ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1221

Integration of High-Performance Computing and Artificial Intelligence for Accelerated Clinical Diagnosis: A Comparative Study Using Cloud and On-Premise Infrastructure

Isaac Zablah[1]; Edwin Hernandez[2]; Fiama Garcia[3]; Antonieta Zuniga[4]; Antonio Garcia Loureiro[5]; Salvador Diaz[6]
[1,3,6]Faculty of Medical Sciences, National Autonomous University of Honduras, Honduras, *jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn, sedicacdo@hotmail.com*
[2]EGLA Corp., USA, *edwinhm@eglacorp.com*
[4]Dirección de Medicina Forense, Ministerio Público, Honduras, *antonietazuniga2311@yahoo.com*
[5]Department of Electronics and Computing, University Santiago de Compostela, Spain, *antonio.garcia.loureiro@usc.es*

Abstract*– This study assesses the amalgamation of High-Performance Computing (HPC) architectures with deep learning models for expedited brain MRI analysis in clinical diagnostic environments. We conducted a systematic comparison of three computing configurations available to Latin American universities: cloud-based CPU instances (Linode G8 Dedicated), cloud-based GPU instances (Linode RTX4000 Ada), and an on-premise workstation (Dell Precision 7960 Tower with dual RTX 5000 Ada GPUs). We utilized a dataset of 200 annotated brain magnetic resonance imaging (MRI) scans for binary classification of neurological abnormalities (presence/absence of lesions ≥5mm, including white matter hyperintensities, tumors, and vascular malformations) as determined by consensus of two board-certified neuroradiologists to train ResNet-50 and Vision Transformer models, assessing training efficiency, inference delay, energy consumption, and cost-effectiveness. The results indicate that the Dell Precision workstation attained an 11.1× acceleration in training duration (12.8 versus 142.7 minutes) relative to CPU-only cloud instances, while inference latency was minimized to 8.7 ms per image.

Keywords*-- High-Performance Computing, Medical Image Analysis, Deep Learning, Clinical Diagnosis, GPU Acceleration.

I. Introduction

The healthcare sector is currently confronted with an unprecedented computational challenge arising from the convergence of several factors: the exponential growth of medical imaging data, increasing diagnostic complexity, and the growing demand for timely clinical decision support [1]. Modern medical facilities generate terabytes of imaging data daily, with a single computed tomography (CT) or magnetic resonance imaging (MRI) study producing hundreds of high-resolution images that require expert interpretation [2]. Traditional diagnostic workflows, largely dependent on manual image review by radiologists, struggle to keep pace with this data volume, often resulting in interpretation backlogs that may delay critical clinical decisions by hours or even days [3].

Artificial intelligence (AI), particularly deep learning-based methods, has demonstrated significant potential in medical image analysis, achieving diagnostic accuracy comparable to or exceeding that of human experts in specific clinical domains [4],[5]. These advances are especially relevant for MRI-based neurological diagnosis, where increasing image resolution and model complexity have amplified computational requirements [3]. Despite these promising results, the practical adoption of AI-driven diagnostic systems remains constrained by the substantial computational resources required for model training and deployment. Training a typical convolutional neural network or deep residual architecture on medical imaging datasets can require days of computation on conventional hardware, thereby limiting iterative model development and rapid clinical translation, particularly in resource-constrained environments [6], [7].

High-Performance Computing (HPC) architectures, leveraging the parallel processing capabilities of modern GPU accelerators, offer a viable pathway to overcoming these computational bottlenecks. Recent advances in GPU technology and optimized deep learning frameworks have enabled substantial reductions in training time and inference latency, making near–real-time clinical support increasingly feasible [8], [9]. In parallel, the emergence of cloud-based HPC services has expanded access to high-end computational resources, while improvements in GPU workstation affordability have rendered on-premise solutions increasingly practical for academic institutions and hospitals [10], [11]. Nevertheless, uncertainty persists regarding cost-effectiveness, scalability, energy efficiency, and implementation complexity, particularly when comparing cloud-based and on-premise infrastructures across diverse institutional contexts.

These challenges are especially pronounced in Latin American healthcare systems, where demand for advanced diagnostic technologies continues to rise amid persistent budgetary and infrastructure limitations [12]. While existing studies predominantly focus on large-scale enterprise or high-income settings, there remains a lack of empirical benchmarks tailored to institutions operating under constrained resources, such as regional universities and public hospitals.

