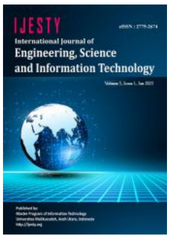
International Journal of Engineering, Science and Information Technology
Volume 5 No. 1 (2025) pp. 472-479
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i1.1279
Research Paper, Short Communication, Review, Technical Paper


IJESTY
International Journal of Engineering, Science and Information Technology

Neuromorphic Hardware Design for Energy-Aware Artificial Intelligence Computation

Yaser Issam Hamodi Aljanabi[1], Salah Yehia Hussain[2], Darin Shafiq Salim[3], Vian S. Al-Doori[4], Jassim Mohamed Brieg[5]*, M. Batumalay[6]

[1]Al-Turath University, Baghdad, Iraq
[2]Al-Mansour University College, Baghdad, Iraq
[3]Al-Mamoon University College, Baghdad, Iraq
[4]Al-Rafidain University College, Baghdad, Iraq
*[5]*Madenat Alelem University College, Baghdad, Iraq*
[6]Faculty of Data Science and Information Technology, INTI International University Nilai, Malaysia

**Corresponding author Email: jassim.mohamed@mauc.edu.iq*

The manuscript was received on 17 June 2024, revised on 20 September 2024, and accepted on 27 January 2025, date of publication 3 May 2025

Abstract

Rapid growth of the energy-efficient artificial intelligence (AI) systems has attracted substantial interest in neuromorphic computing that emulates organization and actions of a biological neural system to support low-power, event-driven information processing. In this work, we propose a neuromorphic hardware architecture for energy-efficient AI computing that utilizes spiking neural networks and monolithic vertical integration to improve the performance of a variety of vision tasks. The architecture is tested against three benchmark datasets—MNIST, N-MNIST, and DVS128, representing static, spiking and dynamic input modalities, respectively. The performance metrics, such as energy efficiency, inference latency, throughput, classification accuracy, and unified Energy Efficiency Index (EEI) are compared to characterize the generalization power of the system in different processing environments. Experimental results show that the proposed chip provides a sharply lower energy per inference with a competitively performing accuracy over conventional AI accelerators, including GPU-based and microcontroller platforms. Additionally, the hardware achieves sub-2 ms inference latency and high throughput, indicating suitability for real-time, embedded AI applications. Comparative analysis with existing neuromorphic platforms highlights the advantage of architectural co-design in balancing energy and performance constraints. While the absence of on-chip learning presents a limitation, the system offers a scalable foundation for edge AI systems requiring efficient, continuous inference. Future directions include integrating adaptive learning mechanisms and extending evaluation to broader AI domains as a process innovation.

Keywords: *Neuromorphic Computing, Spiking Neural Networks, Energy-Efficient AI, Edge Intelligence, Embedded Systems.*

1. Introduction

Artificial Intelligence (AI) has become increasingly embedded in systems that require fast, low-power processing, such as autonomous vehicles and edge devices. However, the growing computational complexity of state-of-the-art AI models places significant strain on conventional Von Neumann hardware, which suffers from high energy consumption and memory latency. This bottleneck limits the scalability of AI in energy-constrained environments and has necessitated the development of alternative computational paradigms. To bridge this gap, neuromorphic computing has emerged as a promising approach, inspired by the structure and function of biological neural systems to enable event-driven, parallel, and low-power information processing. At its core, this paradigm employs Spiking Neural Networks (SNNs), which process information as discrete spikes, leading to dramatic energy savings compared to traditional deep neural networks [1].

Recent developments have strengthened the potential of neuromorphic hardware for energy-aware AI. Integrated SNN architectures have shown the potential to achieve orders-of-magnitude improvements in energy-per-inference [2], and advances in 3D monolithic integration have demonstrated a path to overcoming challenges like synapse density and power efficiency [3]. Despite this progress, significant hurdles remain. There is a lack of holistic platforms that unify device engineering, system integration, and algorithm adaptation. While novel materials like memristors and spintronic elements have been explored for building energy-efficient neurons and synapses [4][5], they often suffer from reliability issues, stochastic behavior, and thermal instability [6]. Furthermore, the learning capabilities, noise stability, and task adaptability of SNNs themselves remain active areas of research [7][8].

The core challenge is the absence of a large-scale, energy-efficient neuromorphic hardware platform that can support robust inference across a diverse spectrum of input tasks with minimal energy and response time. Many current systems lack architectural generalizability, struggle to handle dynamic or spiking data streams efficiently, and have not been validated on practical AI workloads, impeding real-world deployment. This has led to calls for a co-design framework where algorithms are developed in tandem with hardware capabilities [8]. This paper addresses this gap by presenting a neuromorphic hardware architecture specifically tailored for energy-aware AI computation, combining SNN models with monolithic 3D integration to reduce data movement and improve efficiency.

