Comparative Analysis of Anomaly Detection Techniques for Classification of Southern Ocean Sound Data From Elephant Seals

Sara Norouzinia[1], Kenny Saint Fleur[1], Louis Vozzola[1], Anatole Gros-Martial[2], and Cédric Pradalier[1,3]

[1] Georgia Institute of Technology, USA
{snorouzinia3, kfleur3, lvozzola3}@gatech.edu
[2] Centre d'Etudes Biologiques de Chizé, France
[3] CNRS IRL2958 GeorgiaTech-CNRS, France

Abstract. The rapid decline of marine biodiversity due to anthropogenic climate change necessitates the development of scalable, non-intrusive monitoring systems. While using "marine sentinels" such as elephant seals equipped with acoustic bio-loggers offers a promising solution, the manual processing of thousands of hours of recordings remains a critical bottleneck. This paper presents a comparative analysis of three machine learning paradigms for automated Anomalous Sound Detection (ASD) in the Southern Ocean: a distance-based approach using Perch 2.0 embeddings, a supervised ConvNeXt-based CNN, and a self-supervised autoregressive model (AR-BEATs). Our results demonstrate that the Perch 2.0 pipeline achieves the highest performance with a PR-AUC of 0.954 and an F1-score of 91.0%, followed by AR-BEATs (F1: 87.0%), which prove highly efficient for low-resource deployment requiring only ≈12 min of confirmed normal audio with no anomalous labels. In contrast, the supervised PyTorch CNN yielded an F1-score of 74.7%. Expert validation further highlights a significant limitation: high-frequency biological signals, such as sperm whale vocalizations (8–25 kHz), are consistently misclassified due to sampling rate constraints and terrestrial-biased pretraining.

Keywords: Anomaly Detection · Machine Learning · Bioacoustics · Southern Ocean · Elephant Seals

1 Introduction

The Earth's oceans harbour the vast majority of our planet's biodiversity and are currently facing an unprecedented crisis. While marine life has historically survived various geological fluctuations, anthropogenic climate change is placing ecosystems under extreme pressure. Penn and Deutsch (2022) [13] suggest that under "business-as-usual" scenarios, we are on track for a mass extinction comparable to the Permian-Triassic event.

Regarding cetaceans, this pressure stems from a century-long whaling industry that severely depleted baleen and sperm whale populations. Following the

cessation of commercial whaling, research has focused on conservation status, distribution, and habitat use [16], with efforts recently expanding to delphinids [9] and odontocete feeding ecology [15]. Ongoing work seeks to estimate population abundance and assess recovery trajectories under changing environmental conditions.

To address monitoring at scale, Dr. Gros-Martial [7] proposed utilising elephant seals as marine "sentinels," equipped with data loggers integrating a hydrophone, salinity, temperature, chlorophyll, accelerometer, and magnetometer sensors. Mel-spectrogram analysis was used to correlate acoustic data with animal presence and behaviour without human interference.

Despite its potential, manual labelling of acoustic data remains a critical bottleneck: processing thousands of hours of recordings is time-consuming, labour-intensive, and prone to inter-annotator discrepancies [4]. This paper evaluates and compares three ML architectures for autonomously labelling anomalies and identifying animal presence in marine records.

2 Related Work

The field of acoustic anomaly detection is defined by three primary technical paradigms.

Reconstruction-based Architectures. Auto-Encoders (AE) trained on mel-spectrogram features [2] learn a compressed representation of "normal" ambient noise [14]. Acoustic events that cannot be accurately reconstructed yield a high reconstruction error, signalling an anomaly. Researchers have enhanced this approach through noise suppression and denoising layers [10] to isolate target signals from background hydrophone noise (flow noise, wave action). A fundamental limitation is that these models may inadvertently learn to reconstruct consistent background patterns, reducing discrimination between normal and anomalous signals.

Autoregressive Density Estimation. Autoregressive (AR) methods [8] model the joint distribution of an acoustic representation as a product of conditionals, flagging observations that deviate from the learned normal likelihood. Erdil et al. [5] extended this paradigm to structured patch embeddings from a frozen self-supervised encoder, using a masked dilated CNN to model conditional dependencies across a 2D token grid. Dilated convolutions extend the receptive field without proportional cost, capturing long-range temporal structure across the full clip with sensitivity to localised deviations such as brief biological vocalizations.

