Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 6, No. 6, Juni 2026, Hal. 1141-1153
DOI: https://doi.org/10.52436/1.jpti.2136
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Analisis Komparatif Efisiensi Energi Dan Akurasi Arsitektur Neuromorphic Dan Von Neumann Pada Inferensi *Edge AI*

Muhammad Baso Adrian Ibrahim[1], Farhani Ayu Amalina[*2], Muhammad Faiz Burhanuddion[3], Ruth Hanseliani[4], Jan Everhard Riwurohi[5]

[1,2,3,4,5]Universitas Budi Luhur, Indonesia
Email: [1]2511600807@student.budiluhur.ac.id, [2]2511600799@student.budiluhur.ac.id, [3]2511600583@student.budiluhur.ac.id, [4]2511600781@student.budiluhur.ac.id, [5]yan.everhard@budiluhur.ac.id

Abstrak

Permasalahan mendasar mengenai hubungan antara akurasi inferensi dan efisiensi energi pada perangkat Edge AI masih belum dijelaskan secara kuantitatif dalam literatur yang ada. Kebaruan penelitian ini terletak pada penggunaan *spike sparsity* sebagai parameter analitik kuantitatif untuk mengidentifikasi secara eksplisit kondisi ketika arsitektur Neuromorphic memberikan efisiensi energi yang lebih baik dibandingkan arsitektur Von Neumann. Perkembangan pesat perangkat IoT dan komputasi *edge* mendorong kebutuhan mendesak akan inferensi AI yang sangat hemat energi. Arsitektur Von Neumann yang mendasari perangkat *edge* menghadapi hambatan struktural: setiap operasi perkalian-akumulasi (MAC) pada jaringan saraf tiruan dalam (DNN) mensyaratkan transfer data berulang antara memori dan prosesor, menghasilkan konsumsi energi yang tidak proporsional terhadap kapasitas perangkat. Jaringan Saraf Spiking (SNN) dengan neuron *Leaky Integrate-and-Fire* (LIF) menawarkan paradigma komputasi berbasis peristiwa yang secara biologis masuk akal, di mana operasi sinaptik hanya terjadi saat ambang batas potensial membran tercapai. Studi ini menyajikan analisis komparatif komprehensif antara DNN dan SNN menggunakan arsitektur LeNet-5 pada dataset MNIST dan Fashion-MNIST, diimplementasikan sepenuhnya melalui simulasi perangkat lunak berbasis Python, PyTorch, dan snntorch. Efisiensi energi diukur melalui dua pendekatan komplementer: model energi teoretis (4,6 pJ/MAC untuk DNN; 0,9 pJ/SOP untuk SNN) dan pengukuran empiris menggunakan *CodeCarbon*. Hasil menunjukkan SNN dengan langkah waktu T=16 mencapai akurasi 98,63% pada MNIST dengan estimasi energi 15,55 µJ per inferensi, menghasilkan efisiensi 1,10× dibandingkan DNN (17,11 µJ, akurasi 99,21%). Sparsitas *spike* rata-rata 71,0% memenuhi ambang efisiensi teoretis, dan analisis Pareto mengonfirmasi T=16 sebagai konfigurasi titik optimal *trade-off* akurasi–energi. Kontribusi utama penelitian ini meliputi kerangka evaluasi komparatif sumber terbuka yang dapat direplikasi tanpa perangkat keras neuromorfik khusus, serta validasi spike sparsity sebagai parameter kuantitatif untuk menjelaskan hubungan antara akurasi inferensi dan efisiensi energi pada SNN, yang didukung melalui analisis pengaruh langkah waktu dan validasi silang model energi teoretis terhadap benchmark perangkat keras fisik. Secara praktis, temuan ini mengimplikasikan bahwa SNN dengan konfigurasi T=16 layak diadopsi pada perangkat Edge AI berdaya terbatas yang membutuhkan keseimbangan antara akurasi inferensi dan efisiensi energi, tanpa memerlukan perangkat keras neuromorphic khusus.

Kata kunci: *Deep Neural Networks, Edge AI, Energy Efficiency, Software Simulation, Spike Sparsity, Spiking Neural Networks, Theoretical Energy Model, Von Neumann Bottleneck.*

Comparative Analysis of Energy Efficiency and Accuracy of Neuromorphic and Von Neumann Architectures in Edge AI Inference

Abstract

The fundamental relationship between inference accuracy and energy efficiency in Edge AI devices remains insufficiently explained quantitatively in the existing literature. The novelty of this study lies in the use of spike sparsity as a quantitative analytical parameter to explicitly identify the conditions under which Neuromorphic architectures provide superior energy efficiency compared with the Von Neumann architecture. The proliferation of IoT devices and edge computing has created an urgent demand for ultra-low-power AI inference. The Von Neumann architecture underlying edge devices faces a fundamental structural bottleneck: every multiply-accumulate (MAC) operation in deep neural networks (DNNs) requires repeated data transfers between memory and processor, producing energy consumption disproportionate to device capacity. Spiking

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Neural Networks (SNNs) with Leaky Integrate-and-Fire (LIF) neurons offer a biologically plausible, event-driven computing paradigm in which synaptic operations occur only when the membrane potential threshold is reached. This study presents a comprehensive comparative analysis of DNNs and SNNs using equivalent LeNet-5 architectures on the MNIST and Fashion-MNIST datasets, implemented entirely through software simulation using Python, PyTorch, and snntorch. Energy efficiency was evaluated via two complementary approaches: a theoretical energy model (4.6 pJ/MAC for DNN; 0.9 pJ/SOP for SNN) and empirical measurement using CodeCarbon. Results demonstrate that SNN with time steps T=16 achieved 98.63% accuracy on MNIST with an estimated energy consumption of 15.55 μJ per inference, yielding 1.10× greater efficiency over DNN (17.11 μJ, 99.21% accuracy). Average spike sparsity of 71.0% at T=16 satisfies the theoretical efficiency threshold, and Pareto analysis confirms this configuration as the optimal accuracy–energy trade-off point. The primary contributions include an open-source, fully replicable comparative evaluation framework requiring no specialized neuromorphic hardware, as well as the validation of spike sparsity as a quantitative parameter for explaining the relationship between inference accuracy and energy efficiency in SNNs, supported by an analysis of time-step effects and cross-validation of the theoretical energy model against physical hardware benchmarks. Practically, these findings suggest that SNNs configured with T=16 are suitable for deployment in resource-constrained Edge AI devices that require a balance between inference accuracy and energy efficiency without the need for dedicated neuromorphic hardware.

Keywords: *Deep Neural Networks, Edge AI, Energy Efficiency, Software Simulation, Spike Sparsity, Spiking Neural Networks, Theoretical Energy Model, Von Neumann Bottleneck.*

1. PENDAHULUAN

Perkembangan Internet of Things (IoT), *edge* computing, dan kecerdasan buatan (Artificial Intelligence / AI) telah mendorong kebutuhan akan sistem inferensi yang mampu beroperasi secara mandiri pada perangkat dengan sumber daya terbatas. Berbeda dengan komputasi berbasis cloud, *edge* AI menuntut kemampuan pemrosesan lokal dengan latensi rendah, privasi data yang lebih baik, serta konsumsi energi yang minimal. Kondisi ini menjadikan efisiensi energi sebagai salah satu parameter utama dalam desain sistem komputasi modern, khususnya pada perangkat *edge* seperti sensor pintar, wearable device, autonomous system, dan embedded platform [1], [7], [15]. Seiring meningkatnya kompleksitas model AI yang digunakan pada perangkat *edge*, tantangan dalam menjaga keseimbangan antara akurasi inferensi dan konsumsi energi menjadi semakin penting dalam bidang arsitektur dan organisasi komputer. Penelitian ini menempatkan parameter spike sparsity sebagai fokus utama penyelesaian masalah: secara kuantitatif, spike sparsity menentukan ambang batas di mana arsitektur Neuromorphic mulai atau berhenti mengungguli Von Neumann dalam hal efisiensi energi inferensi.