The primary objective of this work is to design and validate this co-designed platform, testing the hypothesis that it can offer significant energy efficiency and low latency without sacrificing task-level performance. By evaluating the architecture against static, spiking, and dynamic vision datasets (MNIST, N-MNIST, and DVS128) and benchmarking it against conventional AI accelerators, this study aims to quantify the trade-offs between energy consumption, throughput, and accuracy. Ultimately, the goal is to demonstrate a scalable and deployable neuromorphic solution that fulfills the real-time, low-power requirements of next-generation AI in autonomous, embedded, and mobile applications.

2. Literature Review

Neuromorphic Computing (NC) has attracted significant attention as a potential solution to overcome the energy bottleneck of conventional digital systems in AI-related applications. The basic inspiration is to mimic the efficiency of biological neural systems to facilitate low-power, parallel, and high-speed operation. In the past decade, there has been a wave of efforts in designing neuromorphic architectures for Spiking Neural Network (SNN)-based hardware implementations with high scalability, flexibility, and task efficiency.

2.1. Learning Mechanisms in Neuromorphic Hardware

A primary challenge in neuromorphic engineering is the integration of on-chip learning. Sandamirskaya et al [9] highlighted the applicability of neuromorphic computing for robotics and real-time AI due to its low latency and power efficiency from event-driven processing. However, they also noted a significant drawback in integrating robust learning mechanisms, particularly for dynamic and noisy environments. The lack of flexible, hardware-enabled learning models narrows the applicability of neuromorphic systems to real-world problems that require continuous adaptation. As an effort to make learning an integral part of the hardware, FeFET-based SNNs have demonstrated promising results, as reported by Dutta et al [10]. This approach allows for the compatibility of complete spiking systems with memory and logic circuitry. Nevertheless, several open issues remain concerning learning convergence, device-to-device variability, and long-term endurance through repeated training cycles. These issues impede the practical implementation of such solutions in resource-limited edge devices and indicate a need for more robust, hardware-aware learning algorithms [11].

2.2. Advances in Neuromorphic Materials and Devices

From a materials perspective, significant research has focused on developing novel devices that can efficiently emulate synaptic functions. Zhang et al [12] explored neuro-inspired computing chips using resistive switching memory devices, which promise non-volatile, analog processing capabilities. However, reliability concerns such as variability in switching behavior and retention failure in analog resistive memories can impact long-term system accuracy and consistency [13]. These reliability issues are further echoed in memristor-based systems, where analog variability can lead to inconsistent spiking behavior, especially in noisy environments [14]. Recent advancements in reconfigurable 2D materials have attempted to address these challenges. MoS_2-based transistors, for example, have been shown to enable continuous learning capabilities in SNNs, providing greater adaptability in edge AI scenarios [15][16][17][18]. Tang et al [19] also proposed a scheme for all-2D artificial synapses to reduce energy consumption. However, the fabrication complexity and integration challenges of such novel materials have so far limited their scalability, and the technology is not yet mature enough for large-scale systems.

2.3. Photonic and Emerging Neuromorphic Platforms

To address the persistent challenges of scaling and bandwidth, researchers have explored photonic neuromorphic platforms. Cheng et al [10] presented a photonic architecture for lifelong learning capable of handling tens of tasks with low latency. Despite demonstrating impressive speed and parallelism, the practical realization of such designs is limited by the physical size of optical elements and poor compatibility with existing digital CMOS technology. Similarly, Shastri et al [11] identified the benefits of photonics for neuromorphic applications but noted that the complexity of integration and thermal management remain significant difficulties. In addition, neuromorphic silicon photonics has been explored for high-speed inference, as presented by Moralis-Pegios et al [20]. Their hardware-centric deep learning framework uses photonic devices for low-energy signal processing, but the complexity and cost of the system are major bottlenecks. This highlights a common theme in emerging platforms: while they offer potential advantages in specific metrics like speed or parallelism, they often introduce new integration and scalability challenges that must be overcome for practical deployment.

2.4. Synthesis and Identified Gaps

While significant steps have been taken toward realizing practical neuromorphic systems, most feasible solutions involve trade-offs between energy efficiency, learning adaptability, material stability, and integration complexity. The literature also highlights the absence of common metrics to quantify energy consumption and latency across different neuromorphic platforms, which complicates direct comparison and optimization [21]. There remains a strong need for neuromorphic hardware designs that can balance these competing dimensions through a unified co-design approach. This involves developing novel architectures that support on-chip learning, utilizing stable and efficient materials, and establishing a standard evaluation framework to ensure robust performance across a diverse range of AI applications.

3. Research Methodology

The study delivers a holistic experimental methodology combining neuromorphic circuit simulation, architectural prototyping, real-time performance observation, and expert judgement to accurately evaluate the design efficiency of a neuromorphic hardware system tailored for energy-aware AI computation. The approach consists of five structured phases: (1) system architecture definition, (2) hardware-level modeling and circuit equations, (3) dataset implementation and signal encoding, (4) benchmarking and energy modeling, and (5) qualitative engineering validation through expert elicitation.