Transformer-based Detection. Transformers [6] treat mel-spectrogram patches as tokens, using self-attention to model long-range dependencies across a recording. Perch 2.0 [11] represents the current state-of-the-art, combining a large-scale

linear classification head (14,795 species), a prototype-based similarity head, and a large-vocabulary source classification head using low-rank (512-dimensional) projections for scalable self-supervised learning. Given its demonstrated reliability, it serves as our reference baseline.

Existing literature focuses on incremental enhancements of single methods on standardised datasets. This paper addresses the need for cross-paradigm benchmarking on a custom marine dataset derived from elephant seal bio-loggers.

3 Methodology

This section describes the three methods: acoustic feature extraction via Perch 2.0, supervised fine-tuning via a ConvNeXt-based CNN, and autoregressive density estimation via AR-BEATs.

Shared Data Pipeline. The dataset [7] consists of Southern Ocean soundscape recordings from tagged elephant seals (32 kHz sampling, 20 Hz–16 kHz effective range). Animal movement introduces low-frequency flow noise, but during drift phases—passive gliding periods when the seal sleeps—recordings are high-quality and correlated with foraging activity. An unsupervised algorithm isolates these phases into a subdataset of ≈90% ambient noise and ≈10% anomalous events. Preprocessing involves: (1) **Segmentation** into 5-second windows with gain amplification; (2) **Indexing** via temporal metadata; and (3) **Ground Truth Isolation** from manually verified clean recordings. 100 samples were independently validated by Dr. Gros-Martial as the gold-standard test set; the remaining 1,202 samples serve as training and validation.

3.1 Perch 2.0

Each 5-second clip x_i is mapped to a 1,536-dimensional embedding $\boldsymbol{e}_i \in \mathbb{R}^{1536}$ by the frozen Perch 2.0 encoder. A reference centroid is computed from $N = 777$ validated normal samples:

$$\boldsymbol{V}_{\text{ref}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{e}_i, \tag{1}$$

under the assumption of a multivariate Gaussian distribution in latent space, making $\boldsymbol{V}_{\text{ref}}$ the maximum likelihood estimator of the nominal acoustic state. Anomaly scoring uses cosine distance:

$$D_{\cos}(\boldsymbol{V}_{\text{unk}}, \boldsymbol{V}_{\text{ref}}) = 1 - \frac{\boldsymbol{V}_{\text{unk}} \cdot \boldsymbol{V}_{\text{ref}}}{\|\boldsymbol{V}_{\text{unk}}\|_2 \|\boldsymbol{V}_{\text{ref}}\|_2}, \tag{2}$$

chosen for amplitude invariance. A clip is flagged anomalous if $D_{\cos} > \xi$, with $\xi = 0.3278$ selected by F1 maximisation on the self-annotated training set.

3.2 PyTorch CNN

A fully unfrozen ConvNeXt backbone [12] with a logistic regression head is fine-tuned end-to-end on mel-spectrograms. For logistic regression, training uses BCEWithLogitsLoss to combine Sigmoid activation with binary cross-entropy loss. The model captures the highest performing parameters by utilizing early stopping (patience = 5) and ReduceLROnPlateau scheduling (patience = 2, factor 0.6). Lastly, we incorporate Adam Optimizer L2 Regularization (lr = 5×10^{-5}, weight decay 10^{-4}). The architecture leverages ConvNeXt's established classification performance [3, 12] while adapting to the high-variance Southern Ocean environment through pooling layers for translation robustness.

3.3 AR-BEATs

AR-BEATs adapts the encoder–autoregressive framework of Erdil et al. [5] to underwater bioacoustics, replacing the vision encoder with BEATs [1]. Trained exclusively on 130 normal clips (10.8 min), it detects clips whose token-level density deviates from the learned distribution with no anomalous labels required.

Raw 5 s waveforms at 16 kHz are encoded into a 248-token grid of 768-dimensional embeddings (128-bin log-mel filterbank, 16×16 patches). The AR model factorises the joint distribution as:

$$p(E) = \prod_{i,j} \mathcal{N}\big(E_{i,j};\, \mu_{i,j}, \mathrm{diag}(\sigma^2_{i,j})\big)\,, \tag{3}$$

predicted by a 6-layer masked dilated CNN (kernel 3×3, dilation 8, hidden dim 512). Training minimises the negative log-likelihood with AdamW (lr=10^{-3}, weight decay 10^{-2}, batch 64). Each clip is scored as:

$$A_{\text{clip}} = \bar{A} + 2\sqrt{\tfrac{1}{N}\textstyle\sum_{i,j}(A_{i,j} - \bar{A})^2}, \tag{4}$$

capturing average surprise and spatial unevenness. The deployable threshold ξ_{p60} is the 60th percentile of validation scores.