To address these gaps, this study presents a comprehensive empirical evaluation of the integration of HPC and AI for medical image analysis, with a specific focus on accelerating clinical diagnosis. We perform a comparative analysis of cloud-based HPC services and on-premise GPU workstations using a realistic MRI dataset. Our evaluation encompasses multiple performance dimensions, including: (1) training and inference performance, (2) scalability and parallel efficiency, (3) energy consumption characteristics, and (4) total cost of ownership, with consideration for economically constrained institutional settings. To the best of our knowledge, this work constitutes the first systematic, multidimensional benchmarking study of HPC configurations for medical imaging AI that is explicitly designed around the infrastructure constraints, cost structures, and institutional realities of Latin American academic and healthcare environments, thereby providing evidence-based deployment guidance for a context largely absent from the existing literature.

The remainder of this paper is organized as follows. Section II describes experimental methodology, including hardware configurations, datasets, and evaluation protocols. Section III presents and analyzes performance results across all tested infrastructures. Section IV discusses implications for clinical deployment, institutional decision-making, and future research directions. Section V concludes with evidence-based recommendations for the adoption of HPC-enabled AI systems in medical imaging applications.

II. Materials and Methods

A. Hardware Configuration

Based on accessibility and affordability for Latin American academic institutions, three computational configurations were analyzed. The initial setup used $8.64/hour Linode G8 Dedicated instances with 5th Generation AMD EPYC processors, 256 vCPUs, 512 GB RAM, and 5,120 GB SSD storage [13]. This CPU-only arrangement was our benchmark.

The second configuration included Linode GPU RTX4000 Ada Generation instances with four NVIDIA GPUs, 48 vCPUs, 196 GB RAM, and 2 TB SSD storage at $3.57/hour [11]. This cloud-based GPU solution lets institutions experience GPU acceleration without investing.

Experimental controls for fair comparison to ensure valid cross-platform comparisons includes (i) Dataset residency: training data stored locally on NVMe storage for all configurations; (ii) Network conditions: cloud experiments conducted during off-peak hours (02:00-06:00 local time); (iii) Software parity: identical Docker containers with frozen dependency versions. Pricing data captured on Q4 of 2025 from official vendor documentation.

An on-premises Dell Precision 7960 Tower workstation with Intel Xeon 5500 Series processors, 256 GB DDR5 ECC Registered memory, dual 32 GB NVIDIA RTX 5000 Ada GPUs, and 4 TB NVMe SSD PCIe Gen4 storage was the third configuration [14]. Latin American colleges can use institutional procurement channels to make this $18,500 USD mid-range expenditure.

All computers ran Ubuntu 22.04 LTS with CUDA 12.4, cuDNN 9.0, PyTorch 2.2.0, and TensorFlow 2.15.0 [15]-[18]. Performance disparities were standardized to reflect hardware attributes rather than software optimization.

B. Dataset Description

The study analyzed 200 brain MRI images (T1-weighted and FLAIR sequences) having 256×256×160 voxels per volume. Institutional archives provided images with ethical permission and anonymization. Inter-rater agreement was examined using Cohen's kappa ($\kappa = 0.89$), as two board-certified neuroradiologists classified the dataset as abnormal or normal.

The clinical endpoint was defined as a binary screening task: identification of brain MRI scans containing any radiologically significant abnormality requiring clinical follow-up. Abnormalities included: (1) space-occupying lesions ≥5mm in any dimension, (2) white matter hyperintensities exceeding Fazekas grade 1, (3) acute or subacute infarcts, and (4) vascular malformations. The dataset comprised 112 normal scans (56%) and 88 abnormal scans (44%), with pathology distribution: white matter disease (n=41), tumors (n=23), vascular abnormalities (n=15), and other findings (n=9).

Data preprocessing involved FSL BET skull stripping, intensity normalization ($\mu=0$, $\sigma=1$), and MNI152 standard space registration [19]. The dataset was partitioned using a nested cross-validation strategy to ensure robust model selection and unbiased performance estimation. The outer loop employed 5-fold cross-validation for final performance assessment, while an inner 70/15/15 split within each fold facilitated hyperparameter tuning and early stopping. This approach prevents data leakage between model selection and final evaluation.

