MANTECH Publications

Journal of Converging Technologies in Intelligent
Volume 1, Issue 3, September-December 2025

Neuromorphic Computing and Brain-Inspired AI Architectures for Energy-Efficient Intelligent Systems: A Review of Spiking Models, Hardware Platforms, and Edge Applications

Vishal R. Sharma[1], Neha V. Tyagi[2], Suraj S. Pal[3]

ABSTRACT

Artificial intelligence workloads are increasingly limited not by arithmetic capability but by the energy cost of moving data between separated memory and processing units. Neuromorphic computing addresses this limit by emulating the brain's event-driven, massively parallel, memory-centric organisation through spiking neural networks (SNNs) and co-located memory and compute. This review synthesises landmark peer-reviewed work on spiking neuron models, learning algorithms, digital and analogue neuromorphic hardware, and edge applications. Five hardware platforms are compared within a common framework, and a first-order analytical model is derived to identify when spike-based computation saves energy relative to dense artificial neural networks (ANNs). Using published 45 nm operation energies, the model gives a break-even of about 5.1 spike events per synapse per inference, with roughly tenfold savings at 0.5 events. Reported platforms consume about 24–26 pJ per synaptic event, whereas a 32-bit off-chip memory access costs about 640 pJ. Training maturity, benchmarking and programming tools remain the principal barriers to adoption.

KEYWORDS: *Neuromorphic computing, Brain-inspired AI, Spiking neural networks, Energy efficiency, In-memory computing, Event-based sensing, Edge intelligence, Memristor, Loihi, TrueNorth*

INTRODUCTION

Modern artificial intelligence delivers remarkable accuracy, yet its energy demand is growing faster than the value it creates at the point of deployment. Training one large natural-language

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

model with neural architecture search was estimated to emit about 626,000 lb of carbon dioxide equivalent, roughly five times the lifetime emissions of an average car [1], and inference at scale compounds the burden. The root cause is architectural. In a von Neumann machine, weights and activations reside in off-chip DRAM, and a single 32-bit DRAM access consumes about 640 pJ, nearly two orders of magnitude more than a 32-bit floating-point multiplication (3.7 pJ) at 45 nm [2]. Arithmetic is cheap; communication is expensive. The human brain, by contrast, performs perception, planning and motor control within roughly 20 W by combining sparse, asynchronous spiking activity with synapses that store and process information in the same physical location [3], [4].

The problem addressed here is therefore stated as follows: how can brain-inspired architectures be designed, trained and deployed so that intelligent systems retain useful accuracy while consuming far less energy, and under what conditions is that advantage real rather than nominal? Neuromorphic computing, a term introduced by Mead [3], answers with three principles: information carried by discrete spikes, state and memory co-located within neurons and synapses, and computation triggered only by events [5]. Figure 1 contrasts the two organisations. This paper reviews models, hardware and applications, and contributes an analytical model that makes the sparsity condition for energy savings explicit.

The urgency is sharpened by the shift of intelligence towards the network edge. Sensors in wearables, vehicles and industrial plants generate continuous streams that cannot always be sent to the cloud because of latency, bandwidth and privacy limits, and they must be processed within milliwatt budgets. Scaling transistor density no longer lowers energy per operation as quickly as it once did, so further gains must come from architectural change rather than from process shrinkage alone [2]. The remainder of the paper follows the structure of a review: background, gap, objectives, methodology, results, discussion, limitations, future scope and conclusions.

