HARD-NEGATIVE MINING, MEAN-TEACHER, AND CLASS-SPECIALIST ENSEMBLING FOR ANTARCTIC BALEEN WHALE CALL DETECTION

Technical Report

Matthias Nagl, Kevin Eberl, Pablo Guamán, Omar Yasser Mostafa

Institute of Computational Perception
Johannes Kepler University Linz, Austria

ABSTRACT

We present our system for BioDCASE Task 2, which revolves around the temporal detection of Antarctic blue and fin whale calls in passive acoustic recordings. Systems must predict both the class and the temporal extent of every call. We build on WhaleVAD, a lightweight CNN-BiLSTM detector, and improve it through class-aware training and an ensembling strategy. To this end, we collect confident false positives through per-class hard-negative mining, add Mean-Teacher self-training on unlabelled recordings to strengthen the rarer classes, and route the final decision for each class to the model group that detects it best. The resulting ensemble reaches a F1 of 0.501 on the official development set. Our leave-one-site-out estimate, which avoids tuning on the held-out site and matches the protocol of WhaleVAD-BPN, is 0.483. This is an improvement over the WhaleVAD baseline, which scored 0.422, and competitive with the stronger WhaleVAD-BPN system at 0.475. We observe the largest gains on the difficult D-call class, and attribute them to the semi-supervised training procedure and class-specialist routing.

***Index Terms*—** Sound event detection, bioacoustics, semi-supervised learning, hard-negative mining, model ensembling

1. INTRODUCTION

Passive acoustic monitoring is an important tool for studying Antarctic blue and fin whales, which are difficult to observe directly over large areas and long time spans. However, automatic detection of whale calls and classification of call types remains difficult. Calls are sparse, recording conditions vary strongly between sites, and the same call type can look different depending on distance, noise, and local recording conditions. These issues make the BioDCASE Task 2 a useful benchmark for whale-call sound event detection [1, 2].

The task is a strongly labelled multi-label detection problem. Systems must return both the class and the temporal bounds of each detected event. The original annotations contain seven call labels, which are grouped into three evaluation classes: BMABZ for blue-whale A-, B-, and Z-call types (*bma*/*bmb*/*bmz*), D for D-calls (*bmd*/*bpd*), and BP for fin-whale 20 Hz pulse calls (*bp20*/*bp20plus*). The development set is based on the Acoustic Trends Blue Fin Library (ATBFL) [3], a collection of annotated underwater recordings from Antarctic blue and fin whale monitoring sites [2].

We analysed the dataset to identify properties that were likely to make detection difficult. Two aspects were most relevant for our later experiments. First, the data are highly imbalanced in time: only a small fraction of the audio is covered by call annotations. As a result, false positives can quickly dominate the event counts, so reducing them is almost as important as detecting the calls themselves. Second, the merged class distributions vary strongly across site-years, as shown in Table 1.

Table 1: Per-class event distribution across development site-years, as a percentage of each class's events by split.

Site-year	BMABZ	D	BP
ballenyisland2015	4.1%	0.7%	7.1%
casey2014	25.4%	3.6%	0.1%
elephantisland2013	18.8%	63.0%	35.9%
elephantisland2014	32.9%	27.1%	51.0%
greenwich2015	4.2%	0.6%	0.0%
kerguelen2005	5.0%	4.7%	5.6%
maudrise2014	9.3%	0.4%	0.2%
rosssea2014	0.4%	0.0%	0.0%
casey2017	22.7%	21.2%	5.6%
kerguelen2014	50.1%	25.2%	70.4%
kerguelen2015	27.1%	53.7%	24.0%

In particular, the D class is concentrated in only a few training sites: around 90% of its examples come from elephantisland2013 and 2014. This motivated treating classes separately in hard-negative mining and in the final ensemble.

We build on WhaleVAD, a CNN-BiLSTM for baleen whale vocalisation activity detection [4], as our baseline detector. We do not introduce a new backbone ourselves, but focus on training specialized models for different call types and combining these specialists. This was also a practical choice: the reproduced models already capture the main time-frequency structure of the calls, but their errors are not evenly distributed over the three classes with D-calls remaining the weakest. The main components of our system are:

- per-class hard-negative mining, where confident false positives are collected separately and used for fine-tuning.
- Mean-Teacher self-training on unlabelled recordings from the Australian Antarctic Data Centre (AADC), used together with the supervised hard-negative objective [5].
- class-specialist ensembling, where the final prediction for each class is taken from the model group that performed best.