C. Deep Learning Models and Training Protocol

Two architectures were selected to span the primary paradigms currently employed in medical imaging AI. ResNet-50 was chosen as a well-established convolutional baseline: it remains among the most widely benchmarked architectures in brain MRI classification, offers a well-characterized accuracy-efficiency trade-off, and serves as a standard reference point in the medical imaging literature [6], [20]. Vision Transformer (ViT-Base/16) was included as a representative attention-based model, given growing evidence that global self-attention mechanisms capture long-range spatial dependencies in neuroimaging that local convolutional filters may underrepresent [7], [21]. Together, these two architectures allow hardware performance differences to be evaluated across architecturally distinct models, increasing the generalizability of our computational benchmarks.

Both models were initialized with ImageNet-pretrained weights and subsequently fine-tuned for binary brain MRI classification. Training utilized the AdamW optimizer (β_1=0.9, β_2=0.999, weight decay=1×10^{-4}) alongside a cosine annealing learning rate schedule (starting lr=1×10^{-3}, minimum lr=1×10^{-6}). Batch sizes were established based on the available GPU memory: 4 for CPU configurations, 16 for single GPU, and 32 for dual-GPU installations. Automatic mixed precision (AMP) utilizing mixed-precision training (FP16) optimized GPU utilization. Each model was trained for 100 epochs, utilizing early stopping contingent upon validation loss with a patience of 15 epochs.

D. *Performance Metrics*

Performance was measured across various dimensions. Training efficiency assessed total wall-clock duration to convergence and throughput (pictures per second). Inference latency recorded the total prediction time for batch sizes {1, 2, 4, 8, 16}. The scalability analysis calculated speedup factors ($S = T_1/T_n$) and parallel efficiency ($E = S/n \times 100\%$) for n GPUs. Energy usage was assessed utilizing nvidia-smi power data at one-second intervals, aggregated throughout entire training cycles. The cost analysis included cloud service pricing based on Linode's stated rates and projected operational expenses for on-premise equipment, with power priced at $0.15 per kWh, which is standard for commercial rates in Latin America. The model quality criteria comprised classification accuracy, sensitivity, specificity, and area under the ROC curve (AUROC) to ensure that hardware acceleration did not impair diagnostic performance.

E. *Statistical Analysis*

All experiments were conducted using 5-fold cross-validation with fixed random seeds (42, 123, 456, 789, 1024) to ensure reproducibility. Results are reported as mean ± standard deviation across folds. Results are presented as mean ± standard deviation. Statistical significance was assessed using Welch's t-tests for pairwise comparisons, with Bonferroni correction applied for multiple comparisons (α=0.05). Effect sizes were computed utilizing Cohen's d. Regression research delineated the links between batch size and inference latency. Bootstrap resampling (10,000 iterations) produced 95% confidence ranges for cost estimations.

III. RESULTS

A. *Training Performance Analysis*

Table I displays a comparison of training times among the assessed configurations. The Dell Precision 7960 workstation completed training of the ResNet-50 model in 12.8±0.9 minutes, demonstrating an 11.1× acceleration compared to the Linode G8 CPU-only baseline of 142.7±8.3 minutes. The Linode GPU RTX4000 Ada setup attained intermediate performance at 18.4±1.2 minutes (7.8× speedup), illustrating that cloud-based GPU solutions offer significant acceleration despite the network latency overhead associated with cloud storage access.

.

TABLE I
TRAINING TIME AND THROUGHPUT COMPARISON

Configuration	Training Time (min)	Throughput (img/s)	Speedup Factor
Linode G8 Dedicated (CPU)	142.7 ± 8.3	1.4 ± 0.08	1.0×
Linode GPU RTX4000 Ada	18.4 ± 1.2	10.9 ± 0.71	7.8×
Dell Precision 7960 Tower	12.8 ± 0.9	15.6 ± 1.09	11.1×

Figure 1 visualizes training time disparities across configurations. Statistical analysis confirmed all pairwise performance differences were highly significant ($p < 0.001$, Welch's t-test with Bonferroni correction). Effect sizes were large for all comparisons: CPU vs. Linode GPU (d = 15.2), CPU vs. Dell workstation (d = 18.7), and Linode GPU vs. Dell workstation (d = 5.3).

