Scalable and Efficient Deep Learning for Diabetic Retinopathy Classification on ARM-Based Architectures*

Thiago da Silva Araújo[1] **and Philippe O. A. Navaux**[1]

[1]Informatics Institute, Federal University of Rio Grande do Sul, Brazil

{tsaraujo, navaux}@inf.ufrgs.br

Abstract. *Diabetic retinopathy (DR) is a leading cause of preventable blindness and a public health problem in Brazil, where access to ophthalmological specialists is often limited. Automated DR classification using deep learning (DL) can support screening programs, but deployment requires accurate, efficient, and scalable models. This work evaluates model selection and distributed scalability of convolutional neural networks (CNNs) for binary DR classification on ARM-based systems, using datasets from Federal University of São Paulo (UNIFESP), Hospital de Clínicas de Porto Alegre (HCPA), and TeleOftalmo (TelessaúdeRS-UFRGS). Among 38 evaluated CNN architectures, MobileNet performed best, consuming 77% less energy, training 83% faster, and producing an 85% smaller model than InceptionV3, while improving AUC by 3%. A preliminary distributed scalability evaluation across two NVIDIA Grace Superchips using data parallelism achieved a 1.93× speedup and 96.5% scaling efficiency. These results indicate that ARM-based distributed training is a feasible, energy-efficient option for scalable DR screening in Brazilian healthcare.*

1. Introduction

Diabetic retinopathy (DR) is one of the leading causes of vision loss in adults worldwide, affecting millions of diabetic patients [Flaxman et al. 2017]. In Brazil, the Brazilian Unified Health System (SUS) faces significant barriers to providing timely ophthalmological care: excess demand, unnecessary referrals, and severe geographic disparities in specialist distribution. These factors cause dangerous delays between diagnosis and treatment, particularly in rural and low-resource areas [Massuda et al. 2018]

Convolutional neural networks (CNNs) have shown diagnostic accuracy above 90% in DR classification, making automated screening more feasible and easing the workload for specialists [Nagpal et al. 2022]. However, the variety of available architectures can make model selection challenging. Researchers need to weigh considerations such as energy consumption, model size, and training duration alongside predictive accuracy.

This work evaluates deep learning (DL) models for DR classification with an emphasis on efficiency and scalability. For model selection, we evaluate 38 CNN architectures for binary DR classification using the Federal University of São Paulo (UNIFESP) dataset, identifying models that balance accuracy, energy consumption, and computational efficiency. For the distributed scalability analysis, we combine all three Brazilian

*This work was supported by CAPES - Funding Code 001, , by CNPq/MCTI/FNDCT - Universal under grants 408755/2025-3, CIARS RITEs/FAPERGS and CI-IA FAPESP-MCTIC-CGI-BR projects. Some experiments in this work used the PCAD infrastructure, http://gppd-hpc.inf.ufrgs.br, at INF/UFRGS.

clinical datasets, UNIFESP, Hospital de Clínicas de Porto Alegre (HCPA), and TeleOftalmo (TelessaúdeRS-UFRGS) to provide a larger, more heterogeneous training set. We then investigate distributed training on the NVIDIA Grace Superchip using data parallelism, as a preliminary step toward scalable DR screening pipelines.

Our results show that MobileNet is the best candidate among the 38 evaluated architectures, consuming 77% less energy, training 83% faster, and producing an 85% smaller model than InceptionV3, while improving AUC by 3%. The distributed training evaluation on two Grace Superchips achieved a 1.93× speedup and 96.5% scaling efficiency, halving training time while increasing energy consumption by less than 1%.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the dataset and the experimental methodology, followed by the results in Section 4. Section 5 concludes the paper with directions for future work.

2. Related Work

Prior studies on DL for DR classification have primarily focused on diagnostic accuracy across a limited set of architectures. Kallel and Echtioui (2024) evaluated EfficientNet variants, both using accuracy as the sole metric. Saproo et al. (2024) and Abushawish et al. (2024) extended this to 20 and 26 pre-trained models, respectively, yet still measured only diagnostic performance. Dos Reis et al. (2024) evaluated InceptionV3 for DR diagnosis in the Brazilian population, obtaining strong diagnostic accuracy and motivating its use as a baseline for Brazilian clinical datasets. To the best of our knowledge, no work has jointly evaluated energy consumption, model size, and training time alongside diagnostic accuracy for DR classification on ARM-based systems.