Overall Metrics. F1-score is the primary metric given class imbalance, as it jointly accounts for precision and recall; PR-AUC provides threshold-independent comparison across methods.

4 Results and Evaluation

Evaluation conditions. The three methods were developed independently and evaluated under heterogeneous conditions: Perch 2.0 on 1,302 self-annotated samples (positive rate 59.4%); PyTorch CNN on 384 balanced samples randomly drawn with a fixed seed; and AR-BEATs on 1,159 samples (positive rate 66.8%). Threshold selection procedures also differ — Perch 2.0 and the CNN

use annotated data to maximise F1, placing them on equivalent footing with the AR-BEATs oracle rather than its label-free deployable threshold. Consequently, the broad comparison in Table 4 should be interpreted as indicative rather than definitive. The 100-sample expert-validated set, independently annotated by Dr. Gros-Martial and evaluated under identical conditions for all three methods, constitutes the single unified benchmark of this study and is the basis for the conclusions drawn in Table 5.

4.1 Perch 2.0

A reference background vector was established using 777 validated noise samples (mean per-dimension variance 0.1270). The optimal threshold $\xi = 0.3278$ was identified by sweeping candidates against 1,202 self-annotated samples, achieving PR-AUC 0.954 and F1 90.98% (precision 92.06%, recall 89.92%). The model's performance was then tested against the 100-sample gold-standard set validated by Dr. Gros-Martial (Fig 1), analysed under two criteria: *Exploration* (any acoustically unusual event) and *Human Validation* (confirmed animal vocalisation).

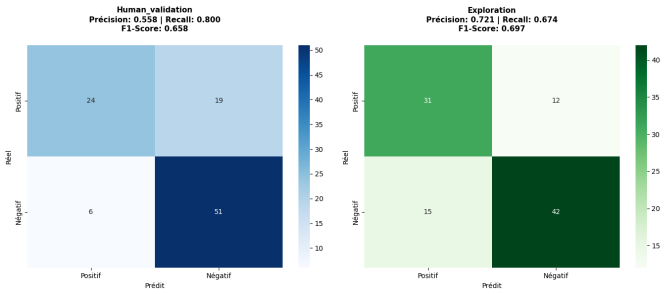

Fig. 1. Perch 2.0 confusion matrix on expert-validated set.

The results demonstrate robust precision during exploration and high recall for human validation (0.80). However, the model exhibits clear limitations for high-frequency signals: Dr. Gros-Martial identified four false negatives corresponding to sperm whale vocalizations (average $D_{\cos} = 0.24 \ll \xi$). As illustrated in Fig. 2, sperm whale clicks range from 8 to 25 kHz. With data sampled at 32 kHz, the Nyquist limit ($f_s/2 = 16$ kHz) aliases or filters the upper register, while Perch 2.0's terrestrial pre-training lacks the spectral weights to represent these high-frequency marine signals.

4.2 PyTorch CNN

The model was tested on 384 samples (190 Void, 194 Anomaly) using a dynamically calculated threshold ξ, resulting in a weighted **F1-score of 74.69%** (precision 76.08%, recall 75%) as shown in Table 1. Anomaly recall (85.56%)

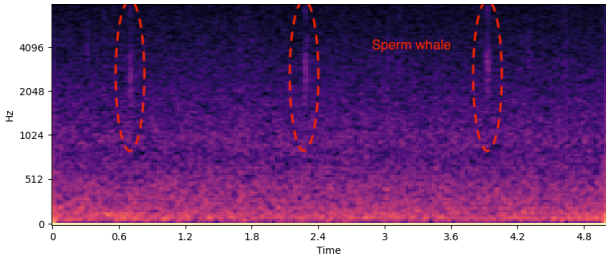

Fig. 2. Mel-spectrogram of a sperm whale click (ping).

exceeds void recall (64.21%), reflecting the supervised training's prioritisation of rare anomaly capture, at the cost of 83 false positives.

Table 1. PyTorch CNN classification report (384 test samples).

Class	Prec.	Rec.	F1	Support
Void (noise)	0.8133	0.6421	0.7176	190
Anomaly	0.7094	0.8556	0.7757	194
Wtd. avg	0.7608	0.7500	0.7469	384

4.3 AR-BEATs Performance

AR-BEATs is evaluated under two complementary label sets on the 100-sample expert-validated set: *Exploration* (any acoustically unusual event) and *Human Validation* (confirmed animal only). All 48 human-validated animal clips are contained within the 62 exploration-labelled clips; the remaining 14 were acoustically unusual but not confirmed as biological in origin.