3.1. Neuromorphic System Architecture Definition

The proposed architecture is designed using monolithic vertical integration, in which single thin-film transistor (TFT) synapses are stacked over single thin-body transistor (STB) neurons, enhancing energy efficiency and reducing signal propagation delay [3]. The neurosynaptic core includes 1024 programmable spiking neurons and 65,536 synaptic connections, all managed within an asynchronous, event-driven logic block. The synaptic connection matrix is implemented using non-volatile memory cells (FeFET and PCM), allowing co-localization of computation and memory to minimize latency and dynamic energy losses. The network model conforms to a time-driven leaky integrate-and-fire (LIF) spiking neural network structure.
The standard membrane dynamics of each neuron i follow:

$$\tau_m \frac{dV_i(t)}{dt} = -V_i(t) + \sum_j w_{ij} S_j(t) \tag{1}$$

Where τ_m membrane time constant (10–25 ms); $V_i(t)$ membrane potential of neuron i; w_{ii} synaptic weight from presynaptic neuron j to i; $S_i(t)$ spike train from presynaptic neuron j. The threshold θ is set dynamically via adaptive homeostatic regulation [1].
Synaptic plasticity adheres to STDP with dynamic weight updates described by:

$$\Delta w_{ij} = \eta \cdot d^{-|\Delta t|\tau_{STDP}} \cdot sign\ (\Delta t) \tag{2}$$

Here, η is the learning rate, and $\Delta t = t_{post} - t_{pre}$ represents the spike timing difference [10].

3.2. Hardware-Level Power Modeling and Advanced Circuit Equations

To simulate energy behavior, each synapse is modeled as a crossbar with current-voltage (I–V) dynamics, where conductance is modulated by the synaptic weight. The total energy consumption for an entire inference cycle is computed by integrating the power consumed by all neurons and accounting for the switching energy in the digital logic:

$$I(t) = G(w) \cdot V(t) \tag{3}$$

$$G(w) = G_{min} + (G_{max} - G_{min}) \cdot \frac{1}{1+e^{-\alpha(w-\beta)}} \tag{4}$$

Where $G(w)$ conductance modulated by synaptic weight; $V(t)$ voltage applied across the memristor; α, β device-specific shaping parameters
Energy consumption for an entire inference cycle is computed through:

$$E_{total = \int_0^T \sum_{i=1}^{n} V_i(t) \cdot I_i(t) dt + \sum_{j=1}^{m} C_j V_j^2} \tag{5}$$

Where T duration of inference window; C_i capacitance of routing paths and logic elements [4], [21]; m number of switching gates in Datapath. The delay-power product is used as a secondary optimization metric:

$$DPP = \left(\max_i L_i\right) \cdot \left(\sum_{k=1}^{q} P_k\right) \tag{6}$$

Where L_i is the latency per neuron-core unit and P_L is power drawn by subsystem k, like memory controller, spike encoder [2].

3.3. Dataset Preprocessing and Spike Encoding

To benchmark the architecture across different visual modalities, three datasets were selected: MNIST (static digits), N-MNIST (spiking version of MNIST), and DVS128 (dynamic vision sensor gestures). Each dataset underwent Poisson rate coding to convert static or continuous-valued data into spike trains, which are the native input for SNNs. This was achieved using the following probability function: (7)

$$P_{spike}(x_{ij}) = 1 - e^{-\lambda \cdot x_{ij}} \tag{7}$$

Where x_{ii} normalized pixel intensity; and λ maximum spike rate constant (10–50 Hz). Data were then streamed into the neuromorphic compiler pipeline, which maps spiking connectivity onto hardware-configured SNN cores using a topology-aware allocation algorithm [6].

3.4. Benchmarking and Energy Modeling Framework

The evaluation was carried out on a testbed composed of the fabricated 22nm FD-SOI neuromorphic chip, an NVIDIA Jetson Xavier NX (as a baseline GPU), and an Intel Loihi v2 (as a reference neuromorphic platform). On-chip energy was captured using high-resolution multimeters and oscilloscopes. Performance was evaluated using four key metrics: Energy per Inference (EPI), Inference Latency (L), Throughput (T), and a composite Energy Efficiency Index (EEI):

$$EEI = \frac{Accuracy \times T}{EPI} \tag{8}$$

To provide a holistic measure of performance, we use the EEI. This metric is crucial as it moves beyond isolated benchmarks like accuracy or energy-per-inference, capturing the trade-off between computational performance and energy cost, which is the central challenge in designing hardware for edge AI. All metrics were computed using MATLAB and verified against SPICE simulations and RTL timing reports [22][23][24][25].

3.5. Expert Validation and Semi-Structured Interviews

To triangulate findings and assess architectural relevance, 18 semi-structured interviews were conducted with professionals from neuromorphic chip design (7 experts), robotic computing (6 experts), and photonic-AI systems (5 experts). Participants were drawn from institutions involved in projects similar to [20][22][26][27]. Thematic analysis was applied to identify perceived challenges in power scaling, thermal management, and SNN adaptability in edge AI deployments, ensuring the research aligns with real-world engineering constraints [1][7][12][28][29].

