Siska Fitriani

Template dan Pedoman Penulisan Artikel PROGRESIF_siska.pdf

Assignment

Document Details

Submission ID
trn:oid:::3618:125738030

Submission Date
Jan 7, 2026, 9:31 AM GMT+7

Download Date
Jan 7, 2026, 9:32 AM GMT+7

File Name
unknown_filename

File Size
229.5 KB

10 Pages

4,759 Words

27,913 Characters

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text

Top Sources

5% Internet sources
4% Publications
7% Submitted works (Student Papers)

Top Sources

5% Internet sources
4% Publications
7% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
University of Southern Mississippi on 2023-02-21		2%
2	Publication	
"The Endocrine Society's 99th Annual Meeting and Expo", Endocrine Reviews, 2017		1%
3	Student papers	
Christchurch Polytechnic Institute of Technology on 2017-11-16		<1%
4	Internet	
netapps.internetworks.my		<1%
5	Internet	
repository.uin-suska.ac.id		<1%
6	Internet	
publications.polymtl.ca		<1%
7	Student papers	
itera on 2025-03-20		<1%
8	Publication	
Rida Adhari Yanti, Novaliyosi Novaliyosi. "Systematic Literature Review: Model Pe...		<1%
9	Internet	
www.mdpi.com		<1%
10	Publication	
Fika Trisnawati, Sigit Doni Ramdan, Muhammad Anwar Sadat Faidar, Jaka Persad...		<1%
11	Student papers	
Higher Education Commission Pakistan on 2024-11-15		<1%

12 Internet
eprints.undip.ac.id <1%

13 Internet
hj.diva-portal.org <1%

14 Internet
talenta.usu.ac.id <1%

15 Internet
doaj.org <1%

16 Internet
jurnal.univpgri-palembang.ac.id <1%

17 Internet
qspace.qu.edu.qa <1%

18 Internet
www.coursehero.com <1%

19 Publication
"Intelligent Sustainable Systems", Springer Science and Business Media LLC, 2023 <1%

20 Publication
Christ Mario, Ryan Randy Suryono. "Public Sentiment Analysis on Dirty Vote Movi... <1%

Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Systematic Literature Review: GPU, TPU, FPGA dalam Akselerasi AI

Siska Fitriani, Ega Budiman, Nitami Evita Inonu, Amarudin
Fakultas Teknik dan Ilmu Komputer, Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
siskafitriani@teknokrat.ac.id, ega_budiman@teknokrat.ac.id,
nitami_evita_inonu@teknokrat.ac.id, amarudin@teknokrat.ac.id
* siskafitriani@teknokrat.ac.id

Abstract

The increasing complexity of artificial intelligence (AI) models has raised the demand for efficient hardware accelerators. A key challenge is selecting an accelerator that aligns with application needs, as mismatches can affect energy efficiency, inference speed, and system scalability. GPU, TPU, and FPGA are the most commonly used accelerators in AI deployment, each with specific advantages and limitations. This study aims to systematically evaluate the utilization of these three accelerators across various AI domains. A Systematic Literature Review (SLR) was conducted using the Kitchenham framework, analyzing 20 scientific articles. Results show that GPUs are used in 90% of studies, TPUs in 50%, and FPGAs in 65%. In terms of energy efficiency, FPGAs are superior in 78% of the relevant articles, while GPUs dominate inference performance in 85% of cases. This study concludes that selecting an AI accelerator should be guided by power efficiency, system architecture, and domain-specific requirements. The findings offer practical implications for graduate students in selecting appropriate accelerators that align with their research topics, experimental goals, and resource constraints in AI-driven thesis projects.

Keywords: AI Accelerator, GPU, TPU, FPGA, Systematic Review.

Abstrak

Perkembangan model kecerdasan buatan (AI) yang semakin kompleks menimbulkan kebutuhan akan akselerator perangkat keras yang efisien. Masalah utama yang sering dihadapi adalah pemilihan akselerator yang tidak sesuai dengan kebutuhan aplikasi, yang berdampak pada efisiensi daya, kecepatan inferensi, dan skalabilitas sistem. GPU, TPU, dan FPGA merupakan tiga jenis akselerator yang paling banyak digunakan dalam implementasi AI, namun masing-masing memiliki kelebihan dan keterbatasan. Tujuan utama dari penelitian ini adalah untuk mengevaluasi secara sistematik pemanfaatan ketiga akselerator tersebut dalam berbagai domain aplikasi AI. Metode yang digunakan adalah Systematic Literature Review (SLR) berbasis pendekatan Kitchenham. Data terdiri dari 20 artikel ilmiah yang dipilih secara sistematis. Hasil analisis menunjukkan bahwa GPU digunakan dalam 90% studi, TPU dalam 50%, dan FPGA dalam 65%. Dalam hal efisiensi energi, FPGA dinyatakan unggul dalam 78% studi yang relevan, sementara GPU unggul dalam performa inferensi dalam 85% kasus. Penelitian ini menyimpulkan bahwa pemilihan akselerator AI perlu mempertimbangkan efisiensi daya, arsitektur sistem, dan kebutuhan domain spesifik. Implikasinya, hasil studi ini dapat menjadi pedoman praktis bagi mahasiswa magister dalam memilih akselerator yang sesuai dengan topik, tujuan eksperimen, dan keterbatasan sumber daya dalam penelitian berbasis AI.