On the hardware side, recent work has characterised the NVIDIA Grace Superchip for High Performance Computing (HPC) workloads. Banchelli et al. (2024) reported speedups of 1.3×–4.28× against Intel Skylake across five scientific applications, and Ruhela et al. (2024) demonstrated competitive performance with lower power draw than x86 systems. However, neither study extends to DL workloads.

In their research, Patil and Chickerur (2023) compared data and model parallelism for DR classification using a DGX-1 system. Meanwhile, Shanthala and Kundur (2025) introduced DR-EfficientNet-L, which incorporates gradient compression across 16 nodes,

Table 1. Representative studies and their evaluation focus.

Focus	Studies	Evaluation Criteria
CNN benchmarking for DR	[Kallel and Echtioui 2024, Arora et al. 2024, Saproo et al. 2024, Abushawish et al. 2024]	Accuracy, AUC
InceptionV3 for Brazilian population	[Dos Reis et al. 2024]	Accuracy, diagnostic performance
ARM-based HPC platforms	[Banchelli et al. 2024, Ruhela et al. 2024]	Performance, energy efficiency
Distributed DL training	[Patil and Chickerur 2023, Shanthala and Kundur 2025]	Speedup, scalability
This work	38 CNNs on ARM Grace, Brazilian datasets	Accuracy, energy, model size, distributed scalability

resulting in a remarkable 14.8× speedup. Both studies concentrate on algorithmic strategies but do not address energy efficiency or the performance of ARM-based platforms.

As summarised in Table 1, no prior work combines multi-criteria CNN benchmarking with distributed training on ARM for DR using Brazilian clinical data. This work addresses that gap.

3. Methodology

This work follows a two-phase experimental methodology to evaluate deep learning models for DR classification, focusing on predictive performance, computational efficiency, and distributed scalability on ARM-based platforms.

3.1. Datasets and Preprocessing

Phase 1 used the UNIFESP dataset [Nakayama et al. 2023], comprising 2,579 retinal fundus photographs from 1,252 diabetic patients, with 550 images (21.3%) labelled as referable DR. A 70/30 stratified split was applied at the patient level to prevent data leakage. Phase 2 combined all three Brazilian clinical datasets, UNIFESP, HCPA, and TeleOftalmo (TelessaúdeRS-UFRGS) [Lutz de Araujo et al. 2020], to increase training volume and heterogeneity across disease severities and acquisition protocols. All images were resized, center-cropped, and converted to TFRecord format. Data augmentation included random rotations and horizontal flips. Full dataset details, preprocessing scripts, and energy monitoring tools are available at Github[1].

3.2. Model Architecture

All 38 architectures shared a common classification head to ensure a fair comparison. Each model receives a 299×299 RGB retinal fundus image as input. The convolutional feature extractor is initialised with ImageNet-pretrained weights; its layers are frozen during an initial fine-tuning pass, after which all layers are set as trainable. A Global Average Pooling layer reduces the spatial feature maps to a fixed-length vector, and a single Dense neuron with a configurable activation function produces the binary output (referable vs. non-referable DR).

