**Progresif:** Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
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

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

DOI: http://dx.doi.org/10.35889/progresif.v22i2.3613



**Siska Fitriani[1*], Ega Budiman[2], Muhammad Fadli[3], Amarudin[4]**
[1,2,3,4]Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
[3]Jurusan Ekonomi dan Bisnis, Politeknik Negeri Lampung, Bandar Lampung, Indonesia
**e-mail Corresponding Author:* 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 70%. 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:*** *Artificial Intelligence Accelerator; Graphics Processing Unit; Tensor Processing Unit; Field Programmable Gate Array; Systematic Literature 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, Penelitian ini bertujuan mengevaluasi pemanfaatan ketiga akselerator dalam berbagai domain AI menggunakan *Systematic Literature Review* (SLR) berbasis pendekatan Kitchenham. Sebanyak 20 artikel diseleksi dari lima basis data ilmiah terkemuka. Hasil menunjukkan GPU digunakan dalam 90% studi, FPGA dalam 70%, dan TPU dalam 50%. FPGA unggul dalam efisiensi energi (78% studi), sementara GPU dominan dalam performa inferensi (85% kasus). Penelitian menyimpulkan pemilihan akselerator AI harus mempertimbangkan efisiensi daya, arsitektur sistem, dan kebutuhan domain. Temuan ini memberikan panduan praktis bagi mahasiswa magister dalam memilih akselerator sesuai topik, tujuan eksperimen, dan keterbatasan sumber daya.

**Kata Kunci:** *Akselerator Artificial Intelligence; Graphics Processing Unit; Tensor Processing Unit; Field Programmable Gate Array; 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. Akselerator ini digunakan secara luas dalam berbagai domain AI, mulai dari pengenalan citra dan pemrosesan bahasa alami, hingga sistem *edge computing* dan IoT. Pendekatan berbasis bukti

*(evidence-based approach)* kini diadopsi dalam ilmu komputer untuk membantu pengambilan keputusan arsitektural, sebagaimana sebelumnya diterapkan dalam bidang medis dan sosial [1]. Pemilihan akselerator yang tepat tidak hanya berdampak pada performa inferensi dan pelatihan model, tetapi juga pada efisiensi energi, skalabilitas sistem, dan kelayakan implementasi di berbagai lingkungan operasional.

GPU, TPU, dan FPGA masing-masing memiliki keunggulan dan keterbatasan dalam menangani beban kerja AI. GPU unggul dalam fleksibilitas framework dan pelatihan skala besar, TPU menawarkan inference yang cepat dan hemat daya, sedangkan FPGA memberikan efisiensi tinggi untuk sistem embedded dan edge [2] [3]. Meskipun sejumlah studi telah membandingkan akselerator secara parsial, belum terdapat tinjauan sistematis yang secara komprehensif mengevaluasi ketiga akselerator tersebut sekaligus mencakup aspek arsitektur, framework, performa, efisiensi energi, dan tantangan implementasi di berbagai domain aplikasi AI yang beragam. Kesenjangan ini menghambat pengambilan keputusan berbasis bukti, khususnya bagi peneliti dan mahasiswa yang memerlukan panduan dalam memilih akselerator yang paling sesuai dengan kebutuhan spesifik riset mereka.

Penelitian ini bertujuan melakukan *Systematic Literature Review (SLR)* menggunakan kerangka kerja *Kitchenham* [1] terhadap 20 artikel ilmiah terpilih yang membahas penggunaan GPU, TPU, dan FPGA dalam akselerasi AI. Secara khusus, studi ini bertujuan: (1) mengidentifikasi arsitektur, framework, dan tools yang digunakan pada masing-masing akselerator; (2) membandingkan performa training dan inferensi; (3) mengevaluasi efisiensi energi dan kinerja komputasi; serta (4) menganalisis tantangan dan keterbatasan teknis dalam deployment akselerator AI di berbagai domain. Dengan demikian, studi ini tidak hanya membandingkan performa dari sisi kecepatan dan akurasi, tetapi juga mempertimbangkan efisiensi energi, fleksibilitas arsitektur, dan kesesuaian terhadap domain aplikasi tertentu seperti keamanan siber, sistem otonom, dan jaringan komunikasi cerdas.

Studi ini didasarkan pada empat pertanyaan penelitian *(Research Questions/RQ)* yang saling melengkapi. Pertama, studi ini menyelidiki arsitektur, *framework*, dan *tools* yang digunakan pada GPU, TPU, dan FPGA dalam akselerasi AI (RQ1). Kedua, dikaji bagaimana performa masing-masing akselerator dalam proses training dan inferensi model AI (RQ2). Ketiga, studi ini mengevaluasi efisiensi energi dan kinerja komputasi dari ketiga akselerator tersebut secara komparatif (RQ3). Keempat, dianalisis tantangan dan keterbatasan teknis yang dihadapi dalam deployment akselerator AI di berbagai domain aplikasi (RQ4). Keempat pertanyaan ini secara kolektif membentuk kerangka analisis yang sistematis dan berbasis bukti untuk mengevaluasi kesesuaian masing-masing akselerator terhadap kebutuhan spesifik implementasi AI.

Studi ini memberikan kontribusi berupa sintesis berbasis bukti mengenai perbandingan GPU, TPU, dan FPGA yang dapat dijadikan referensi ilmiah bagi peneliti, praktisi, dan mahasiswa dalam memilih akselerator yang paling sesuai dengan kebutuhan riset dan implementasi sistem AI mereka. Secara khusus, mahasiswa magister ilmu komputer akan memperoleh panduan praktis dalam menentukan akselerator berdasarkan fokus topik tesis, tujuan eksperimen, dan keterbatasan sumber daya yang tersedia. Selain itu, temuan SLR ini diharapkan 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 aplikasi.