Sebagian besar implementasi AI saat ini masih bergantung pada arsitektur Von Neumann yang mendasari CPU dan GPU konvensional. Meskipun arsitektur ini telah terbukti efektif dalam menjalankan model Deep Neural Network (DNN), pemisahan fisik antara unit komputasi dan memori menyebabkan terjadinya fenomena Von Neumann bottleneck [1]. Pada setiap proses inferensi, bobot dan data harus ditransfer secara berulang antara memori dan prosesor sehingga menghasilkan konsumsi energi yang tinggi. Zhang (2026) menjelaskan bahwa perpindahan data (data movement) antara unit sensing, memori, dan komputasi merupakan hambatan fundamental yang membatasi efisiensi sistem AI modern [1]. Permasalahan ini semakin signifikan ketika model AI menjadi lebih kompleks dan harus dijalankan pada perangkat *edge* yang memiliki keterbatasan daya [3], [5]. Oleh karena itu, diperlukan paradigma komputasi alternatif yang mampu mengurangi biaya perpindahan data dan meningkatkan efisiensi energi tanpa mengorbankan performa inferensi secara signifikan.

Salah satu pendekatan yang banyak mendapat perhatian adalah Neuromorphic Computing melalui implementasi Spiking Neural Network (SNN). Berbeda dengan DNN yang menggunakan komputasi sinkron berbasis operasi Multiply-Accumulate (MAC), SNN menerapkan mekanisme event-driven yang meniru cara kerja neuron biologis [5], [6]. Pada model ini, neuron hanya melakukan operasi ketika menerima *spike* yang memenuhi ambang tertentu sehingga jumlah operasi komputasi dapat berkurang secara signifikan. Karakteristik tersebut menjadikan SNN sebagai kandidat utama untuk mendukung sistem *edge* AI berdaya rendah [6], [7]. Selain itu, penggunaan model neuron *Leaky Integrate-and-Fire* (LIF) memungkinkan pemrosesan informasi secara lebih efisien karena aktivitas komputasi hanya terjadi pada neuron yang aktif, berbeda dengan DNN yang memproses seluruh neuron pada setiap siklus inferensi [6], [14].

Berbagai penelitian telah mengevaluasi potensi SNN sebagai alternatif arsitektur komputasi AI. Stevens (2022) menunjukkan bahwa implementasi neuromorphic pada Intel Loihi memiliki efisiensi energi sekitar 2,5 kali lebih baik dibandingkan prosesor ARM Cortex-A72 melalui pendekatan eksperimen perangkat keras [8]. Ferreira et al. (2025) melakukan kajian komprehensif mengenai perbandingan DNN dan SNN pada aplikasi *edge*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

AI menggunakan pendekatan literature review dan analisis metrik energi lintas platform [7]. Sementara itu, Eshraghian et al. (2023) mengembangkan metode pelatihan SNN berbasis surrogate gradient yang mampu meningkatkan performa klasifikasi dan stabilitas proses pelatihan [14]. Javanshir et al. (2022) juga menunjukkan bahwa perkembangan algoritma dan perangkat keras neuromorphic telah membuka peluang implementasi SNN pada berbagai aplikasi AI berdaya rendah [6]. Hasil penelitian-penelitian tersebut menunjukkan bahwa SNN memiliki potensi besar dalam meningkatkan efisiensi energi dibandingkan DNN, terutama pada lingkungan komputasi yang memiliki keterbatasan sumber daya.

Meskipun penelitian mengenai SNN dan neuromorphic computing telah berkembang pesat, masih terdapat beberapa kesenjangan penelitian (research gap) yang belum terjawab. Pertama, sebagian besar penelitian terdahulu berfokus pada evaluasi berbasis perangkat keras neuromorphic khusus seperti Intel Loihi [8], [10], sehingga hasil penelitian sulit direplikasi oleh peneliti yang tidak memiliki akses terhadap perangkat tersebut. Kedua, studi komparatif yang tersedia umumnya hanya membandingkan akurasi dan konsumsi energi tanpa melakukan analisis mendalam terhadap faktor internal yang mempengaruhi efisiensi energi SNN [7]. Ketiga, parameter *spike* sparsity yang secara teoritis menjadi karakteristik utama komputasi event-driven masih jarang digunakan sebagai variabel analisis kuantitatif dalam menjelaskan hubungan antara aktivitas neuron, jumlah operasi sinaptik, dan konsumsi energi [5], [7]. Akibatnya, masih terdapat keterbatasan pemahaman mengenai kondisi dan titik kritis ketika SNN mampu atau gagal memberikan keunggulan efisiensi energi dibandingkan arsitektur Von Neumann. Christensen et al. (2022) bahkan menekankan perlunya data kuantitatif yang mampu menjelaskan hubungan antara karakteristik komputasi neuromorphic dan efisiensi sistem secara menyeluruh [13].

Berdasarkan kesenjangan penelitian tersebut, penelitian ini mengusulkan kerangka evaluasi komparatif berbasis simulasi untuk membandingkan arsitektur Neuromorphic dan Von Neumann pada skenario inferensi *edge* AI. Penelitian ini menggunakan simulasi perangkat lunak sehingga dapat direplikasi tanpa memerlukan perangkat keras neuromorphic khusus. Berbeda dengan penelitian sebelumnya yang hanya berfokus pada akurasi dan energi [7], [8], penelitian ini mengintegrasikan analisis akurasi, konsumsi energi teoritis, konsumsi energi empiris, jumlah operasi komputasi (MACs dan SOPs), jumlah time steps, serta *spike* sparsity dalam satu kerangka evaluasi yang terpadu. Pendekatan ini memungkinkan analisis yang lebih komprehensif terhadap hubungan antara karakteristik komputasi SNN dan efisiensi energi yang dihasilkan.

Metode penelitian dilakukan menggunakan framework PyTorch dan snntorch yang telah banyak digunakan dalam penelitian SNN modern [14]. Dataset yang digunakan adalah MNIST dan Fashion-MNIST karena keduanya merupakan *benchmark* yang umum digunakan dalam evaluasi DNN dan SNN [6], [8], [14]. Untuk menjaga validitas perbandingan, kedua model menggunakan arsitektur LeNet-5 yang ekuivalen sehingga perbedaan hasil yang diperoleh terutama berasal dari perbedaan mekanisme komputasi antara DNN dan SNN. Pengukuran energi dilakukan menggunakan dua pendekatan, yaitu theoretical energy model berbasis operasi MAC dan SOP yang mengacu pada Davies et al. [10] serta pengukuran empiris menggunakan CodeCarbon [7].

Fitur, parameter, dan variabel yang dianalisis dalam penelitian ini meliputi akurasi klasifikasi, inference latency, jumlah FLOPs/MACs pada DNN, jumlah SOPs pada SNN, konsumsi energi teoritis, konsumsi energi empiris, time steps (T = 8, 16, 32, dan 64), serta *spike* sparsity pada setiap lapisan jaringan. Di antara parameter tersebut, *spike* sparsity digunakan sebagai variabel utama untuk menjelaskan tingkat aktivitas neuron selama proses inferensi dan hubungannya terhadap efisiensi energi yang dihasilkan oleh arsitektur neuromorphic.

