Rainfall Acoustic Classification in Semi-Urban and Forest Soundscapes with Machine Learning

G. G. R. Milan [1], *Student Member, IEEE*, A. F. Fernandes [1], *Student Member, IEEE*, R. S. Xavier [2], T. F. Maciel [1], *Senior Member, IEEE*, M. T. R. Chaves [3], M. Gosset [4]

[1] Laboratory of Applied Signal Processing, UFC, Fortaleza, Brazil, [2] Mamirauá Institute for Sustainable Development, Tefé, Brazil, [3] Hydraulic and Environmental Engineering Dept., UFC, Fortaleza, Brazil, [4] Géosciences Environnement Toulouse, OMP, Toulouse, France

Global Challenges

- **Mechanical Failures:** Traditional rain gauges are prone to clogging and maintenance issues in tropical regions.
- **High-Resolution Gaps:** Coverage is limited in the Southern Hemisphere, especially in Amazon and urban centers [1].
- **Opportunistic Sensing:** Utilizing ecoacoustic recorders offers a scalable, low-cost alternative for climate monitoring [1].

Our Contributions

- **Soundscape analysis:** Acoustic rainfall classification via Machine Learning in two contrasting environments: a semi-urban campus (Ceará State University (UECE)) and a secondary rainforest (Mamirauá Sustainable Development Institute (IDSM)).
- **Feature optimization:** Extraction of 91 off-the-shelf acoustic metrics, followed by empirical dimensionality reduction to isolate an optimal, lightweight subset per site.
- **Robust data pipeline:** 10-second audio segmentation [2], strict class balancing, and stochastic augmentation to mitigate the *accuracy paradox* in rare events.
- **Dimensionality scale:** Comprehensive comparison between the standard 5-class precipitation scale and a condensed 3-class scheme.

Study Areas

- **Secondary rainforest (IDSM):** Located in Tefé, Amazonas, Brazil. Dominated by intense and diverse biophony (e.g., vocalizations of birds, insects and amphibians) and characterised by a dense, multilayered canopy that fundamentally shapes the acoustic signature of raindrops.
- **Semi-urban campus (UECE):** Located in Fortaleza, Brazil. Dominated by anthropogrnic background noise from vehicles and people, and wind-inducted geophony, as well some biophony from birds, making the acoustic signature of rain much closer to background noise.

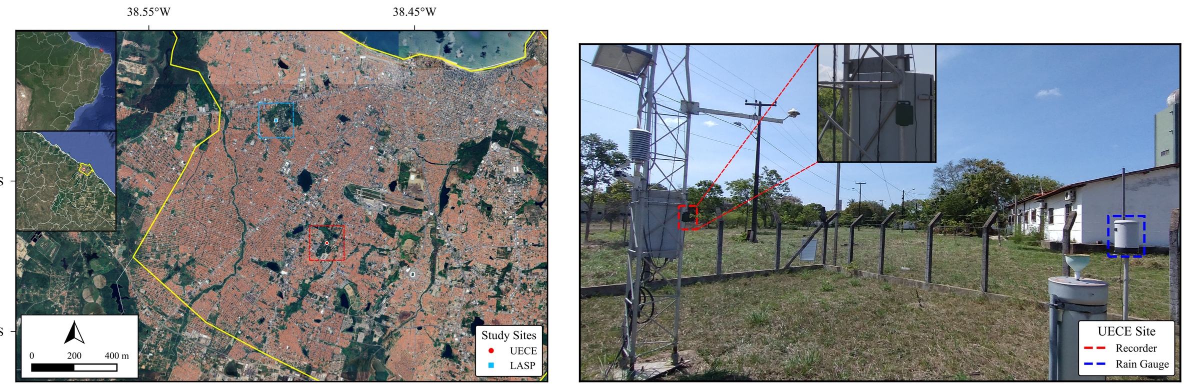

Fig. 1: Location and experimental setup of the primary study area at UECE. The Laboratory of Applied Signal Processing (LASP), where the data analyses were conducted is also indicated. The IDSM measurement site location is available in [1].