## 2. Tinjauan Pustaka

Kajian mengenai akselerasi AI berbasis perangkat keras telah berkembang pesat seiring meningkatnya kebutuhan komputasi untuk model *deep learning*. Sebagai fondasi metodologis, Kitchenham et al. [1] memperkenalkan kerangka kerja *Systematic Literature Review* (SLR) dalam rekayasa perangkat lunak melalui tinjauan sistematis terhadap berbagai studi primer yang telah dipublikasikan. Penelitian tersebut menetapkan prosedur baku SLR yang mencakup perencanaan tinjauan, pelaksanaan pencarian literatur secara sistematis, serta sintesis temuan menggunakan kriteria inklusi dan eksklusi yang terstruktur. Kerangka kerja ini kemudian diadopsi secara luas di bidang ilmu komputer, termasuk dalam kajian akselerator AI, karena kemampuannya menghasilkan sintesis yang komprehensif, transparan, dan dapat direproduksi.

Alzubaidi et al. [5] melakukan tinjauan komprehensif mengenai konsep, arsitektur, tantangan, aplikasi, dan arah pengembangan *deep learning*, dengan cakupan pembahasan yang mencakup implementasi berbasis GPU dan FPGA dalam domain *medical imaging* dan pengenalan citra. Tinjauan tersebut dilakukan menggunakan pendekatan *narrative review* yang menghimpun ratusan artikel dari berbagai basis data ilmiah seperti IEEE, ACM, dan Springer. Meskipun kajian ini

memberikan gambaran menyeluruh mengenai arsitektur CNN dan tantangan komputasinya, pendekatan yang digunakan bersifat deskriptif tanpa protokol seleksi yang sistematis dan tidak membandingkan ketiga jenis akselerator (GPU, TPU, FPGA) secara bersamaan dalam konteks efisiensi energi maupun performa inferensi.

Blott et al. [6] meneliti perbandingan performa akselerator GPU, TPU, dan FPGA dalam tugas inferensi model jaringan saraf tiruan pada *benchmark* standar seperti ImageNet, CIFAR-10, dan MNIST. Prosedur peninjauan dilakukan melalui pengujian empiris menggunakan platform *QuTiBench*, yang memungkinkan evaluasi kuantitatif terhadap berbagai akselerator secara konsisten. Studi ini berkontribusi penting dalam menyediakan data performa komparatif; namun, cakupannya terbatas pada tolok ukur klasifikasi citra dan belum menjangkau domain aplikasi yang lebih beragam seperti sistem otonom, keamanan siber, atau penginderaan jauh. Selain itu, metode yang digunakan bersifat *benchmarking* eksperimental, bukan tinjauan sistematis berbasis literatur.
Alam et al. [7] melakukan survei sistematis mengenai penggunaan akselerator GPU, TPU, dan FPGA dalam ekosistem *Edge* AI, IoT, NLP, dan *Computer Vision*. Metode peninjauan yang digunakan adalah survei literatur dengan pendekatan tematik menggunakan *framework* TensorFlow, ONNX, PyTorch, dan MetaTF sebagai acuan ekosistem perangkat lunak. Studi ini memberikan gambaran yang cukup luas mengenai tren adopsi akselerator di lingkungan komputasi terdistribusi; namun, prosedur seleksi artikel tidak mengacu pada protokol SLR yang baku sehingga transparansi dan reproduktivitas hasil tinjauannya masih terbatas. Evaluasi kualitas literatur juga tidak dilakukan secara eksplisit.

Rech [8] meninjau keandalan jaringan saraf tiruan yang diimplementasikan pada GPU, TPU, dan FPGA dalam aplikasi *safety critical* di bidang antariksa dan otomotif. Tinjauan dilakukan dengan mengkompilasi hasil eksperimen injeksi kesalahan (*fault injection*) dari berbagai studi yang telah dipublikasikan, menggunakan pendekatan *integrative review* yang berfokus pada aspek keandalan dan toleransi kesalahan perangkat keras. Meski kajian ini memperkaya pemahaman tentang keterbatasan akselerator dalam lingkungan kritis, cakupannya tidak membahas efisiensi energi, performa *training*, maupun aspek *framework* dan *toolchain* yang lazim digunakan dalam pengembangan model AI secara umum.

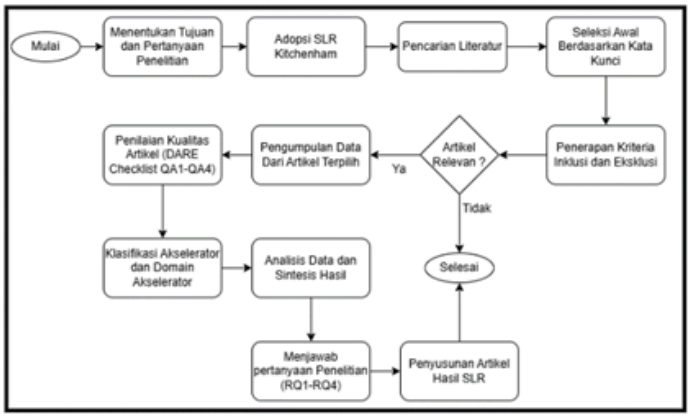
Vaithianathan et al. [9] melakukan studi komparatif antara GPU dan FPGA dalam konteks *High-Performance Computing* (HPC) untuk tugas *training* dan inferensi AI. Metode yang digunakan adalah perbandingan eksperimental berbasis implementasi model dengan *tools* CUDA, TensorFlow, HDL, dan OpenCL. Studi ini mengungkap bahwa GPU unggul dalam fleksibilitas dan kecepatan pelatihan, sementara FPGA lebih efisien dalam hal konsumsi daya. Namun, kajian tersebut tidak menyertakan TPU sebagai objek perbandingan, tidak menggunakan protokol SLR yang terstandar, dan tidak membahas konteks domain aplikasi yang beragam sehingga generalisasi temuannya masih terbatas.