On the large self-annotated test set (1,159 clips, positive rate 66.8%), AR-BEATs achieves **AUROC** 0.852 and **PR-AUC** 0.906 against a random baseline of 0.668. The 60th-percentile threshold ξ_{p60} achieves F1 = 0.870, within 0.0007 of the oracle upper bound, confirming near-perfect calibration from a single minute of validation audio (Table 2).

High anomaly recall (92.3%) at the cost of moderate void recall (60.0%) is inherent to one-class training: any deviation from the learned ambient distribution is flagged, naturally favoring sensitivity. On the expert-validated set, both label sets are evaluated at the same deployable threshold $\xi_{p90} = -411.47$, set as the 90th percentile of the 38 exploration normal scores (Table 3). Both criteria produce the exact same 60 flagged clips, confirming that the threshold is stable across evaluation conditions.

The 60 flagged clips contain *all* 48 confirmed animal vocalisations, reducing expert review workload by 40% on this subset, with all 48 expert-confirmed an-

Table 2. AR-BEATs at ξ_{p60} (1,159 test clips). PR-AUC = 0.906, AUROC = 0.852. Trained on 130 clips (10.8 min), no anomalous labels. Oracle-to-deployable F1 gap: 0.0007.

Class	Prec.	Rec.	F1	Support
Void (noise)	0.7938	0.6000	0.6834	385
Anomaly	0.8226	0.9225	0.8697	774
Wtd. avg	0.8134	0.8154	0.8078	1 159

Table 3. AR-BEATs on expert-validated set ($N = 100$, $\xi_{p90} = -411.47$, flags 60/100 clips).

Criteria	TP	FP	FN	TN	Prec.	Rec.	F1
Exploration	56	4	6	34	0.93	0.90	**0.918**
Human Val.	42	18	6	34	0.70	0.88	**0.778**

$FN = 6$ is identical across both criteria — the same 6 clips are missed regardless of label set. The 14-clip gap between $FP = 4$ (exploration) and $FP = 18$ (human validation) corresponds exactly to the 14 unusual but non-biological events. All 48 confirmed animal clips are recovered.

imal vocalisations recovered — noting that this claim is bounded by the scale of the expert-validated set. The 14 clips flagged by AR-BEATs as acoustically unusual but not confirmed as biological events warrant further discussion. As a one-class model, AR-BEATs flags any deviation from the learned ambient distribution regardless of its origin: these false positives under the human validation criterion likely correspond to transient non-biological sources such as anthropogenic noise (shipping, sonar), hydrophone flow artefacts during brief movement phases, or other marine environmental sounds outside the normative acoustic profile of the training recordings. From a monitoring perspective, flagging such events is not necessarily an error — it reflects the model operating as designed, alerting to any acoustic anomaly and leaving the biological interpretation to the expert reviewer. A dedicated classifier for non-biological transients could be applied as a post-processing step to further reduce false alarms without modifying the core detection pipeline.

4.4 Comparative Analysis

Tables 4 and 5 present the broad and expert-validated comparisons. Fig. 3 illustrates detection behaviour on a representative ≈7-minute recording: Perch 2.0 (orange) closely tracks human annotations; the CNN (green) misses several event clusters; AR-BEATs (red) produces the most temporally faithful detections with few spurious flags.

Table 4. Broad performance comparison. †: supervised.

Method	PR-AUC	F1 (deploy.)
Perch 2.0	**0.954**	0.910
PyTorch CNN†	0.876	0.747
AR-BEATs	0.906	0.870

Table 5. Performance on expert-validated set ($N = 100$). †: supervised. Exploration: any unusual event. Human Val.: confirmed animal vocalisation only.

