhanced layout quality by cutting down on space and wire length and accelerating up layout times [46].

6. RESULTS AND DISCUSSION: BENCHMARKING AND PERFORMANCE EVALUATION

With rapid growth in analog VLSI NN topologies, it is necessary to have robust benchmarking frameworks to evaluate their practicality [47]. Analog systems' distinct trade-offs are captured by conventional metrics because of basic differences in computation, memory, and noise behavior from digital equivalents [48]. Effective benchmarking must take into account both algorithmic accuracy and hardware-specific figures of merit to provide balanced comparisons [49]. Commonly used evaluation metrics include:

- Energy per inference: for edge jobs, analog accelerators usually aim for sub-micro joule levels per inference.
- Area efficiency: measures computational density for chip integration in resource-constrained devices.
- Throughput: determines processing rate, often influenced by analog signal settling times.
- Latency: inference delay from input to output, crucial for real-time applications.
- Robustness to noise: indicates model degradation under environmental or device noise.
- Calibration overhead: captures the time and resources required to maintain system performance over time, such as compensation for drift and temperature variation.

6.1. Task-oriented benchmarks

Typically, analog accelerators are designed for particular edge AI applications including sensor fusion, keyword identification, and image classification. Task-specific benchmarks are therefore more beneficial than general ones. Although caution must be taken when comparing results across different hardware settings, benchmarks like MNIST and Google Speech Commands are still often employed [50].

6.2. Cross-platform comparisons

Cross-platform evaluation remains challenging due to the hybrid nature of analog-digital systems and differences in device technologies and architectures. Compound metrics like the energy-accuracy product (EAP) or throughput per watt are commonly used to capture overall system efficiency by considering accuracy and hardware cost factors [51]. Additionally, essential metrics like dynamic range, noise margins, and gain linearity for analog building blocks are evaluated with the help of circuit-level simulations. Calibration routines rely on on-chip circuit techniques like DAC tuning and analog memory compensation to ensure linearity and accuracy in inference [52]. These evaluation frameworks are crucial to enabling the reliable deployment of analog AI systems across diverse hardware platforms and edge computing environments.

6.3. Benchmarking case study

MNIST classification on an OTA-based accelerator is taken into consideration [53]. With 10 μs latency, the system derived ~0.5 μJ/inference at same accuracy (96%). In-memory crossbar systems for CIFAR-10 [54] showed greater throughput but needed to be calibrated. Furthermore, recent research has shown that analog neuromorphic circuitry may approach digital baselines with ~98% accuracy on MNIST while preserving energy gains [45]. On the CIFAR-10, hybrid analog synapse circuits [46] have also demonstrated competitive performance, underscoring the trade-off between density and calibration overhead.

7. CONCLUSION

Exploring analog VLSI designs for NN has created possibilities for low-latency, energy-efficient AI computation for edge deployment. This review has examined the trade-offs involved in analog AI implementations, focusing on architectural classifications, hardware primitives, training challenges, system-level integration, and performance evaluation. Several key areas are likely to shape the future trajectory of analog VLSI systems for AI: i) device scalability and reliability: although emerging memory technologies such as RRAM and PCM enable dense analog weight storage, they still face challenges like drift, limited endurance, and variability. Future work must address these issues through robust encoding schemes and more stable material innovations; ii) reconfigurability and on-chip learning: because of their restricted programmability, the majority of analog accelerators on the market today prefer inference-only modes. Their versatility could be enhanced through the development of efficient on-device learning techniques and customizable hardware; and iii) hybrid system design: it will continue to be crucial to integrate digital logic seamlessly, including with ADC/DAC interfaces and calibration devices. A practical solution is provided by hybrid analog-digital systems that can dynamically share computation across domains.

Analog VLSI systems offer ultra-low power consumption, high computational density, and bio-inspired architectures that suit edge-AI use cases. Although the field remains specialized, it is evolving rapidly. Transforming current prototypes into scalable, deployable AI systems will require continued collaboration across device physics, circuit design, AI algorithms, and system integration. Ensuring circuit-level robustness through layout symmetry, noise isolation, and analog calibration will be essential for building reliable, low-power analog AI systems for real-world deployment.

ACKNOWLEDGMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kanchan S. Gorde	✓	✓		✓		✓			✓	✓	✓	✓	✓	
Sonali M. Sonavane		✓					✓			✓	✓			
Sonal Hutke				✓	✓		✓		✓					
Ankush Hutke					✓	✓				✓			✓	

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] G. Menghani, “Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–37, Dec. 2023, doi: 10.1145/3578938.