Berdasarkan tinjauan terhadap penelitian-penelitian terdahulu, teridentifikasi bahwa sebagian besar kajian yang ada hanya membandingkan satu atau dua jenis akselerator, menggunakan metode *narrative review*, *benchmarking* eksperimental, atau survei tematik tanpa protokol SLR yang baku, serta memiliki cakupan domain aplikasi yang terbatas. Penelitian yang dilakukan oleh Fitriani et al. saat ini menghadirkan kebaruan (*novelty*) dalam beberapa aspek. Pertama, dari sisi metode, studi ini secara eksplisit mengadopsi protokol SLR Kitchenham yang terstruktur mencakup perencanaan, pencarian sistematis, penilaian kualitas berbasis DARE, dan sintesis naratif yang belum diterapkan secara konsisten dalam kajian akselerator AI sebelumnya. Kedua, dari sisi cakupan, studi ini mengevaluasi ketiga akselerator (GPU, TPU, dan FPGA) secara simultan dalam satu kerangka analisis yang terpadu, tidak secara parsial. Ketiga, dari sisi domain, cakupan tinjauan mencakup delapan domain aplikasi yang beragam, mulai dari penginderaan jauh, sistem otonom, keamanan siber, hingga komunikasi 6G, sehingga menghasilkan peta performa akselerator yang lebih komprehensif dan aplikatif. Keempat, studi ini secara khusus menghasilkan rekomendasi akselerator yang dipersonalisasi untuk kebutuhan riset mahasiswa magister ilmu komputer, sebuah kontribusi yang belum ditemukan dalam kajian-kajian sebelumnya.

## 3. Metodologi

Pendekatan yang digunakan dalam penelitian ini mengacu pada kerangka kerja *Systematic Literature Review (SLR)* yang diusulkan oleh Kitchenham et al. [1]. Meskipun dipublikasikan lebih dari 10 tahun yang lalu, framework Kitchenham [1] masih menjadi acuan utama dalam penelitian SLR di bidang rekayasa perangkat lunak dan ilmu komputer, sebagaimana tercermin dari tingginya frekuensi sitasi hingga saat ini. 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 bersifat komprehensif, transparan, dan dapat direproduksi. Gambar 1 berikut menunjukkan alur penelitian berdasarkan pendekatan SLR yang digunakan dalam studi ini




**Gambar 1**. Alur Penelitian

### 3.1 Strategi Pencarian Literatur

Pencarian literatur dalam studi ini dilakukan secara sistematis mengacu pada protokol SLR yang dikembangkan oleh Kitchenham et al. (2020) [1]. Basis data ilmiah yang digunakan meliputi IEEE *Xplore*, ACM Digital Library, *ScienceDirect* (Elsevier), MDPI, dan Springer Link. Pemilihan kelima basis data tersebut didasarkan pada reputasinya sebagai sumber publikasi *peer-reviewed* terkemuka di bidang ilmu komputer, rekayasa perangkat keras, dan kecerdasan buatan. Rentang waktu publikasi dibatasi dari tahun 2020 hingga 2025 guna memastikan relevansi dan keterbaruan temuan, mengingat pesatnya perkembangan teknologi akselerator AI dalam lima tahun terakhir.
Teknik pencarian dilakukan menggunakan kombinasi kata kunci dengan operator Boolean (AND, OR, NOT) secara terstruktur. Kelompok kata kunci yang digunakan mencakup: (*"GPU"* OR *"Graphics Processing Unit"*) AND (*"TPU"* OR *"Tensor Processing Unit"*) AND (*"FPGA"* OR *"Field Programmable Gate Array"*) AND (*"AI Accelerator"* OR *"Hardware Acceleration"* OR *"Deep Learning"*). Selain itu, kata kunci tambahan seperti *"energy efficiency"*, *"inference performance"*, *"edge computing"*, dan *"neural network deployment"* digunakan untuk memperluas cakupan hasil pencarian. Pencarian dilakukan pada judul, abstrak, dan kata kunci artikel.

Prosedur seleksi literatur dilakukan dalam empat tahap. Tahap pertama adalah identifikasi awal, di mana hasil pencarian dari seluruh basis data dikompilasi dan diperiksa duplikasinya, menghasilkan 50 artikel kandidat. Tahap kedua adalah *screening* berdasarkan judul dan abstrak dengan menerapkan kriteria inklusi: (1) artikel membahas GPU, TPU, dan/atau FPGA secara eksplisit dalam konteks akselerasi AI; (2) diterbitkan antara tahun 2020–2025; (3) tersedia dalam bahasa Inggris atau Indonesia; dan (4) telah melalui proses *peer-review*. Kriteria eksklusi meliputi artikel yang hanya membahas konsep AI tanpa keterkaitan akselerator perangkat keras, serta sumber yang tidak memiliki kedalaman analisis komparatif. Tahap ketiga adalah *full-text review* untuk memverifikasi kesesuaian isi artikel dengan fokus kajian. Tahap keempat adalah penilaian kualitas menggunakan standar *Database Abstracts of Effects* (DARE) dengan empat kriteria: kejelasan dan relevansi kriteria inklusi/eksklusi (QA1), kelengkapan pencarian literatur (QA2), evaluasi validitas penelitian (QA3), dan kecukupan penjelasan data (QA4). Dari 50 artikel awal, setelah penghapusan 14 duplikat dan 11 artikel yang tidak memenuhi kriteria inklusi, diperoleh 25 artikel final, di antaranya 20 artikel digunakan sebagai sumber utama sintesis ([10 - 13],[2],[5],[8],[6],[14],[7],[15],[9],[16 - 23]).