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

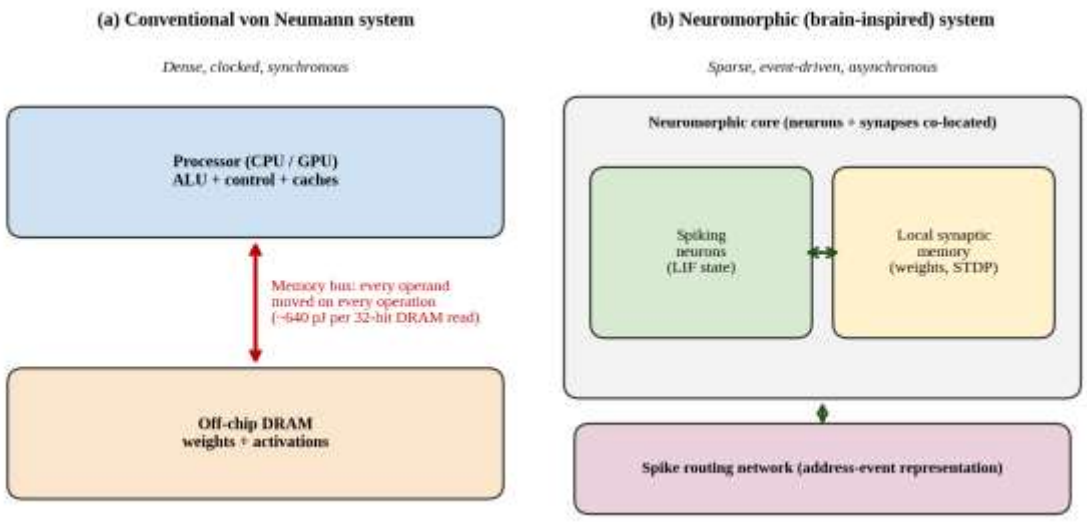


Figure 1. Architectural contrast between (a) a conventional von Neumann system with a memory-bus bottleneck and (b) a neuromorphic core with co-located neurons and synaptic memory connected by spike routing

LITERATURE REVIEW

Spiking Neuron Models And Learning Rules

Maass classified SNNs as the third generation of neural models, showing that they are at least as computationally powerful as sigmoidal networks while encoding information in spike timing [6]. Learning follows three routes. Local plasticity rules such as spike-timing-dependent plasticity (STDP), observed experimentally by Bi and Poo [7], adjust synapses according to the relative timing of pre- and postsynaptic spikes and suit unsupervised on-chip learning. Surrogate-gradient methods replace the non-differentiable spike function with a smooth proxy during backpropagation, allowing deep SNNs to be trained end to end [8]. ANN-to-SNN conversion maps trained weights onto rate-coded spiking equivalents with small accuracy loss, but typically needs many time steps [9]. Both Roy et al. and Schuman et al. identify the training–latency–energy trade-off as the central open algorithmic question [4], [5]. The leaky integrate-and-fire (LIF) neuron is the dominant choice in hardware because it needs only one state variable and one threshold comparison per update, whereas richer models capture bursting and adaptation at higher cost [4]. Designers therefore trade biological fidelity against area and energy, and most reported energy figures refer to LIF-class neurons.

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Neuromorphic Hardware Platforms

IBM's TrueNorth integrated 4,096 cores, one million neurons and 256 million synapses in 28 nm CMOS at about 65 mW [10]. Intel's Loihi added programmable on-chip learning, with 130,000 neurons per chip and 23.6 pJ per synaptic spike operation [11]; a later survey documented large energy-delay advantages on sparse, event-driven workloads such as optimisation, sparse coding and adaptive control [12]. SpiNNaker took a software-defined route, combining ARM cores with a multicast packet-routing fabric to support flexible neuron models at million-core scale [13]. Tianjic showed that heterogeneous cores can run ANNs and SNNs concurrently [14]. Indiveri and Liu argue that the decisive design choice is how memory is organised close to computation [15]. Hardware–algorithm co-design is increasingly emphasised: neuron precision, synaptic bit width, connectivity limits and routing capacity constrain which network topologies map efficiently, and a mismatch can erase theoretical sparsity gains [5], [15].

Emerging Devices And Event-Based Sensing

Memristive crossbars, made credible by the physical demonstration of the memristor [16], perform analogue multiply–accumulate operations inside memory [17]; device variability, conductance drift and limited endurance remain obstacles. On the sensing side, event cameras emit asynchronous brightness-change events with microsecond resolution and high dynamic range, forming a natural front end for SNNs [18].

RESEARCH GAP

Existing surveys treat algorithms [4], [5], hardware [15], [12] or devices [17] largely in isolation, and energy claims are usually quoted per operation without stating the activity conditions under which a spiking system actually beats a dense accelerator. Three gaps follow: (i) no unified comparison linking algorithmic sparsity to hardware energy; (ii) no explicit break-even criterion expressed in time steps and firing probability; and (iii) limited discussion of system-level overheads, including memory, routing and sensing, that erode headline gains.