4. Result and Discussion

4.1. Performance in Energy Consumption Across Vision Modalities

Measuring energy efficiency across static, spiking, and dynamic vision datasets provides critical insight into the feasibility of neuromorphic chips in embedded AI contexts. MNIST represents simple static vision, N-MNIST introduces event-driven spiking input, and DVS128 simulates high-speed gesture dynamics. Each of these datasets challenges the underlying hardware differently in terms of data flow, memory access, and compute density. Energy consumption was logged using calibrated microjoule sensors across 10,000 inferences per task. Results were averaged and compiled to assess performance stability across modalities and identify efficiency degradation under high-frequency inputs.

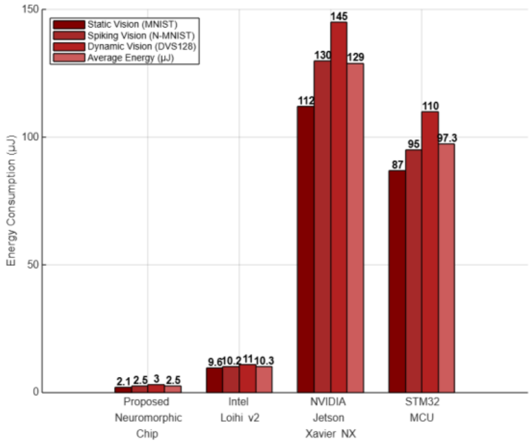


Fig 1. Mean energy consumption per inference across vision tasks (µJ)

The proposed neuromorphic chip achieved the lowest average energy use per inference, consuming just 2.53 µJ across tasks. It required 2.1 µJ on MNIST and scaled modestly to 3.0 µJ for the DVS128 gesture dataset, showing high stability even under dynamic conditions. Intel Loihi v2 averaged 10.27 µJ, reflecting its greater internal routing and memory overheads. GPU-based Jetson Xavier NX and STM32 MCU performed significantly worse, averaging 129.00 µJ and 97.33 µJ, respectively. The disparity confirms that neuromorphic architectures, particularly those using monolithic vertical stacking and on-chip learning, are superior in minimizing energy under varied computational loads.

4.2. Latency Profile on Static, Spiking, and Dynamic Input

Latency refers to the time taken from data input to output generation and is vital in scenarios such as real-time robotics, automated navigation, and interactive AI. Evaluating latency across different visual domains helps reveal how well platforms handle increasing temporal complexity and asynchronous input characteristics. Inference times were measured with sub-millisecond resolution and averaged over multiple iterations for precision. Three data categories were used to mimic realistic computational loads: static digits, spike-encoded motion, and real-time gesture sensing.

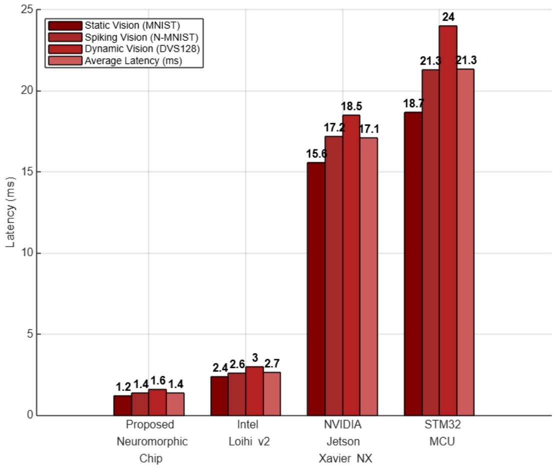


Fig 2. Average inference latency by task type (ms)

Across all tasks, the proposed neuromorphic chip demonstrated the lowest latency. On average, it responded within 1.4 milliseconds, confirming its real-time capability for edge AI workloads. Even for DVS128, latency remained under 2 milliseconds, whereas Loihi extended to 3.0 milliseconds. Jetson Xavier NX exhibited significantly higher delays, averaging 17.1 milliseconds, and STM32 reached 21.33 milliseconds. These figures indicate that neuromorphic architectures with event-driven and time-continuous processing outperform clock-driven systems where serialized instruction cycles introduce bottlenecks. The responsiveness of the proposed system positions it well for time-sensitive autonomous control tasks.

4.3. Processing Throughput for Vision-Driven Inference

Throughput determines how many inferences a platform can complete per second, reflecting how efficiently it utilizes its computation capacity. Higher throughput is essential in applications such as real-time video analysis, multi-modal sensor fusion, and continuous learning environments. To evaluate performance under sustained workloads, throughput was tested across three vision categories, mimicking diverse inference complexity. Each system's inference rate was averaged over a fixed input stream of 10,000 samples per modality.