### 3.2 Teknik Analisis Data

Analisis data dalam penelitian ini dilakukan melalui kombinasi *thematic analysis* dan *content analysis* yang diterapkan secara bertahap terhadap 20 artikel yang telah lolos seleksi. Tahap pertama adalah ekstraksi data terstruktur, di mana informasi kunci dari setiap artikel dikumpulkan ke dalam formulir ekstraksi yang memuat: identitas artikel (penulis, tahun, jurnal/prosiding, basis data sumber), jenis akselerator yang dikaji (GPU, TPU, FPGA, atau kombinasi), domain aplikasi, *framework* dan *tools* yang digunakan, metrik performa yang dilaporkan (latensi, *throughput*, konsumsi daya, akurasi), serta tantangan dan solusi yang diidentifikasi. Proses ekstraksi dilakukan secara independen untuk menjaga konsistensi dan objektivitas data.

Tahap kedua adalah kategorisasi dan pengkodean tematik. Proses dimulai dengan *open coding*, di mana temuan utama dari setiap artikel diberi kode berdasarkan tema yang muncul, seperti "efisiensi energi FPGA", "performa *training* GPU", atau "keterbatasan operator TPU". Selanjutnya dilakukan *axial coding* untuk mengelompokkan kode-kode tersebut ke dalam tema-tema utama yang merespons masing-masing RQ. *Content analysis* diterapkan untuk menganalisis frekuensi dan pola kemunculan tema lintas artikel, sehingga menghasilkan gambaran distribusi topik secara kuantitatif, misalnya proporsi artikel yang membahas GPU (90%), TPU (50%), dan FPGA (70%). Hasil pengkodean kemudian ditabulasikan dalam tabel perbandingan yang mencakup parameter efisiensi daya, latensi, *throughput*, dan skalabilitas untuk memfasilitasi analisis lintas studi.
Tahap ketiga adalah sintesis naratif, di mana temuan-temuan yang telah dikategorikan diintegrasikan untuk menjawab seluruh pertanyaan penelitian (RQ1–RQ4) secara terpadu. Sintesis dilakukan dengan mendeskripsikan pola-pola konsisten yang ditemukan lintas studi, mengidentifikasi perbedaan dan kontradiksi antar temuan, serta mengaitkan hasil analisis dengan konteks domain aplikasi yang relevan. Pendekatan ini memungkinkan penarikan kesimpulan yang didukung oleh bukti dari keseluruhan korpus literatur, sekaligus menghasilkan rekomendasi praktis mengenai pemilihan akselerator yang disesuaikan dengan karakteristik spesifik riset di bidang kecerdasan buatan.

## 4. 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. Tahun publikasi berada di luar rentang yang ditentukan (2020–2025)

Selanjutnya, dilakukan pemeriksaan duplikasi dan ditemukan 14 artikel duplikat yang kemudian dihapus dari daftar. Setelah melalui seluruh tahapan seleksi, diperoleh sebanyak 25 artikel final yang digunakan untuk analisis lanjutan dalam studi ini. Dari jumlah tersebut, 20 artikel digunakan sebagai sumber utama dalam sintesis literatur (ditandai sebagai J1–J20), 4 artikel pendukung digunakan untuk memperkuat latar belakang dan diskusi tematik, serta 1 artikel diadopsi sebagai acuan metode SLR [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) [10] | GPU | *land cover classification* menggunakan dataset BigEarthNet. | PyTorch, ONNX, dan PyTorch *Lightning* untuk pelatihan dan inferensi. |
| J2 | S. Costa, et al. (2023) [11] | GPU, TPU, FPGA | *Agro Robotic* | TensorFlow 2.8, Vitis-AI, TF-TRT, Edge TPU *Compiler* |
| J3 | Yu et al (2023) [12] | GPU | Deteksi fase *seismic* | PyTorch, EQTransformer |
| J4 | Huang et al (2022) [13] | GPU, FPGA | NLP, *Transformer* | FPGA, ASIC, LUT |
| J5 | Bertazzoni (2024) [2] | FPGA | *Edge* AI, *Embedded* CNN | MATLAB HDL, Vivado, ZC706 |
| J6 | Alzubaidi (2021) [5] | GPU, FPGA | Umum, *Medical Imaging*, CNN | TensorFlow, MATLAB, PyTorch |
| J7 | Paolo Rech (2024) [8] | GPU, TPU, FPGA | *Aerospace, automotive, safety-critical* | TensorFlow Lite, CUDA |
| J8 | Blott et. al. (2021)[6] | GPU, TPU, FPGA | ImageNet, CIFAR-10, MNIST | QuTiBench, Caffe, TensorRT, FINN, DNNDK |
| J9 | Pilsung and Jongmin (2021) [14] | TPU | *Edge* AI, CNN *Inference* | TensorFlow Lite, CUDA, TensorRT |
| J10 | Alam et.al. (2024) | GPU, TPU, | *Edge*, IoT, NLP, *Vision* | TensorFlow, ONNX, |

| Kode | Penulis | Akselerator Dibahas | Domain/Studi kasus | Framework/Tools |
|---|---|---|---|---|
| | [7] | FPGA | | PyTorch, MetaTF |
| J11 | Aguiar et. al. (2020) [15] | TPU, GPU | Deteksi batang pohon anggur, SLAM | TensorFlow Lite, Edge TPU Compiler |
| J12 | Vaithianathan et. al. (2024) [24] | 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 |
| 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*.

