A Systematic Review of Post-Training Optimization for Embedded Deep Learning to Support Biodiversity Monitoring in the Pantanal

Isis Milena Daron[1], Julia Milioranza Gomes[1], Thiago Meirelles Ventura[1], Lucas Korzune[2], Allan Gonçalves de Oliveira[1]

[1]Universidade Federal de Mato Grosso

[2]Instituto Nacional de Pesquisas do Pantanal

{isis.daron, julia.gomes1}@sou.ufmt.br, {thiago, allan}@ic.ufmt.br,

lucaskorzune@gmail.com

Abstract. *Deploying deep learning models in remote ecosystems like the Pantanal requires autonomous on-device processing, which is challenging because embedded devices demand lightweight architectures that most current models fail to provide. This systematic review investigated post-training optimization techniques for embedded devices across 1,036 studies, synthesizing 32 that met quality criteria. Quantization was most prevalent (62.5%), followed by knowledge distillation (50.0%) and pruning (46.9%), with 43.8% using hybrid approaches. Results showed substantial size reductions with accuracy improvements. These findings support autonomous acoustic monitoring systems for conservation in infrastructure-limited regions.*

1. Introduction

Biodiversity plays a fundamental role as an ecosystem service by maintaining natural ecological processes [Alho and Sabino 2011], particularly in unique biomes such as the Pantanal. As the largest tropical wetland in the world, the Pantanal is under serious threat from climate change, agricultural expansion, and other anthropogenic pressures that affect the biodiversity and ecological dynamics of this environment [Justino et al. 2025]. Given these threats, conservation of the Pantanal is fundamental to maintaining ecological balance and the ecosystem services that depend on its resources [Justino et al. 2025].

Acoustic monitoring of birds has emerged as an effective tool for biodiversity assessment, with deep learning models enabling automated species identification from vocalizations [Kahl et al. 2021]. In the Pantanal, initiatives have demonstrated the potential of these approaches for monitoring regional fauna [Ventura et al. 2020]. However, deploying such models in remote regions remains challenging, as deep neural networks typically require substantial computational and memory resources [Zhao et al. 2022], while areas like the Pantanal lack reliable connectivity for cloud-based processing. This highlights the need for local, autonomous processing on resource-constrained devices.

Addressing this deployment challenge requires understanding post-training optimization techniques applicable to deep learning models across diverse domains. This systematic literature review synthesizes the current state of post-training optimization techniques for deep learning models on embedded devices. By identifying prevalent techniques, effectiveness measures, evaluation metrics, and deployment platforms, this review establishes the foundation for developing optimized systems capable of autonomous operation in resource-constrained environments where connectivity cannot be assumed.

2. Methodology

The review was conducted in accordance with systematic literature review guidelines [Kitchenham 2004], following a predefined protocol encompassing research question formulation, search strategy, selection criteria, quality assessment, and data extraction, supported by the Parsifal platform.

The protocol was structured according to the PICOC framework [Wohlin et al. 2012]. The population (P) consists of deep learning models. The intervention (I) includes post-training optimization techniques, specifically quantization and knowledge distillation. The comparison (C) is performed against baseline non-optimized models. The outcomes (O) include improvements in accuracy, inference time, memory consumption, and computational efficiency. The context (C) corresponds to embedded systems and edge computing platforms. Based on this framework, the following research questions (RQs) were defined: RQ01 – Which post-training optimization techniques are most commonly applied to deep learning models? RQ02 – How effective are these techniques in improving performance on embedded or resource-constrained devices? RQ03 – Which performance metrics are most frequently used to evaluate optimization effectiveness? RQ04 – On which types of embedded platforms have these techniques been tested and validated?

Inclusion and exclusion criteria were defined a priori and applied during study selection. Studies were included if published from 2023 onwards, available in full text, and presenting post-training optimization techniques for deep learning models. Excluded were non-English papers, duplicates, out-of-scope studies, reviews, purely theoretical works, studies lacking methodological transparency, and those addressing only lightweight architectures, which are designed from the ground up for efficiency rather than through post-training optimization [Li et al. 2024].

Quality assessment followed five criteria: (QA01) clear description of the optimization technique; (QA02) experimental comparison with baseline models; (QA03) evaluation on embedded or resource-constrained platforms; (QA04) reporting of performance metrics; and (QA05) availability of models, code, or artifacts. Each criterion was scored as Yes (1), Partially (0.5), or No (0), with a maximum score of five points, and only studies scoring above four were included. The search strategy relied on a predefined query string *("deep learning") AND (quantization OR distillation) AND ("embedded devices")*, which was consistently applied across IEEE Xplore, Scopus, and ScienceDirect.

Concerning the threats to validity inherent to the review process, certain limitations must be acknowledged. The search string was intentionally restricted to a limited set of terms, which may have excluded relevant studies employing alternative terminology such as model compression or TinyML, potentially reducing the breadth of the final corpus. Additionally, the temporal restriction to studies published from 2023 onward, while ensuring relevance to current ecosystems, excludes seminal works that might better contextualise the evolution of the techniques analysed.