Kata Kunci: Akselerator AI, GPU, TPU, FPGA, Tinjauan Sistematik

1. Pendahuluan

Seiring meningkatnya kompleksitas model Artificial Intelligence (AI), pemilihan akselerator perangkat keras yang tepat seperti Graphics Processing Unit (GPU), Tensor Processing Unit (TPU), dan Field Programmable Gate Array (FPGA) menjadi krusial dalam menjamin efisiensi komputasi. Pendekatan berbasis bukti (evidence-based approach) kini diadopsi dalam ilmu komputer untuk membantu pengambilan keputusan arsitektural, sebagaimana sebelumnya diterapkan dalam bidang medis dan sosial.

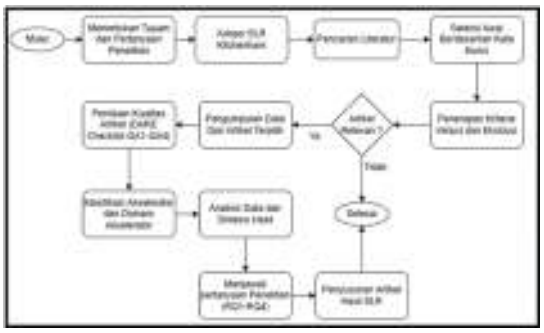
GPU, TPU, dan FPGA masing-masing memiliki keunggulan dan keterbatasan dalam menangani beban kerja AI, baik untuk pelatihan maupun inferensi. Faktor seperti efisiensi waktu, konsumsi daya, serta skalabilitas sistem menjadi pertimbangan utama. Oleh karena itu, kajian berbasis bukti sangat dibutuhkan untuk menghindari kesalahan strategis dalam investasi teknologi.

Penelitian ini menggunakan pendekatan Systematic Literature Review (SLR) [1], untuk menyusun dan mengevaluasi temuan dari 20 artikel terpilih mengenai akselerasi AI menggunakan GPU, TPU, dan FPGA. Huang et al. (2022) [2], menekankan pentingnya algorithm-hardware co-design untuk mengoptimalkan kinerja akselerator. Bertazzoni et al. (2024) [3] menunjukkan bahwa FPGA unggul dalam efisiensi daya di lingkungan edge, sementara Martín-Martín et al. (2024) [4] membuktikan bahwa FPGA dapat mengungguli GPU hingga 20× dan TPU hingga 3× dalam inferensi model CNN.

Dalam konteks GPU, Kaur et al. (2025) [5], membahas strategi penjadwalan seperti adaptive warp scheduling dan prefetching untuk meningkatkan efisiensi eksekusi, sementara itu Hazman et al. (2024) [6] menunjukkan bahwa implementasi intrusion detection system (IDS) berbasis TPU menghasilkan klasifikasi cepat (6 ms) dan akurasi tinggi (99.94%) dalam domain smart city dan IoT.

2. Metodologi

Pendekatan yang digunakan dalam penelitian ini mengacu pada kerangka kerja (SLR) yang diusulkan oleh Kitchenham [1]. SLR merupakan metode penelitian yang sistematis dan terstruktur untuk mengidentifikasi, mengevaluasi, serta mensintesis bukti-bukti yang relevan dari kumpulan literatur ilmiah yang tersedia. Metode ini membantu memastikan bahwa penelitian yang dilakukan komprehensif, transparan, dan dapat direproduksi. Gambar 1 berikut menunjukkan alur penelitian berdasarkan pendekatan SLR yang digunakan dalam studi ini.



Gambar 1 Alur Penelitian

2.1 Pertanyaan Penelitian (RQ)

Untuk menjaga fokus dan kejelasan penelitian, beberapa pertanyaan penelitian telah dirumuskan sebagai berikut:

a. RQ1: Apa saja arsitektur, framework, dan tools yang digunakan pada GPU, TPU, dan FPGA dalam akselerasi AI?
b. RQ2: Bagaimana performa akselerator GPU, TPU, dan FPGA dalam hal training dan inferensi?
c. RQ3: Bagaimana efisiensi dan kinerja dari GPU, TPU, dan FPGA dalam akselerasi AI?
d. RQ4: Apa tantangan dan keterbatasan teknis dalam deployment akselerator AI tersebut di berbagai domain aplikasi?

Setiap pertanyaan penelitian dirancang untuk menjawab aspek-aspek kritis dari topik yang ditinjau, sehingga hasilnya bermanfaat baik untuk kalangan akademisi maupun industri.

2.2 Proses Pencarian Literatur

Proses pencarian dalam studi ini mengacu pada pendekatan sistematis yang diadaptasi dari metodologi (SLR) yang dikembangkan oleh Kitchenham. Pencarian dilakukan secara manual melalui sejumlah jurnal dan prosiding konferensi terkemuka, seperti IEEE, ACM, Springer, dan Elsevier, dengan rentang waktu publikasi dari tahun 2020 hingga 2025. Sumber literatur dipilih berdasarkan reputasi dan relevansi topik dengan kajian tentang konsep deep learning, arsitektur CNN, tantangan, aplikasi, dan pendekatan komputasional seperti penggunaan FPGA, GPU, dan CPU. Kata kunci utama yang digunakan dalam pencarian meliputi kelompok kata seperti: ("Deep Learning") , ("Convolutional Neural Network" ATAU "CNN") , ("Architecture" ATAU "Model Compression" ATAU "Hardware Acceleration") , ("FPGA" ATAU "GPU" ATAU "CPU") , serta kombinasi lainnya yang relevan, seperti ("Medical Image Analysis") , ("Edge Computing") , dan ("Transfer Learning") . Proses pencarian bertujuan untuk mengidentifikasi artikel yang membahas aspek teoritis maupun implementasi praktis dalam penerapan deep learning, khususnya yang berkaitan dengan efisiensi komputasi dan optimasi model.