Data Processing & Classification

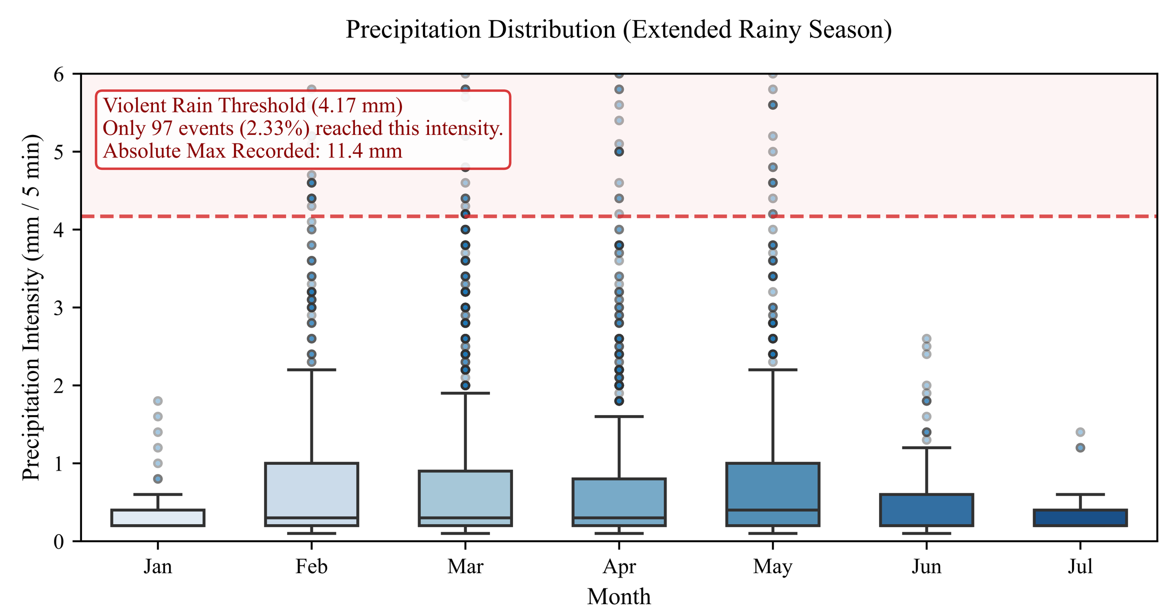

Fig. 2: Fortaleza's (2023–2025) 5-minute precipitation intensity distributions from the IFCE platform [3].

- **The Imbalance Challenge:** Raw data has "No Rain" class dominance, while extreme events (e.g., the "Violent") are extremely rare.
- **Curation & Balancing:** Strict data curation and balancing (1:1 No Rain/Rain) was applyed to prevent models from collapsing toward the majority label.
- **Segmentation & Augmentation:** Audio is segmented into 10-second windows to provide high temporal resolution [2]. Concurrently, stochastic data augmentation on training sets ensured the models could learn and perform well on minority classes.
- **Feature Extraction:** A multidimensional vector was extracted (incorporating Power Spectrum Density (PSD), Mel-Frequency Cepstral Coefficients (MFCCs) and ecoacoustic metrics) and optimized via dimensionality reduction.
- **Classification Schemes:** We evaluate linear (Stochastic Gradient Descent (SGD), Logistic Regression (LR)) and tree-based (Random Forest (RF), Extreme Gradient Boosting (XGB)) algorithms across two complexity levels: 5-class and 3-class schemes.