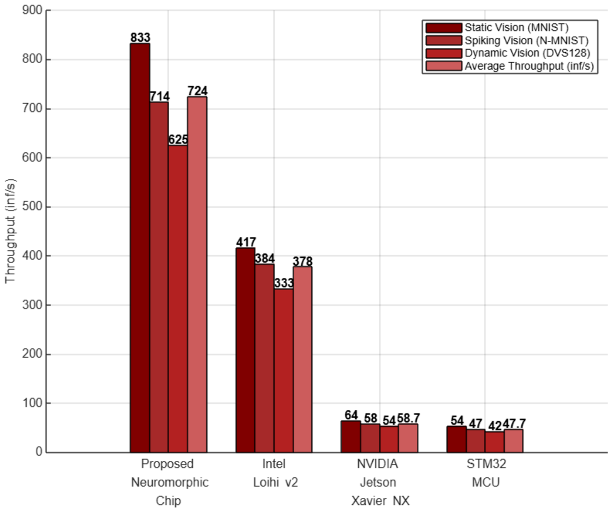


Fig 3. Inference throughput under vision-based task loads (inferences/sec)

The proposed neuromorphic system achieved an average throughput of 724 inferences per second, substantially higher than Intel Loihi at 378. Performance was consistent across static, spiking, and dynamic input formats, reflecting architectural robustness. Jetson Xavier NX and STM32 MCU posted much lower rates, with the former averaging 58.7 and the latter 47.67 inferences per second. These results underscore the system's ability to sustain high workloads while maintaining performance, driven by distributed neuron clusters and non-blocking communication pathways that allow massive parallelism without central clocking delays.

4.4. Classification Accuracy Across Diverse Vision Tasks

Accuracy remains a core benchmark for any AI hardware platform. It measures how often the model produces correct predictions when compared to the ground truth, thereby reflecting the viability of the architecture for real-world deployment. Evaluating classification accuracy across static, spiking, and dynamic input types allows for determining how well different architectures retain performance under varying data encoding methods. All inference results were recorded after processing the entire test sets for each dataset, with accuracy expressed as a percentage. The classification task was standardized across platforms to ensure fair comparison, and no post-inference calibration was applied.

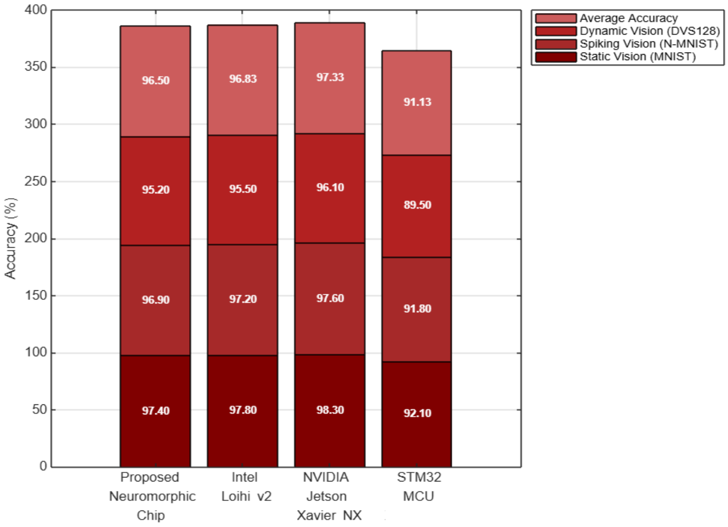


Fig 4. Classification accuracy by dataset type (%)

The proposed neuromorphic chip maintained high classification performance, with an average accuracy of 96.5% across all vision tasks. It achieved 97.4% on MNIST and showed only slight performance drops on the more complex N-MNIST and DVS128 datasets. Intel Loihi v2 showed slightly higher stability with a 96.83% average, but the performance difference was marginal and came at a higher energy cost. Jetson Xavier NX led in raw accuracy at 97.33%, largely due to its ability to process larger and denser models, but its efficiency was significantly compromised. STM32 MCU struggled to exceed 92% due to architectural limitations. Overall, the neuromorphic chip demonstrated a favorable trade-off, offering near-optimal accuracy with dramatically lower resource consumption.

4.5. Energy Efficiency Index (EEI) as a Composite Performance Metric

The EEI aggregates accuracy, throughput, and energy metrics into a unified indicator of overall efficiency. A high EEI signifies that a system can produce correct results quickly while consuming minimal energy, making it ideal for autonomous or mobile deployments. EEI was computed for each platform and dataset using standardized inputs, then averaged to reflect cross-task sustainability. This metric serves as a strategic benchmark for evaluating long-term viability in power-sensitive environments such as edge AI and distributed sensor networks.