3. Results and discussion

A total of 1,036 studies were retrieved, of which 32 met temporal, format, scope, and quality criteria and were included in the final synthesis (Figure 1).

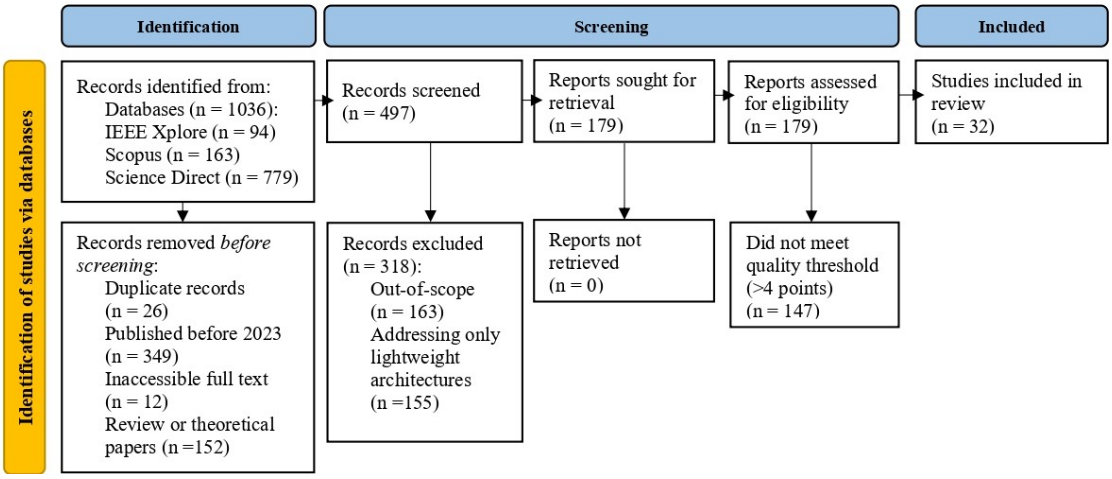

Figure 1. PRISMA flow diagram (adapted from [Haddaway et al. 2022]).

The systematic analysis of these 32 studies revealed quantization as the most prevalent post-training optimization technique, applied in 20 studies (62.5%), followed by knowledge distillation in 16 studies (50.0%) and pruning in 15 studies (46.9%). While the search string targeted quantization and knowledge distillation specifically, pruning frequently appeared as a complementary technique in the retrieved studies, reflecting its common integration with the primary techniques of interest. All three techniques demonstrated substantial effectiveness across diverse applications and hardware platforms, with quantization proving particularly effective for severely resource-constrained devices, knowledge distillation excelling in accuracy preservation during aggressive compression, and pruning showing strong results predominantly within hybrid approaches. Notably, 14 studies (43.8%) employed hybrid approaches combining multiple techniques, reflecting the field's evolution toward multi-technique strategies.

These effectiveness patterns manifest across diverse deployment scenarios. [Gragnaniello et al. 2025] achieved 70% size reduction (1.2 MB to 347 kB) with 89.52% accuracy maintained through quantization for wearable diabetes detection. [He et al. 2023] demonstrated a 170-fold size reduction while preserving 94.37% accuracy for RV reducer fault diagnosis. [Gutiérrez-Zaballa et al. 2024] combined pruning and quantization for semantic segmentation, achieving 99.74% size reduction (118.77 MB to 0.31 MB) with merely 0.28% IoU degradation.

Hybrid approaches deserve further attention, as they produced some of the most compelling results in the reviewed literature. [Zhao et al. 2025] demonstrated the synergistic potential of combining techniques: their IDD-YOLOv7 system integrated knowledge distillation and pruning, achieving a 69% model size reduction (18.1 MB to 5.6 MB) while improving performance from 83.9% to 92.8% mAP. Similarly, [Huang et al. 2023] combined quantization, pruning, and knowledge distillation in a CycleGAN model for infrared video colorization, reducing parameters by 98.8% (11.38M to 0.137M). These results suggest that the techniques address complementary aspects of efficiency: quantization optimizes weight representation, pruning removes structural redundancy, and knowledge distillation preserves learned representations, highlighting the importance of hybrid strategies for resource-constrained deployments.

Quantitative analysis across all studies revealed substantial optimization effectiveness. Model size reduction averaged 78.2%, ranging from 29.4% to 99.86%. Overall performance slightly improved after optimization, with a mean increase of 0.8%. Importantly, 13 studies (40.6%) reported metric gains post-optimization, indicating that compression can positively affect model effectiveness. Inference time improvements were also significant: [He et al. 2024] achieved 66.6% latency reduction for tilapia segmentation on Jetson Orin Nano, while [Huang et al. 2023] demonstrated 84.8% reduction enabling real-time video processing at nearly 60 fps.