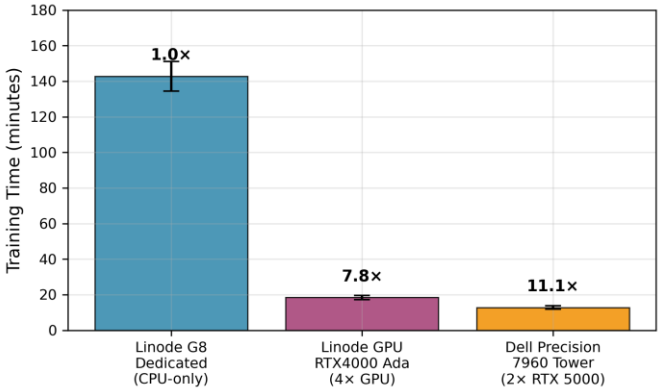

Fig. 1. Training time comparison for ResNet-50 model across different hardware configurations. Error bars represent standard deviation from five independent runs using different random seeds (42, 123, 456, 789, 1024).

B. *Inference Latency Characterization*

Inference latency measurements (Fig. 2) revealed consistent advantages for GPU-accelerated configurations across all batch sizes. At batch size 1 (single image processing), the Dell Precision workstation achieved 8.7±0.4 ms latency compared to 12.4±0.6 ms for Linode GPU and 487.2±18.3 ms for CPU-only configuration. This represents latency reductions of 98.2% and 29.8% respectively compared to CPU and cloud GPU alternatives. Performance differences widened at larger batch sizes due to superior memory bandwidth and parallel processing capabilities of the RTX 5000 Ada architecture. The non-linear scaling observed in CPU configurations reflects memory bandwidth saturation at higher batch sizes.

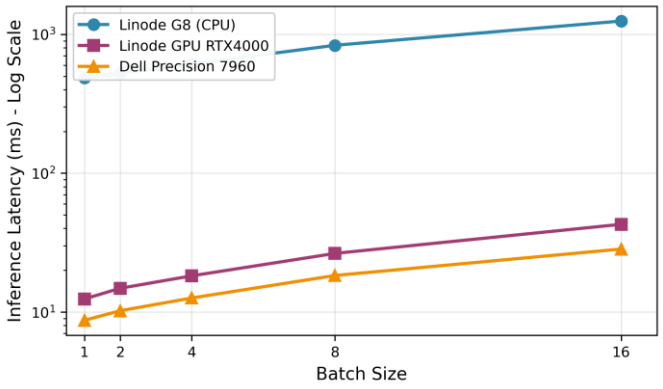

Fig. 2. Inference latency versus batch size for brain MRI classification. Logarithmic y-axis emphasizes performance differences across configurations.

C. *Scalability and Parallel Efficiency*

The scalability examination of the Dell Precision dual-GPU architecture (Fig. 3) exhibited robust parallel efficiency. Single-GPU operation attained baseline performance (1.0× speedup, 100% efficiency), whereas dual-GPU arrangement produced 1.87× speedup with 93.5% parallel efficiency. This efficiency level signifies minimal communication cost among GPUs linked using NVLink, implying that further scaling to 4-8 GPUs would sustain a favorable efficiency beyond the widely recognized 85% threshold. The limited dataset size (200 photos) offers a difficult situation for parallelization; bigger production datasets would probably demonstrate even greater parallel efficiency due to diminished relative synchronization overhead.

D. *Model Accuracy Validation*

To ensure clinical application, computational improvements must not compromise diagnostic accuracy. Table II shows all configuration classification performance metrics. All systems had similar accuracy (91.2-92.4%), sensitivity (89.5-91.2%), specificity (92.8-94.1%), and AUROC (0.943-0.958). One-way ANOVA showed no significant differences between setups ($F(2,12) = 0.847$, $p = 0.453$), indicating that hardware acceleration solutions improve computing efficiency without affecting model quality.