OBJECTIVES

- To synthesise spiking neuron models, learning rules and training strategies relevant to energy-efficient inference.

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

- To compare representative neuromorphic hardware platforms on scale, process technology, energy and programmability.
- To derive and evaluate a first-order analytical model of the energy break-even between spiking and dense networks.
- To map demonstrated application domains and identify the barriers to deployment.

METHODOLOGY

Review Design And Selection Criteria

A structured narrative review was conducted. Sources were drawn from peer-reviewed journals and proceedings published between 1990 and 2022 and were retained if they reported quantitative algorithmic or hardware results, established a foundational design principle, or provided an authoritative survey. Unreviewed vendor claims were excluded from quantitative comparison. Each retained study was assessed against the evaluation framework in Table 1.

Table 1. Evaluation Framework Applied to Reviewed Studies

Dimension	Metric	Evidence base
Learning	Rule, time steps T, accuracy loss	[8], [9]
Hardware scale	Neurons, synapses, node	[10], [11], [13]
Energy	pJ per event, power (W)	[2], [10]
Programmability	On-chip learning, neuron models	[11], [14]
Applications	Task, benchmark, reported gain	[12], [18]

Analytical Energy Model

To isolate the conditions for spike-based savings, a first-order model was formulated for one layer with N synaptic connections. A dense ANN performs one multiply–accumulate (MAC) per synapse, whereas an SNN performs one accumulate (AC) per arriving spike over T time steps with mean spike probability r:

$$E_{ANN} = N \times E_{MAC}, \quad E_{SNN} = N \times T \times r \times E_{AC}, \quad R = E_{SNN} / E_{ANN} = (T \times r \times E_{AC}) / E_{MAC} \qquad (1)$$

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Spiking is advantageous when $T \times r < E_{MAC}/E_{AC}$. Using 45 nm, 32-bit floating-point energies of 3.7 pJ (multiply) and 0.9 pJ (add) [2], E_{MAC} = 4.6 pJ and E_{AC} = 0.9 pJ, giving a break-even of $T \times r \approx 5.1$. Memory access, spike routing and leakage are excluded, so the model indicates an upper bound on achievable savings. Sensitivity was examined by varying T from 5 to 50 time steps and r from 10^{-3} to 1, spanning long-window conversion regimes through to low-latency directly trained networks. Platform energy values were taken as reported by the original authors and were not rescaled between process nodes, because normalisation requires circuit-level details that are rarely published. All derived ratios were computed directly from the quoted values.

RESULTS AND FINDINGS

Hardware Platform Comparison

Table 2 compares representative platforms. Two design philosophies emerge. Fixed-function digital designs such as TrueNorth minimise energy at the cost of flexibility, whereas programmable designs such as Loihi and SpiNNaker support on-chip learning and diverse neuron models at higher cost per event. Tianjic trades some specialisation for hybrid ANN–SNN execution.

Table 2. Comparison of Representative Neuromorphic Platforms

Platform	Node	Scale	Energy	Key feature
TrueNorth [10]	28 nm	1 M neurons, 256 M synapses	≈65 mW; 26 pJ/event	Fixed digital cores
Loihi [11]	14 nm	130 k neurons, 130 M synapses	23.6 pJ/synaptic op	On-chip learning
SpiNNaker [13]	130 nm	18 ARM cores per chip	Software-defined	Flexible neuron models
Tianjic [14]	28 nm	≈40 k neurons, ≈10 M synapses	Not compared	Hybrid ANN–SNN
Memristive array [16], [17]	Research	Array level	Analogue in-memory MAC	Compact; variable devices

Energy Per Operation

Figure 2 contrasts operation energies. A 32-bit DRAM read (640 pJ) costs 24.6 times a TrueNorth synaptic event (26 pJ) and 27.1 times a Loihi synaptic operation (23.6 pJ) [2], [10],

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

[11]. The comparison is not like-for-like, since a synaptic event includes a local memory fetch while the multiplier figure excludes operand movement, yet it shows that eliminating off-chip traffic, rather than faster arithmetic, drives the saving.