[2] N. P. Jouppi *et al.*, “Ten lessons from three generations shaped Google’s TPUv4i: Industrial product,” in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, Jun. 2021, pp. 1–14, doi: 10.1109/ISCA52012.2021.00010.

[3] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017, doi: 10.1109/JSSC.2016.2616357.

[4] S. O. Park *et al.*, "Linear conductance update improvement of CMOS-compatible second-order memristors for fast and energy-efficient training of a neural network using a memristor crossbar array," *Nanoscale Horizons*, vol. 8, no. 10, pp. 1366–1376, 2023, doi: 10.1039/d3nh00121k.

[5] C. Lammie *et al.*, "Improving the Accuracy of Analog-Based In-Memory Computing Accelerators Post-Training," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2024, pp. 1–5, doi: 10.1109/ISCAS58744.2024.10558540.

[6] Y. Wu *et al.*, "Bulk-Switching Memristor-Based Compute-In-Memory Module for Deep Neural Network Training," *Advanced Materials*, vol. 35, no. 46, pp. 1–13, 2023, doi: 10.1002/adma.202305465.

[7] R. M. Kumar, C. J. Sree, G. R. K. Reddy, P. N. K. Reddy, and T. B. Kumar, "Design of Low-Power OTA for Bio-medical Applications," in *Advances in Cognitive Science and Communications*, 2023, pp. 99–103, doi: 10.1007/978-981-19-8086-2_10.

[8] Y. Yin *et al.*, "An Ultra-Low-Voltage Transconductance Stable and Enhanced OTA for ECG Signal Processing," *Micromachines*, vol. 15, no. 9, pp. 1–14, Aug. 2024, doi: 10.3390/mi15091108.

[9] A. Sengupta, Y. Shim, and K. Roy, "Proposal for an All-Spin Artificial Neural Network: Emulating Neural and Synaptic Functionalities through Domain Wall Motion in Ferromagnets," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 6, pp. 1152–1160, Dec. 2016, doi: 10.1109/TBCAS.2016.2525823.

[10] M. Zolfagharinejad, U. Alegre-Ibarra, T. Chen, S. Kinge, and W. G. van der Wiel, "Brain-inspired computing systems: a systematic literature review," *European Physical Journal B*, vol. 97, no. 6, pp. 1–23, Jun. 2024, doi: 10.1140/epjb/s10051-024-00703-6.

[11] X. Wang, M. A. Zidan, and W. D. Lu, "A Crossbar-Based In-Memory Computing Architecture," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4224-4232, Dec. 2020, doi: 10.1109/TCSI.2020.3000468.

[12] I. Nagaoka *et al.*, "A High-Throughput Multiply-Accumulate Unit With Long Feedback Loop Using Low-Voltage Rapid Single-Flux Quantum Circuits," *IEEE Transactions on Applied Superconductivity*, vol. 33, no. 3, pp. 1–8, 2023, doi: 10.1109/TASC.2023.3239329.

[13] M. Hellenbrand, I. Teck, and J. L. MacManus-Driscoll, "Progress of emerging non-volatile memory technologies in industry," *MRS Communications*, vol. 14, no. 3, pp. 1099–1112, 2024, doi: 10.1557/s43579-024-00660-2.

[14] M. Monjur, J. Calzadillas, and Q. Yu, "Hardware Security Risks and Threat Analyses in Advanced Manufacturing Industry," *ACM Transaction on Design Automation of Electronic System*, vol. 28, no. 5, pp. 1-22, 2023, doi: 10.1145/3603502.

[15] O. Krestinskaya, K. N. Salama, and A. P. James, "Analog Backpropagation Learning Circuits for Memristive Crossbar Neural Networks," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, 2018, pp. 1–5, doi: 10.1109/ISCAS.2018.8351344.

[16] S. M. Sze and K. K. Ng, *IGFET and Related Surface Field Effects*, Physics of Semiconductor Devices, 3rd ed., New York: Hoboken, NJ: John Wiley & Sons, 2006, pp. 505–550.

[17] Y. Li *et al.*, "In situ parallel training of analog neural network using electrochemical random-access memory," *Frontiers in Neuroscience*, vol. 15, p. 636127, 2021, doi: 10.3389/fnins.2021.636127.

[18] S. Puneeth *et al.* "Performance Evaluation of Ultra-Low Power ADCs in Energy Harvesting Systems," *SSRG International Journal of Electrical and Electronics Engineering (SSRG-IJEEE)*, vol. 11, no. 8, pp. 1–10, Aug. 2024, doi: 10.14445/23488379/IJEEE-V11I8P120.