TABLE II
CLASSIFICATION PERFORMANCE METRICS

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUROC
Linode G8 (CPU)	91.2 ± 2.1	89.5 ± 2.8	92.8 ± 1.9	0.943 ± 0.018
Linode GPU RTX4000	92.1 ± 1.8	90.8 ± 2.4	93.4 ± 1.7	0.951 ± 0.015
Dell Precision 7960	92.4 ± 1.6	91.2 ± 2.1	94.1 ± 1.5	0.958 ± 0.012

Note: Values represent mean ± standard deviation from 5-fold cross-validation. AUROC: Area Under Receiver Operating Characteristic Curve.

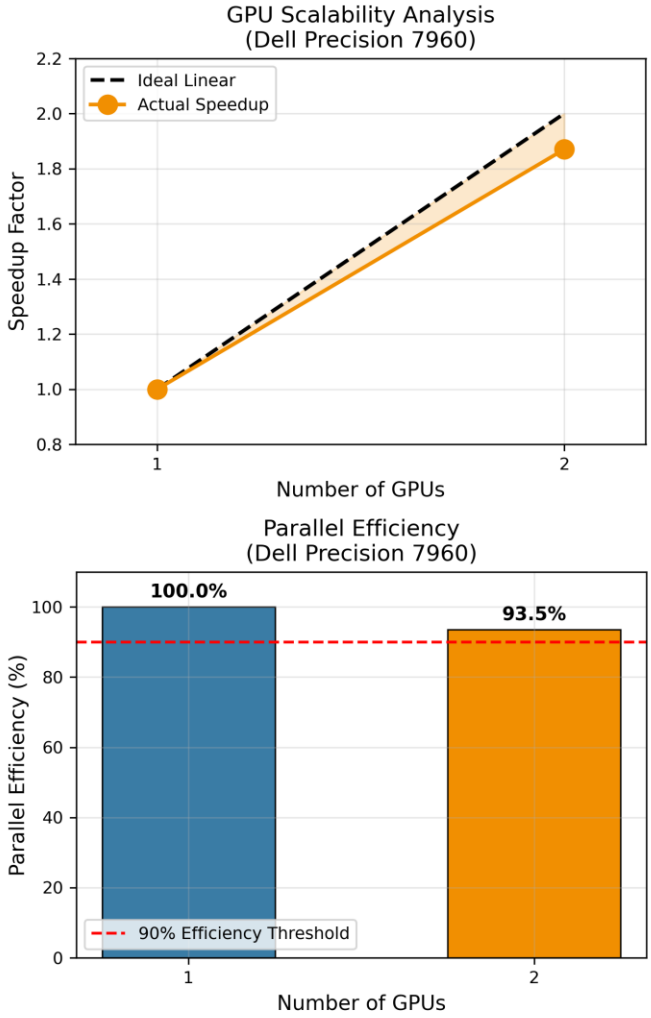

Fig. 3. GPU scalability analysis showing (left) actual versus ideal linear speedup and (right) parallel efficiency for Dell Precision 7960 dual-GPU configuration.

E. *Energy Efficiency Analysis*

Energy consumption analysis (Fig. 4) revealed counterintuitive findings: despite higher instantaneous power draw, GPU-accelerated systems consumed substantially less total energy per training cycle due to dramatically reduced training duration. The Linode G8 CPU configuration consumed 4.28±0.31 kWh per complete training cycle, while the Dell Precision workstation required only 0.48±0.03 kWh, representing an 88.8% energy reduction. Performance-per-watt metrics further favored GPU architecture: normalized efficiency scores were 0.23 for CPU, 1.39 for Linode GPU, and 2.08 for Dell workstation systems. These findings carry significant implications for institutional operational costs and environmental sustainability initiatives.