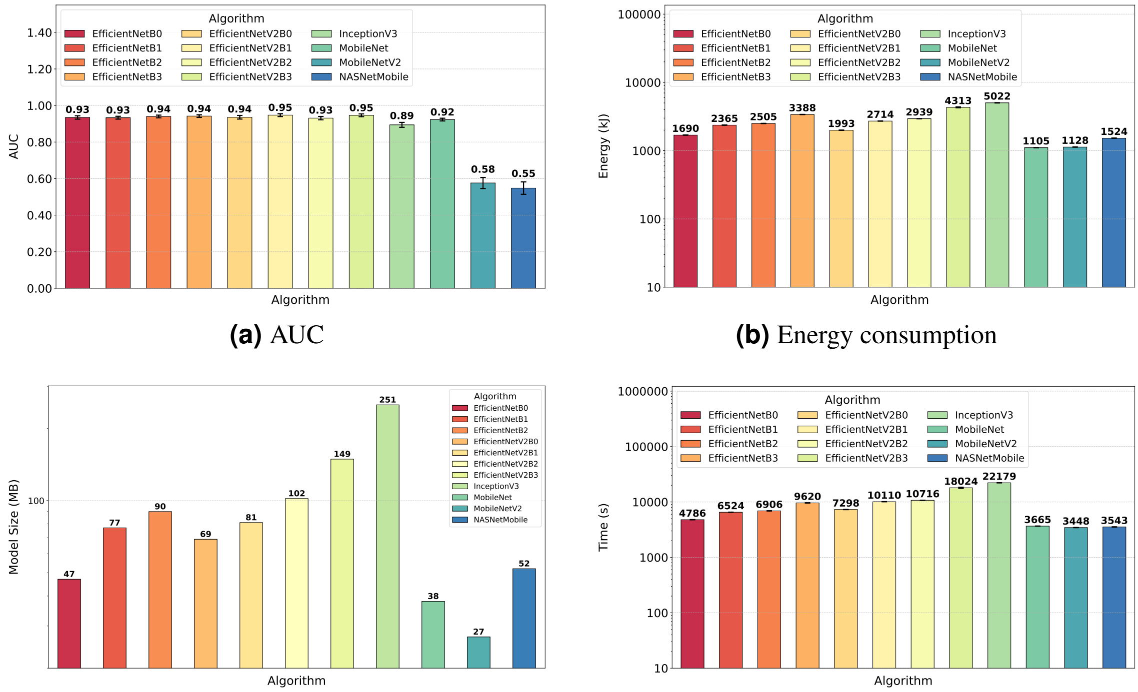
3.3. Phase 1: Model Selection

We evaluated 38 CNN architectures using transfer learning with ImageNet-pretrained weights and a consistent training protocol. Experiment 1 filtered models over five epochs by energy consumption, model size, and training time relative to the InceptionV3 baseline [Dos Reis et al. 2024]. Candidates that outperformed InceptionV3 on all three metrics advanced to Experiment 2, where models trained for 200 epochs and were also assessed on AUC. Models were compared across four criteria: AUC, training time, model size, and energy consumption.

3.4. Phase 2: Distributed Scalability on ARM-Based Systems

Phase 2 evaluated the scalability of the selected model on the NVIDIA Grace Superchip, a dual-socket ARM Neoverse V2 system with 144 cores and 480 GB of LPDDR5X memory. Multi-node experiments used TensorFlow's `MultiWorkerMirroredStrategy`

[1] `https://github.com/tsilvaaraujo/sbac-pad2025`.



(a) AUC

(b) Energy consumption

(c) Model size

(d) Training time

Figure 1. Performance and efficiency metrics over 10 runs for models outperforming the baseline.

for synchronous data-parallel training across two Grace nodes, measuring speedup, scaling efficiency, and energy overhead. Energy was measured via hardware-level power telemetry sampled at 50 ms intervals. All experiments were repeated ten times; results report mean and standard deviation.

4. Results

This section summarizes the main experimental findings of this work, focusing on model selection and distributed scalability, which represent the primary contributions of the dissertation.

4.1. Model Selection and Efficiency Analysis

We evaluated 38 convolutional neural network (CNN) architectures for diabetic retinopathy classification, considering classification performance (AUC), training time, model size, and energy consumption. Figure 1 presents the final comparison of the most competitive models.

MobileNet achieved a strong classification performance (AUC of 0.93) while significantly reducing computational cost. Compared to the widely used InceptionV3 baseline, MobileNet reduced energy consumption by 77%, training time by 83%, and model size by 85%, while improving AUC by approximately 3%. Although EfficientNet variants achieved slightly higher AUC values, they required substantially higher computational and energy resources, making them less suitable for resource-constrained healthcare environments. These results demonstrate that lightweight CNNs can provide clinically competitive accuracy while enabling sustainable deployment in large-scale screening systems.

Table 2. Comparison of single-node and two-node training results using the MultiWorkerMirroredStrategy.

Number of nodes	Mean Energy (kJ)	Mean AUC	Mean Time (s)
1	6,886.00 $\pm$ 55.21	0.937 $\pm$ 0.003	21,544.26 $\pm$ 153.86
2	6,954.11 $\pm$ 85.22	0.936 $\pm$ 0.004	11,172.74 $\pm$ 203.51

4.2. Distributed Training Scalability on ARM Systems

To evaluate scalability, we conducted distributed training experiments on two NVIDIA Grace ARM-based nodes. Table 2 summarizes the results.

The two-node configuration achieved a $1.93\times$ speedup compared to single-node execution, with a scaling efficiency of 96.5%. Total energy consumption increased by less than 1%, while training time was nearly halved, resulting in improved energy efficiency per training epoch. Model accuracy remained stable across configurations (AUC $\approx$ 0.936), indicating that distributed synchronization did not affect convergence. These results demonstrate the feasibility of scalable and energy-efficient distributed deep learning for medical imaging on ARM-based platforms. However, the evaluation is limited to two nodes due to infrastructure constraints, and future work will explore larger-scale deployments to assess communication bottlenecks and scalability limits.

