Efficiency-First Bioacoustic Audio Event Detection Under Resource-Constrained Systems *

André M. S. Magalhães[1], Willian D. de Oliveira[1], Cesar A. P. Garbossa[2], Ricardo V. Ventura[2], Elaine P. M. de Sousa[1]

[1] Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos – SP – Brazil

andre.moreira.souza@usp.br, {willian,parros}@icmc.usp.br

[2]Faculdade de Medicina Veterinária e Zootecnia (FMVZ)
Universidade de São Paulo (USP) – Pirassununga – SP – Brazil

{cgarbossa,rvventura}@usp.br

Abstract. *Passive acoustic monitoring enables wildlife biodiversity assessment and precision livestock management, yet accurate detectors remain difficult to deploy on resource-limited systems. We present a controlled evaluation of* AST*,* SSAMBA*, and* MAMBASPEC *on two IoT-relevant bioacoustic datasets (aSwine, AnuraSet).* MAMBASPEC *matches* AST *within 0.7 pp ROC-AUC$_w$ on AnuraSet with* $80\times$ *fewer parameters and* $1{,}617\times$ *fewer FLOPs, while* SSAMBA *achieves statistically indistinguishable ROC-AUC$_w$ at a* $12.8\times$ *smaller footprint. Friedman tests ($p<10^{-3}$) and Pareto analysis indicate that* MAMBASPEC *is the efficiency-first choice under resource constraints, whereas* SSAMBA *is the accuracy-oriented choice when modest extra latency is acceptable.*

1. Introduction

Passive acoustic monitoring (PAM) on low-power Internet of Things (IoT) nodes is increasingly used for biodiversity assessment, precision livestock management, and broader environmental monitoring tasks [Stowell 2022, Vuilliomenet et al. 2026]. A typical edge node continuously captures audio at 8–22 kHz, buffers short segments (1–3 s), and must classify each in near-real time within constrained memory and compute budgets. However, the prevailing deep learning approach, Audio Spectrogram Transformers (AST) (and variants), employs Vision Transformer (ViT)-Base backbones pretrained on ImageNet and AudioSet, resulting in 86.9 M parameters and 8.77 GFLOPs per inference for 3-second clips [Gong et al. 2021]. Both the model size and activation memory of such architectures exceed the on-chip storage and SRAM of typical microcontroller-class IoT nodes by two to three orders of magnitude [Lin et al. 2020].

Recent State Space Models (SSMs), particularly MAMBA [Gu and Dao 2023], match Transformer accuracy at $\mathcal{O}(N)$ time and space complexity via hardware-aware selective scans, and SSAMBA [Shams et al. 2024] extends this to audio via self-supervised

*Acknowledgments: This research was supported by FAPESP (grants 2016/17078-0 and 2020/07200-9) and CAPES (grants PROEX-9778985/M and 88887.893807/2023-00). AI Usage: Grammatical revision of this manuscript was assisted by Grammarly; all content was written and verified by the authors.

AudioSet pretraining. However, no prior work has systematically evaluated spectrogram-patch SSMs against AST on IoT-deployed bioacoustic datasets with rigorous Hyperparameter Optimization (HPO) and inference-efficiency analysis. This is fundamentally a data-management problem: the detector acts as a stream operator that performs data reduction at the source, filtering high-rate raw audio into sparse event records before they enter downstream storage and event-detection pipelines, so model efficiency sets the ingestion rate an edge node can sustain [Lima et al. 2022]. Therefore, the following research question arises: Can SSMs achieve accuracy comparable to AST for bioacoustic Audio Event Detection (AED) while enabling deployment on IoT devices through reduced parameter count and inference latency?

Our contribution is methodological rather than a new architecture: we define a reproducible, efficiency-first evaluation protocol for selecting bioacoustic detectors under edge data-management constraints, and instantiate it on two datasets. The protocol couples (i) controlled HPO under a shared search space (Optuna Tree-structured Parzen Estimator (TPE), 30 trials per model) with multi-seed evaluation (5 seeds) for robust estimates; (ii) Friedman and Nemenyi statistical validation of the observed differences; and (iii) a Pareto-efficiency analysis over accuracy, parameter count, and GFLOPs that maps hardware constraints to architecture choice. Applied to *AnuraSet* (wildlife monitoring, 42 classes) and *aSwine* (precision livestock farming, 7 classes), it yields concrete, statistically grounded deployment guidance for IoT practitioners.