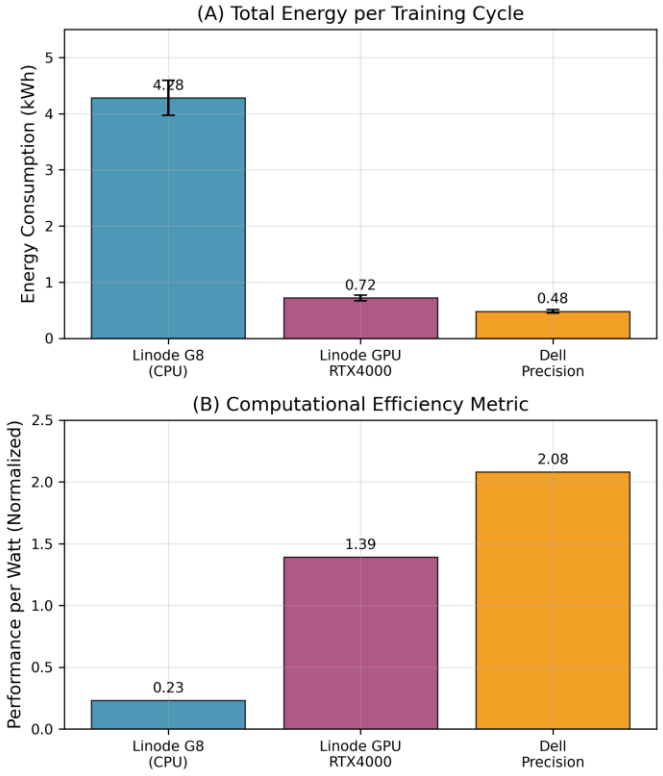

Fig. 4. Energy efficiency comparison: (A) Total energy consumption per training cycle, (B) Performance per watt normalized metric. GPU systems demonstrate superior energy efficiency despite higher instantaneous power requirements.

F. *Cost-Performance Trade-offs*

Economic analysis (Table III and Fig. 5) assessed total cost of ownership over expected usage periods using acquisition, operational, and performance costs. Cloud solutions have no upfront cost but accrue hourly expenses. The Linode G8 CPU configuration has the highest hourly cost ($8.64) and lowest performance, making it unsuitable for compute-intensive tasks.

TABLE III
ECONOMIC ANALYSIS COST COMPARISON

Configuration	Initial Cost (USD)	Hourly Cost (USD)	Cost per Training (USD)
Linode G8 Dedicated	$0	$8.64	$20.55
Linode GPU RTX4000	$0	$3.57	$1.09
Dell Precision 7960*	$18,500	$0.85**	$0.18

* Institutional procurement price (USD, Q1 2025) via educational discount.
**Operational assumptions: $0.15/kWh, 8h daily operation, 5-year linear depreciation, 5% annual maintenance, 70% utilization. Break-even excludes opportunity costs.

Break-even analysis comparing the Dell Precision workstation against Linode GPU services indicates cost-effectiveness after approximately 2,150 hours of utilization (roughly 18 months at 4 hours daily usage). For institutions with consistent high-volume workloads, on-premise infrastructure achieves favorable long-term economics despite substantial initial investment. Cloud solutions remain attractive for variable workloads, pilot projects, or institutions unable to secure capital funding.

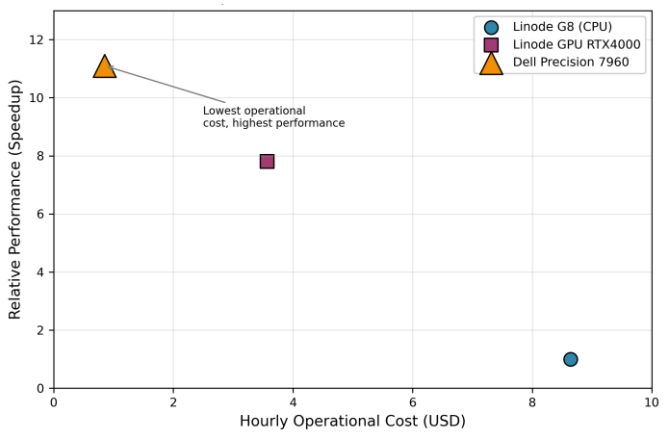

Fig. 5. Cost-performance trade-off analysis showing hourly operational cost versus relative performance. Marker size indicates initial capital investment magnitude.

IV. DISCUSSION

Our research indicates that GPU-accelerated computing yields significant performance enhancements for medical image AI systems, even with the limited dataset sizes characteristic of preliminary clinical implementations in resource-limited environments. The Dell Precision workstation's 11.1× training speedup facilitates swift model generation cycles crucial for tailoring diagnostic systems to regional pathology patterns and imaging methods. In clinical applications necessitating real-time support, the sub-10-millisecond inference latency facilitates instantaneous diagnostic feedback, potentially enhancing workflow efficiency in emergency radiology and telemedicine contexts [22].