Method / Criteria	TP	FP	FN	TN	F1
Perch 2.0 (Exploration)	31	12	15	42	0.70
Perch 2.0 (Human Val.)	24	19	6	51	0.66
CNN† (Exploration)	57	25	5	13	0.67
CNN† (Human Val.)	43	39	5	13	0.51
AR-BEATs (Exploration)	56	4	6	34	**0.918**
AR-BEATs (Human Val.)	42	18	6	34	**0.778**

AR-BEATs: $\xi_{p90} = -411.47$*, applied identically to both label sets. Perch 2.0:* $\xi = 0.3278$*. PyTorch CNN F1 from large-scale evaluation.*

On the expert-validated set, Perch 2.0 misses 15 exploration anomalies (F1 0.70); AR-BEATs and the CNN achieve higher recall (5–6 misses), but AR-BEATs produces far fewer false positives (4 vs. 25), yielding the best exploration F1 (0.918). The CNN's human-validation F1 collapses to 0.51 (39 FPs), reflecting poor generalisation to the stricter expert criterion. On the large self-annotated dataset, Perch 2.0 leads (PR-AUC 0.954, F1 91.0%), while AR-BEATs reaches F1 87.0% using only 130 normal clips with no anomalous labels, recovering all 48 expert-confirmed animal vocalisations in this subset, across its 60 flagged clips and reducing expert review by 40%.

5 Conclusion

This study benchmarked three machine learning paradigms for automated anomalous sound detection in Southern Ocean elephant seal recordings. Perch 2.0 achieves the strongest large-scale performance (PR-AUC 0.954, F1 91.0%), supported by 777 manually validated normal samples. The supervised PyTorch CNN reaches F1 74.7% but requires labelled anomalous data, limiting generalisation to unseen vocalization types.

On this expert-validated subset, AR-BEATs shows higher performance than both alternatives under both criteria, achieving F1 0.918 for general acoustic exploration and F1 0.778 for confirmed animal detection, using a single threshold $\xi_{p90} = -411.47$ calibrated from normal clips with no anomalous labels. Its 60

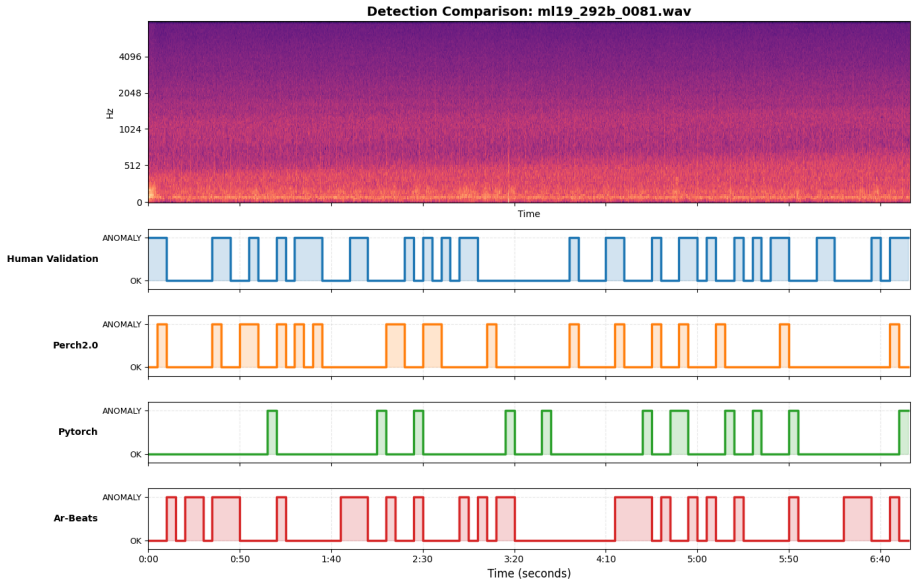

Fig. 3. Detection comparison on `ml19_292b_0081.wav`. *Top:* Mel-spectrogram (128 bins, $f_{max} = 8000$ Hz, 16 kHz sampling, power converted to dB). *Lower panels:* binary anomaly decisions for Human Validation (blue), Perch 2.0 (orange), PyTorch CNN (green), and AR-BEATs (red) segmented in 5-second steps.

flagged clips recover all 48 confirmed animal vocalisations, reducing expert review workload by 40% with zero missed biological events — including the sperm whale clicks that were missed by Perch 2.0 at its fixed threshold ($\xi = 0.3278$, average click cosine distance $0.24 \ll \xi$).

This distinction highlights the core deployment trade-off. AR-BEATs is the method of choice when missing any biological event is unacceptable: its one-class training flags any deviation from ambient conditions, making it robust to unknown or rare vocalization types regardless of frequency content. Perch 2.0 is a strong alternative when higher precision is prioritised, but its threshold must be recalibrated for each new acoustic environment — a fixed ξ tuned on one dataset will not generalise reliably to recordings with different ambient noise profiles or species compositions.

Future work should prioritise marine-specific self-supervised pre-training and higher-rate hydrophone recordings to address the Nyquist aliasing of high-frequency clicks (8–25 kHz at 32 kHz sampling), which remains a shared limitation for feature-extraction-based methods.