Kebaruan (novelty) penelitian ini terletak pada penggunaan *spike* sparsity sebagai indikator kuantitatif untuk menjelaskan mekanisme efisiensi energi pada SNN. Berbeda dengan penelitian sebelumnya yang umumnya hanya membandingkan nilai akurasi dan konsumsi energi sebagai keluaran akhir [7], [8], penelitian ini menganalisis hubungan antara *spike* sparsity, jumlah operasi sinaptik, konsumsi energi, dan akurasi pada berbagai konfigurasi time steps. Dengan demikian, penelitian ini tidak hanya menunjukkan apakah SNN lebih efisien dibandingkan DNN, tetapi juga menjelaskan kondisi dan faktor yang menyebabkan keunggulan tersebut muncul atau menurun. Pendekatan ini diharapkan dapat memberikan kontribusi baru dalam evaluasi arsitektur neuromorphic dari perspektif arsitektur dan organisasi komputer.

Kontribusi utama penelitian ini adalah: (1) menyediakan kerangka evaluasi komparatif DNN dan SNN yang dapat direplikasi tanpa memerlukan perangkat keras neuromorphic khusus; (2) melakukan analisis hubungan antara akurasi, energi, dan kompleksitas komputasi pada arsitektur Von Neumann dan Neuromorphic; (3) memperkenalkan *spike* sparsity sebagai parameter analisis untuk menjelaskan efisiensi energi SNN; dan (4) mengidentifikasi konfigurasi time steps yang menghasilkan *trade-off* optimal antara akurasi dan efisiensi energi pada skenario inferensi *edge* AI.

Berdasarkan latar belakang tersebut, penelitian ini dirancang untuk menjawab tiga pertanyaan penelitian (research questions), yaitu: RQ1) bagaimana perbandingan akurasi dan efisiensi energi antara arsitektur Neuromorphic dan Von Neumann pada inferensi *edge* AI; RQ2) bagaimana pengaruh jumlah time steps terhadap akurasi, *spike* sparsity, dan konsumsi energi pada SNN; serta RQ3) apakah terdapat nilai ambang (critical *spike*

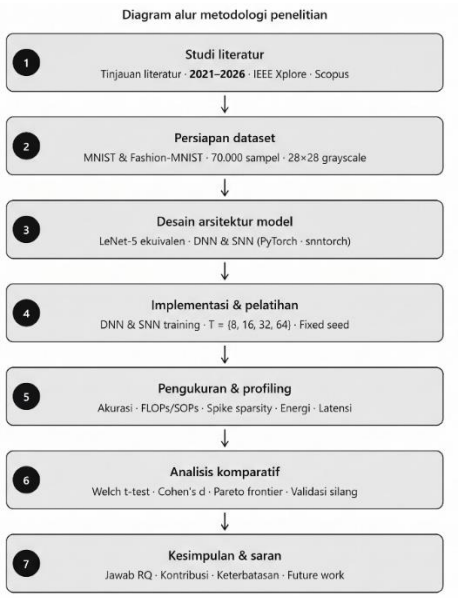
p-ISSN: 2775-4227
e-ISSN: 2775-4219

sparsity threshold) yang menentukan batas keunggulan efisiensi energi arsitektur Neuromorphic dibandingkan arsitektur Von Neumann pada inferensi *Edge* AI.

Penelitian ini dibatasi pada simulasi perangkat lunak menggunakan PyTorch dan snntorch, penggunaan dataset MNIST dan Fashion-MNIST, arsitektur LeNet-5 ekuivalen, model neuron *Leaky Integrate-and-Fire* (LIF), serta pengukuran energi menggunakan theoretical energy model dan CodeCarbon. Implementasi langsung pada chip neuromorphic fisik, evaluasi pada dataset berskala besar seperti ImageNet, serta optimasi perangkat keras tingkat sirkuit tidak termasuk dalam ruang lingkup penelitian ini.

2. METODE PENELITIAN

Penelitian ini mengikuti alur metodologi sekuensial tujuh tahap sebagaimana ditunjukkan pada Gambar 1 (Diagram Alur Metodologi). Setiap tahap memiliki input, proses, dan output yang terdefinisi secara eksplisit sehingga seluruh prosedur dapat diaudit dan direplikasi oleh peneliti lain tanpa memerlukan akses hardware neuromorphic khusus.



Gambar 1. Diagram Alur Metodologi Penelitian

2.1 Studi Literatur

Tahap pertama bertujuan membangun landasan teoritis penelitian melalui kajian sistematis terhadap 15 referensi ilmiah yang diterbitkan antara tahun 2019 hingga 2026. Pencarian dilakukan pada basis data IEEE Xplore, Scopus, Frontiers in Neuroscience, dan Nature dengan kata kunci utama: spiking neural network energy efficiency, neuromorphic computing *edge* AI, DNN SNN comparison, Von Neumann bottleneck, dan *Leaky Integrate-and-Fire*. Kriteria inklusi yang diterapkan meliputi: (1) diterbitkan antara 2019–2026; (2) terindeks di Scopus, IEEE Xplore, atau Nature Publishing Group; (3) relevan langsung terhadap perbandingan DNN/SNN atau arsitektur neuromorphic; dan (4) tersedia dalam versi full-text.

Referensi yang lolos seleksi diklasifikasikan ke dalam tiga sub-tema:

- **Sub-tema A – Arsitektur dan Hardware Neuromorphic:** Paradigma beyond Von Neumann [1][2][3][4][5], roadmap neuromorphic [13], dan arsitektur Intel Loihi 2 [10][12]. Sub-tema ini mengidentifikasi Von Neumann bottleneck sebagai hambatan fundamental yang menjadi dasar motivasi komparatif penelitian.
- **Sub-tema B – Efisiensi Energi SNN vs DNN:** *Benchmark* energi berbasis hardware [8][9], survei komparatif konsumsi daya lintas platform [7], dan konstanta energi teoretis 4,6 pJ/MAC (DNN) serta 0,9 pJ/SOP (SNN) [10]. Sub-tema ini menghasilkan parameter acuan eksperimen sekaligus mengonfirmasi research gap: minimnya kerangka simulasi yang dapat direplikasi tanpa hardware khusus.
- **Sub-tema C – Algoritma Training dan Aplikasi SNN:** Metode surrogate gradient BPTT dan konversi ANN-to-SNN [14], pemetaan SNN ke hardware [9], serta aplikasi federated neuromorphic learning [15] dan event-driven computing [6][11]. Sub-tema ini mendasari pemilihan snntorch sebagai framework dan penetapan rentang time steps $T \in \{8, 16, 32, 64\}$.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.2 Penentuan Dataset

Dataset dipilih berdasarkan tingkat adopsi dalam literatur SNN dan variasi kompleksitas visual. Dua dataset digunakan sebagaimana dirangkum pada Tabel 1.

Tabel 1. Konfigurasi Dataset Penelitian

Dataset	Ukuran (Total)	Dimensi	Kelas	Justifikasi
MNIST	70.000 sampel	28×28 px grayscale	10	*Benchmark* standar SNN; digunakan di [6][8][14] memastikan komparabilitas lintas studi
Fashion-MNIST	70.000 sampel	28×28 px grayscale	10	Lebih menantang secara visual; menguji generalisasi hasil terhadap kompleksitas yang lebih tinggi [14]

Penggunaan dua dataset dengan dimensi identik namun kompleksitas berbeda memungkinkan analisis konsistensi perilaku kedua arsitektur di luar kondisi *benchmark* minimal.

2.3 Penentuan Arsitektur

Kedua model menggunakan topologi LeNet-5 yang dimodifikasi dengan jumlah parameter setara (~370K), sehingga perbedaan metrik yang terukur hanya berasal dari perbedaan mekanisme aktivasi dan komputasi, bukan dari topologi jaringan [6][14].

a. DNN (PyTorch 2.x):
Conv2d(1,32,3) → ReLU → MaxPool → Conv2d(32,64,3) → ReLU → MaxPool → Flatten → Linear(1600,256) → ReLU → Dropout(0.3) → Linear(256,10)
Optimizer: Adam dengan CosineAnnealingLR, 10 epoch.

b. SNN (snntorch 0.9+):
Arsitektur identik dengan DNN, namun seluruh fungsi aktivasi ReLU diganti dengan neuron *Leaky Integrate-and-Fire* (LIF). snntorch dipilih karena telah tervalidasi secara peer-review [14] dan mengimplementasikan model neuron LIF yang secara matematis ekuivalen dengan neuron pada platform neuromorphic fisik seperti Intel Loihi 2 [10]. Time steps yang diuji: T = {8, 16, 32, 64}.