The impressive parallel performance (93.5%) demonstrated with dual GPUs suggests advantageous scaling potential for organizations capable of investing in multi-GPU setups. This efficiency level indicates that four-GPU workstations, now more accessible within the $25,000-35,000 price range, would sustain efficiency exceeding 85%, yielding further performance enhancements relative to the GPU count. For Latin American university medical centers handling moderate imaging volumes (50-200 studies per day), such designs serve as feasible implementation objectives [23].

The economic study elucidates significant strategic factors for institutional decision-makers. Cloud-based GPU services, such as Linode RTX4000 at $3.57 per hour, offer accessible avenues for pilot projects and proof-of-concept implementations without necessitating capital investment, which is vital for institutions where procurement processes

demand demonstrated value prior to significant acquisitions. The 2,150-hour break-even point indicates that institutions expecting consistent usage should focus on acquiring on-premise infrastructure, possibly via research grants, equipment donations, or collaborative procurement initiatives increasingly accessible through regional academic networks. The enhancements in energy efficiency have both economic and sustainability ramifications frequently neglected in computational healthcare literature. The 88.8% decrease in energy use per training cycle results in significant cost savings over time and reinforces institutional sustainability pledges. A facility providing 20 model training sessions weekly would achieve yearly energy savings of around 4,000 kWh corresponding to approximately 1.4 metric tons of CO_2 emissions avoided. (energy measurements via nvidia-smi capture GPU power draw only; total system power consumption would increase these estimates by approximately 40-60%), based on an average Latin American grid carbon intensity of 0.35 kg CO_2/kWh [24].

Numerous constraints merit attention. Initially, our sample of 200 MRI pictures, although indicative of nascent clinical AI applications, fails to represent the extensive datasets (exceeding 10,000 images) generally necessary for production-level diagnostic systems. More extensive datasets would likely demonstrate distinct scaling traits and potentially enhanced parallel efficiency owing to diminished relative synchronization overhead. Secondly, our assessment concentrated on classification problems; segmentation and detection applications may exhibit distinct computational characteristics [25].

Future studies must examine longitudinal factors such as model drift, retraining frequency necessities, and overall lifespan expenses. The examination of federated learning methodologies may facilitate the construction of multi-institutional models by utilizing distributed computational resources while safeguarding patient privacy across institutional borders. Furthermore, novel AI accelerator architectures and cloud service provisions may transform the cost-performance dynamics, necessitating regular reevaluation of deployment guidelines [26], [27].

V. Conclusion

This paper presents actual evidence that GPU-accelerated computing significantly enhances performance for medical imaging AI applications in settings available to Latin American academic and healthcare organizations. The Dell Precision 7960 workstation, equipped with two RTX 5000 Ada GPUs, attained an 11.1× acceleration in training performance and sub-10-millisecond inference latency relative to CPU-only options, while preserving 93.5% parallel efficiency and achieving an 88.8% reduction in energy consumption.

Economic analysis indicates that on-premise GPU workstations attain cost-effectiveness compared to cloud alternatives after roughly 2,150 hours of use. Institutions expecting ongoing AI development and implementation should invest in local infrastructure, as it offers advantageous long-term economics despite significant initial financial demands. Cloud-based solutions are suitable for fluctuating workloads, pilot projects, or organizations lacking access to capital funding.

All assessed configurations exhibited comparable diagnostic accuracy (91-92%), so affirming that computational acceleration techniques sustain model integrity. This discovery alleviates apprehensions about performance-accuracy trade-offs that could otherwise hinder HPC use in healthcare settings where diagnostic reliability is critical.

Our findings suggest that Latin American institutions may consider initiating using cloud-based GPU services for preliminary exploratory tasks and proof-of-concept validation. Institutions with ongoing computational demands could evaluate weighing local factors of dual-GPU workstations within the $15,000-20,000 range to achieve an ideal mix of performance and affordability. Collaborative agreements among regional institutions could facilitate shared access to advanced computational resources, optimizing utilization and dispersing expenses among various research initiatives. As AI-assisted medical diagnostics evolves towards standard clinical implementation, computational infrastructure will shift from a supplementary enhancement to a critical necessity, rendering strategic investment planning vital for institutional competitiveness in healthcare delivery and medical research.