2.3 Inklusi dan Ekslusi Literatur

Penelitian ini akan menyertakan literatur yang membahas penggunaan GPU, TPU, dan FPGA dalam akselerasi (AI), termasuk perkembangan terkini, arsitektur, framework, serta tools yang digunakan. Selain itu, sumber daya yang membahas efisiensi, kinerja, tantangan, dan solusi implementasi dari masing-masing perangkat keras juga akan dimasukkan. Artikel yang relevan harus membahas aspek teknis penerapan AI, seperti training dan inferensi model machine learning atau deep learning, serta komparasi performa antar perangkat keras.

Sebaliknya, penelitian ini akan mengeksklusi literatur yang hanya membahas konsep AI tanpa fokus pada akselerasi perangkat keras, studi kasus yang tidak melibatkan GPU/TPU/FPGA secara eksplisit, artikel yang tidak tersedia dalam bahasa Inggris atau Indonesia, serta makalah yang tidak melalui proses peer-review atau publikasi di jurnal/sumber tidak terpercaya. Sumber yang hanya memberikan informasi sekilas tanpa analisis mendalam mengenai perbandingan antara GPU, TPU, dan FPGA juga akan dikeluarkan dari tinjauan ini.

2.4 Penilaian Kualitas

Setiap artikel dalam tinjauan ini dinilai berdasarkan standar Database Abstracts of Effects (DARE), yang dikembangkan di York University [1]. Empat komponen utama digunakan untuk menilai kualitas. Mereka adalah kejelasan dan relevansi kriteria inklusi dan eksklusi (QA1), cakupan pencarian literatur untuk semua studi yang relevan (QA2), evaluasi kualitas dan validitas penelitian yang disertakan oleh penulis (QA3), dan kecukupan penjelasan tentang data dasar atau studi yang dievaluasi (QA4).

2.5 Data Collection

Data termasuk sumber, referensi, klasifikasi studi, ruang lingkup, topik utama, penulis, dan ringkasan. Penilaian artikel didasarkan pada pemeriksaan kualitas, pedoman praktis, dan jumlah penelitian primer.

2.6 Data Analysis

Analisis data dalam penelitian ini dilakukan melalui pendekatan sistematis dengan memanfaatkan tabulasi hasil dari literatur yang telah dikumpulkan. Proses dimulai dengan open coding, di mana kutipan atau temuan utama dari setiap sumber dikelompokkan berdasarkan kata kunci dan catatan penting terkait penggunaan GPU, TPU, dan FPGA dalam akselerasi AI. Setelah itu dilanjutkan dengan axial coding, di mana kode-kode tematik dikelompokkan ke dalam beberapa tema utama. Setiap tema dicocokkan dengan artikel yang relevan untuk memberikan gambaran menyeluruh mengenai distribusi topik dalam literatur.

Selanjutnya, dibuat tabel ringkasan perbandingan yang mencakup informasi seperti asal jurnal, tema utama, serta perbandingan performa antara GPU, TPU, dan FPGA berdasarkan parameter seperti efisiensi daya, latensi, throughput, dan skalabilitas. Tabel ini membantu menjawab RQ1, RQ3, dan RQ4 dengan menampilkan sintesis data secara kuantitatif dan kualitatif. Terakhir, dilakukan sintesis naratif yang merangkum kesimpulan dari seluruh analisis sebelumnya, dengan fokus pada jawaban terhadap semua Research Questions (RQ), terutama

RQ1 terkait framework dan tools yang digunakan pada tiap platform, serta tantangan implementasi dan solusi yang diajukan dalam literatur.

3. Hasil dan Pembahasan

Dari proses pencarian awal, ditemukan sebanyak 50 artikel yang dianggap relevan secara judul dan abstrak. Setelah melalui proses penyaringan (screening), sebanyak 11 artikel dihapus karena tidak memenuhi kriteria inklusi, seperti

1. Tidak membahas akselerator GPU, TPU, atau FPGA secara eksplisit
2. Tahunnya tidak dalam range yang ditentukan (2020-2025)

Selanjutnya, dilakukan pemeriksaan duplikasi dan ditemukan 15 artikel duplikat. Artikel duplikat ini dihapus dari daftar. Setelah melalui seluruh tahapan seleksi, diperoleh sebanyak 24 artikel akhir yang digunakan untuk analisis lanjutan dalam studi ini. Dari jumlah tersebut, 20 artikel digunakan sebagai sumber utama dalam sintesis literatur (ditandai sebagai J1–J20), sedangkan 4 artikel pendukung digunakan untuk memperkuat latar belakang dan diskusi tematik. Dan 1 artikel digunakan sebagai metode SLR adopsi [1]. Tabel 1 berikut merangkum identitas umum masing-masing artikel, termasuk tahun terbit, domain aplikasi, framework atau tool yang digunakan, serta jenis akselerator yang dibahas.

