BioDCASE: Towards standardised data challenges supporting robust development of ML models for underwater acoustic event detection

Parcerisas C; Jean-Labadye L; Schall E; Dubus G; Carvaillo P; Miller B; Moummad I; and Cazau D

Clea Parcerisas
clea.parcerisas@vliz.be
Flanders Marine Institute (VLIZ) / Ghent University, Belgium

With the current rise of the application of Machine Learning (ML) models to detect specific sound events in underwater passive acoustics monitoring recordings, the need for standardised and scalable comparison between developed models emerges. For this reason Task 2 in the bioDCASE (2025) data challenge was proposed, with the objectives of (1) bringing together ML, marine science, and bioacoustics communities (2) improving rigour and fairness in comparing performance of proposed solutions (3) helping to bridge the gap between the requirements of operational monitoring projects and existing ML evaluation protocols (4) engaging entry-level researchers into model development, and (5) providing an open-source benchmark for further developments. The designed data challenge asked participants to develop automatic multilabel detectors using strongly labelled events. The dataset selected for the task was the Miller et al. (2020) dataset, comprising recordings from 11 location-year pairs. The annotation effort focused on calls from Antarctic blue whales (Balaenoptera musculus) and fin whales (Balaenoptera physalus), split into seven labels, which were grouped into three final labels for the task. A curated version of the dataset was made available on Zenodo. The task proposed a fixed training, validation and evaluation split where the ground truth of the evaluation set remained unknown for participants. The splitting was designed to evaluate the model's ability to generalise across unseen locations and years using a hold-out dataset approach. In the selected split, the validation set and the evaluation set included site-year combinations which were not included in the training set regarding year, location or both. The evaluation metric was based on one-dimensional intersection over union along the temporal axis and penalised models that generate multiple detections for a single ground-truth event to support applications such as density estimation. Considering the imbalance of occurrences between classes, the comparison metric was chosen to be macro-averaged. In the 2025 bioDCASE edition, two baselines were provided by the organising team using YOLOv11 and ResNet, respectively, and the task had 12 submissions from 6 different teams. The best performing model achieved an F-score of 0.5, with only one model beating the baseline YOLOv11 performance, showcasing the difficulty of the task. The same task will be released in 2026 with a supplementary evaluation dataset.