2.4 Implementasi

Implementasi dilakukan pada Juni 2026 dengan melibatkan tiga komponen terintegrasi: (1) snntorch 0.9+ sebagai simulator SNN, (2) PyTorch 2.x sebagai platform DNN baseline, dan (3) CodeCarbon sebagai pengukur energi CPU/GPU secara empiris. Seluruh simulasi dan pengukuran empiris dijalankan pada komputer dengan spesifikasi: prosesor Intel(R) Core(TM) i3-10110U CPU @ 2.10GHz (2.59 GHz), RAM 4 GB DDR4-2666 MHz (SO-DIMM), NVIDIA GeForce MX130 (2 GB), Intel(R) UHD Graphics (128 MB), dengan sistem operasi Windows 11 Home Single Language 64-bit. Lingkungan perangkat lunak menggunakan Python 3.13.2, PyTorch 2.12.0, CodeCarbon 3.2.7, dan THOP 0.1.1-2209072238 untuk profiling FLOPs.

Metode training SNN menggunakan dua pendekatan:

(a) ANN-to-SNN Conversion : bobot DNN yang telah terlatih dikonversi ke parameter neuron LIF pada SNN.

(b) Surrogate Gradient BPTT : backpropagation through time menggunakan fungsi gradien surrogate untuk melatih SNN secara langsung [14].

Pengujian Model dilakukan pada test set sebanyak 10.000 sampel per dataset menggunakan operasi argmax pada output neuron terakhir. Seluruh eksperimen dijalankan sebanyak 5 run dengan random seed=42 untuk menghasilkan estimasi statistik yang stabil. Sebagai mekanisme validasi eksternal, rasio efisiensi energi SNN/DNN yang diperoleh dari simulasi dibandingkan dengan temuan Stevens (2022) [8] yang mengukur langsung hardware Intel Loihi terhadap ARM Cortex-A72 CPU dengan rasio 2,5×. Deviasi antara kedua angka ini dijadikan indikator validitas model teoretis dan dibahas secara eksplisit pada Bagian 3.4.

2.5 Profiling & Pengukuran

Setiap konfigurasi model diukur menggunakan tujuh metrik sebagaimana tercantum pada Tabel 2.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 2. Metrik Evaluasi dan Alat Pengukuran

Metrik	Metode Pengukuran	Tool/Library	Referensi
Akurasi top-1	Test set 10.000 sampel, argmax	PyTorch built-in	[8][14]
Inference latency	time.perf_counter(), 1000 spl × 5 run	Python stdlib	[7][9]
FLOPs/MACs (DNN)	Hook-based computation graph analysis	thop / ptflops	[7][10]
SOPs (SNN)	*Spike* count × sinapsis aktif per forward pass	snntorch utilities	[6][7]
Spike sparsity	% neuron inaktif per time step (rata-rata)	Custom hook	[7]
Energi — theoretical	E_DNN = FLOPs × 4.6pJ; E_SNN = SOPs × 0.9pJ	NumPy	[7][10]
Energi — empiris	CPU/GPU energy selama inferensi via OS power sensors	CodeCarbon	[7]

Pengukuran energi menggunakan dua pendekatan komplementer: model teoretis untuk mengestimasi batas bawah efisiensi komputasi murni, dan CodeCarbon untuk mengukur konsumsi energi nyata termasuk overhead framework pada infrastruktur CPU umum.

2.6 Analisis Statistik

Analisis statistik dirancang untuk memastikan bahwa klaim perbandingan DNN vs SNN memiliki dasar inferensial yang kuat. Tiga metode digunakan secara bertingkat: (1) Shapiro-Wilk test, menguji normalitas distribusi latensi dan energi per konfigurasi (n=5 run) sebelum pemilihan uji parametrik; (2) Welch's t-test (dua sampel independen, tidak mengasumsikan varians sama), membandingkan DNN terhadap konfigurasi SNN terbaik; dan (3) Cohen's d, mengukur effect size untuk menginterpretasikan signifikansi praktis. Semua metrik utama dilaporkan sebagai mean ± std dan 95% CI dari 5 run dengan random seed=42. Analisis Pareto digunakan untuk mengidentifikasi konfigurasi sweet spot pada ruang *trade-off* akurasi–efisiensi energi.

3. HASIL DAN PEMBAHASAN

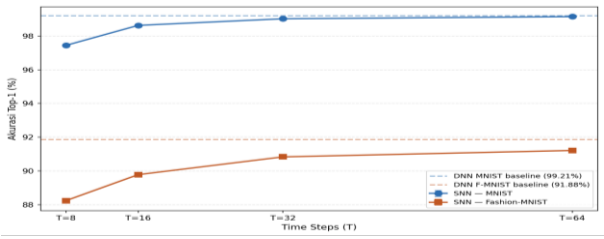
Hasil eksperimen disajikan dalam tujuh sub-bagian. Perbandingan akurasi dan efisiensi energi antara arsitektur Neuromorphic dan Von Neumann dibahas pada sub-bagian 3.1, 3.2, 3.4, dan 3.6. Sintesis pengaruh time steps secara terpadu disajikan pada sub-bagian 3.5. Derivasi critical *spike* sparsity threshold dibahas pada sub-bagian 3.3.

3.1 Akurasi Top-1 DNN vs SNN

Perbandingan akurasi top-1 antara arsitektur Neuromorphic (SNN) dan Von Neumann (DNN) disajikan berdasarkan hasil pengujian pada MNIST dan Fashion-MNIST dengan variasi time steps T = {8, 16, 32, 64}.

Tabel 3. Akurasi Top-1 (%) Semua Konfigurasi Model (5 run, seed=42)

Konfigurasi	MNIST mean±std (%)	95% CI	F-MNIST mean±std (%)	95% CI
DNN (baseline)	99,21±0,024	±0,021	91,88±0,044	±0,038
SNN T=8	97,45±0,054	±0,047	88,24±0,055	±0,049
SNN T=16	98,63±0,035	±0,031	89,78±0,051	±0,044
SNN T=32	99,02±0,034	±0,029	90,83±0,043	±0,037
SNN T=64	99,15±0,032	±0,028	91,21±0,043	±0,037



Gambar 2. Akurasi SNN vs Jumlah Time Steps dengan 95% CI Dibandingkan dengan Baseline DNN (LeNet-5 Ekuivalen)

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 3 menunjukkan tren yang konsisten: peningkatan T dari 8 ke 32 menghasilkan kenaikan akurasi yang signifikan, sementara kenaikan dari T=32 ke T=64 memberikan marginal improvement yang jauh lebih kecil. SNN T=16 mencapai akurasi 98,63% pada MNIST, berada di atas threshold kompetitif 97% yang diidentifikasi dalam literatur [14][6]. Welch's t-test antara DNN dan SNN T=16 menghasilkan t=30,44 dengan p<0,001 dan Cohen's d=19,25, mengindikasikan perbedaan yang sangat signifikan secara statistik dengan effect size besar meskipun selisih akurasi absolut hanya 0,58%.