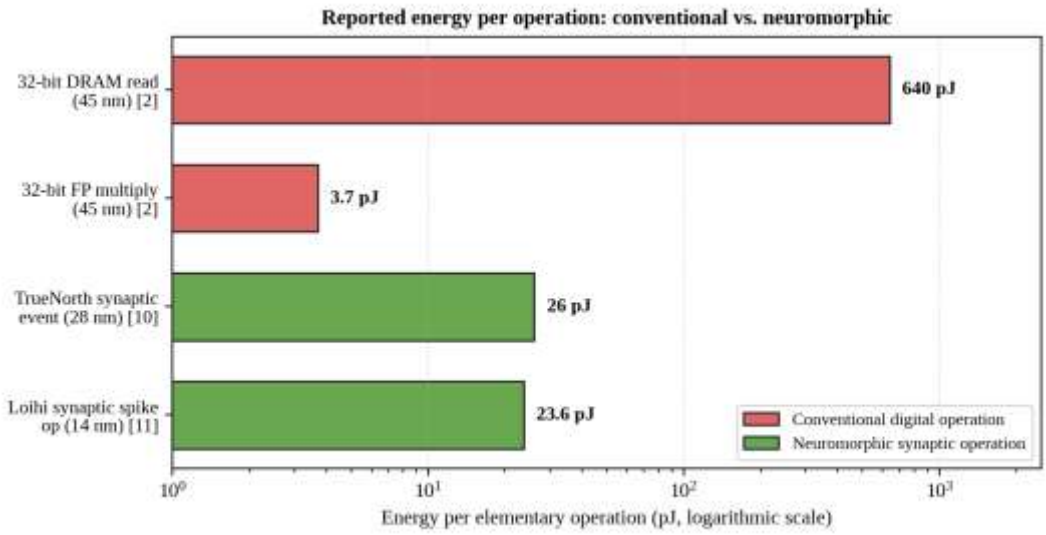


Figure 2. Reported energy per elementary operation: conventional digital operations (red) versus neuromorphic synaptic operations (green), logarithmic scale

Break-Even Analysis

Figure 3 plots the energy ratio R against r for T = 5, 10, 20 and 50. At T = 10 and r = 0.05 (T × r = 0.5), R = 0.098, a 10.2-fold saving. At T = 20 and r = 0.10, R = 0.39, a 2.6-fold saving. At T = 50 and r = 0.20, R = 1.96, and the SNN consumes almost twice the energy of the dense network. Savings therefore require short inference windows and sparse activity, which explains why conversion methods needing long windows weaken the benefit, whereas surrogate-gradient training with few time steps strengthens it [9], [8].