[19] Y. Wu, F. Ye, and J. Ren, "A Calibration-Free, 16-Channel, 50-MS/s, 14-Bit, Pipelined-SAR ADC with Reference/Op-Amp Sharing and Optimized Stage Resolution Distribution," *Electronics*, vol. 11, no. 5, pp. 1-17, 749, 2022, doi: 10.3390/electronics11050749.

[20] N. Afroz, A. Sayem, G. Volanis, D. Maliuk, H. Stratigopoulos, and Y. Makris, "On the Sensitivity of Analog Artificial Neural Network Models to Process Variation," in *2024 IEEE 42nd VLSI Test Symposium (VTS)*, IEEE, Apr. 2024, pp. 1–7, doi: 10.1109/VTS60656.2024.10538718.

[21] R. K. Vartak, V. Saraswat, and U. Ganguly, "Robustness to Variability and Asymmetry of In-Memory On-Chip Training," in *International Conference on Artificial Neural Networks*, Cham: Springer Nature Switzerland, 2023, vol. 14262, pp. 249–257, doi: 10.1007/978-3-031-44201-8_21.

[22] L.-Y. Song, C.-Y. Chou, T.-C. Kuo, C.-N. Liu, and J.-D. Huang, "Machine Learning Assisted Circuit Sizing Approach for Low-Voltage Analog Circuits with Efficient Variation-Aware Optimization," *ACM Transactions on Design Automation of Electronic Systems*, vol. 28, no. 2, Mar. 2023, doi: 10.1145/3567422.

[23] L. Zhang and J. Sitte, "Hardware In-the-Loop Training of Analogue Neural Network Chip," *International Symposium on Neural Networks*, Berlin, Heidelberg: Springer, 2006, vol. 3973, pp. 182–189, doi: 10.1007/11760191_194.

[24] H. Sun *et al.*, "Reliability-Aware Training and Performance Modeling for Processing-In-Memory Systems," in *ASPDAC '21: Proceedings of the 26th Asia and South Pacific Design Automation Conference*, New York, NY, USA, 2021, pp. 847–852, doi: 10.1145/3394885.3431633.

[25] M. Abedin *et al.*, "Material to system-level benchmarking of CMOS-integrated RRAM with ultra-fast switching for low power on-chip learning," *Scientific Reports*, vol. 13, no. 14963, 2023, doi: 10.1038/s41598-023-42214-x.

[26] S. Wei, X. Lin, F. Tu, Y. Wang, L. Liu, and S. Yin, "Reconfigurability, Why It Matters in AI Tasks Processing: A Survey of Reconfigurable AI Chips," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 3, pp. 1228-1241, Mar. 2023, doi: 10.1109/TCSI.2022.3228860

[27] D. N. Zadeh and M. B. Elamien, "Generative AI for Analog Integrated Circuit Design: Methodologies and Applications," in *IEEE Access*, vol. 13, pp. 58043-58059, 2025, doi: 10.1109/ACCESS.2025.3553743.

[28] X. Meng *et al*., "Digital-analog hybrid matrix multiplication processor for optical neural networks," *Nature Communications*, vol. 16, art. no. 7465, 2025, doi: 10.1038/s41467-025-62586-0.

[29] J. Ivković and J. L. Ivković, "Exploring the potential of new AI-enabled MCU/SOC systems with integrated NPU/GPU accelerators for disconnected Edge computing applications: towards cognitive SNN Neuromorphic computing," in *LINK IT & EdTech International Scientific Conference*, Belgrade, Serbia, 2023, pp. 26–27, doi: 10.1109/VLSI58301.2023.10191.

[30] A. Shafiee *et al*., "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, Seoul, Korea (South), 2016, pp. 14-26, doi: 10.1109/ISCA.2016.12.

[31] I. Barraj, H. Mestiri, and M. Masmoudi, "Overview of Memristor-Based Design for Analog Applications," *Micromachines*, vol. 15, no. 4, pp. 1–25, Apr. 2024, doi: 10.3390/mi15040505.

[32] S.-T. Wei, B. Gao, D. Wu, J.-S. Tang, H. Qian, and H.-Q. Wu, "Trends and challenges in the circuit and macro of RRAM-based computing-in-memory systems," *Chip*, vol. 1, no. 1, 2022, doi: 10.1016/j.chip.2022.100004.