### 4.1 Karakteristik Studi yang Diseleksi

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, yaitu tidak membahas akselerator GPU, TPU, atau FPGA secara eksplisit, serta tahun publikasi yang berada di luar rentang yang ditentukan (2020–2025). Selanjutnya, dilakukan pemeriksaan duplikasi dan ditemukan 14 artikel duplikat yang kemudian dihapus dari daftar. Setelah melalui seluruh tahapan seleksi, diperoleh 25 artikel final, di mana 20 artikel digunakan sebagai sumber utama sintesis literatur ([2 - 6] [9], - [23]), 5 artikel sebagai pendukung latar belakang dan diskusi tematik, dan 1 artikel sebagai acuan metode SLR.

Dari 20 artikel yang digunakan sebagai sumber utama sintesis, sebagian besar berasal dari lembaga riset dan perguruan tinggi di berbagai negara, mencerminkan sebaran geografis yang luas. Penulis-penulis berafiliasi dengan institusi dari Amerika Serikat ([6], [16], [17], [22]), Italia ([17], [23]), Inggris ([5]), Korea Selatan ([14], [18]), Brasil ([15]), Malaysia ([19], [20]), dan Portugal ([11]), antara lain. Keragaman negara asal ini mengindikasikan bahwa kajian akselerator AI merupakan topik yang berkembang secara global dan tidak terpusat pada satu kawasan tertentu. Dari sisi jenis publikasi, 11 artikel bersumber dari jurnal ilmiah internasional bereputasi (IEEE *Transactions*, MDPI *Electronics*, *Remote Sensing*, dan lainnya), 5 artikel berasal dari prosiding konferensi internasional (ACM, IEICE), dan 4 artikel lainnya diterbitkan dalam jurnal lintas disiplin seperti *Engineering Applications of Artificial Intelligence* dan *Journal of Big Data*. Distribusi jenis publikasi ini menunjukkan bahwa topik akselerator AI banyak dikaji baik dalam forum akademis formal maupun ajang konferensi riset terapan yang kompetitif.

Evaluasi kualitas literatur dilakukan dengan pendekatan DARE, yang mengukur kualitas setiap artikel berdasarkan empat kriteria utama (QA1–QA4). 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 dijadikan rujukan utama.

### 4.2 Analisis Hasil

Analisis hasil dilakukan mengacu pada empat pertanyaan penelitian (RQ1–RQ4) yang telah ditetapkan. Setiap RQ dijawab berdasarkan sintesis temuan dari 20 artikel yang telah diseleksi, sebagaimana diuraikan pada sub-sub bagian berikut.

1) RQ1 — Arsitektur, *Framework*, dan *Tools* yang Digunakan

Dari hasil analisis terhadap dua puluh artikel ([9-12],[2],[4-7],[6],[14],[7],[15],[9],[15-19],[20-22]), ditemukan bahwa arsitektur *Convolutional Neural Network* (CNN) merupakan yang paling dominan, digunakan dalam berbagai domain seperti pengenalan objek ([10], [2], [7],), klasifikasi citra ([12], [15],), dan komunikasi nirkabel ([13], [22]). Beberapa studi menggunakan varian CNN seperti UNet dan UNet++ ([10], [2]) untuk segmentasi citra satelit dan medis. Arsitektur LSTM digunakan dalam konteks deteksi intrusi dan keamanan siber ([5], [16]), terutama saat diterapkan pada TPU. Sementara itu, arsitektur *Transformer* dan ViT muncul dalam studi-studi yang menyoroti kompleksitas komputasi tinggi pada GPU ([10], [14]), dan model ringan seperti RNN, EQT, serta LPPN digunakan pada FPGA dalam konteks sistem tertanam ([8], [6], [20]).

Dari sisi *framework*, *TensorFlow* digunakan secara luas ([11], [12], [5], [7], [19]), termasuk dalam implementasi *Edge* TPU dengan *Edge* TPU *Compiler* ([11], [5]). PyTorch hadir di sejumlah studi pelatihan dan *benchmarking* ([10], [2], [14], [17]), menunjukkan fleksibilitasnya pada GPU. FPGA memiliki ekosistem *tools* yang lebih khusus, seperti Vitis-AI ([2], [6], [21], [23]) serta MathWorks HDL Toolbox ([13], [2]) untuk eksplorasi *hardware-aware design*. ONNX juga digunakan untuk interoperabilitas model ([16], [2]), dan TF-TRT dipakai untuk optimasi *inference* pada GPU dan TPU ([11],[7]). *Pipeline* yang digunakan dalam *deployment* bervariasi tergantung akselerator. 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.

2) RQ2 — 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 ([10], [12], [2], [14]). Beberapa artikel menunjukkan bahwa GPU masih unggul dalam kecepatan *training* model besar seperti CNN dan ViT, meskipun konsumsi daya cenderung lebih tinggi ([7], [17]). Di sisi lain, TPU menunjukkan efisiensi tinggi untuk inferensi *real-time* pada sistem seperti deteksi intrusi dan IoT ([14], [16]), dengan latensi hanya beberapa milidetik dan konsumsi daya rendah ([11]). FPGA sering kali menjadi yang tercepat dalam inferensi, seperti ditunjukkan pada [11] dan [21], dengan *throughput* mencapai 14–25 FPS dan efisiensi daya jauh di atas GPU/TPU. Studi [2] dan [23] juga melaporkan bahwa FPGA unggul dalam aplikasi *edge*, dengan kombinasi latensi rendah dan *footprint* memori kecil. Dari 20 artikel yang dianalisis, tren menunjukkan bahwa GPU paling unggul untuk pelatihan, FPGA unggul dalam inferensi dengan efisiensi energi, dan TPU berada di tengah dengan performa *inference* cepat namun fleksibilitas terbatas.