Tabel 1 Identitas Literatur

Kode	Penulis	Akselerator Dibahas	Domain/Studi kasus	Framework/Tools
J1	Papoutsis et al. (2023) [7]	GPU	*land cover classification* menggunakan dataset BigEarthNet.	PyTorch, ONNX, dan PyTorch *Lightning* untuk pelatihan dan inferensi.
J2	Magalhães et al. (2023) [8]	GPU, TPU, FPGA	*Agro Robotic*	TensorFlow 2.8, Vitis-AI, TF-TRT, Edge TPU *Compiler*
J3	Yu et al (2023) [9]	GPU	Deteksi fase *seismic*	PyTorch, EQTransformer
J4	Huang et al (2022) [2]	GPU, FPGA	NLP, *Transformer*	FPGA, ASIC, LUT
J5	Bertazzoni (2024) [3]	FPGA	*Edge* AI, *Embedded* CNN	MATLAB HDL, Vivado, ZC706
J6	Alzubaidi (2021) [10]	GPU, FPGA	Umum, *Medical Imaging*, CNN	TensorFlow, MATLAB, PyTorch
J7	Paolo Rech (2024) [11]	GPU, TPU, FPGA	*Aerospace, automotive, safety-critical*	TensorFlow Lite, CUDA
J8	Blott et. al. (2021)	GPU, TPU, FPGA	ImageNet, CIFAR-10, MNIST	QuTiBench, Caffe, TensorRT, FINN, DNNDK
J9	Pilsung and Jongmin (2021) [12]	TPU	*Edge* AI, CNN *Inference*	TensorFlow Lite, CUDA, TensorRT
J10	Alam et.al. (2024) [13]	GPU, TPU, FPGA	*Edge*, IoT, NLP, *Vision*	TensorFlow, ONNX, PyTorch, MetaTF
J11	Aguiar et. al. (2020) [14]	TPU, GPU	Deteksi batang pohon anggur, SLAM	TensorFlow Lite, Edge TPU Compiler
J12	Vaithianathan et. al. (2024) [15]	GPU, FPGA	HPC, AI *training* & *inference*	CUDA, TensorFlow, HDL, OpenCL
J13	Liu et. al (2023) [16]	GPU	*Super-resolution (image upscaling)*	PyTorch, CUDA
J14	Zhang et. al. (2022) [17]	TPU	EfficientNet, BERT, OCR, ResNet	Timeloop, XLA, Google Vizier

Kode	Penulis	Akselerator Dibahas	Domain/Studi kasus	Framework/Tools
J15	Pilsung and Somtham (2022) [18]	GPU, TPU	*Object Detection on Edge*	TensorRT, PyTorch, TFLite
J16	Munanday et. al. (2023) [19]	GPU, TPU	*Facial Expression Recognition*	TensorFlow, Keras, Google Colab
J17	Pacini et.al (2024) [20]	GPU, FPGA	EEG BCI, *Motor Imagery*	TensorFlow, TF-Lite, FPG-AI
J18	Lu and Tan (2024) [21]	GPU, TPU	*Thermal Estimation, Hardware*	TensorFlow, FLIR IR, EdgeTPU
J19	Rao et. al. (2022) [22]	FPGA, GPU, TPU	*Neural Rendering* (NeRF, SLF)	Synopsys, HAPS-80, MLP
J20	Rapuano (2021) [23]	FPGA	EO *Satellites, Hyperspectral Cloud Detection*	Keras, Vivado, VHDL, NCSKD

Berdasarkan jumlah artikel tiap akselerator akan ditampilkan dalam tabel 2, dan data tren akselerator per tahun artikel dapat dilihat pada tabel 3. Database literatur akan ditampilkan pada tabel 4.

Tabel 2 Distribusi Jumlah Akselerator

No	Nama Akselerator	Jumlah Artikel
1	GPU	18/20
2	TPU	10/20
3	FPGA	14/20

Tabel 3 Distribusi Akselerator Per Tahun

No	Tahun	GPU	TPU	FPGA
1	2020	1	1	0
2	2021	2	0	2
3	2022	5	2	3
4	2023	6	4	6
5	2024	4	3	5

Tabel 4 Distribusi Sumber Database

No	Sumber *Database*	Jumlah Artikel
1	IEEE Xplore	5
2	ScienceDirect	4
3	MDPI	6
4	ACM Digital Library	2
5	Lainnya	3

Dari total 20 artikel, sebagian besar membahas GPU sebagai baseline akselerasi AI, namun terdapat peningkatan fokus terhadap efisiensi energi FPGA dan keterjangkauan TPU untuk aplikasi edge.

3.1 Evaluasi Kualitas Literatur

Evaluasi kualitas literatur dilakukan dengan pendekatan DARE, yang mengukur kualitas setiap artikel berdasarkan empat kriteria utama (QA1–QA4). Tabel 5 akan menampilkan kualitas 20 artikel yang dikumpulkan.