2. METHODS

Figure 1 gives an overview of the complete system. We first describe the base SED model, then the stages built on top of it: hard-negative mining, Mean-Teacher self-training, post-processing, and class-specialist ensembling.

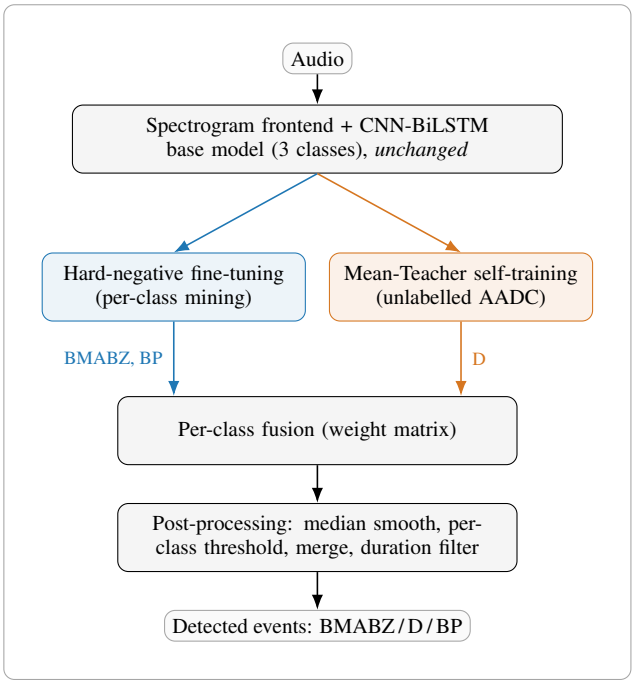

Figure 1: System overview of the final class-specialist ensemble. The unchanged WhaleVAD backbone is refined by complementary training branches, fused per class, and post-processed into event detections.

2.1. Base SED model

WhaleVAD is a lightweight CNN-BiLSTM detector for baleen whale calls [4]. It takes a three-channel phase-aware STFT (magnitude, and the sine and cosine of the phase), from which a convolutional front-end extracts per-frame features. A BiLSTM adds temporal context, and a final linear layer with a sigmoid activation produces one independent posterior sequence per class, one decision per 20 ms of audio. We adopt this architecture unchanged, so every contribution below acts on how the model is trained or on how the trained models outputs are combined at inference, never on the network itself.

2.2. Hard-negative mining

The base model has high recall but a high false-positive rate, which is the principal cause of its low precision and is most harmful for the minority classes [4, 6]. We address this error mode directly by mining the model's own confident false positives on the training set and treating them as hard negatives for a subsequent fine-tuning stage.

Mining is performed independently for each of the three classes. We run the converged base model over all eight training sites in 30 s tiles with 2 s overlap, and average the overlapping tiles into one per-frame, three-class posterior stream per file. To collect candidates for a given class we run the shared multi-class event extractor with that class's frame threshold set to a low 0.2 and the other two set to 0.999. The high thresholds suppress almost all events from the other classes, and only proposals carrying the target label are kept, so the candidate set is the target class alone. Before thresholding, the posteriors are smoothed with a 500 ms per-class median filter, and contiguous above-threshold frames are merged into events. The deliberately low threshold surfaces borderline detections that the tuned operating point would suppress. Each proposed event is matched against the target-class ground-truth annotations in the same recording using one-dimensional temporal intersection-over-union (IoU), and any proposal whose largest IoU with a true call falls below 0.1 is recorded as a false positive. The false positives are ranked by event confidence, the mean posterior over the event span, and the 1500 most confident per class are kept as the hard-negative pool.

We then fine-tune the base model on a single pooled training set: all positive segments, an equal number of randomly sampled negatives that are redrawn every five epochs, and the mined pool replicated five times. The set is shuffled, so batches are not balanced by a fixed ratio. The set is shuffled, not balanced per batch.

2.3. Semi-supervised learning with Mean Teacher

Beyond the annotated training set, the AADC provides long-term unlabelled recordings from further Southern Ocean sites [5]. From the subset listed in Section 3 we draw 10 000 unlabelled 30 s clips per epoch as a self-training stream. To exploit this audio, we add a mean-teacher consistency term [7] on top of the hard-negative fine-tuning, optimising the two objectives jointly rather than in separate stages. This is similar to the consistency-based self-training scheme that underlies the DCASE Task 4 sound-event-detection baseline [8], applied here to exploit unlabelled audio for the sparse D-class.