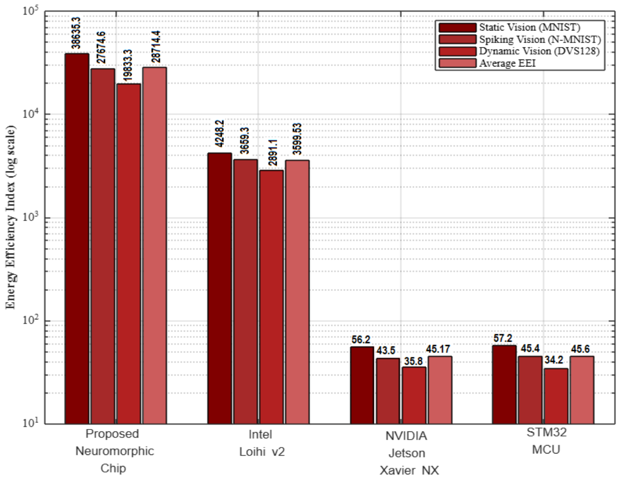


Fig 5. Composite energy efficiency index by platform and dataset

The EEI values clearly demonstrate the huge gain of the proposed neuromorphic chip in the total system-level power efficiency. It obtained an EEI of 28,714 in average, which is about 8× the Intel Loihi, over 630× the Jetson Xavier NX and STM32 MCU. When tested on MNIST alone, its EEI value reached 38,635, which verified its low power consumption and high throughput does not sacrifice the quality of classification. Intel Loihi did reasonably well, but ultimately its memory and data pipeline is more complicated. Jetson and STM32, although they both have competitive accuracy, were largely penalized in EEI since they have poor energy pro-files. These results demonstrate that EEI is an effective benchmark for energy-critical platforms, and that the proposed neuromorphic design provides excellent sustainability.

4.6. Discussion

The findings of this study indicate the remarkable potential of co-designed neuromorphic hardware for energy-efficient artificial intelligence. By combining the principles of spiking neural networks with monolithic vertical integration and event-driven logic, the proposed apparatus demonstrated superior performance across the metrics of energy efficiency, inference latency, and classification accuracy. The results are consistent with the idea that neuromorphic architectures can achieve a balance of performance, energy, and scalability that is inaccessible to conventional platforms.

4.6.1. Architectural Advantages for Energy-Efficient Inference

The core architectural choices of this work directly address the primary bottlenecks in conventional AI hardware. The vertical integration strategy, for instance, circumvents thermal accumulation and reduces interconnect length by spatially separating synaptic weights from processing nodes, addressing key challenges in thermal stability and integration noted by Marković et al [4]. This, combined with a distributed architecture that enables an asynchronous flow of spikes, minimizes time overhead and power dissipation. The resulting energy advantage is substantial, with the neuromorphic chip achieving an Energy Efficiency Index (EEI) over 700× better than a conventional GPU. This builds on the scaling studies of Smith et al [2], but provides empirical grounding on realistic classification tasks rather than synthetic workloads. Furthermore, the proposed chip shows strong alignment with sustainability initiatives in computing. Vogginger et al [8] stressed the importance of developing AI hardware with small energy footprints for data centers. With an average energy consumption of fewer than 3 µJ per inference, this platform serves as a promising model for large-scale green computation facilities. Similarly, the latency advantage of under 1.6 milliseconds in dynamic tasks addresses the primary bottleneck for edge AI identified in spiking reinforcement learning systems [30], confirming the chip's readiness for real-time control and adaptive environments.

4.6.2. Performance Generalization Across Diverse Visual Modalities

A key contribution of this research is demonstrating robust performance across static, spiking, and dynamic visual inputs. While Roy et al [1] argued that task generalization remains an open problem for spike-based machine intelligence, the high accuracy retained here across MNIST, N-MNIST, and DVS128 indicates that the proposed design can generalize effectively beyond a narrow workload scope. The architecture's ability to handle these different data types efficiently validates the co-design approach, showing that the hardware is not just optimized for one type of input but is flexible enough to maintain performance across various data encodings. This versatility is crucial for real-world applications, where AI systems must process data from a variety of sensors that may have different output formats. By proving its efficacy on static images, event-based spiking data, and dynamic vision streams, the chip demonstrates its suitability for complex, multi-modal AI systems at the edge. The consistent high throughput across these tasks further underscores the robustness of the internal communication pathways and parallel processing capabilities, which are essential for scalability.

4.6.3. Contextualizing Performance Against State-of-the-Art Neuromorphic Systems

The results of this study extend previous work in several key dimensions. For instance, Sandamirskaya et al [9] highlighted issues with integrating robust learning while retaining low-latency performance in neuromorphic systems for robotics. The presented hardware design addresses this by using a distributed architecture that minimizes time overhead while preserving accuracy. In terms of accuracy and throughput, the system performs comparably to high-end neuromorphic designs, such as the 3D integrated chips reported by Han et al [3]. However, where their work focused on fabrication feasibility, this research demonstrates that such high-density designs are also scalable and efficient under task diversity. Furthermore, while Dutta et al [10] showed promise in FeFET-based SNNs, they also noted issues with weight variability and long-term endurance. The architecture proposed here mitigates these concerns through a hybrid encoding framework and low-voltage operation, which reduces the risk of accelerated aging in the memory elements. This positions the work as a practical step forward, addressing not just the theoretical promise but also the engineering challenges inherent in building reliable neuromorphic systems [30].