3) RQ3 — 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 [11], [2], dan [21], dengan efisiensi energi mencapai 5–10W bahkan pada performa inferensi tinggi. Studi [23] 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 inferensi berbasis LSTM untuk sistem deteksi intrusi ([5], [17]), dengan latensi hanya 6 ms dan akurasi tinggi. Sementara itu, GPU masih cenderung lebih boros daya, terutama pada versi *embedded* seperti Jetson TX2 ([11], [7]), meskipun tetap menjadi pilihan utama dalam skenario pelatihan karena kemampuannya memproses *batch* besar dengan *throughput* tinggi ([10], [14], [17]). Secara keseluruhan, 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.

### 4.3 Analisis Gap

Meskipun studi-studi yang ditinjau memberikan kontribusi yang signifikan, masih terdapat sejumlah kesenjangan (gap) yang belum terjawab secara memadai dalam literatur yang ada. Selain tantangan teknis yang melekat pada masing-masing akselerator sebagaimana diuraikan dalam sub-bab sebelumnya, analisis lintas-studi mengidentifikasi empat *gap* penelitian yang bersifat struktural. Gap pertama adalah ketiadaan standar evaluasi yang seragam. Mayoritas studi menggunakan metrik dan dataset yang berbeda-beda, sehingga perbandingan langsung antar akselerator menjadi sulit dilakukan secara objektif. Hanya [6] yang menggunakan platform *benchmarking* terstandar (*QuTiBench*), sementara studi-studi lainnya mengandalkan konfigurasi eksperimen yang sangat spesifik terhadap domain dan model yang digunakan. Gap kedua adalah keterbatasan cakupan TPU dalam studi komparatif. Dari 20 artikel, hanya 10 yang menyertakan TPU sebagai objek kajian, dan sebagian besar berfokus pada *Edge* TPU (Google Coral) bukan TPU generasi terbaru dari Google Cloud sehingga menciptakan *blind spot* mengenai performa TPU untuk pelatihan model skala besar. Gap ketiga adalah minimnya kajian mengenai akselerator *hybrid*. Hanya sedikit studi ([11], J19) yang mengeksplorasi kombinasi lebih dari satu jenis akselerator dalam satu *pipeline* sistem, padahal pendekatan *heterogeneous computing* merupakan tren yang semakin relevan dalam arsitektur sistem AI modern. Gap keempat adalah absennya panduan berbasis bukti untuk peneliti akademis. Hampir seluruh studi ditujukan untuk kalangan industri atau peneliti tingkat lanjut, dan tidak ada yang secara eksplisit menyusun rekomendasi yang disesuaikan dengan keterbatasan sumber daya yang umum dihadapi oleh mahasiswa atau peneliti di negara berkembang. Keempat *gap* ini secara kolektif memperkuat urgensi dan relevansi penelitian yang dilakukan oleh Fitriani et al. saat ini.

Setiap akselerator juga memiliki tantangan teknis dan keterbatasan yang khas. GPU, meskipun fleksibel dan didukung banyak *framework* ([10], [12], [14]), menghadapi kendala pada efisiensi daya dan latensi, terutama pada perangkat *embedded* seperti Jetson TX2 ([11], [7]). Manajemen *thread* dan *bottleneck* memori menjadi isu umum pada GPU, seperti dilaporkan dalam studi penjadwalan pada [2] dan [17]. TPU memiliki keterbatasan dalam hal kompatibilitas model; beberapa artikel ([11], [5], [16]) melaporkan bahwa TPU hanya mendukung *subset* operator tertentu, sehingga diperlukan penyesuaian arsitektur dan *quantization* model agar dapat dikompilasi, yang menyulitkan penggunaan TPU untuk arsitektur *custom* seperti *Transformer* ([14]). FPGA, meskipun

sangat efisien secara daya dan latensi ([2], [21], [23]), memerlukan proses kompilasi yang kompleks dan waktu pengembangan yang lebih lama, serta kinerjanya sangat sensitif terhadap pengaturan paralelisme dan ukuran *batch* ([8], [6], J17).

### 4.4 Klasifikasi Topik

Berdasarkan hasil analisis tematik terhadap 20 artikel ([9 - 13],[2 - 8], [14 - 22]), teridentifikasi lima klaster topik utama yang sedang dikaji dalam literatur akselerator AI. Klaster pertama adalah inferensi model di lingkungan edge, yang mencakup studi-studi tentang implementasi CNN, LSTM, dan model ringan pada FPGA dan TPU untuk aplikasi IoT, pertanian cerdas, dan kendaraan otonom ([11], [2], [14], [15], [21], [23]). Klaster ini merupakan yang paling dominan, muncul dalam 12 dari 20 artikel yang dikaji. Klaster kedua adalah efisiensi energi dan optimasi daya, yang membahas perbandingan konsumsi daya antara GPU, TPU, dan FPGA dalam berbagai skenario beban kerja ([11], [2], [8], [24], [21], [23]). Klaster ketiga adalah arsitektur dan framework akselerator, yang meninjau kompatibilitas model AI dengan ekosistem perangkat keras tertentu, termasuk penggunaan TensorFlow, PyTorch, Vitis-AI, dan HDL ([10], [13], [5], [6], [7]). Klaster keempat adalah performa *training dan benchmarking*, yang membandingkan kecepatan pelatihan dan *throughput* inferensi antar akselerator menggunakan dataset standar ([12], [6], [16], [17], [18], [19]). Klaster kelima adalah domain aplikasi khusus, yang mencakup penerapan akselerator pada bidang-bidang spesifik seperti penginderaan jauh ([10], [23]), komunikasi 6G ([13], J19), sistem keamanan siber ([5], [16]), dan aplikasi *safety-critical* di bidang antariksa dan otomotif ([8]). Tren yang paling menonjol adalah meningkatnya jumlah studi pada klaster pertama dan kedua sejak tahun 2022, yang mencerminkan pergeseran fokus riset dari eksperimen laboratorium menuju implementasi nyata di lingkungan terbatas daya.