The rest of this paper covers related work (Section 2), methods (Section 3), results (Section 4), and conclusions (Section 5). [1]

2. Background and Related Work

The Audio Spectrogram Transformer (AST) [Gong et al. 2021] recasts audio classification as a vision task, converting raw waveforms to log-mel spectrograms and feeding 16×16 patches to a ViT encoder pretrained on ImageNet. Fine-tuned on AudioSet, AST achieves 0.459 mAP on the AudioSet evaluation set and has been widely adopted for ecoacoustic tasks. Its main limitation is 86.9 M trainable parameters and quadratic self-attention cost.

In bioacoustic AED, deep learning has become the dominant paradigm across ecological and agricultural contexts. A comprehensive review of the field [Stowell 2022] identifies CNNs and Transformer-based architectures as the primary approaches, highlighting data scarcity and domain shift as persistent challenges for field-deployable systems. To reduce evaluation fragmentation, standardized benchmarks such as BEANS [Hagiwara et al. 2023], BirdSet [Rauch et al. 2025], and the few-shot Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Task 5 [Liang et al. 2024] consolidate dozens of bioacoustic datasets and tasks, yet all emphasize server-side accuracy over edge efficiency. Despite this progress, evaluations are predominantly conducted on server-grade hardware; the efficiency-first perspective required for IoT edge deployment remains underexplored in the bioacoustic AED literature.

MAMBA [Gu and Dao 2023] is a State Space Model (SSM) with hardware-aware selective scans that scales linearly with sequence length. SSAMBA [Shams et al. 2024] adapts it to audio via masked spectrogram modeling on AudioSet, yielding 6.8 M parameters, 0.19 GFLOPs, and a $12.8\times$ smaller footprint than AST.

[1] Code publicly available at `https://github.com/andremsouza/ssm-bioaed-iot`.

BioMamba [Tang and Baskiyar 2025] (unpublished) is a Mamba2-based audio representation model pretrained via self-supervised learning and evaluated on the BEANS benchmark [Hagiwara et al. 2023]; it reports accuracy comparable to the Transformer-based Animal Vocalization Encoder based on Self-Supervision (AVES) model with substantially lower VRAM consumption. That study assesses neither latency, FLOPs, nor parameter budgets under IoT constraints, nor performs rigorous HPO.

3. Methods

Our evaluation protocol comprises two bioacoustic datasets representing distinct IoT sensing scenarios, three architectures spanning a wide range of efficiency and accuracy, and a hyperparameter optimization procedure with evaluation metrics for robust, fair comparisons. All inputs are log-mel spectrograms uniformly partitioned into 16×16 patches across the three models.

3.1. Datasets

AnuraSet [Cañas et al. 2023] comprises 93,378 3-second segments (27 h) of multi-label bioacoustic data spanning 42 Neotropical anuran species, recorded in situ by IoT acoustic sensors. We use the predefined train/test split, further divided 85/15 for train/val (stratified).

aSwine [Souza et al. 2025] contains 7 classes of weakly labeled audio covering animal events (e.g., stress, coughing/sneezing) and human activities (e.g., cleaning, speech, feeding), collected via a fixed surveillance infrastructure in a non-commercial swine barn.[2] We use the pruned 1-second variant (boundary-invalid segments removed), with the predefined 80/20 train/test split further divided 85/15 for train/val (stratified).

Both datasets use 64-band log-mel spectrograms (10 ms hop, 25 ms window): *AnuraSet* at 22 050 Hz (shape 64×301), *aSwine* at 16 000 Hz (shape 64×101); all three models process the spectrogram as 16×16 patches.

3.2. Models

AST [Gong et al. 2021] is a DeiT-base encoder (86.9 M parameters) that reframes audio classification as a vision task. Raw waveforms are converted to 128-band log-mel spectrograms (25 ms Hamming window, 10 ms hop) and divided into 16×16 patches with a stride of 10 in both axes (6-pixel overlap); the patches are fed to a Transformer pretrained on ImageNet and AudioSet, then fine-tuned on the target dataset. At inference, the outputs of the CLS and distillation tokens are averaged before the classification head.

SSAMBA [Shams et al. 2024] replaces the Transformer encoder with a bidirectional MAMBA encoder (6.8 M parameters, Tiny variant) pretrained on AudioSet via masked spectrogram patch modeling. The full model is fine-tuned end-to-end on each target dataset; predictions are obtained by mean-pooling over all non-CLS output tokens.

MAMBASPEC is a compact MAMBA encoder ($\approx$1.1 M parameters) we train from randomly initialized weights. Unlike existing pretrained SSM audio encoders such as SSAMBA [Shams et al. 2024] and Audio Mamba [Erol et al. 2024], MAMBASPEC carries no pretraining and its architecture (depth, embedding dimension, SSM state dimension, and patch size) is searched jointly with training via HPO, yielding depth = 6, embed = 128 on *AnuraSet* and depth = 8, embed = 192 on *aSwine*.