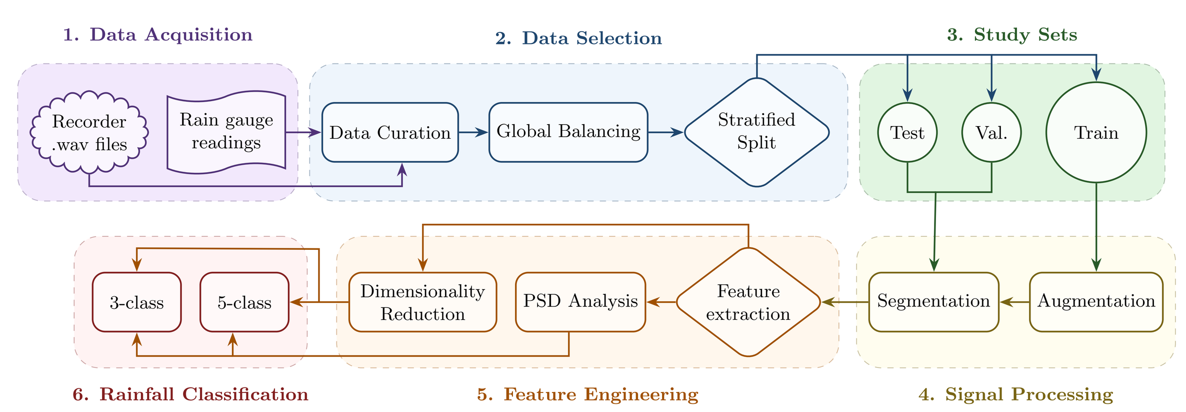

Fig. 3: End-to-end data curation, signal processing, feature engineering and rainfall classification data flow.

Complete datasets information, source code, and extended high-resolution results are openly available in our repository [4].

Results and Discussion

- **Baseline Failure (PSD):** Models trained exclusively on PSD failed across all setups (F1-Score (F1S) between 17.44% and 34.80%). Separability plots confirm that rain classes distributions severely overlap.
- **Pipeline Effectiveness:** Data curation minimized cross-confusion between rainfall and "No Rain". Temporal segmentation (10s windows) and stochastic augmentation prevented majority-label bias.
- **Sensor Limitations:** Boundary confusion and feature outliers are driven by the mechanical inertia of the tipping-bucket rain gauge.
- **Forest Soundscape (IDSM):** RF peaked in the 5-class scheme (71.88% Balanced Accuracy (B-ACC)), while LR dominated the 3-class setup (83.65% B-ACC), maintaining a low False Alarm Rate (FAR) (≈ 10%).