Tabel 5 Kualitas Literatur

Kode	Q1	Q2	Q3	Q4	Total
J1	Y	Y	Y	Y	4/4
J2	Y	Y	Y	Y	4/4
J3	Y	P	P	P	2.5/4
J4	P	P	P	N	1.5/4
J5	Y	Y	Y	Y	4/4
J6	Y	P	Y	P	3/4
J7	Y	Y	Y	Y	4/4
J8	P	P	P	P	2/4
J9	Y	Y	Y	Y	3/4
J10	Y	P	P	P	2.5/4
J11	Y	Y	Y	Y	4/4
J12	P	P	P	P	2/4
J13	Y	Y	P	P	3/4
J14	Y	Y	Y	Y	4/4
J15	Y	P	Y	P	3/4
J16	P	Y	P	P	2.5/4
J17	Y	Y	Y	Y	4/4
J18	Y	Y	Y	Y	4/4
J19	Y	Y	Y	P	3.5/4
J20	Y	Y	Y	Y	4/4

Sebagian besar artikel memiliki kualitas metodologis yang baik, terutama dalam menjelaskan tujuan dan metodologi (QA1) serta cakupan studi (QA2). Namun, masih ditemukan beberapa artikel dengan evaluasi eksperimen dan pelaporan data yang kurang rinci, khususnya pada QA3 dan QA4. Ini menjadi pertimbangan saat menyusun sintesis akhir, di mana bobot temuan dari artikel yang memenuhi seluruh kriteria DARE sebaiknya dijadikan rujukan utama.

3.2 Arsitektur, Framework, dan Tools yang Digunakan

Dari hasil analisis terhadap dua puluh artikel (J1–J20), ditemukan bahwa arsitektur Convolutional Neural Network (CNN) merupakan yang paling dominan, digunakan dalam berbagai domain seperti pengenalan objek (J1, J5, J10), klasifikasi citra (J3, J11), dan komunikasi nirkabel (J4, J19). Beberapa studi menggunakan varian CNN seperti UNet dan UNet++ (J1, J5) untuk segmentasi citra satelit dan medis. Arsitektur LSTM digunakan dalam konteks deteksi intrusi dan keamanan siber (J6, J13), terutama saat diterapkan pada TPU. Sementara itu, arsitektur Transformer dan ViT muncul dalam studi-studi yang menyoroti kompleksitas komputasi tinggi pada GPU (J1, J9), dan model ringan seperti RNN, EQT, serta LPPN digunakan pada FPGA dalam konteks sistem tertanam (J7, J8, J17).

Dari sisi framework, TensorFlow digunakan secara luas (J2, J3, J6, J10, J13), termasuk dalam implementasi Edge TPU dengan Edge TPU Compiler (J2, J6). PyTorch hadir di sejumlah studi pelatihan dan benchmarking (J1, J5, J9, J14), menunjukkan fleksibilitasnya pada GPU. FPGA memiliki ekosistem tools yang lebih khusus, seperti Vitis-AI (J5, J8, J18, J20), serta MathWorks HDL Toolbox (J3, J5) untuk eksplorasi hardware-aware design. ONNX juga digunakan untuk interoperabilitas model (J13, J18), dan TF-TRT dipakai untuk optimasi inference pada GPU dan TPU (J2, J10).

Pipeline yang digunakan dalam deployment bervariasi, tergantung akselerator. Pada J5 dan J18, dilakukan quantization aware training sebelum dikompilasi dengan Vitis-AI. J2 menunjukkan integrasi TensorFlow, TF-TRT, Edge TPU Compiler untuk inference pada Jetson TX2 dan Coral Dev. J20 menyajikan deployment model CNN ke FPGA melalui high level synthesis untuk aplikasi satelit.

Secara umum, GPU unggul dari sisi fleksibilitas framework dan eksperimen, TPU menawarkan performa inference cepat dengan konsumsi daya rendah, namun terbatas pada operator tertentu, dan FPGA memberikan efisiensi tinggi tetapi memerlukan proses kompilasi dan optimasi model yang lebih kompleks.

3.3 Performa Training dan Inferensi

Performa akselerator dalam proses pelatihan dan inferensi menjadi fokus utama di banyak studi yang ditinjau. GPU secara umum digunakan sebagai baseline pelatihan karena memiliki dukungan framework luas dan paralelisme tinggi (J1, J3, J5, J9). Beberapa artikel menunjukkan bahwa GPU masih unggul dalam kecepatan training model besar seperti CNN dan ViT, meskipun konsumsi daya cenderung lebih tinggi (J10, J14). Disisi lain, TPU menunjukkan efisiensi tinggi untuk inferensi real-time pada sistem seperti deteksi intrusi dan IoT (J6, J13), dengan latensi hanya beberapa milidetik dan konsumsi daya rendah (J2).

FPGA sering kali menjadi yang tercepat dalam inferensi, seperti ditunjukkan pada J2 dan J18, dengan throughput mencapai 14–25 FPS dan efisiensi daya jauh di atas GPU/TPU. Studi J5 dan J20 juga melaporkan bahwa FPGA unggul dalam aplikasi edge, dengan kombinasi antara latensi rendah dan footprint memori kecil. Sementara itu, pada J4, J8, dan J12, perbandingan antar akselerator menunjukkan bahwa performa sangat bergantung pada optimasi model dan konfigurasi toolchain, bukan hanya jenis perangkat keras.