[2]The *aSwine* dataset is publicly available at `https://github.com/andremsouza/aswine`.

3.3. Hyperparameter Optimization

We use Optuna with the TPE sampler (30 trials per model×dataset, 180 total). The shared search space covers learning rate (log-uniform; 10^{-5}–10^{-2} for randomly initialized, 10^{-5}–5×10^{-4} for pretrained), weight decay (10^{-5}–10^{-1}), scheduler type ({cosine, step}), positive-class weighting ({True, False}), and augmentation parameters. For MAMBASPEC, architecture hyperparameters are additionally searched: depth $\in \{2, 4, 6, 8\}$, embedding dimension $\in \{64, 128, 192, 256\}$, SSM state dimension $\in \{8, 16, 32, 64\}$, and patch size $\in \{8, 16, 32\}$. Early stopping (patience = 10 epochs) is applied during HPO; best hyperparameters are then evaluated across five seeds ($\{0, 1, 2, 3, 4\}$). All models use AdamW with cosine scheduling (the HPO-selected scheduler in all runs), mixed-precision BF16, and up to 100 epochs, on an NVIDIA RTX 5070 Ti.

3.4. Metrics and Statistical Tests

Predictive performance. mAP and weighted ROC-AUC (ROC-AUC$_w$) are computed on the held-out test set; weighted averaging is used throughout to account for class imbalance in both datasets.

Efficiency. Inference throughput (samples/s), latency (ms), and FLOPs (via `ptflops`) are profiled at batch = 1 using the *AnuraSet* input shape as an IoT-scale proxy. Parameter count and GFLOPs are the primary hardware-agnostic efficiency axes for deployment planning [Lin et al. 2020].

Statistical significance. Following the standard protocol for comparing classifiers across multiple datasets [Demšar 2006], we apply the Friedman omnibus test across 5 seeds × 2 datasets (10 blocks, 3 models), with the Nemenyi post-hoc test for pairwise comparisons. Pareto frontiers are computed over (ROC-AUC$_w$, #params) and (ROC-AUC$_w$, GFLOPs) pairs per dataset.

4. Results and Discussion

4.1. Accuracy

Table 1 reports mAP and ROC-AUC$_w$ for all three models across both datasets (mean ± std over 5 seeds). Self-supervised pretraining on AudioSet transfers positively to in-domain bioacoustic tasks: SSAMBA surpasses MAMBASPEC by 3.2 pp mAP on *AnuraSet* and 2.3 pp on *aSwine* despite being only $6.2\times$ larger in parameter count. This asymmetry indicates that *aSwine*'s seven event classes benefit more from large-scale pretraining than the 42 anuran species, consistent with the larger gap between SSAMBA and MAMBASPEC on *aSwine*.

Table 1. Accuracy metrics: mean ± std over 5 seeds. Best results in bold.

	AnuraSet (42 cls)		*aSwine* (7 cls)	
Model	**mAP**	**ROC-AUC$_w$**	**mAP**	**ROC-AUC$_w$**
AST	$\mathbf{0.8771 \pm 0.0025}$	$\mathbf{0.9812 \pm 0.0007}$	$\mathbf{0.6144 \pm 0.0065}$	$\mathbf{0.8695 \pm 0.0027}$
SSAMBA	0.8623 ± 0.0013	0.9804 ± 0.0008	0.5637 ± 0.0043	0.8496 ± 0.0023
MAMBASPEC	0.8307 ± 0.0062	0.9743 ± 0.0013	0.5409 ± 0.0019	0.8456 ± 0.0022

4.2. Efficiency

Table 2 profiles all three architectures. SSM complexity savings do not reduce latency at single-segment scale: SSAMBA (116 s/s) is marginally slower than AST (133 s/s) at batch=1, as FLOPs savings amortize only under batched inference. MAMBASPEC breaks this pattern through compactness: with $80\times$ fewer parameters and $1\,617\times$ fewer FLOPs, it achieves $8.9\times$ higher single-segment throughput and $55\times$ the batch-128 throughput of AST (10,689 s/s).

Table 2. Efficiency metrics (AnuraSet, RTX 5070 Ti). Latency is end-to-end batch time; batch = 1 simulates single-segment IoT inference and batch = 128 server-side throughput.