A student copy of the model is trained by gradient descent, while a teacher is maintained as an exponential moving average (EMA) of the student weights, updated at each step and never by backpropagation. We found that the consistency loss alone degrades the converged warm start, so we keep the hard-negative fine-tuning of Section 2.2 as a supervised anchor. Each step then minimises a supervised term plus a consistency term weighted by λ. The supervised term is the weighted binary cross-entropy of the hard-negative stage, with positive frames weighted above negatives, on a labelled batch of positives, oversampled hard negatives, and random negatives. For the consistency term the student sees a batch of unlabelled AADC clips under strong augmentation, namely SpecAugment-style time and frequency masking [9] together with volume perturbation, additive noise, and cross-site no-call mixing, while the teacher sees only lightly perturbed views of the same clips, and the student is trained to match the teacher's detached per-frame probabilities. The consistency weight λ is ramped from zero over the first few epochs so the supervised objective shapes the model first, and the EMA decay is ramped towards its final value over a similar warm-up.

Because the AADC stream comes from a different acoustic domain than the labelled sites, the batch-normalisation running statistics are frozen throughout: letting them update would drift them towards a mixture of the two domains and break the alignment between the features and the classifier at validation. The unlabelled

audio also contains no validation or test recording. The consistency loss is an asymmetric mean squared error between student and teacher probabilities, up-weighting confident positive teacher predictions to target the sparse D-class. At inference we use the teacher, the more stable of the two models.

2.4. Post-processing

Per-frame posteriors are converted to discrete events by a fixed pipeline. The per-window predictions, from 60 s windows with 2 s overlap, are stitched into one per-file stream, smoothed with a 500 ms median filter, and thresholded per class. The resulting same-class events are merged when separated by less than 0.5 s and kept only if their duration lies between 0.5 and 30 s. The per-class thresholds are the only tuned component, optimised on the validation set (Section 3). The smoothing window, merge gap, and duration bounds are fixed and shared across classes. A broader search using Optuna [10] over those remaining parameters did not improve on tuning the thresholds alone, so we kept the pipeline fixed.

2.5. Class-specialist ensembling

The hard-negative fine-tuning of Section 2.2 and the Mean-Teacher self-training of Section 2.3 yield two families of models with complementary strengths. The hard-negative models are the stronger detectors of the abundant BMABZ and BP calls, and the Mean-Teacher models are the better detectors of the rare D-class. Averaging all members uniformly would dilute each family with members that are weaker on a given class, so we instead route each class to the family that detects it best.

We fuse the member posteriors through a per-class weight matrix. For each class c, the ensemble posterior on a frame is

$$p_c \;=\; \sum_m w_{c,m}\, p_c^{(m)}, \tag{1}$$

where $p_c^{(m)}$ is member m's posterior for class c and the weights $w_{c,\cdot}$ are non-negative, sum to one, and are non-zero only on the members that specialise in class c: BMABZ and BP are drawn from the hard-negative models and D from the Mean-Teacher models. The fused posteriors are thresholded per class and post-processed as in Section 2.4. At inference the members are run on 60 s tiles rather than the 30 s training tiles, a longer evaluation window we found to improve the ensemble.

3. EXPERIMENTAL SETUP

All models are trained on the eight training sites and evaluated on the three validation site-years as described in Section 1. The four test site-years are held out and used for neither model selection nor threshold tuning. The Mean-Teacher stage draws its unlabelled audio from five additional AADC site-years (casey2018, ddu2018, kerguelen2018, kerguelen2021, and kerguelen2023). These recording stations also appear in the validation or test sets at different years, but no validation or test recording itself is used for training. Every configuration is trained with five seeds and the numbers we report are the five-seed ensembles of Section 2.5, not per-seed averages.

Table 2 lists the hyperparameters for the three training stages. The weighted BCE scales each class's positive term by $w_c = N/P_c$, the total number of positive training segments over the class-c count, so the rarer classes contribute more. Across all stages the checkpoint is selected by the validation F1 (Eq. 2), with the per-class detection thresholds re-tuned each epoch by a coordinate-descent grid search.

Table 2: Training configuration. Mining hyperparameters are given in Section 2.2.