[33] Y. H. Wei, Z. Wan, B. Crafton, S. Spetalnick, and A. Raychowdhury, “Characterization and Mitigation of ADC Noise by Reference Tuning in RRAM-Based Compute-In-Memory,” in *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, London, United Kingdom, 2025, pp. 1–5, doi: 10.1109/ISCAS56072.2025.11044056.

[34] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov, “4K-memristor analog-grade passive crossbar circuit,” *Nature Communications*, vol. 12, no. 5198, 2021, doi: 10.1038/s41467-021-25455-0.

[35] A. Antolini *et al.*, “The Role of Phase-Change Memory in Edge Computing and Analog In-Memory Computing: An Overview of Recent Research Contributions and Future Challenges,” *Sensors*, vol. 25, no. 12, pp. 1-18, 2025, doi: 10.3390/s25123618.

[36] L. Fick, S. Skrzyniarz, M. Parikh, M. B. Henry, and D. Fick, “Analog Matrix Processor for Edge AI Real-Time Video Analytics,” in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 260–262, doi: 10.1109/ISSCC42614.2022.9731773.

[37] Syntiant, “Syntiant unveils NDP200 Neural Decision Processor for always-on vision edge AI applications,” Business Wire. [Online]. Available: https://www.syntiant.com/news/syntiant-unveils-ndp200-neural-decision-processor-for-always-on-vision-edge-ai-applications. (Accessed: Sep. 22, 2021).

[38] S. Ambrogio *et al*., “An analog-AI chip for energy-efficient speech recognition and transcription,” *Nature*, vol. 620, pp. 768–775, 2023, doi: 10.1038/s41586-023-06337-5.

[39] S. Uemura, Y. Hiraoka, T. Kai, and S. Dosho, “Isolation Techniques Against Substrate Noise Coupling Utilizing Through Silicon Via (TSV) Process for RF/Mixed-Signal SoCs,” in *IEEE Journal of Solid-State Circuits*, vol. 47, no. 4, pp. 810-816, Apr. 2012, doi: 10.1109/JSSC.2012.2185169.

[40] M. J. Rasch, F. Carta, O. Fagbohungbe, and T. Gokmen, “Fast and robust analog in-memory deep neural network training,” *Nature Communications*, vol. 15, no. 1, pp. 1–15, Aug. 2024, doi: 10.1038/s41467-024-51221-z.

[41] M. Emani *et al*., “Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture,” in *Computing in Science & Engineering*, vol. 23, no. 2, pp. 114-119, 2021, doi: 10.1109/MCSE.2021.3057203.

[42] T. You, M. Zhao, Z. Fan, and C. Ju, “Emerging Memtransistors for Neuromorphic System Applications: A Review,” *Sensors*, vol. 23, no. 12, pp. 1–41, Jun. 2023, doi: 10.3390/s23125413.

[43] Aspinity, “Aspinity redefines always-on power efficiency with first analog machine learning chip,” Business Wire. [Online]. Available: https://www.businesswire.com/news/home/20220215005315/en/Aspinity-Redefines-Always-on-Power- Efficiency-with-First-Analog-Machine-Learning-Chip. (Accessed: Feb. 15, 2022).

[44] R. Martins, “Closing the gap between electrical and physical design steps with an analog IC placement optimizer enhanced with machine-learning-based post-layout performance regressors,” *Electronics*, vol. 13, no. 22, pp. 1-22, 2024, doi: 10.3390/electronics13224360.

[45] D. Basso, L. Bortolussi, M. Videnovic-Misic, and H. Habal, “Fast ML-driven Analog Circuit Layout using Reinforcement Learning and Steiner Trees,” in *Proceedings - 2024 20th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design, SMACD 2024*, Volos, Greece, 2024, pp. 1–4, doi: 10.1109/SMACD61181.2024.10745442.

[46] R. M. F. Martins, “A survey of machine and deep learning techniques in analog integrated circuit layout synthesis,” *Microelectronics*, vol. 1, no. 1, pp. 1-19, 2025, doi: 10.3390/microelectronics1010002.

[47] M. V. Örnhag, P. Güler, D. Knyaginin, and M. Borg, “Accelerating AI using next-generation hardware: Possibilities and challenges with analog in-memory computing,” *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, 2023, pp. 488-496, doi: 10.1109/WACVW58289.2023.00054.

[48] J. S. Seo *et al*., “Digital Versus Analog Artificial Intelligence Accelerators: Advances, trends, and emerging designs,” *IEEE Solid-State Circuits Magazine*, vol. 14, no. 3, pp. 65–79, 2022, doi: 10.1109/MSSC.2022.3182935.