Studi-studi yang dianalisis (J1–J20) mencakup beragam domain aplikasi, mulai dari penginderaan jauh ([10], [23]), pertanian cerdas ([11]), hingga komunikasi nirkabel 6G ([13], [22]), sistem deteksi intrusi IoT ([5], [16]), dan sistem tertanam otonom ([2], [8], [21]). 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 klasifikasi citra medis atau *training* model skala besar ([12], [14], [17]). Temuan ini menunjukkan bahwa pemilihan akselerator tidak hanya bergantung pada performa teknis, tetapi juga sangat dipengaruhi oleh konteks aplikasi dan batasan lingkungan operasional.

### 4.5 Pembahasan

Dari sisi metodologi, studi ini mengadopsi protokol SLR Kitchenham yang terstruktur, berbeda dari pendekatan yang digunakan oleh sebagian besar penelitian terdahulu. Alzubaidi et al. [5] menggunakan *narrative review* yang bersifat deskriptif tanpa prosedur seleksi yang terstandar; Blott et al. [6] mengandalkan *benchmarking* eksperimental tanpa sintesis literatur; sementara Alam et al. [7] melakukan survei tematik tanpa penilaian kualitas artikel secara eksplisit. Studi saat ini mengatasi keterbatasan-keterbatasan tersebut dengan menerapkan seleksi empat tahap, penilaian kualitas DARE, dan sintesis naratif yang dipandu oleh RQ, sehingga menghasilkan temuan yang lebih transparan dan dapat diverifikasi.

Dari sisi cakupan, studi ini mengevaluasi ketiga akselerator (GPU, TPU, dan FPGA) secara simultan dalam satu kerangka analisis, mencakup delapan domain aplikasi yang beragam. Ini merupakan perbedaan mendasar dibandingkan Vaithianathan et al. yang hanya membandingkan GPU dan FPGA tanpa TPU, serta Rech yang membatasi kajian pada domain *safety-critical* saja. Studi saat ini memperluas temuan kedua penelitian tersebut dengan menunjukkan bahwa profil performa akselerator sangat bergantung pada domain aplikasi dan bukan semata-mata pada spesifikasi perangkat keras.

Kontribusi studi ini terhadap kumpulan pengetahuan yang ada dapat dilihat dari tiga aspek. Pertama, studi ini *mengkonfirmasi dan memperkuat* temuan Alzubaidi et al. mengenai dominasi arsitektur CNN dalam implementasi AI berbasis perangkat keras, sekaligus memperluas cakupannya ke domain penginderaan jauh dan komunikasi 6G yang belum dicakup dalam kajian tersebut. Kedua, studi ini memperbarui dan memvalidasi temuan Alam et al. mengenai tren adopsi akselerator di ekosistem *edge* AI dengan data yang lebih mutakhir (hingga 2025) dan metodologi yang lebih ketat. Ketiga, studi ini mengisi *gap* yang ditinggalkan oleh Blott et al. [6] dan Vaithianathan et al. [9] dengan menyediakan sintesis lintas-domain yang komprehensif, termasuk analisis *toolchain* dan implikasi praktis bagi peneliti akademis. Berdasarkan temuan dari 20 artikel yang ditinjau, pemilihan

akselerator dalam penelitian tesis mahasiswa magister ilmu komputer sebaiknya mempertimbangkan kompleksitas model, ketersediaan perangkat keras, domain aplikasi, serta kebutuhan akan efisiensi daya atau latensi. GPU disarankan untuk eksperimen arsitektur model skala besar; TPU untuk *deployment real-time* dan hemat daya; dan FPGA untuk aplikasi khusus dengan kendali penuh atas *pipeline* komputasi. Dengan demikian, temuan-temuan studi ini terintegrasi secara koheren ke dalam korpus literatur yang ada dan memberikan landasan bukti yang lebih solid untuk pengembangan riset akselerator AI ke depan, khususnya dalam eksplorasi arsitektur *hybrid* dan strategi *algorithm-hardware co-design*.

## 5. Simpulan

Penelitian ini melakukan Systematic Literature Review (SLR) berbasis kerangka kerja Kitchenham terhadap 20 artikel ilmiah (2020–2025) yang mengkaji penggunaan GPU, TPU, dan FPGA sebagai akselerator dalam sistem kecerdasan buatan. Hasil sintesis menunjukkan bahwa GPU merupakan akselerator yang paling dominan digunakan dalam 90% studi, terutama untuk pelatihan model berskala besar berkat fleksibilitas framework dan kemampuan paralelisme tinggi, meskipun cenderung lebih boros daya dibanding akselerator lainnya.