Dari 20 artikel yang dianalisis, tren menunjukkan bahwa GPU paling unggul untuk pelatihan (training), FPGA unggul dalam inferensi dengan efisiensi energi, dan TPU berada di tengah dengan performa inference cepat namun fleksibilitas terbatas. Masing-masing memiliki keunggulan kontekstual tergantung pada skenario aplikasi, arsitektur model, dan kebutuhan daya komputasi.

3.4 Efisiensi Energi dan Kinerja Akselerator

Efisiensi energi menjadi perhatian utama dalam implementasi akselerator AI, khususnya untuk sistem edge dan real-time. Berdasarkan literatur yang dianalisis, FPGA secara konsisten menunjukkan konsumsi daya yang paling rendah, seperti pada J2, J5, dan J18, dengan efisiensi energi mencapai 5–10W bahkan pada performa inferensi tinggi. Studi J20 menunjukkan bahwa FPGA juga unggul dalam menjaga performa tetap stabil meski dijalankan dalam lingkungan dengan daya terbatas, seperti sistem satelit.

TPU juga menunjukkan efisiensi energi yang tinggi, terutama dalam inference berbasis LSTM untuk sistem deteksi intrusi (J6, J13). Latensi dan konsumsi dayanya sangat rendah, misalnya hanya 6 ms dengan akurasi tinggi di J6. Namun, fleksibilitas model yang didukung TPU lebih terbatas dibanding GPU atau FPGA.

Sementara itu, GPU masih cenderung lebih boros daya, terutama pada versi embedded seperti Jetson TX2 (J2, J10). Meskipun demikian, GPU tetap menjadi pilihan utama dalam skenario pelatihan karena kemampuannya memproses batch besar dengan throughput tinggi (J1, J9, J14). Beberapa artikel (J3, J12, J16) menekankan bahwa efisiensi GPU sangat tergantung pada konfigurasi model, framework, dan optimasi memori.

Secara keseluruhan, dari J1–J20 terlihat bahwa FPGA paling efisien dalam aspek energi dan latensi, TPU efisien untuk inferensi ringan dan spesifik, sementara GPU kuat dalam komputasi berat namun membutuhkan strategi manajemen daya yang lebih hati-hati.

3.5 Tantangan dan Keterbatasan Akselerator

Setiap akselerator memiliki tantangan teknis dan keterbatasan yang khas. GPU, meskipun fleksibel dan didukung banyak framework (J1, J3, J9), menghadapi kendala pada efisiensi daya dan latensi, terutama pada perangkat embedded seperti Jetson TX2 (J2, J10). Selain itu, manajemen thread dan bottleneck memori menjadi isu umum pada GPU, seperti dilaporkan dalam studi penjadwalan pada J5 dan J14.

TPU memiliki keterbatasan dalam hal kompatibilitas model. Beberapa artikel (J2, J6, J13) melaporkan bahwa TPU hanya mendukung subset operator tertentu, sehingga diperlukan penyesuaian arsitektur dan quantization model agar dapat dikompilasi. Hal ini menyulitkan penggunaan TPU untuk model kompleks atau arsitektur custom seperti Transformer (J9).

FPGA, meskipun sangat efisien secara daya dan latensi (J5, J18, J20), memerlukan proses kompilasi yang kompleks dan waktu pengembangan yang lebih lama. Studi J7, J8, dan J17 menunjukkan bahwa penyesuaian desain low-level serta penggunaan tool seperti Vitis-AI atau HDL Toolbox menjadi hambatan adopsi di kalangan noneksperimen. Selain itu, kinerja FPGA sangat sensitif terhadap pengaturan paralelisme dan ukuran batch, sehingga optimasi menjadi proses yang teknis dan tidak selalu generalizable.

Dari keseluruhan J1–J20, tantangan terbesar mencakup kompatibilitas model (TPU), efisiensi daya dan manajemen memori (GPU), serta kompleksitas desain dan kompilasi

(FPGA), yang semuanya perlu dipertimbangkan saat memilih akselerator untuk sistem AI tertentu.

3.6 Perbandingan Lintas Domain Aplikasi

Studi-studi yang dianalisis (J1–J20) mencakup beragam domain aplikasi, mulai dari penginderaan jauh (J1, J20), pertanian cerdas (J2), hingga komunikasi nirkabel 6G (J4, J19), sistem deteksi intrusi IoT (J6, J13), dan sistem tertanam otonom (J5, J7, J18). Di domain edge computing seperti pertanian dan sistem otonom, FPGA sering dipilih karena efisiensi daya dan latensi rendah. Untuk aplikasi smart city dan keamanan siber, TPU digunakan karena inference yang cepat dan footprint kecil. Sementara itu, GPU dominan dalam domain yang memerlukan pelatihan intensif dan eksperimen arsitektur, seperti pada klasifikasi citra medis atau training model skala besar (J3, J9, J14). Temuan ini menunjukkan bahwa pemilihan akselerator tidak hanya bergantung pada performa teknis, tetapi juga sangat dipengaruhi oleh konteks aplikasi dan batasan lingkungan operasional.