References

1. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: BEATs: Audio pre-training with acoustic tokenizers. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 5178–5193 (2023)
2. Chinnasamy, M.D., Sumbwanyambe, M., Hlalele, T.S.: Acoustic Anomaly Detection of Machinery using Autoencoder based Deep Learning pp. 1–6 (Jan 2024). https://doi.org/10.1109/SAUPEC60914.2024.10445053, https://ieeexplore.ieee.org/document/10445053/
3. Chunhachatrachai, P., Bahrudin, Lin, C.Y.: Comparative analysis of pre-trained cnn backbones and machine learning techniques for high similarity in circular object image classification. arXiv preprint / (or conference/journal name if known) (2024), national Taiwan University of Science and Technology, National Research and Innovation Agency
4. Dubus, G., Torterotot, M., Duc, P.N.H., Beesau, J., Cazau, D., Adam, O.: Better quantifying inter-annotator variability: A step towards citizen science in underwater passive acoustics. In: OCEANS 2023 - Limerick. pp. 1–8 (2023). https://doi.org/10.1109/OCEANSLimerick52467.2023.10244502
5. Erdil, E., Schulthess, N., Tombak, G., Konukoglu, E.: Spatial autoregressive modeling of DINOv3 embeddings for unsupervised anomaly detection. In: arXiv preprint arXiv:2603.02974 (2026)
6. Ferreira, A.I.S., Da Silva, N.F.F., Mesquita, F.N., Rosa, T.C., Buchmann, S.L., Mesquita-Neto, J.N.: Transformer Models improve the acoustic recognition of buzz-pollinating bee species. Ecological Informatics **86**, 103010 (May 2025). https://doi.org/10.1016/j.ecoinf.2025.103010, https://linkinghub.elsevier.com/retrieve/pii/S1574954125000196
7. Gros-Martial, A.: Assessment of wind, rain, and cetacean biodiversity in the southern ocean using passive acoustic monitoring from biologged elephant seals . https://doi.org/10.70675/229eae1az6371z4a48z81a1zccd267eda349, https://theses.fr/2025BRUN0090
8. Hayashi, T., Komatsu, T., Kondo, R., Toda, T., Takeda, K.: Anomalous Sound Event Detection Based on WaveNet. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 2494–2498. IEEE (Sep 2018). https://doi.org/10.23919/EUSIPCO.2018.8553423, https://ieeexplore.ieee.org/document/8553423/
9. Janik, V.M.: Acoustic communication in delphinids. Advances in the Study of Behavior **40**, 123–157 (2009)
10. Kim, S.M., Soo Kim, Y.: Enhancing Sound-Based Anomaly Detection Using Deep Denoising Autoencoder. IEEE Access **12**, 84323–84332 (2024). https://doi.org/10.1109/ACCESS.2024.3414435, https://ieeexplore.ieee.org/document/10557607/
11. van Merriënboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., Denton, T.: Perch 2.0: The bittern lesson for bioacoustics. https://doi.org/10.48550/ARXIV.2508.04665, https://arxiv.org/abs/2508.04665, version Number: 2
12. Pant, K., Pramanik, P.K.D., Zhao, Z.: A robust convnext-based framework for efficient, generalizable, and explainable brain tumor classification on mri. Bioengineering (Basel, Switzerland) **13**(2), 157 (2026). https://doi.org/10.3390/bioengineering13020157
13. Penn, J.L., Deutsch, C.: Avoiding ocean mass extinction from climate warming. Science **376**(6592), 524–526 (2022). https://doi.org/10.1126/science.abe9039, https://www.science.org/doi/10.1126/science.abe9039

14. Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., Kawaguchi, Y.: MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. https://doi.org/10.48550/arXiv.1909.09347, http://arxiv.org/abs/1909.09347
15. Richard, G., Bonnel, J., Beesau, J., Calvo, E., Cassiano, F., Dramet, M., Glaziou, A., Korycka, K., Guinet, C., Samaran, F.: Passive acoustic monitoring reveals feeding attempts at close range from soaking demersal longlines by two killer whale ecotypes. Marine Mammal Science **38**(1), 304–325 (2022)
16. Savoca, M.S., Kumar, M., Sylvester, Z., Czapanskiy, M.F., Meyer, B., Goldbogen, J.A., Brooks, C.M.: Whale recovery and the emerging human-wildlife conflict over antarctic krill. Nature Communications **15**(1), 7708 (2024)