5. Conclusion

This work presented a systematic evaluation of deep learning models for diabetic retinopathy classification, with an emphasis on computational efficiency and scalability for healthcare deployment. By benchmarking 38 CNN architectures, we demonstrated that lightweight models such as MobileNet can achieve competitive diagnostic performance while significantly reducing energy consumption, training time, and model size compared to commonly used architectures. In addition, we evaluated the scalability of the selected model on ARM-based NVIDIA Grace systems. Distributed training across two nodes achieved near-linear speedup with minimal energy overhead and stable classification performance, indicating that ARM-based platforms are a promising solution for scalable and sustainable medical imaging workloads.

A limitation of this study is that the distributed scalability analysis was conducted using only two nodes due to infrastructure constraints. Although the results indicate high scaling efficiency, larger-scale experiments are required to evaluate communication overheads, network contention, and scalability limits in multi-node clusters. Future work will extend the evaluation to larger ARM clusters and explore deployment in real-world screening pipelines.

References

Abushawish, I. Y., Modak, S., Abdel-Raheem, E., Mahmoud, S. A., and Hussain, A. J. (2024). Deep learning in automatic diabetic retinopathy detection and grading systems: a comprehensive survey and comparison of methods. *IEEE Access*, 12:84785–84802.

Arora, L., Singh, S. K., Kumar, S., Gupta, H., Alhalabi, W., Arya, V., Bansal, S., Chui, K. T., and Gupta, B. B. (2024). Ensemble deep learning and efficientnet for accurate diagnosis of diabetic retinopathy. *Scientific Reports*, 14(1):30554.

Banchelli, F., Vinyals-Ylla-Catala, J., Pocurull, J., Clascà, M., Peiro, K., Spiga, F., Garcia-Gasulla, M., and Mantovani, F. (2024). Nvidia grace superchip early evaluation for hpc applications. In *Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region Workshops*, pages 45–54.

Dos Reis, M. A., Künas, C. A., et al. (2024). Advancing healthcare with artificial intelligence: diagnostic accuracy of machine learning algorithm in diagnosis of diabetic retinopathy in the brazilian population. *Diabetology & Metabolic Syndrome*, 16(1):209.

Flaxman, S. R., Bourne, R. R., Resnikoff, S., et al. (2017). Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *The Lancet Global Health*, 5(12):e1221–e1234.

Kallel, F. and Echtioui, A. (2024). Retinal fundus image classification for diabetic retinopathy using transfer learning technique. *Signal, image and video processing*, 18(2):1143–1153.

Lutz de Araujo, A., Moreira, T. d. C., Varvaki Rados, D. R., Gross, P. B., Molina-Bastos, C. G., Katz, N., Hauser, L., Souza da Silva, R., Gadenz, S. D., Dal Moro, R. G., et al. (2020). The use of telemedicine to support brazilian primary care physicians in managing eye conditions: The teleoftalmo project. *PloS one*, 15(4):e0231034.

Massuda, A., Hone, T., et al. (2018). The brazilian health system at crossroads: progress, crisis and resilience. *BMJ Global Health*, 3(4):e000829.

Nagpal, D., Panda, S. N., Malarvel, M., Pattanaik, P. A., and Khan, M. Z. (2022). A review of diabetic retinopathy: Datasets, approaches, evaluation metrics and future trends. *Journal of King Saud University-Computer and Information Sciences*, 34(9):7138–7152.

Nakayama, L. F., Goncalves, M., Zago Ribeiro, L., Santos, H., Ferraz, D., Malerbi, F., Celi, L. A., and Regatieri, C. (2023). A brazilian multilabel ophthalmological dataset (brset). *PhysioNet https://doi.org/10.13026/xcxw-8198*, 13026.

Patil, M. S. and Chickerur, S. (2023). Study of data and model parallelism in distributed deep learning for diabetic retinopathy classification. *Procedia Computer Science*, 218:2253–2263.

Ruhela, A., Cazes, J., McCalpin, J., del Castillo-Negrete, C., Li, J., Liu, H., Chen, H., Lu, C.-Y., Milfeld, K., Zhang, W., et al. (2024). Performance analysis of scientific applications on an nvidia grace system. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 558–566. IEEE.

Saproo, D., Mahajan, A. N., and Narwal, S. (2024). Deep learning based binary classification of diabetic retinopathy images using transfer learning approach. *Journal of Diabetes & Metabolic Disorders*, 23(2):2289–2314.

Shanthala, K. and Kundur, N. C. (2025). Dr-efficientnet-l: A distributed deep learning architecture for efficient detection and grading of diabetic retinopathy. *Engineering, Technology & Applied Science Research*, 15(5):28362–28367.