3.7 Implikasi Pemilihan Akselerator untuk Penelitian Mahasiswa Magister Ilmu Komputer

Berdasarkan temuan dari 20 artikel yang ditinjau, pemilihan akselerator dalam penelitian tesis mahasiswa magister ilmu komputer sebaiknya mempertimbangkan beberapa aspek utama: kompleksitas model, ketersediaan perangkat keras, domain aplikasi, serta kebutuhan akan efisiensi daya atau latensi.

Jika fokus penelitian berada pada eksperimen arsitektur model atau pelatihan skala besar, seperti CNN, Transformer, atau ViT, maka GPU menjadi pilihan utama karena fleksibilitas framework (seperti PyTorch dan TensorFlow) dan dukungan library yang luas (J1, J9, J14). Untuk penelitian berbasis sistem tertanam atau edge computing, khususnya yang berkaitan dengan IoT, smart city, atau kendaraan otonom, TPU dan FPGA lebih direkomendasikan. TPU cocok untuk deployment real-time dan hemat daya (J6, J13), sedangkan FPGA unggul untuk aplikasi khusus dengan kendali penuh atas pipeline komputasi dan efisiensi tinggi (J5, J18, J20).

Namun, perlu dicatat bahwa penggunaan FPGA memerlukan pemahaman tambahan tentang toolchain seperti Vitis-AI atau HDL Toolbox, serta proses kompilasi low-level, sehingga mungkin lebih cocok untuk mahasiswa dengan latar belakang atau bimbingan di bidang sistem tertanam atau rekayasa perangkat keras. Sebaliknya, GPU dan TPU lebih ramah untuk eksperimen di tingkat aplikasi dan perangkat lunak, dan umumnya didukung oleh cloud platform seperti Google Colab atau Edge TPU Dev Board.

Dengan mempertimbangkan hasil literatur ini, mahasiswa dapat merancang strategi penelitian yang lebih tepat sasaran sesuai kebutuhan sistem, ketersediaan alat, dan arah topik tesis. Tabel 6 akan memberikan saran akselerator berdasarkan fokus riset mahasiswa magister ilmu komputer.

Tabel 6 Saran Akselerator Berdasarkan Fokus Riset

Kode	Q1	Q2	Q3	Q4	Total
J1	Y	Y	Y	Y	4/4
J2	Y	Y	Y	Y	4/4
J3	Y	P	P	P	2.5/4
J4	P	P	P	N	1.5/4
J5	Y	Y	Y	Y	4/4
J6	Y	P	Y	P	3/4
J7	Y	Y	Y	Y	4/4
J8	P	P	P	P	2/4
J9	Y	Y	Y	Y	3/4
J10	Y	P	P	P	2.5/4
J11	Y	Y	Y	Y	4/4
J12	P	P	P	P	2/4
J13	Y	Y	P	P	3/4
J14	Y	Y	Y	Y	4/4
J15	Y	P	Y	P	3/4

J16	P	Y	P	P	2.5/4
J17	Y	Y	Y	Y	4/4
J18	Y	Y	Y	Y	4/4
J19	Y	Y	Y	P	3.5/4
J20	Y	Y	Y	Y	4/4

4. Kesimpulan

Penelitian ini melakukan kajian sistematis terhadap 20 artikel terkini guna mengevaluasi pemanfaatan akselerator perangkat keras GPU, TPU, dan FPGA dalam mendukung komputasi AI. Hasil menunjukkan bahwa GPU unggul dalam proses pelatihan karena fleksibilitas framework dan dukungan komunitas luas. TPU menonjol dalam inference real-time dengan konsumsi daya rendah, khususnya pada aplikasi berbasis IoT dan keamanan siber. Sementara itu, FPGA menjadi akselerator yang paling efisien dari segi energi dan latensi, sangat cocok untuk deployment di lingkungan edge dan sistem tertanam, meskipun memerlukan konfigurasi teknis yang lebih kompleks.

Temuan ini memperlihatkan bahwa pemilihan akselerator sangat dipengaruhi oleh konteks aplikasi, kompleksitas model, dan ketersediaan perangkat serta toolchain pendukung. GPU banyak digunakan dalam domain seperti klasifikasi citra dan benchmark model besar, TPU dalam deteksi intrusi dan sistem cerdas real time, sementara FPGA banyak digunakan dalam komunikasi 6G, kendaraan otonom, dan sistem energi terbatas. Framework seperti TensorFlow, PyTorch, Vitis-AI, dan Edge TPU Compiler menjadi bagian penting dalam mendukung kompatibilitas dan efisiensi proses pelatihan maupun inferensi.

Bagi mahasiswa magister ilmu komputer, hasil ini dapat dijadikan acuan dalam memilih akselerator sesuai fokus tesis. GPU disarankan bagi penelitian eksperimental di level model, TPU untuk riset berbasis inference ringan dan efisien, dan FPGA untuk topik sistem embedded yang menuntut optimasi hardware. Namun demikian, aspek ketersediaan perangkat, kesiapan alat bantu teknis, dan arah bimbingan juga perlu dipertimbangkan. Studi ini mendorong penelitian lanjutan dalam eksplorasi akselerator hybrid serta pengembangan strategi co-design antara algoritma dan perangkat keras guna menghasilkan sistem AI yang lebih adaptif dan optimal di berbagai domain.

Daftar Referensi

[1] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009, doi: 10.1016/j.infsof.2008.09.009.