Hyperparameter	Value
Base model	
Optimiser	AdamW
Learning rate	5×10^{-5}
Batch size	32
Epochs	50
Loss	weighted BCE
Random negatives	resampled each epoch
Hard-negative fine-tune	
Learning rate	1×10^{-5}
Epochs (early stop)	15 (8)
Hard-negative oversampling	$5\times$
Random-negative resample	every 5 epochs
LR schedule	plateau ($\times 0.5$, patience 4)
Gradient clipping	1.0
Mean-Teacher	
Epochs	20
Unlabelled clips / epoch	10 000
Consistency weight λ	$0 \rightarrow 1.0$ over 3 epochs
EMA decay α	$0.99 \rightarrow 0.999$ over 5 epochs
Consistency loss	Asymmetric MSE

Our headline metric is the F1 used by the challenge. Detections are matched to the annotations by the challenge's greedy one-dimensional temporal intersection-over-union at 0.3, with at most one detection per ground-truth call counted as a true positive [1]. Per-class precision P_c and recall R_c are computed from the resulting true positives, false positives, and false negatives pooled over the three validation sites, and the score is the harmonic mean of their macro-averages:

$$\mathrm{F1} = \frac{2\,\bar{P}\,\bar{R}}{\bar{P}+\bar{R}}, \qquad \bar{P} = \frac{1}{C}\sum_{c=1}^{C} P_c, \quad \bar{R} = \frac{1}{C}\sum_{c=1}^{C} R_c. \tag{2}$$

This is the convention under which WhaleVAD (0.422) and WhaleVAD-BPN (0.475) are reported [4, 6].

A single CNN-BiLSTM base model is refined along two paths that share the hard-negative objective: one applies hard-negative fine-tuning alone, the other runs the same fine-tuning jointly with Mean-Teacher self-training on unlabelled AADC audio. The first is the stronger detector of the abundant BMABZ and BP calls, the second of the rare D class. Their outputs are combined by a per-class fusion matrix and then converted into final detections using the fixed post-processing pipeline.

4. RESULTS

4.1. Development

We first measure each component on the development split, with the per-class thresholds tuned on that split. This is internally consistent for ranking design choices but optimistic in absolute terms, and we

do not compare it to the baselines here. The cascade lifts the F1 from a reproduced baseline of 0.439 to 0.501 (Table 3).

Table 3: Effect of each component on the validation split (five-seed ensembles, 60 s tiles, thresholds tuned on the split). Per-class F1 and the challenge F1 of Eq. 2. Bold marks the best in each column.

System	BMABZ	D	BP	F1
Baseline (no HNM)	0.598	0.159	0.528	0.439
Hard-negative ensemble	**0.607**	0.222	**0.543**	0.464
Mean-Teacher ensemble	0.549	**0.332**	0.532	0.483
Two-family blend (uniform)	0.607	0.289	0.540	0.489
Class-specialist routing	0.607	0.332	0.543	**0.501**

The improvement is carried by the D class, the rarest and most recall-limited call type, which the baseline detects at only 0.159. Hard-negative fine-tuning lifts this to 0.222, but the larger gain comes from the Mean-Teacher family, whose ensemble reaches 0.332, roughly double the baseline, while remaining weaker on BMABZ (0.549 against 0.607). At the D operating point the gain is recall-driven and clean: Mean-Teacher more than doubles the true positives over the baseline (369 to 807) while lowering false positives (1424 to 1205), whereas hard-negative fine-tuning raises both. A uniform blend keeps most of the D recall but reintroduces false positives from the D-weaker hard-negative members, lowering the D F1 from 0.332 to 0.289. Class-specialist routing avoids this dilution by taking D only from the Mean-Teacher family, recovering the full 0.332 and the best split F1. Because the routed ensemble and the uniform blend use the same members, the gain is attributable to the routing itself rather than to the number of models.

4.2. Generalisation and comparison

The development figures are optimistic, because the thresholds are tuned on the same split they are scored on. The model is trained only on the eight training sites and never sees an evaluation site, so the optimism is in the threshold selection, not the model. For an honest estimate, and to compare against the baselines on the protocol they use, we re-select the thresholds under three-fold leave-one-site-out cross-validation over the three evaluation sites, tuning on two sites and scoring the held-out third, with every site held out once and the three scores averaged. The trained model is identical in every fold, so this isolates how an operating point transfers to an unseen site, and it matches the cross-validation WhaleVAD and WhaleVAD-BPN use to select their post-processing.

Table 4: Leave-one-site-out comparison on the evaluation site-years. Per-class and challenge F1 (Eq. 2). WhaleVAD + opt. is WhaleVAD with optimised post-processing, and the baselines are the published figures [6].