Performance evaluation showed clear metric preferences reflecting the multidimensional nature of embedded optimization. Model size was nearly universal (29 studies, 90.6%), followed by inference time (27 studies, 84.4%) and number of parameters (25 studies, 78.1%). Task-specific accuracy metrics appeared in 21 studies (65.6%), with precision in 17 studies (53.1%) and recall in 16 studies (50.0%). Computational efficiency metrics included FPS (16 studies, 50.0%) and GFlops (12 studies, 37.5%).

Platform analysis revealed NVIDIA Jetson family dominance, with Jetson Nano appearing in 7 studies, Jetson Orin Nano in 5 studies, and Jetson Xavier NX in 4 studies, reflecting the platform's balance between GPU acceleration, power efficiency, and mature deep learning ecosystem. FPGA platforms appeared in 3 studies for hardware acceleration, microcontroller-class devices in 3 studies for extremely resource-constrained environments, and smartphone-based devices in 4 studies for mobile edge deployment.

4. Conclusion

This systematic literature review examined post-training optimization techniques for deploying deep learning models on embedded and resource-constrained devices, motivated by autonomous biodiversity monitoring in ecosystems such as the Brazilian Pantanal.

Quantization, knowledge distillation, and pruning were the most used methods, reducing model size, computational cost, and inference latency while preserving accuracy. Individually, quantization proved particularly suitable for highly constrained hardware, knowledge distillation helped retain performance under strong compression, and pruning was frequently combined with other methods. NVIDIA Jetson platforms were the most common deployment targets due to their balance between performance and energy efficiency, while microcontrollers and FPGA accelerators demonstrated feasibility in highly constrained environments. These results support the development of on-device acoustic monitoring systems capable of real-time species recognition without cloud dependence.

Future work will apply and evaluate these optimization strategies in deep learning models for Pantanal bird species recognition, comparing optimized and baseline models in terms of accuracy, inference time, and model size reduction to assess their effectiveness for autonomous acoustic monitoring in connectivity-limited environments.

5. Acknowledgments

We thank the Computational Bioacoustics Research Unit (CO.BRA) at the Federal University of Mato Grosso for their support and guidance throughout this research.

References

Alho, C. J. R. and Sabino, J. (2011). A conservation agenda for the pantanal's biodiversity. *Brazilian Journal of Biology*, 71:327–335.

Gragnaniello, M., Marrazzo, V. R., Borghese, A., Maresca, L., Breglio, G., and Riccio, M. (2025). Edge-ai enabled wearable device for non-invasive type 1 diabetes detection using ecg signals. *Bioengineering*, 12.

Gutiérrez-Zaballa, J., Basterretxea, K., and Echanobe, J. (2024). Evaluating single event upsets in deep neural networks for semantic segmentation: An embedded system perspective. *Journal of Systems Architecture*, 154.

Haddaway, N. R., Page, M. J., Pritchard, C. C., and McGuinness, L. A. (2022). Prisma2020 : An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, 18(2).

He, F., Liu, C., Wang, M., Yang, E., and Liu, X. (2023). Network lightweight method based on knowledge distillation is applied to rv reducer fault diagnosis. *Measurement Science and Technology*, 34.

He, Y., Xu, W., Liu, D., Qu, M., Fu, J., and Cheng, F. (2024). Real-time tilapia fillet defect segmentation on edge device for robotic trimming. *Computers and Electronics in Agriculture*, 225.

Huang, R., Wang, H., Huang, X., Tai, Y., Cheng, F., and Shi, J. (2023). Edge device-based real-time implementation of cyclegan for the colorization of infrared video. *Future Generation Computer Systems*, 149:402–415.

Justino, S. T. P., Silva, R. B., Guerrini, I. A., da Silva, R. B. G., and Simões, D. (2025). Monitoring environmental degradation and spatial changes in vegetation and water resources in the brazilian pantanal. *Sustainability*, 17(1):51.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele Univ.*, 33.

Li, J., Li, X., Ju, M., Zhao, X., Wang, Y., and Yang, F. (2024). Lightweight implementation of the signal enhancement model for early wood-boring pest monitoring. *Forests*, 15.

Ventura, T. M., de Oliveira, A. G., Ganchev, T. D., Marques, M. I., and Schuchmann, K.-L. (2020). Inteligência artificial a serviço da biodiversidade do pantanal. *Computação Brasil*, pages 23–26.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer Berlin Heidelberg.

Zhao, T., Xie, Y., Wang, Y., Cheng, J., Guo, X., Hu, B., and Chen, Y. (2022). A survey of deep learning on mobile devices: Applications, optimizations, challenges, and research opportunities. *Proceedings of the IEEE*, 110:334–354.

Zhao, Y., Zhang, G., Luo, W., Tang, R., Sun, Y., Wang, P., Liu, J., and Mei, K. (2025). Idd-yolov7: A lightweight and efficient feature extraction method for insulator defect detection. *Energy Reports*, 13:1467–1487.