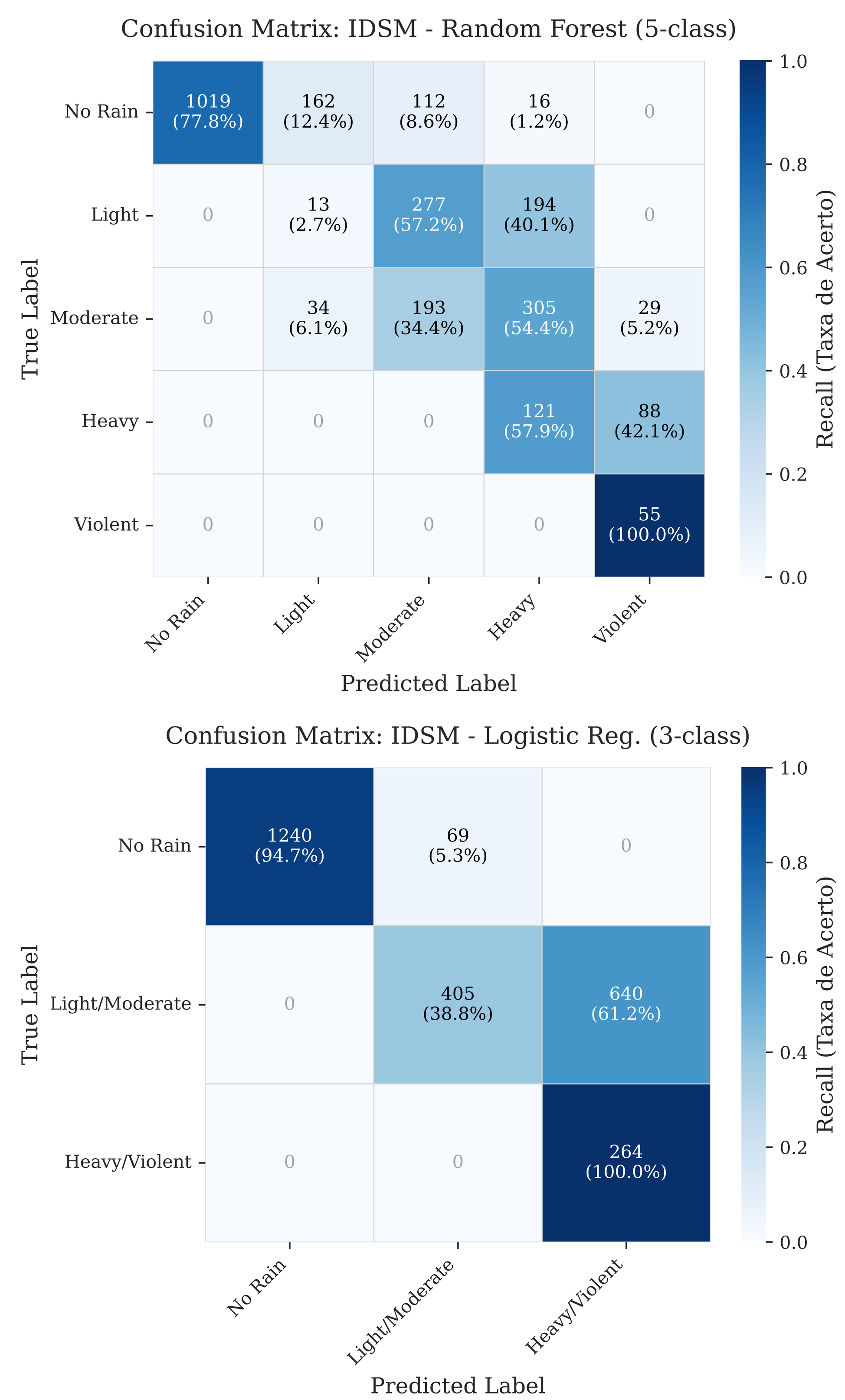

Fig. 4: Confusion matrices for the top-performing models in IDSM.