Model	Params	GFLOPs	Throughput (s/s)	Latency (ms)	Peak Mem. (MB)
batch = 1 (single-segment IoT inference)					
AST	86.9 M	8.77	133	7.5	346
SSAMBA	6.8 M	0.19	116	8.7	369
MAMBASPEC	**1.1 M**	**0.0054**	**1 178**	**0.85**	**41**
batch = 128 (server-side batch processing)					
AST	—	—	193	663	959
SSAMBA	—	—	1 035	124	544
MAMBASPEC	—	—	**10 689**	**12**	**215**

4.3. Pareto Analysis and Statistical Tests

All three models are non-dominated: Figure 1 shows MAMBASPEC, SSAMBA, and AST tracing a smooth accuracy-efficiency Pareto frontier consistent across both datasets. Marginal accuracy gains diminish as cost grows: moving from MAMBASPEC to SSAMBA adds $+0.6$ pp on *AnuraSet* at $6.2\times$ more parameters, while moving from SSAMBA to AST adds only $+0.1$ pp at $12.8\times$ more parameters.

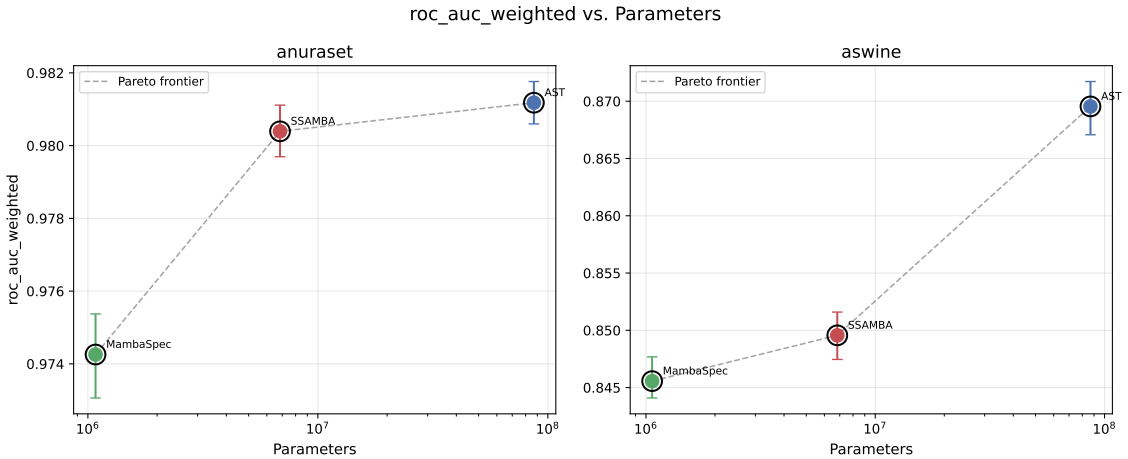

Figure 1. Pareto frontier per dataset: ROC-AUC$_w$ vs. parameter count (log scale). All three models are non-dominated. Error bars show 95 % bootstrap Confidence Interval (CI).

The key statistical result is that SSAMBA matches AST: Nemenyi post-hoc tests (Figure 2) show $p = 0.37$ between them, confirming that a pretrained SSM reaches

Transformer-level ROC-AUC$_w$ at $12.8\times$ lower parameter cost. MAMBASPEC ranks significantly below both ($p = 0.020$ vs. SSAMBA; $p < 0.001$ vs. AST), and the Friedman omnibus test confirms a global effect (ROC-AUC$_w$: $p = 2.25 \times 10^{-4}$; mAP: $p = 4.5 \times 10^{-5}$).

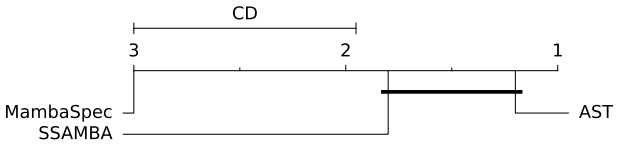

Figure 2. Critical-difference diagram for ROC-AUC$_w$ (3 models across 10 seed $\times$ dataset blocks; Nemenyi, $\alpha = 0.05$). Models joined by a bar are not significantly different.