[2] S. Huang, E. Tang, S. Li, X. Ping, and R. Chen, “Hardware-friendly compression and hardware acceleration for transformer : A survey,” vol. 30, no. July, pp. 3755–3785, 2022, doi: 10.3934/era.2022192.

[3] S. Bertazzoni *et al.*, “Design Space Exploration for Edge Machine Learning Featured by MathWorks FPGA DL Processor : A Survey,” vol. 12, no. November 2023, 2024.

[4] A. Martín-martín *et al.*, “Hardware Implementations of a Deep Learning Approach to Optimal Configuration of Reconfigurable Intelligence Surfaces,” pp. 1–21, 2024.

[5] R. Kaur, A. Asad, S. Al, A. Wahid, and F. Mohammadi, “A Survey of Advancements in Scheduling Techniques for Efficient Deep Learning Computations on GPUs,” pp. 1–30, 2025.

[6] C. Hazman, A. Guezzaz, S. Benkirane, and M. Azrour, “Enhanced IDS with Deep Learning for IoT-Based Smart Cities Security,” vol. 29, no. 4, pp. 929–947, 2024.

[7] I. Papoutsis, N. Ioannis, A. Zavras, D. Michail, and C. Tryfonopoulos, “ISPRS Journal of Photogrammetry and Remote Sensing Benchmarking and scaling of deep learning models for land cover image classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 195, no. July 2022, pp. 250–268, 2023, doi: 10.1016/j.isprsjprs.2022.11.012.

[8] S. Costa, F. Neves, P. Machado, P. Moreira, and J. Dias, “Engineering Applications of Artificial Intelligence Benchmarking edge computing devices for grape bunches and trunks detection using accelerated object detection single shot multibox deep learning models,” *Eng. Appl. Artif. Intell.*, vol. 117, no. October 2022, p. 105604, 2023, doi: 10.1016/j.engappai.2022.105604.

[9] Z. Yu, W. Wang, and Y. Chen, “Benchmark on the accuracy and efficiency of several neural network based phase pickers using datasets from China Seismic Network,” *Earthq. Sci.*, vol. 36, no. 2, pp. 113–131, 2023, doi: 10.1016/j.eqs.2022.10.001.
[10] L. Alzubaidi *et al.*, *Review of deep learning : concepts , CNN architectures , challenges , applications , future directions*. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.
[11] P. Rech, “Artificial Neural Networks for Space and Safety-Critical Applications : Reliability Issues and Potential Solutions,” *IEEE Trans. Nucl. Sci.*, vol. 71, no. 1, pp. 377–404, 2024, doi: 10.1109/TNS.2024.3349956.
[12] P. Kang and J. Jo, “Benchmarking Modern Edge Devices for AI Applications ∗ SUMMARY,” no. 3, pp. 394–403, 2021.
[13] S. Alam, C. Yakopcic, Q. Wu, M. Barnell, S. Khan, and T. M. Taha, “Survey of Deep Learning Accelerators for Edge and Emerging Computing,” *Electronics*, vol. 13, 2024.
[14] A. S. Aguiar *et al.*, “Visual Trunk Detection Using Transfer Learning and a Deep Learning-Based Coprocessor,” vol. 8, 2020, doi: 10.1109/ACCESS.2020.2989052.
[15] M. Vaithianathan, M. Patil, F. S. Ng, and S. Udkar, “Comparative Study of FPGA and GPU for High-Performance Computing and AI,” *ESP Int. J. Adv. Comput. Technol.*, vol. 1, no. May 2023, pp. 37–46, 2024, doi: 10.56472/25838628/IJACT-V1I1P107.
[16] Y. Liu, M. Yue, H. Yan, and L. Zhu, “Single-image super-resolution using lightweight transformer-convolutional neural network hybrid model,” *Inst. Enginering Technol.*, no. May, pp. 2881–2893, 2023, doi: 10.1049/ipr2.12833.
[17] D. Zhang *et al.*, “A Full-Stack Search Technique for Domain Optimized Deep Learning Accelerators,” *Assoc. Comput. Mach.*, pp. 27–42, 2022, doi: 10.1145/3503222.3507767.
[18] P. Kang and A. Somtham, “An Evaluation of Modern Accelerator-Based Edge Devices for Object Detection Applications,” *Mathematics*, vol. 10, pp. 1–14, 2022.
[19] A. P. Munanday, N. Sazali, W. Sharuzi, and W. Harun, “Analysis of Convolutional Neural Networks for Facial Expression Recognition on GPU , TPU and CPU,” vol. 3, no. 3, pp. 50–67, 2023.
[20] F. Pacini, T. Pacini, G. Lai, A. M. Zocco, and L. Fanucci, “Deployment of a CNN for Motor Imagery Classification in Brain-Computer Interfaces,” 2024.
[21] J. Lu and S. X. Tan, “Thermal Map Dataset for Commercial Multi / Many Core CPU / GPU / TPU”, doi: 10.1145/3670474.3685963.
[22] C. Rao *et al.*, “ICARUS : A Specialized Architecture for Neural Radiance Fields Rendering,” vol. 41, no. 6, 2022, doi: 10.1145/3550454.3555505.
[23] E. Rapuano *et al.*, “An FPGA-Based Hardware Accelerator for CNNs Inference on Board Satellites : Benchmarking with Myriad 2-Based Solution for the CloudScout Case Study,” 2021.