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

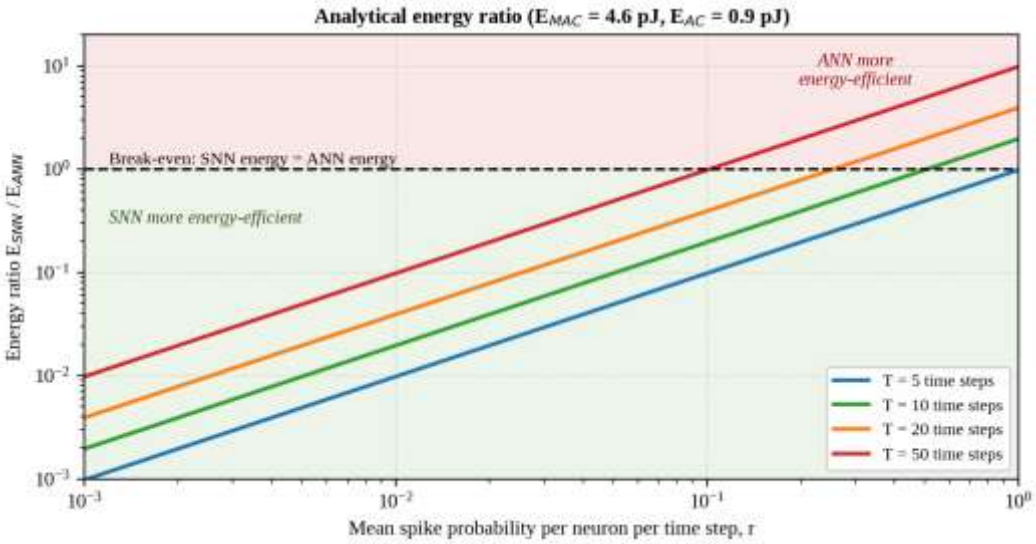


Figure 3. Analytical energy ratio E_{SNN}/E_{ANN} versus mean spike probability r for four inference window lengths T; dashed line marks break-even (R = 1)

Application Synthesis

Table 3 maps application domains to reported evidence. Benefits are strongest where data are naturally sparse in time, as with event-camera streams, always-on audio and closed-loop control.

Table 3. Application Domains, Reported Benefits and Challenges

Domain	Evidence	Reported benefit	Challenge
Event-based vision	Event cameras with SNNs [18]	Low latency, high dynamic range	Few labelled event datasets
Adaptive control	Loihi robotics [12]	Low-power closed-loop learning	Task-specific tuning
Keyword spotting	Always-on audio [12]	Low standby energy	Accuracy gap to ANNs
Sparse optimisation	Loihi solvers [12]	Large energy-delay gains	Problem mapping
Edge inference	Spike-based edge AI [4]	Battery-friendly operation	Immature toolchains

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Summary Of Findings

Three findings emerge. First, energy per synaptic event is now in the tens of picojoules on both fixed-function and programmable digital platforms, more than an order of magnitude below an off-chip memory access [10], [11]. Second, the sparsity condition $T \times r < 5.1$ is met by event-driven data but not by dense frame-based inputs, so the input representation matters as much as the processor. Third, programmability and efficiency remain in tension: the most energy-efficient platforms offer the least flexibility, while the most flexible depend on learning rules and toolchains that are still maturing [13].

DISCUSSION

The central finding is that neuromorphic advantage is conditional, arising from three mechanisms acting together: removal of off-chip memory traffic, activity sparsity, and event-driven sensing. Figure 1 summarises the architectural contrast. Memory co-location addresses the dominant energy term identified in Figure 2, while sparsity determines whether accumulate-only arithmetic outweighs the repeated time-step cost quantified in Figure 3. A design that achieves only one of these, for example a dense SNN with long inference windows, gains little over a quantised ANN accelerator.

The engineering significance lies in always-on, battery-constrained systems: wearable health monitors, implantable devices, drones, industrial condition monitoring and in-sensor processing, where milliwatt budgets exclude conventional accelerators. Scientifically, local learning rules such as STDP offer a route to continual, on-device adaptation that backpropagation-based systems struggle to provide [7], [11]. The principal obstacle is no longer silicon but software: fragmented toolchains, the absence of shared benchmarks and an accuracy gap on large-scale tasks continue to restrict adoption [5], [4].

Comparison with quantised dense accelerators is essential. Reduced-precision ANN hardware lowers E_{MAC} and removes much of the arithmetic penalty captured in Equation (1), so the neuromorphic case rests increasingly on data movement and sensing rather than on arithmetic substitution alone. Hybrid approaches, such as the heterogeneous cores of Tianjic [14], may therefore prove more practical than purely spiking systems: dense layers handle feature extraction while spiking layers handle sparse temporal streams. For sustainability, lowering

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

inference energy at the edge reduces the aggregate footprint of billions of deployed devices, complementing efforts to curb training emissions [1].

LIMITATIONS

- The analytical model uses 45 nm, 32-bit floating-point energies and omits memory, routing and leakage; quantised accelerators alter the E_{MAC}/E_{AC} ratio and therefore the break-even point.
- Reported energy figures originate from different process nodes and workloads, so cross-platform comparisons are indicative rather than definitive.
- The review is narrative rather than a formal meta-analysis, may favour high-profile platforms, and includes no new hardware measurements.
- Application evidence derives mainly from research-laboratory demonstrations on small benchmarks; performance on production-scale workloads and under real-world noise is not yet established.