4.4. Deployment Implications

The Pareto frontier maps hardware constraints to architecture choice. MAMBASPEC (1.1 M parameters, 5.4 MFLOPs/segment) is the efficiency-first choice, suiting the most constrained nodes. We hypothesize, based on its resource footprint, that MAMBASPEC is deployable in real time on microcontroller-class hardware: its ≈4.4 MB weight file fits within a small fraction of a Raspberry Pi Zero 2 W's 512 MB RAM, and 5.4 MFLOPs per inference is small relative to the board's CPU budget. This is a projection from parameter and FLOPs proxies [Lin et al. 2020], not a measured result; [Vuilliomenet et al. 2026] report that Raspberry Pi-class hardware can sustain real-time bioacoustic inference in the field, but on-device validation of MAMBASPEC remains future work. SSAMBA occupies the intermediate tier: AudioSet pretraining lifts its ROC-AUC$_w$ to statistical parity with AST at $12.8\times$ lower parameter count, making it well-suited for edge servers or single-board computers where accuracy outweighs frugality. AST remains the accuracy-first ceiling for applications with unconstrained memory and compute. *Limitations and future work.* Efficiency was profiled on a GPU (RTX 5070 Ti), so FLOPs and parameter count serve as hardware-agnostic transfer proxies [Lin et al. 2020] rather than measured edge performance. On-device benchmarking on real IoT hardware (e.g., Raspberry Pi, Jetson Nano) is left to future work, including energy-per-inference (as critical as FLOPs for battery-powered nodes) and extending the protocol beyond two domains to birds, cetaceans, and bats to test generalization.

5. Conclusion

Addressing whether SSMs can match AST accuracy while enabling IoT deployment, our reproducible, efficiency-first evaluation protocol (controlled HPO, five seeds, statistical validation, and Pareto deployment mapping) answers affirmatively: MAMBASPEC and SSAMBA form a smooth Pareto frontier with AST, each mapping to a distinct hardware tier. The protocol itself is domain-agnostic and applies to any data-management pipeline where continuous IoT sensor streams must be classified and reduced under strict memory and latency budgets, including environmental, agricultural, and industrial monitoring. MAMBASPEC runs at $80\times$ lower parameter cost and sub-millisecond latency, while SSAMBA matches AST ($p = 0.37$, Nemenyi) at a $12.8\times$ smaller footprint. Validating these projections on embedded hardware and broadening the protocol to further PAM domains are the primary directions for future work.

References

Cañas, J. S., Toro-Gómez, M. P., Sugai, L. S. M., Benítez Restrepo, H. D., Rudas, J., Posso Bautista, B., Toledo, L. F., Dena, S., et al. (2023). AnuraSet: A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring. *Scientific Data*, 10:771.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Erol, M. H., Senocak, A., Feng, J., and Chung, J. S. (2024). Audio Mamba: Bidirectional State Space Model for Audio Representation Learning. *IEEE Signal Processing Letters*, 31:2975–2979.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Proc. Interspeech*, pages 571–575.

Gu, A. and Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Hagiwara, M., Hoffman, B., Liu, J.-Y., Cusimano, M., Effenberger, F., and Zacarian, K. (2023). BEANS: The Benchmark of Animal Sounds. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Liang, J., Nolasco, I., Ghani, B., Phan, H., Benetos, E., and Stowell, D. (2024). Mind the domain gap: A systematic analysis on bioacoustic sound event detection. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1257–1261.

Lima, J., Salles, R., Escobar, L., Géa, C., Fernandes, P. A., Pacitti, E., Porto, F., Coutinho, R., and Ogasawara, E. (2022). Towards a cloud-based framework for online and integrated event detection. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 199–202. SBC.

Lin, J., Chen, W.-M., Lin, Y., cohn, j., Gan, C., and Han, S. (2020). MCUNet: Tiny deep learning on IoT devices. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11711–11722. Curran Associates, Inc.

Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., Lange, J., Kahl, S., Sick, B., Tomforde, S., and Scholz, C. (2025). BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R., editors, *International Conference on Learning Representations*, pages 29482–29520.

Shams, S., Dindar, S. S., Jiang, X., and Mesgarani, N. (2024). SSAMBA: Self-Supervised Audio Representation Learning with Mamba State Space Model. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 1053–1059.

Souza, A. M., Kobayashi, L. L., Tassoni, L. A., Garbossa, C. A. P., Ventura, R. V., and Machado de Sousa, E. P. (2025). Deep learning solutions for audio event detection in a swine barn using environmental audio and weak labels. *Applied Intelligence*, 55(7).

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152.

Tang, C. and Baskiyar, S. (2025). State space models for bioacoustics: A comparative evaluation with Transformers. *arXiv:2512.03563*.

Vuilliomenet, A., Martínez Balvanera, S., Mac Aodha, O., Jones, K. E., and Wilson, D. (2026). acoupi: An open-source python framework for deploying bioacoustic AI models on edge devices. *Methods in Ecology and Evolution*, 17(1):67–76.