FPGA hadir dalam 70% studi dan unggul dalam efisiensi energi pada 78% artikel relevan, serta menunjukkan latensi inferensi yang rendah sehingga menjadi pilihan utama untuk aplikasi edge computing, sistem tertanam, dan lingkungan dengan keterbatasan daya. TPU digunakan dalam 50% studi dan menawarkan inferensi cepat dengan konsumsi daya rendah, cocok untuk aplikasi real-time berbasis IoT dan keamanan siber, namun memiliki keterbatasan kompatibilitas model karena hanya mendukung subset operator tertentu.

Studi ini menyimpulkan bahwa pemilihan akselerator AI tidak dapat didasarkan pada satu parameter tunggal, melainkan harus mempertimbangkan secara holistik efisiensi daya, arsitektur sistem, kebutuhan domain aplikasi, serta ketersediaan sumber daya. GPU direkomendasikan untuk eksperimen arsitektur model besar, TPU untuk deployment real-time yang hemat daya, dan FPGA untuk aplikasi khusus yang membutuhkan kontrol penuh atas pipeline komputasi. Temuan ini memberikan panduan praktis bagi mahasiswa magister ilmu komputer dalam menentukan akselerator yang sesuai dengan topik tesis, tujuan eksperimen, dan keterbatasan sumber daya yang tersedia. Selain itu, studi ini mengidentifikasi empat kesenjangan penelitian yang masih terbuka, yakni ketiadaan standar evaluasi yang seragam, minimnya kajian akselerator hybrid, keterbatasan cakupan TPU dalam studi komparatif, serta absennya panduan berbasis bukti yang ditujukan khusus bagi peneliti akademis, yang mendorong urgensi penelitian lanjutan di bidang akselerator AI.

## Daftar Referensi

[1] B. Kitchenham, L. Madeyski, & P. Brereton, "Meta-analysis for families of experiments in software engineering: A systematic review and reproducibility and validity assessment. Empirical Software Engineering, Vol. 25, no. 1, pp. 353–401, 2020. https://doi.org/10.1007/s10664-019-09747-0

[2] S. Bertazzoni *et al.*, “Design Space Exploration for Edge Machine Learning Featured by MathWorks FPGA DL Processor : A Survey,” vol. 12, no. 11, pp. 157482–157504, 2024, doi: 10.1109/ACCESS.2024.3349128.

[3] A. Martín-martín *et al.*, “Hardware Implementations of a Deep Learning Approach to Optimal Configuration of Reconfigurable Intelligence Surfaces,” *IEEE Trans. Wireless Commun.,* vol. 23, no. 8, pp. 4521–4541, 2024, doi: 10.1109/TWC.2024.3372610.

[4] 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.

[5] M. Blott *et al.*, “Evaluation of Optimized CNNs on Heterogeneous Accelerators Using a Novel Benchmarking Approach,” *IEEE Trans. Comput.*, vol. 70, no. 10, pp. 1654–1669, 2021, doi: 10.1109/TC.2020.3022318.

[6] 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, doi: 10.3390/electronics13152988.

[7] 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.

[8] M. Vaithianathan, S. Udkar, and A. Micro, “Comparative Study of FPGA and GPU for High-Performance Computing and AI,” *Int. J. Adv. Comput. Technol., v*ol. 1, no. 1, pp. 107–115, 2024,

doi: 10.56472/25838628/IJACT-V1I1P107.
[9] 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.
[10] 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.
[11] 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.
[12] S. Huang, E. Tang, S. Li, X. Ping, and R. Chen, “Hardware-friendly compression and hardware acceleration for transformer : A survey,” *EURASIP J. Adv. Signal Process.,* vol. 30, no. July, pp. 3755–3785, 2022, doi: 10.3934/era.2022192.
[13] P. Kang and J. Jo, “Benchmarking Modern Edge Devices for AI Applications ∗ SUMMARY,” *IEICE Trans. Inf. Syst.*, vol. Vol. E104-, no. 3, pp. 394–403, 2021, doi: 10.1587/transinf.2020EDP7160.
[14] A. S. Aguiar *et al.*, “Visual Trunk Detection Using Transfer Learning and a Deep Learning-Based Coprocessor,” *IEEE Access, 2020, vol. 8, no. 1, pp. 78308–78317,* doi: 10.1109/ACCESS.2020.2989052.
[15] 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.
[16] 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.
[17] P. Kang and A. Somtham, “An Evaluation of Modern Accelerator-Based Edge Devices for Object Detection Applications,” *Mathematics*, vol. 10, pp. 1–14, 2022.
[18] 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.
[19] F. Pacini, T. Pacini, G. Lai, A. M. Zocco, and L. Fanucci, “Deployment of a CNN for Motor Imagery Classification in Brain-Computer Interfaces,” *IEEE Trans. Neural Syst. Rehabil. Eng.,* 2024, vol. 32, no. 1, pp. 1412–1421, doi: 10.1109/TNSRE.2024.3359218.
[20] J. Lu and S. X. Tan, “Thermal Map Dataset for Commercial Multi / Many Core CPU / GPU / TPU”, *in Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED),* 2024, pp. 1–6, doi: 10.1145/3670474.3685963.
[21] C. Rao *et al.*, “ICARUS : A Specialized Architecture for Neural Radiance Fields Rendering,” vol. 41, no. 6, 2022, doi: 10.1145/3550454.3555505.
[22] 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,” *Remote Sens.,* 2021, vol. 13, no. 8, pp. 1519–1536, doi: 10.3390/rs13081519.
[23] 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.*, 2024, vol. 1, no. May 2023, pp. 37–46, doi: 10.56472/25838628/IJACT-V1I1P107.