FUTURE SCOPE

- Standardised benchmarking of accuracy, latency and energy on shared tasks across neuromorphic and conventional accelerators.
- Co-design of low-time-step SNN training with sparsity-aware hardware, including hybrid ANN–SNN execution.
- Variability-tolerant learning algorithms for memristive in-memory computing at array scale.
- Open compilers and toolchains supporting continual, on-device learning in sensor-rich edge systems.
- Lifecycle assessment of neuromorphic deployments that includes fabrication, sensing and software overheads alongside operating energy.

CONCLUSION

Neuromorphic computing offers a credible route to energy-efficient intelligent systems, but its advantage is conditional rather than automatic. Reported platforms reach about 24–26 pJ per synaptic event against roughly 640 pJ for an off-chip memory access, and the analytical model shows that spiking networks save energy only while $T \times r$ remains below about 5.1, with savings near tenfold at $T \times r = 0.5$. Realising these gains in practice depends on short-window training methods, sparsity-aware hardware, standardised benchmarks and mature software

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

tools rather than on further scaling of neuron counts alone. These findings position neuromorphic systems as complementary accelerators for sparse, temporal, always-on workloads rather than universal replacements for dense deep-learning hardware.

REFERENCES

1. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 2019; 3645–3650.
2. Horowitz M. Computing's energy problem (and what we can do about it). *IEEE International Solid-State Circuits Conference (ISSCC).* 2014; 10–14.
3. Mead C. Neuromorphic electronic systems. *Proceedings of the IEEE.* 1990; 78(10): 1629–1636.
4. Roy K, Jaiswal A, Panda P. Towards spike-based machine intelligence with neuromorphic computing. *Nature.* 2019; 575: 607–617.
5. Schuman CD, Kulkarni SR, Parsa M, Mitchell JP, Date P, Kay B. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science.* 2022; 2: 10–19.
6. Maass W. Networks of spiking neurons: the third generation of neural network models. *Neural Networks.* 1997; 10(9): 1659–1671.
7. Bi G, Poo M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience.* 1998; 18(24): 10464–10472.
8. Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine.* 2019; 36(6): 51–63.
9. Rueckauer B, Lungu IA, Hu Y, Pfeiffer M, Liu SC. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience.* 2017; 11: 682.
10. Merolla PA, Arthur JV, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science.* 2014; 345(6197): 668–673.
11. Davies M, Srinivasa N, Lin TH, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro.* 2018; 38(1): 82–99.

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved

12. Davies M, Wild A, Orchard G, Sandamirskaya Y, Guerra GAF, Joshi P, Plank P, Risbud SR. Advancing neuromorphic computing with Loihi: A survey of results and outlook. *Proceedings of the IEEE.* 2021; 109(5): 911–934.
13. Furber SB, Galluppi F, Temple S, Plana LA. The SpiNNaker project. *Proceedings of the IEEE*. 2014; 102(5): 652–665.
14. Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature.* 2019; 572: 106–111.
15. Indiveri G, Liu SC. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE.* 2015; 103(8): 1379–1397.
16. Strukov DB, Snider GS, Stewart DR, Williams RS. The missing memristor found. *Nature.* 2008; 453: 80–83.
17. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E. Memory devices and applications for in-memory computing. *Nature Nanotechnology*. 2020; 15: 529–544.
18. Gallego G, Delbrück T, Orchard G, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2022; 44(1): 154–180.

***Author for Correspondence:**
Vishal R. Sharma
E-mail: vishal.r.sharma@outlook.com

[1]Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow
[2]Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow

Received Date: September 16, 2025
Accepted Date: September 29, 2025
Published Date: October 14, 2025

Citation: Vishal R. Sharma, Neha V. Tyagi, Suraj S. Pal. Neuromorphic Computing and Brain-Inspired AI Architectures for Energy-Efficient Intelligent Systems: A Review of Spiking Models, Hardware Platforms, and Edge Applications. Journal of Converging Technologies in Intelligent. 2025; 1(3): 162-173.

Page 162-173 © MANTECH PUBLICATIONS 2025. All Rights Reserved