Acknowledgment

This research was supported by the Institute for Research in Medical Sciences and Right to Health (ICIMEDES), the Faculty of Medical Sciences (FCM), and the Directorate of Scientific, Humanistic, and Technological Research (DICIHT) of the National Autonomous University of Honduras (UNAH).

References

[1] P. Hamet and J. Tremblay, “Artificial intelligence in medicine”, *Metabolism*, vol. 69, Suppl., pp. S36–S40, Apr. 2017, doi: 10.1016/j.metabol.2017.01.011.

[2] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare", *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719-731, Oct. 2018, doi: 10.1038/s41551-018-0305-z.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., "A survey on deep learning in medical image analysis", *Med. Image Anal.*, vol. 42, pp. 60-88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115-118, Feb. 2017, doi: 10.1038/nature21056.

[5] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine", *Nat. Med.*, vol. 28, no. 1, pp. 31-38, Jan. 2022, doi: 10.1038/s41591-021-01614-0.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., “An image is worth 16×16 words: Transformers for

image recognition at scale", *arXiv* preprint arXiv:2010.11929v2, Jun. 3, 2021. doi: 10.48550/arXiv.2010.11929.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS'17)*, Long Beach, CA, USA, 2017, pp. 4768-4777.
[9] Y. Liu, A. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, et al., "Detecting cancer metastases on gigapixel pathology images", *arXiv preprint arXiv:1703.02442*, Mar. 2017.
[10] NVIDIA Corporation, "NVIDIA RTX 5000 Ada Generation GPU", 2024. [Online]. Available: https://www.nvidia.com/en-us/design-visualization/rtx-5000/
[11] Linode LLC, "GPU Instances", Akamai Connected Cloud, 2024. [Online]. Available: https://www.linode.com/products/gpu/
[12] P. Otero, W. Hersh, and A. U. Jai Ganesh, "Big data: Are biomedical and health informatics training programs ready?," *Yearb. Med. Inform.*, vol. 23, no. 1, pp. 177–181, 2014, doi: 10.15265/IY-2014-0007.
[13] Linode LLC, "Linode Pricing", *Akamai Connected Cloud*, 2025. [Online]. Available: https://www.linode.com/pricing/
[14] Dell Technologies, "Dell Precision 7960 Tower Workstation", 2024. [Online]. Available: https://www.dell.com/en-us/shop/workstations/precision-7960-tower/spd/precision-7960-workstation
[15] PyTorch Foundation, "PyTorch 2.2.0 Release", 2024. [Online]. Available: https://pytorch.org/
[16] TensorFlow Developers, "TensorFlow 2.15.0", 2024. [Online]. Available: https://www.tensorflow.org/
[17] NVIDIA Corporation, "CUDA Toolkit 12.4", 2024. [Online]. Available: https://developer.nvidia.com/cuda-toolkit
[18] NVIDIA Corporation, "cuDNN 9.0", 2024. [Online]. Available: https://developer.nvidia.com/cudnn
[19] S. M. Smith, "Fast robust automated brain extraction", *Hum. Brain Mapp.*, vol. 17, no. 3, pp. 143-155, Nov. 2002, doi: 10.1002/hbm.10062.
[20] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI", *Z. Med. Phys.*, vol. 29, no. 2, pp. 102–127, 2019, doi: 10.1016/j.zemedi.2018.11.002.
[21] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling", *Curr. Oncol.,* vol. 29, no. 10, pp. 7498–7511, 2022, doi: 10.3390/curroncol29100590.
[22] K. Piliuk and S. Tomforde, "Artificial intelligence in emergency medicine. A systematic literature review", *Int. J. Med. Inform.*, vol. 180, Art. no. 105274, 2023, doi: 10.1016/j.ijmedinf.2023.105274.
[23] World Health Organization, "Global spending on health: Rising to the pandemic's challenges", Geneva, Switzerland, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240064911
[24] International Energy Agency, "Greenhouse Gas Emissions from Energy Highlights", IEA. [Online]. Available: https://www.iea.org/data-and-statistics
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.
[26] J. Xu et al., "Federated learning for healthcare informatics", *J. Healthc. Inform. Res.*, vol. 5, no. 1, pp. 1-19, Mar. 2021, doi: 10.1007/s41666-020-00082-4.
[27] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence", *Nat. Med.*, vol. 25, no. 1, pp. 44-56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.