System	BMABZ	D	BP	F1
WhaleVAD + opt.	0.669	0.222	0.363	0.422
WhaleVAD-BPN	0.615	0.340	0.460	0.475
Ours	0.610	0.327	0.493	**0.483**

The routed system reaches a cross-validated F1 of 0.483, a drop of 0.018 from the split-tuned 0.501 that makes the selection optimism explicit. D stays recall-limited (recall 0.294) but well clear of degeneracy. At 0.483 the system is marginally above WhaleVAD-BPN (0.475) and well above WhaleVAD with optimised post-processing (0.422). The margin over WhaleVAD-BPN comes from BP (0.493 against 0.460), while the two systems are close on BMABZ and WhaleVAD-BPN keeps a slight edge on D (0.340 against 0.327). Both refined systems give up a little of the abundant BMABZ class (0.610 and 0.615 against WhaleVAD's 0.669) for large minority-class gains. The 0.008 margin over WhaleVAD-BPN is on the order of the seed-to-seed noise, so we read the result as matching the stronger baseline with no architectural change. Generalisation is uneven across sites (Table 5). The held-out F1 ranges from 0.427 on casey2017 to 0.530 on kerguelen2015. casey2017 is the hardest site, specifically because BP falls to 0.317 and D to 0.267 there, whereas on kerguelen2015 both recover (0.623 and 0.352). BMABZ is the most stable class across sites.

Table 5: Per-site generalisation under leave-one-site-out, per-class and challenge F1 (Eq. 2).

Held-out site	BMABZ	D	BP	F1
casey2017	0.611	0.267	0.317	0.427
kerguelen2014	0.483	0.323	0.503	0.440
kerguelen2015	0.561	0.352	0.623	0.530

5. CONCLUSION

We have shown that a CNN-BiLSTM whale-call detector can be brought to the level of the stronger WhaleVAD-BPN system without any architectural change, using only interventions on the training data and the ensemble. Hard-negative fine-tuning sharpens the abundant blue-whale and fin-whale classes, while Mean-Teacher self-training on unlabelled AADC recordings roughly doubles the F1 of the rare D class. Routing each class to the family that detects it best combines these gains and reaches a cross-validated F1 of 0.483, above the WhaleVAD baseline and matching WhaleVAD-BPN. The D class remains the weakest and the most variable across sites, and the improvement comes from training-side methods and class routing rather than from any single stronger model. Because all three interventions act on the training data and on how the model outputs are combined, and never on the network itself, they are independent of the backbone and should carry over to other detection architectures. They are also orthogonal to architectural change, so pairing them with a stronger backbone or a learned time-frequency frontend is a natural next step.

6. ACKNOWLEDGMENT

This work was carried out for course 344.032 ("Machine Learning and Audio: A Challenge"). We thank Paul Primus, Florian Schmid, and Tobias Morocutti for their supervision and feedback, and the Institute of Computational Perception at Johannes Kepler University Linz for the compute resources.

7. REFERENCES

[1] L. Jean-Labadye *et al.*, "BioDCASE 2025 task 2: Development set," Zenodo, 2025, version 1 (Erratum).

[2] B. S. Miller, K. M. Stafford *et al.*, "An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors," *Scientific Reports*, vol. 11, no. 1, p. 806, 2021.

[3] B. S. Miller *et al.*, "An annotated library of underwater acoustic recordings for testing and training automated algorithms for detecting Antarctic blue and fin whale sounds," Australian Antarctic Data Centre, 2020, version 1.

[4] C. Geldenhuys, G. Tonitz, and T. Niesler, "Whale-vad: Whale vocalisation activity detection," in *Proceedings of the 10th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2025)*, Barcelona, Spain, October 2025, pp. 165–169.

[5] B. S. Miller, M. Milnes, and S. Whiteside, "Long-term underwater acoustic recordings in the Southern Ocean 2013–2024," Australian Antarctic Data Centre, 2025, version 9.

[6] C. M. Geldenhuys, G. Tonitz, and T. R. Niesler, "Whalevad-bpn: Improving baleen whale call detection with boundary proposal networks and post-processing optimisation," 2025. [Online]. Available: https://arxiv.org/abs/2510.21280

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 1195–1204.

[8] S. Cornell, J. Ebbers, C. Douwes, I. Martín-Morató, M. Harju, A. Mesaros, and R. Serizel, "DCASE 2024 task 4: Sound event detection with heterogeneous data and missing labels," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, 2024.

[9] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019, pp. 2623–2631.