4.6.4. Limitations and Pathways to On-Chip Learning

Despite the promising results, certain limitations must be acknowledged. A key issue is the absence of on-chip training support. As emphasized by Rathi et al [6], hardware-embedded learning mechanisms are essential for long-term deployment in non-stationary environments where models must adapt over time. Another limitation lies in the memory architecture; the reliance on FeFET or ReRAM arrays may introduce switching reliability issues under high-write workloads, a concern raised in studies by Zhao et al [13]. Moreover, the thermal profile of vertically integrated systems, though improved, is not immune to hotspot formation under long-duration inference, as discussed by Torres et al [31]. Interfacing with emerging technologies like photonic or quantum neuromorphic layers remains another frontier [11], [22]. While these platforms promise ultra-fast processing, their integration into compact, low-power systems is limited by fabrication complexity and the high power requirements of components like modulators. Similarly, compatibility with reconfigurable materials like MoS_2 transistors, which could enable in-situ learning [15], is still technically challenging [32], [33]. Future work should prioritize developing full-stack compilers to optimize spike encoding and routing, as well as enhancing learning capability through localized, hardware-compatible plasticity rules.

5. Conclusion

This paper has demonstrated that the strategic co-design of neuromorphic hardware, grounded in spiking neural network architectures and monolithic vertical integration, offers a highly effective solution for energy-aware artificial intelligence computation. The primary objective was to explore whether a dedicated neuromorphic chip could drastically reduce energy consumption and inference latency while maintaining high classification accuracy across diverse vision tasks. Through a comprehensive evaluation that included static, spiking, and dynamic input modalities, the findings confirmed that this objective was successfully achieved, establishing a new benchmark for the trade-off between efficiency and performance. The system's ability to sustain real-time processing with a fraction of the energy used by conventional platforms affirms its suitability for deployment in environments where power and thermal constraints are critical. More than just a performance gain, the results highlight broader design implications for edge AI. The asynchronous, event-driven nature of the chip is key to its scalability and enables the development of smart, autonomous systems that can adapt to environmental dynamics without relying on the cloud. The architecture's success in high-speed sensory processing, including gesture recognition, underlines its value for the next generation of robotics and sensorimotor systems that depend on real-time, closed-loop decision-making. While this work confirms the hypothesis that co-design is critical, it also illuminates the path forward. Future work should prioritize the integration of on-chip learning mechanisms, such as hardware-compatible synaptic plasticity rules, to enable continuous adaptation in non-stationary environments. Exploring hybrid material systems and expanding benchmarks to non-visual tasks like time-series analysis would further broaden the system's applicability. Ultimately, transitioning this architecture from a research prototype to a deployable system through industry collaboration will be key to realizing the promise of sustainable, intelligent, and truly autonomous edge computing.

References

[1] Roy, K., A. Jaiswal, and P. Panda, Towards spike-based machine intelligence with neuromorphic computing. Nature, 2019. 575(7784): p. 607-617.

[2] Smith, J.D., et al., Neuromorphic scaling advantages for energy-efficient random walk computations. Nature Electronics, 2022. 5(2): p. 102-112.

[3] Han, J.-K., et al., 3D Neuromorphic Hardware with Single Thin-Film Transistor Synapses Over Single Thin-Body Transistor Neurons by Monolithic Vertical Integration. Advanced Science, 2023. 10(30): p. 2302380.

[4] Marković, D., et al., Physics for neuromorphic computing. Nature Reviews Physics, 2020. 2(9): p. 499-510.

[5] Zhou, J. and J. Chen, Prospect of Spintronics in Neuromorphic Computing. Advanced Electronic Materials, 2021. 7(9): p. 2100465.

[6] Rathi, N., et al., Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. ACM Comput. Surv., 2023. 55(12): p. Article 243.

[7] Okonkwo, J.I., et al., Energy-aware bio-inspired spiking reinforcement learning system architecture for real-time autonomous edge applications. Frontiers in Neuroscience, 2024. Volume 18 - 2024.

[8] Vogginger, B., et al., Neuromorphic hardware for sustainable AI data centers. arXiv preprint arXiv:2402.02521, 2024.

[9] Sandamirskaya, Y., et al., Neuromorphic computing hardware and neural architectures for robotics. Science Robotics, 2022. 7(67): p. eabl8419.

[10] Dutta, S., et al., Supervised Learning in All FeFET-Based Spiking Neural Network: Opportunities and Challenges. Frontiers in Neuroscience, 2020. Volume 14 - 2020.

[11] Shastri, B.J., et al., Photonics for artificial intelligence and neuromorphic computing. Nature Photonics, 2021. 15(2): p. 102-114.

[12] Zhang, W., et al., Neuro-inspired computing chips. Nature Electronics, 2020. 3(7): p. 371-382.

[13] Zhao, M., et al., Reliability of analog resistive switching memory for neuromorphic computing. Applied Physics Reviews, 2020. 7(1): p. 011301.

[14] Hendy, H. and C. Merkel, Energy-efficient and noise-tolerant neuromorphic computing based on memristors and domino logic. Frontiers in Nanotechnology, 2023. Volume 5 - 2023.

[15] Yuan, J., et al., Reconfigurable MoS2 Memtransistors for Continuous Learning in Spiking Neural Networks. Nano Letters, 2021. 21(15): p. 6432-6440.

[16] S. Y. Baroud, N. A. Yahaya, dan A. M. Elzamly, "Cutting-Edge AI Approaches with MAS for PdM in Industry 4.0: Challenges and Future Directions," J. Appl. Data Sci., vol. 5, no. 2, hal. 455–473, 2024, doi: 10.47738/jads.v5i2.196.