[49] T. P. Xiao *et al.*, “On the Accuracy of Analog Neural Network Inference Accelerators,” *IEEE Circuits and Systems Magazine*, vol. 22, no. 4, pp. 26–48, 2022, doi: 10.1109/MCAS.2022.3214409.

[50] S. Alam, C. Yakopcic, Q. Wu, M. Barnell, S. Khan, and T. M. Taha, “Survey of Deep Learning Accelerators for Edge and Emerging Computing,” *Electronics*, vol. 13, no. 15, pp. 1–44, Jul. 2024, doi: 10.3390/electronics13152988.

[51] H. Liu, Z. Qian, W. Wu, H. Ren, Z. Liu, and L. Ni, “AFPR-CIM: An Analog-Domain Floating-Point RRAM -based Compute- In-Memory Architecture with Dynamic Range Adaptive FP-ADC,” in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Valencia, Spain: IEEE, Mar. 2024, pp. 1–6, doi: 10.23919/DATE58400.2024.10546882.

[52] A. Vasilopoulos *et al*., “A framework for analog-digital mixed-precision neural network training and inference,” *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, London, United Kingdom, 2025, pp. 1-5, doi: 10.1109/ISCAS56072.2025.11043537.

[53] J. Weis *et al.*, “Inference with Artificial Neural Networks on Analog Neuromorphic Hardware,” in *International Workshop on IoT, Edge, and Mobile for Embedded Machine Learning*, vol. 1325, 2020, pp. 201–212, doi: 10.1007/978-3-030-66770-2_15.

[54] J. Victor, C. Wang, and S. K. Gupta, “Comparative evaluation of memory technologies for synaptic crossbar arrays – Part 2: Design knobs and DNN accuracy trends,” *arXiv*, 2024, doi: 10.48550/arXiv.2408.05857.

BIOGRAPHIES OF AUTHORS

Kanchan S. Gorde is an Assistant Professor in the Department of Electronics Engineering at Terna Engineering College, Mumbai. She received her Bachelor's, Master's, and Ph.D. degrees in Electronics and Telecommunication Engineering from Sant Gadge Baba Amravati University, India. She also holds a Diploma in Electronics Engineering from the Maharashtra State Board of Technical Education (MSBTE). Her research interests include analog and mixed-signal VLSI design, biomedical image processing, machine learning, and embedded systems. With 20 years of teaching experience, she has authored or co-authored over 36 research publications, along with one registered copyright and three patents. She has been recognized as a Discipline Star and has received multiple Gold and Silver Medals in NPTEL courses. She was awarded the Best Paper Award at the IEEE-OPJU International Conference (2023). She can be contacted at email: kanchangorde@ternaengg.ac.in.

Sonali M. Sonavane has received her Bachelor's, Master's, and Ph.D. degrees in Computer Science and Engineering. She is currently serving as an Assistant Professor in the Department of Artificial Intelligence and Artificial Intelligence and Machine Learning at G. H. Raisoni College of Engineering and Management, Pune, Maharashtra. With 17 years of teaching experience, she has published over 21 research papers in reputed journals and conferences. Her contributions also include nine registered copyrights and three patents. She has received research grants from Savitribai Phule Pune University and NAAC Bangalore. Her research interests include information security, machine learning, and blockchain technology. She can be contacted at email: sonali.sonavane@raisoni.net.

Sonal Hutke received her Bachelor’s degree in Electronics and Telecommunication Engineering in 2005, followed by a Master's degree in 2009. She was awarded a Ph.D. in 2024 from Sant Gadge Baba Amravati University. Currently, she is serving as an Assistant Professor at SIES Graduate School of Technology, Mumbai. She has 18 years of teaching experience. She has received research grant from University of Mumbai. Her primary research interests include RF antennas and microwave devices, the internet of things (IoT) and artificial intelligence (AI), and FPGA-based system design. She is the author/co-author of 14 research publications and published 2 patents. She can be contacted at email: sonalj@sies.edu.in.

Ankush Hutke obtained his Bachelor’s degree in Computer Engineering in 2004 and completed his Master’s degree in 2013 from Sant Gadge Baba Amravati University. He is currently pursuing a Ph.D. from the University of Mumbai. Presently, he is working as an Assistant Professor at Rajiv Gandhi Institute of Technology, Mumbai with 19 years of teaching experience. His main research interests include machine learning and deep learning. He is the author/co-author of 19 research. He can be contacted at email: ankush.hutke@mctrgit.ac.in.