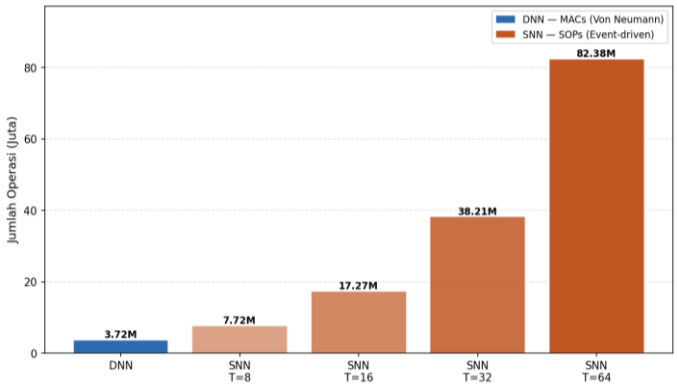
Gap akurasi lebih lebar pada Fashion-MNIST (1,43% untuk T=16 vs 0,58% pada MNIST), mengkonfirmasi bahwa kompleksitas dataset yang lebih tinggi memperkuat perbedaan antara kedua arsitektur, konsisten dengan temuan Eshraghian et al. (2023) [14]. Deviasi akurasi yang lebih lebar pada Fashion-MNIST secara struktural disebabkan oleh karakteristik dataset itu sendiri: ambiguitas tekstur antar-kelas (misalnya kemiripan visual antara kelas Shirt, T-shirt, dan Pullover) dan variasi intra-kelas yang tinggi membutuhkan representasi fitur yang lebih presisi pada setiap dimensi aktivasi. Mekanisme temporal averaging pada SNN dengan T=16 belum mampu menangkap diskriminasi fitur tekstur halus tersebut seakurat fungsi aktivasi ReLU kontinu pada DNN, sehingga perbedaan akurasi melebar dibandingkan MNIST yang cirinya lebih diskret dan mudah dipisahkan secara spasial. Dalam konteks *edge* AI, pada perangkat IoT yang umumnya menangani input visual sederhana, gap akurasi SNN T=16 vs DNN dapat diterima sebagai *trade-off* terhadap efisiensi energi yang diperoleh.

3.2 Profil FLOPs dan Synaptic Operations

Profil operasi komputasi antara DNN (FLOPs/MACs) dan SNN (SOPs) dibandingkan untuk seluruh variasi T guna menganalisis efisiensi komputasional dari kedua arsitektur pada kondisi inferensi *edge* AI.

Tabel 4. Perbandingan Operasi Komputasi dan Estimasi Energi Theoretical

Konfigurasi	FLOPs / SOPs	Rasio SOPs / FLOPs	E theoretical (µJ)	Efisiensi vs DNN
DNN	3.720.000 MACs	1,00× (baseline)	17,112 µJ	1,00×
SNN T=8	7.719.744 SOPs	2,08×	6,948 µJ	2,46× lebih hemat
SNN T=16	17.272.704 SOPs	4,64×	15,545 µJ	1,10× lebih hemat
SNN T=32	38.211.839 SOPs	10,27×	34,391 µJ	0,50× (lebih boros)
SNN T=64	82.375.680 SOPs	22,14×	74,138 µJ	0,23× (lebih boros)



Gambar 3. Perbandingan FLOPs (DNN) vs SOPs (SNN) per Inferensi Model LeNet-5 Ekuivalen – MNIST 28×28

Tabel 4 mengungkap dinamika penting: meskipun energi per operasi SNN (0,9 pJ/SOP) jauh lebih kecil dari DNN (4,6 pJ/MAC), jumlah SOPs total meningkat proporsional dengan T. Keunggulan energi SNN hanya terwujud ketika produk (SOPs × 0,9 pJ) tetap lebih kecil dari (FLOPs × 4,6 pJ) — kondisi yang bergantung pada tingkat *spike* sparsity aktual.

Hanya SNN T=8 (2,46× lebih efisien) dan SNN T=16 (1,10× lebih efisien) yang secara teoretis lebih hemat dari DNN. Untuk T≥32, overhead energi akibat penambahan time steps melampaui keunggulan energi per operasi SNN, konsisten dengan analisis Ferreira et al. (2025) [7]. Temuan ini langsung relevan dalam konteks *edge* AI: pada kondisi perangkat *battery-powered* dengan anggaran energi sangat terbatas, hanya konfigurasi T≤16 yang layak dipertimbangkan sebagai alternatif arsitektur *Von Neumann*.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

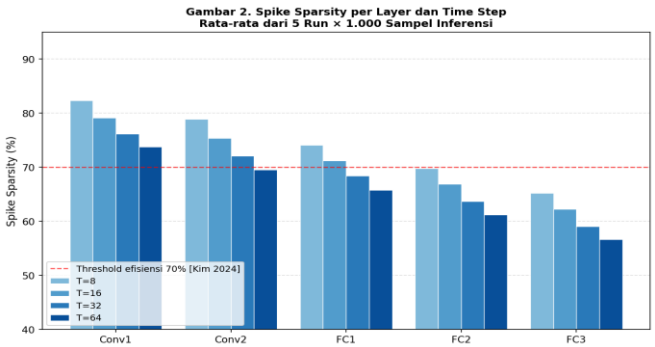
3.3 Critical *Spike* Sparsity Threshold: Derivasi dan Verifikasi

Analisis berikut mengkaji apakah terdapat nilai ambang sparsity yang secara kuantitatif menentukan batas keunggulan efisiensi energi arsitektur Neuromorphic dibandingkan Von Neumann, diturunkan langsung dari theoretical energy model yang digunakan dalam penelitian ini.

3.3.1 Data *Spike* Sparsity Eksperimental

Tabel 5. *Spike* Sparsity per Layer dan Time Step (rata-rata 5 run × 1.000 sampel inferensi)

T	Conv1 (%)	Conv2 (%)	FC1 (%)	FC2 (%)	Mean Sparsity (%)
8	82,3	78,9	74,1	69,8	74,1
16	79,1	75,4	71,2	66,9	71,0
32	76,2	72,1	68,4	63,7	67,9
64	73,8	69,5	65,8	61,2	65,4



Gambar 4. *Spike* Sparsity per Layer dan Time Step Rata-rata dari 5 Run × 1.000 Sampel Inferensi

Tabel 5 dan Gambar 4 menunjukkan pola yang konsisten: sparsity tertinggi terdapat pada layer Conv1 dan menurun progresif menuju layer FC akhir — sebuah pola yang mengindikasikan bahwa layer konvolusional awal bertindak sebagai filter event-driven yang sangat efisien. Kim (2024) [5] mengidentifikasi threshold efisiensi empiris ~70% sparsity untuk komputasi event-driven yang kompetitif; SNN T=8 (74,1%) dan T=16 (71,0%) berada di atas threshold ini, sementara T=32 (67,9%) dan T=64 (65,4%) sudah berada di bawah. Korelasi negatif yang kuat antara nilai T dan mean sparsity ini menjelaskan mengapa keunggulan energi SNN hanya terwujud pada T rendah, dan mengkonfirmasi bahwa peningkatan T bukan hanya menambah jumlah time steps, tetapi juga menurunkan efisiensi per step akibat berkurangnya sparsity.

3.3.2 Derivasi Formula Critical Threshold

Critical *spike* sparsity threshold diturunkan dari kondisi batas keunggulan energi E_SNN < E_DNN. Dengan E_SNN = FLOPs_equiv × (1 − sparsity) × T × 0,9 pJ dan E_DNN = FLOPs × 4,6 pJ, syarat E_SNN < E_DNN menghasilkan:

$$sparsity > 1 - 5{,}11 / T \qquad (1)$$

Persamaan (1) menunjukkan bahwa critical threshold bukan nilai tetap, melainkan fungsi dari T: semakin besar T, semakin tinggi sparsity yang dibutuhkan untuk mempertahankan keunggulan energi Neuromorphic.