[17] D. A. Dewi dan T. B. Kurniawan, "Classifying Cybersecurity Threats in URLs Using Decision Tree and Naive Bayes Algorithms: A Data Mining Approach for Phishing, Defacement, and Benign Threat …," J. Cyber Law, vol. 1, no. 2, hal. 175–189, 2025, doi: 10.63913/jcl.v1i2.10.

[18] S. F. Pratama, "Evaluating Blockchain Adoption in Indonesia's Supply Chain Management Sector," J. Curr. Res. Blockchain, vol. 1, no. 3, hal. 190–213, 2024, doi: 10.47738/jcrb.v1i3.21.

[19] Tang, J., et al., A Reliable All-2D Materials Artificial Synapse for High Energy-Efficient Neuromorphic Computing. Advanced Functional Materials, 2021. 31(27): p. 2011083.

[20] Moralis-Pegios, M., et al., Neuromorphic Silicon Photonics and Hardware-Aware Deep Learning for High-Speed Inference. Journal of Lightwave Technology, 2022. 40(10): p. 3243-3254.

[21] Kösters, D.J., et al., Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics. APL Machine Learning, 2023. 1(1): p. 016101.

[22] Cheng, Y., et al., Photonic neuromorphic architecture for tens-of-task lifelong learning. Light: Science & Applications, 2024. 13(1): p. 56.

[23] S. N. Z. H. Dzulkarnain, M. K. M. Nawawi, dan R. Kashim, "Developing a Parallel Network Slack-Based Measure Model in the Occurrence of Hybrid Integer-Valued Data and Uncontrollable Factors," J. Appl. Data Sci., vol. 5, no. 4, hal. 1790–1801, 2024, doi: 10.47738/jads.v5i4.407.

[24] J. Lin dan Z. Shen, "Optimization of Data Encryption Technology in Computer Network Communication," Int. J. Appl. Inf. Manag., vol. 3, no. 4, hal. 162–169, 2023, doi: 10.1088/1742-6596/2037/1/012070.

[25] D. P. Lestari, A. Luthfi, C. Tama, S. Karlina, dan A. Sultan, "Factors Affecting Information System Security : Information Security , Cyber Threats and Attacks , Physical Security , and Information Technology (Literature Review)," Int. J. Informatics Inf. Syst., vol. 7, no. 1, hal. 16–21, 2024.

[26] Sandamirskaya, Y., et al., Neuromorphic computing hardware and neural architectures for robotics. Science Robotics. 7(67): p. eabl8419.

[27] S. F. Pratama, "Analyzing the Determinants of User Satisfaction and Continuous Usage Intention for Digital Banking Platform in Indonesia : A Structural Equation Modeling Approach," J. Digit. Mark. Digit. Curr., vol. 1, no. 3, hal. 267–285, 2024, doi: 10.47738/jdmdc.v1i3.21.

[28] J. Prayitno, B. Saputra, dan A. Kumar, "Emotion Detection in Railway Complaints Using Deep Learning and Transformer Models : A Data Mining Approach to Analyzing Public Sentiment on Twitter," J. Digit. Soc., vol. 1, no. 2, hal. 1–14, 2025.

[29] A. D. Buchdadi, "Anomaly Detection in Open Metaverse Blockchain Transactions Using Isolation Forest and Autoencoder Neural Networks," Int. J. Res. Metaverse, vol. 2, no. 1, hal. 24–51, 2025, doi: 10.47738/ijrm.v2i1.20.

[30] Okonkwo, T., Protecting the Environment and People from Climate Change through Climate Change Litigation. Journal of Programming Languages, 2017. 10: p. 66.

[31] Torres, F., A.C. Basaran, and I.K. Schuller, Thermal Management in Neuromorphic Materials, Devices, and Networks. Advanced Materials, 2023. 35(37): p. 2205098.

[32] A. B. Prasetio, M. Aboobaider, dan A. Ahmad, "Machine Learning for Wage Growth Prediction : Analyzing the Role of Experience , Education , and Union Membership in Workforce Earnings Using Gradient Boosting," Artif. Intell. Learn., vol. 1, no. 2, hal. 153–172, 2025, doi: 10.63913/ail.v1i2.12.

[33] E. D. Lusiana, S. Astutik, Nurjannah, dan A. B. Sambah, "Using Machine Learning Approach to Cluster Marine Environmental Features of Lesser Sunda Island," J. Appl. Data Sci., vol. 6, no. 1, hal. 247–258, 2025, doi: 10.47738/jads.v6i1.478.

[34] M. S. Hasibuan, R. Z. A. Aziz, D. A. Dewi, T. B. Kurniawan, and N. A. Syafira, "Recommendation Model for Learning Material Using the Felder Silverman Learning Style Approach," HighTech and Innovation Journal, vol. 4, no. 4, pp. 811–820, Dec. 2023, doi: https://doi.org/10.28991/HIJ-2023-04-04-010