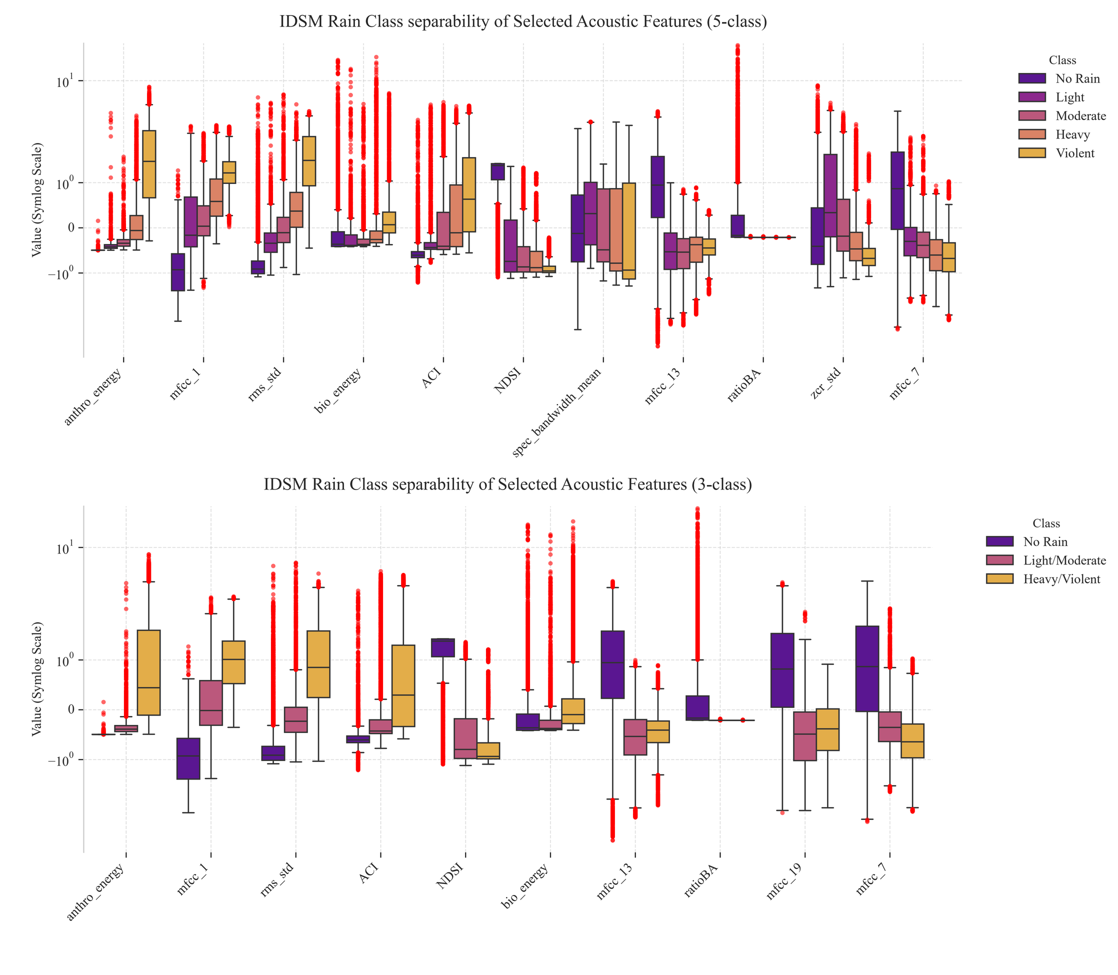

Fig. 5: Selected acoustic features for IDSM in 5-class and 3-class schemes.

- **Urban Open Field (UECE):** Severely hindered by the lack of canopy amplification and strong wind geophony. The algorithm selected 9 MFCC levels, *temp_kurtosis*, and *anthro_energy* to map urban traffic noise. XGB was the only model resilient enough to isolate "Moderate" and "Heavy" rain. All algorithms failed to detect the "Violent" class.

Conclusion

- **Methodology Validation:** The multidimensional pipeline dynamically adapts to contrasting soundscapes by isolating site-specific feature subsets.
- **System Standardization:** Achieving forest-level accuracy in urban settings requires mechanical amplifiers and possibly needs source separation algorithms to isolate rain signatures.
- **Cascade Architecture:** A dual-stage edge computing setup, using an ultra-lightweight SGD for continuous binary detection to trigger an XGB/RF intensity grader, could optimize computational load.

References

[1] R. S. Xavier, M. Gosset, T. F. Maciel, T. Bicudo, L. A. D. Nascimento, E. Ramalho, and A. Fleischmann, "Measuring amazon rainfall intensity with sound recorders," ***Geophysical Research Letters***, vol. 51, no. 18, p. e2024GL108210, 2024.

[2] D. Dunkerley, "Rainfall intensity and intermittency over timescales of minutes: Measurement using low sampling-rate acoustic recording," ***Atmospheric Research***, vol. 290, 2023.

[3] M. T. R. Chaves, F. A. P. Rodrigues, W. M. Eloi, P. H. A. Medeiros, R. d. S. Araújo, and T. R. L. Farias, "Bioretention performance under tropical conditions: Experimental analysis of hydrological controls and design implications," ***Hydrological Processes***, vol. 40, p. e70382, 2025.

[4] G. G. R. Milan, "Github repository: Rainfall acoustic classification in semi-urban and forest soundscapes with machine learning," https://github.com/giovannigremedym/racml.git, accessed in: 09 jun. 2026.

Acknowledgements: This work was partially supported by INCT-Signals under grants INCT-25255-82587.32.41/64 (FUNCAP) and 406517/2022-3 (CNPq), and CAPES (Finance code 001). T. F. Maciel and R. S. Xavier were partially supported by CNPq under grants 312471/2021-1, 305165/2026-7 and 402661/2024-9.