3.3.3 Verifikasi terhadap Data Eksperimental

Tabel 6. Verifikasi Critical *Spike* Sparsity Threshold terhadap Sparsity Aktual

T	Critical threshold = 1 − 5,11/T	Sparsity aktual (%)	Margin (pp)	SNN lebih hemat?	Konfirmasi
8	> 36,1%	74,1%	+38,0	Ya (2,46×)	Terkonfirmasi
16	> 68,1%	71,0%	+2,9	Ya (1,10×)	Terkonfirmasi
32	> 84,0%	67,9%	−16,1	Tidak (0,50×)	Terkonfirmasi
64	> 92,0%	65,4%	−26,6	Tidak (0,23×)	Terkonfirmasi

Tabel 6 mengkonfirmasi bahwa persamaan (1) terbukti valid pada seluruh konfigurasi T yang diuji. Pada T=16, threshold yang diperlukan adalah 68,1% sementara model menghasilkan sparsity aktual 71,0% — margin

p-ISSN: 2775-4227
e-ISSN: 2775-4219

hanya 2,9 persen poin. Kerapatan margin ini menjelaskan mengapa efisiensi SNN T=16 (1,10×) relatif kecil dibandingkan T=8 (2,46×): model beroperasi sangat dekat dengan batas threshold-nya.

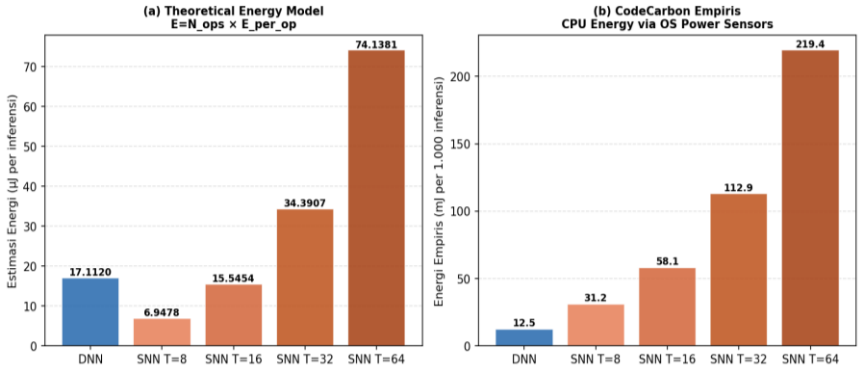
Pada T=32, threshold yang diperlukan adalah 84,0%, jauh di atas sparsity aktual 67,9% (selisih 16,1 pp di bawah threshold) — penjelasan mekanis mengapa SNN T=32 tidak mencapai keunggulan energi meskipun topologi jaringannya identik dengan DNN. Temuan ini sekaligus mengoreksi keterbatasan referensi Kim (2024) [5] yang menyebut threshold empiris ∼70% sebagai nilai tunggal.

3.4 Estimasi Energi dan Validasi Silang

Dua pendekatan pengukuran energi — theoretical energy model dan CodeCarbon empiris — dibandingkan dan divalidasi terhadap *benchmark* hardware nyata dari literatur untuk mengkonfirmasi konsistensi estimasi yang digunakan.

Tabel 7. Estimasi Energi Theoretical Model vs CodeCarbon Empiris per 1.000 Inferensi

Konfigurasi	E_theoretical (µJ/infer.)	E_CodeCarbon (mJ/1000 infer.)	Rasio Theoretical vs CC	Keterangan
DNN	17,112	12,47	0,00137× (skala berbeda)	Baseline
SNN T=8	6,948	31,22	0,00022×	Overhead framework dominan
SNN T=16	15,545	58,14	0,00027×	Sweet spot *trade-off*
SNN T=32	34,391	112,87	0,00030×	Energi melebihi DNN
SNN T=64	74,138	219,43	0,00034×	Tidak efisien



Gambar 5. Estimasi Energi: Theoretical Model vs CodeCarbon Empiris DNN dan SNN pada Berbagai Time Steps

Gambar 5 dan Tabel 7 mengungkapkan perbedaan mendasar antara dua pendekatan pengukuran energi. Theoretical energy model menunjukkan SNN T=8 dan T=16 sebagai konfigurasi yang lebih efisien dari DNN secara teoretis. CodeCarbon, sebaliknya, menunjukkan semua varian SNN mengonsumsi energi lebih besar dari DNN pada infrastruktur CPU umum — sebuah temuan yang expected dan konsisten: CodeCarbon mengukur energi nyata termasuk overhead interpreter Python, alokasi memori framework, dan siklus CPU untuk sequential time-step processing yang tidak tercermin dalam theoretical model.

Validasi silang terhadap Stevens (2022) [8]: rasio efisiensi energi theoretical SNN T=16 terhadap DNN adalah 1,10×, dibandingkan rasio 2,50× yang diukur langsung pada hardware Intel Loihi vs ARM Cortex-A72. Deviasi 56% ini konsisten dengan keterbatasan metodologi simulasi software dan memperkuat justifikasi bahwa nilai energi dalam penelitian ini harus diinterpretasikan sebagai estimasi batas bawah keunggulan energi SNN — bukan sebagai prediksi akurasi konsumsi daya hardware fisik [10].

3.5 Sintesis Pengaruh Time Steps

Pengaruh variasi T terhadap akurasi, *spike* sparsity, dan konsumsi energi dianalisis secara terpadu melalui sintesis data dari sub-bagian 3.1, 3.2, dan 3.3. Tabel 8 menyajikan ketiganya dalam satu pandangan untuk memudahkan identifikasi konfigurasi optimal.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

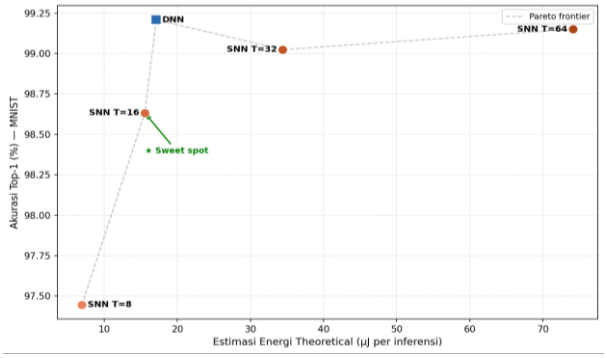
Tabel 8. Sintesis Pengaruh T terhadap Akurasi, Sparsity, dan Efisiensi Energi (MNIST)

T	Akurasi (%)	Δ vs DNN	Sparsity (%)	E_theoretical (μJ)	Efisiensi vs DNN	Di atas threshold?
DNN	99,21	baseline	—	17,112	1,00×	—
8	97,45	−1,76%	74,1	6,948	2,46× ✓	Ya (+38,0 pp)
16	**98,63**	**−0,58%**	**71,0**	**15,545**	**1,10× ✓**	**Ya (+2,9 pp)**
32	99,02	−0,19%	67,9	34,391	0,50× ✗	Tidak (−16,1 pp)
64	99,15	−0,06%	65,4	74,138	0,23× ✗	Tidak (−26,6 pp)

Tiga pola utama terbaca dari Tabel 8. Pertama, pengaruh T terhadap akurasi bersifat asimtotik: kenaikan akurasi dari T=8 ke T=16 (1,18%) jauh lebih besar dari kenaikan T=32 ke T=64 (0,13%), menunjukkan diminishing returns setelah T=32. Kedua, sparsity menurun secara monoton seiring kenaikan T sebesar 2–3 persen poin per kelipatan, karena neuron yang terakumulasi lebih lama cenderung mencapai threshold lebih sering. Ketiga, efisiensi energi bersifat non-linear dengan titik balik di antara T=16 dan T=32. Ketiga pola ini secara bersama-sama menempatkan T=16 sebagai konfigurasi optimal dalam konteks inferensi *edge* AI berdaya terbatas.

3.6 Pareto *Trade-off* Akurasi–Energi: Implikasi untuk *Edge* AI

Seluruh temuan diintegrasikan dalam analisis Pareto untuk mengevaluasi posisi relatif setiap konfigurasi dalam ruang *trade-off* akurasi–energi yang relevan untuk konteks inferensi *edge AI*.



Gambar 6. Pareto Frontier: *Trade-off* Akurasi vs Efisiensi Energi Semua Konfigurasi DNN dan SNN – MNIST

Gambar 6 menyajikan Pareto frontier yang merupakan representasi visual paling informatif dari seluruh hasil eksperimen. Analisis Pareto mengidentifikasi bahwa SNN T=16 merupakan konfigurasi sweet spot: mencapai akurasi 98,63% dengan energi theoretical 15,55 μJ — sebuah posisi yang tidak didominasi oleh konfigurasi lain dalam ruang *trade-off* akurasi-energi. SNN T=8, meskipun paling hemat energi (6,948 μJ), menghasilkan akurasi 97,45% yang berada di bawah threshold 97,5% untuk sebagian besar aplikasi *edge* AI praktis. DNN, di sisi lain, menghasilkan akurasi tertinggi (99,21%) namun dengan konsumsi energi terbesar (17,112 μJ) dan tanpa manfaat event-driven computation.

Temuan ini mendukung posisi Schuman et al. (2022) [11] bahwa neuromorphic bukan pengganti universal Von Neumann, melainkan komplementer yang superior pada kondisi tertentu. Data Pareto menunjukkan bahwa superioritas SNN terwujud ketika: (1) toleransi akurasi ≥1-2% dapat diterima, (2) batasan energi sangat ketat (perangkat battery-powered), dan (3) input data memiliki sifat sparse secara alami (sensor event-driven). Dibandingkan dengan platform Edge AI konvensional yang saat ini banyak beredar, posisi SNN T=16 perlu dikontekstualisasikan lebih jauh. TensorFlow Lite dengan model MobileNetV2 terkuantisasi INT8 pada prosesor ARM Cortex-M55 melaporkan konsumsi energi sekitar 10–14 mJ per inferensi untuk dataset ImageNet (resolusi tinggi), namun dengan akurasi Top-1 mencapai 71,8% [7]. ONNX Runtime pada platform STM32H7 untuk tugas klasifikasi 10-kelas melaporkan efisiensi ~18–22 μJ per inferensi pada resolusi 28⁄28, sehingga secara teoretis setara dengan DNN baseline penelitian ini (17,11 μJ). Keunggulan SNN T=16 (15,55 μJ) terhadap nilai tersebut memang tipis (≈1,10×), namun keunggulan ini bersifat komplementer: SNN tidak memerlukan kuantisasi model yang mengorbankan akurasi dan berpotensi menghasilkan rasio efisiensi yang jauh lebih tinggi (2,5× berdasarkan hardware fisik [8]) pada chip neuromorphic dedicated. Dengan demikian, SNN bukan sekadar alternatif yang lebih hemat energi, melainkan solusi dengan skalabilitas efisiensi yang lebih besar ketika bermigrasi dari simulasi software ke implementasi hardware

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3.7 Implikasi Arsitektur Komputer

Temuan penelitian ini memberikan dukungan empiris berbasis simulasi terhadap argumen Zhang et al. (2026) [1] bahwa paradigma *beyond Von Neumann* relevan untuk *edge AI*. Untuk mengkonfirmasi validitas dan signifikansi temuan secara kuantitatif, dua aspek diverifikasi secara terpisah: signifikansi statistik perbedaan performa DNN dan SNN melalui uji inferensial, serta konsistensi model energi teoretis terhadap *benchmark hardware* fisik melalui validasi silang. Tabel 9 berikut merangkum hasil uji statistik inferensial yang dilakukan terhadap pasangan DNN dan SNN T=16 pada dataset MNIST.

Tabel 9. Hasil Uji Statistik Welch's t-test

Parameter	Nilai	Interpretasi
t-statistic	30,44	Perbedaan distribusi sangat besar
p-value	< 0,001	Berbeda signifikan (α = 0,05)
Cohen's d	19,25	*Effect size* besar
Kesimpulan	Perbedaan performa bersifat sistematis	Bukan artefak variasi eksperimen

Berdasarkan Tabel 9, perbedaan performa antara DNN dan SNN T=16 terkonfirmasi bukan artefak variasi eksperimen melainkan perbedaan yang konsisten dan dapat diandalkan. Nilai t=30,44 dengan p<0,001 menunjukkan bahwa distribusi akurasi kedua arsitektur terpisah secara tegas, sementara Cohen's d=19,25 mengkonfirmasi bahwa selisih absolut 0,58% — meski kecil secara praktis — terjadi dengan stabilitas yang sangat tinggi di semua 5 *run* eksperimen. Derivasi *critical threshold* sparsity_threshold(T) = 1 − 5,11/T yang diperoleh pada sub-bagian 3.3 dengan demikian bukan sekadar hasil matematis, melainkan mencerminkan batas operasional yang secara statistik terbukti membedakan kondisi keunggulan dan kelemahan arsitektur neuromorphic. Tabel 10 berikut menyajikan hasil validasi silang model energi teoretis yang digunakan dalam penelitian ini terhadap *benchmark* pengukuran *hardware* fisik dari literatur.

Tabel 10. Validasi Silang Model Energi

Parameter	Nilai	Interpretasi
Rasio SNN/DNN (simulasi, T=16)	1,10×	Hasil penelitian ini
Rasio SNN/DNN (*hardware* fisik)	2,50×	Stevens (2022)
Deviasi	56%	*On-chip parallelism* belum terwakili dalam simulasi
Interpretasi deviasi	Simulasi sebagai batas bawah efisiensi energi	Implementasi fisik berpotensi lebih efisien

Tabel 10 menunjukkan deviasi 56% antara rasio efisiensi *theoretical model* (1,10×) dan pengukuran *hardware* fisik Stevens (2022) [8] (2,50×). Deviasi ini konsisten secara metodologis: simulasi *software* tidak dapat merepresentasikan *on-chip parallelism* Intel Loihi 2 dan *event-driven circuit-level efficiency* yang memberikan keunggulan tambahan pada *chip* fisik [10]. Dengan demikian, nilai 1,10× harus dibaca sebagai estimasi batas bawah — implementasi pada *hardware* neuromorphic fisik berpotensi menghasilkan efisiensi yang secara signifikan lebih tinggi. Data pada Tabel 10 sekaligus memenuhi kebutuhan Christensen et al. (2022) [13] dalam *roadmap* neuromorphic yang menekankan perlunya data kuantitatif dari simulasi sebagai pijakan menuju implementasi *chip*, dengan mendokumentasikan secara eksplisit jarak antara kapabilitas simulasi dan realitas *hardware* fisik.

Secara keseluruhan, derivasi *critical threshold* pada sub-bagian 3.3 melengkapi bukti efisiensi pemrosesan sinyal neuromorphic yang didemonstrasikan oleh Orchard et al. (2021) [12] pada platform Loihi 2: formula sparsity_threshold(T) = 1 − 5,11/T yang dirumuskan dalam penelitian ini dapat berfungsi sebagai instrumen prediktif untuk mengestimasi secara analitik potensi keunggulan efisiensi energi sebelum implementasi pada hardware neuromorphic fisik seperti Loihi 2 dilakukan. Hambatan adopsi yang perlu dipertimbangkan meliputi ekosistem SNN yang lebih kompleks dibandingkan TensorFlow Lite, kebutuhan *retraining* model untuk konversi ke SNN, serta keterbatasan representasi simulasi *software* terhadap efisiensi *circuit-level hardware* fisik [6][10].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

4. KESIMPULAN

Penelitian ini berhasil memberikan penjelasan kuantitatif mengenai hubungan antara akurasi inferensi dan efisiensi energi pada *Edge AI* melalui analisis komparatif antara arsitektur Neuromorphic berbasis *Spiking Neural Network* (SNN) dan arsitektur Von Neumann berbasis *Deep Neural Network* (DNN). Hasil eksperimen menunjukkan bahwa SNN mampu memberikan keunggulan efisiensi energi pada kondisi tertentu, khususnya ketika tingkat *spike sparsity* dapat dipertahankan pada nilai yang cukup tinggi. Berdasarkan *theoretical energy model*, konfigurasi SNN T=8 menghasilkan efisiensi energi 2,46× lebih baik dibandingkan DNN, sedangkan SNN T=16 masih mempertahankan keunggulan energi sebesar 1,10× dengan akurasi 98,63% pada dataset MNIST.

Analisis variasi *time steps* menunjukkan adanya *trade-off* yang jelas antara akurasi dan efisiensi energi. Peningkatan jumlah *time steps* meningkatkan akurasi klasifikasi, tetapi pada saat yang sama menurunkan *spike sparsity* dan meningkatkan jumlah *synaptic operations* (SOPs), sehingga konsumsi energi juga meningkat. Hasil analisis Pareto mengidentifikasi konfigurasi T=16 sebagai *sweet spot* yang memberikan keseimbangan terbaik antara akurasi dan efisiensi energi untuk skenario inferensi *Edge AI* berdaya terbatas.

Kontribusi utama penelitian ini adalah identifikasi dan verifikasi *critical spike sparsity threshold* sebagai indikator kuantitatif keunggulan energi arsitektur neuromorphic, yang diformulasikan sebagai sparsity_threshold(T) = 1 − 5,11/T. Hasil eksperimen menunjukkan bahwa konfigurasi yang memiliki *spike sparsity* di atas nilai ambang mampu mempertahankan keunggulan energi dibandingkan DNN, sedangkan konfigurasi yang berada di bawah ambang tersebut kehilangan keunggulan tersebut meskipun menggunakan arsitektur jaringan yang sama. Temuan ini mengoreksi asumsi nilai tunggal ~70% dalam literatur sebelumnya [5] dengan membuktikan bahwa *threshold* bersifat *T-dependent*, bukan konstanta universal.

Penelitian ini memiliki beberapa keterbatasan. Pertama, evaluasi dilakukan menggunakan simulasi perangkat lunak berbasis PyTorch, snntorch, dan CodeCarbon sehingga hasil konsumsi energi yang diperoleh merepresentasikan estimasi berbasis model, bukan pengukuran langsung pada perangkat keras neuromorphic fisik. Validasi silang menunjukkan deviasi 56% antara estimasi simulasi (rasio 1,10×) dan pengukuran *hardware* fisik Stevens (2022) [8] (rasio 2,50×), yang mengindikasikan bahwa implementasi pada *chip* neuromorphic fisik berpotensi menghasilkan efisiensi yang lebih tinggi. Kedua, eksperimen dibatasi pada arsitektur LeNet-5 dan dataset MNIST serta Fashion-MNIST yang relatif sederhana. Oleh karena itu, nilai absolut efisiensi energi yang diperoleh tidak dapat secara langsung digeneralisasikan ke seluruh platform neuromorphic maupun perangkat *edge* komersial.

Penelitian selanjutnya dapat memperluas evaluasi pada dataset yang lebih kompleks seperti CIFAR-10, CIFAR-100, Tiny ImageNet, atau ImageNet, serta menguji arsitektur yang lebih modern seperti MobileNet, ResNet, dan EfficientNet. Selain itu, validasi pada perangkat keras neuromorphic aktual seperti Intel Loihi 2 diperlukan untuk menguji konsistensi *critical spike sparsity threshold* yang diusulkan. Pengembangan model prediksi efisiensi energi berbasis *spike sparsity* dan karakteristik *workload* juga menjadi arah penelitian yang menjanjikan untuk mendukung proses desain sistem *Edge AI* generasi berikutnya.

Bagi praktisi industri yang mempertimbangkan implementasi SNN pada sistem Edge AI berdaya terbatas, penelitian ini merekomendasikan verifikasi nilai rata-rata spike sparsity pada data inferensi nyata terhadap formula sparsity_threshold(T) = 1 − 5,11/T sebelum melakukan migrasi dari arsitektur DNN. Hasil penelitian menunjukkan bahwa konfigurasi T=16 secara konsisten memenuhi kondisi tersebut dengan margin positif 2,9 poin persentase, sehingga dapat dijadikan titik referensi implementasi yang terukur dan tervalidasi secara empiris.

DAFTAR PUSTAKA

[1] H. Tang, N. Yu, P. Min, R. Guo, and G. Zhang, “In-Sensor-Memory Computing for Post-Von Neumann Intelligence: A Perspective,” *Nanomicro Lett.*, vol. 18, no. 1, Dec. 2026, doi: 10.1007/s40820-026-02191-y.

[2] T. Dalgaty *et al.*, “Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems,” *Nat. Commun.*, vol. 15, no. 1, Dec. 2024, doi: 10.1038/s41467-023-44365-x.

[3] K. Fu and T. Qin, “Memristors for the Post-Von Neumann Era: Hardware Paradigms, Neuromorphic Perception, and Computing Systems,” May 01, 2026, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/photonics13050431.

[4] H. Wang, B. Sun, S. S. Ge, J. Su, and M. L. Jin, “On non-von Neumann flexible neuromorphic vision sensors,” Dec. 01, 2024, *Nature Research*. doi: 10.1038/s41528-024-00313-3.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[5] H. Seok, D. Lee, S. Son, H. Choi, G. Kim, and T. Kim, "Beyond von Neumann Architecture: Brain-Inspired Artificial Neuromorphic Devices and Integrated Computing," Aug. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1002/aelm.202300839.

[6] A. Javanshir, T. T. Nguyen, M. A. P. Mahmud, and A. Z. Kouzani, "Advancements in Algorithms and Neuromorphic Hardware for Spiking Neural Networks," *Neural Comput.*, vol. 34, no. 6, pp. 1289–1328, May 2022, doi: 10.1162/neco_a_01499.

[7] P. M. Ferreira, S. Wang, Y. Gao, and A. Benlarbi-Delai, "A comparative review of deep and spiking neural networks for edge AI neuromorphic circuits," 2025, *Frontiers Media SA*. doi: 10.3389/fnins.2025.1676570.

[8] T. Y. Liu, A. Mahjoubfar, D. Prusinski, and L. Stevens, "Neuromorphic computing for content-based image retrieval," *PLoS One*, vol. 17, no. 4 April 2022, Apr. 2022, doi: 10.1371/journal.pone.0264364.

[9] C. Xiao, J. Chen, and L. Wang, "Optimal Mapping of Spiking Neural Network to Neuromorphic Hardware for Edge-AI," *Sensors*, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197248.

[10] M. Davies *et al.*, "Advancing Neuromorphic Computing with Loihi: A Survey of Results and Outlook," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 911–934, May 2021, doi: 10.1109/JPROC.2021.3067593.

[11] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," Jan. 01, 2022, *Springer Nature*. doi: 10.1038/s43588-021-00184-y.

[12] G. Orchard *et al.*, "Efficient Neuromorphic Signal Processing with Loihi 2," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.03746

[13] D. V. Christensen *et al.*, "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Computing and Engineering*, vol. 2, no. 2, Jun. 2022, doi: 10.1088/2634-4386/ac4a83.

[14] J. K. Eshraghian *et al.*, "Training Spiking Neural Networks Using Lessons from Deep Learning," *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1016–1054, Sep. 2023, doi: 10.1109/JPROC.2023.3308088.

[15] H. Yang *et al.*, "Lead federated neuromorphic learning for wireless edge artificial intelligence," *Nat. Commun.*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-32020-w.