Supervised detection of baleen whale calls on edge-compute

Astrid van Toor[1]

[1]blueOASIS, Ericeira, Portugal

***Abstract*—This paper presents an edge-optimised approach for baleen whale call detection, addressing both the detection requirements of BioDCASE 2025 Task 2 and deployment constraints similar to Task 3. Common machine learning models contain 4+ million training parameters and use architectures unsuitable for real-time edge deployment. In contrast, our model contains just 35,571 parameters (159KB) and operates efficiently on a 64-bit ARM Cortex-A53 with 512MB RAM. We applied an edge-optimised feature extraction pipeline and a custom CNN model architecture designed for real-time inference in offshore deployments. Classifying on 11.8-second detection windows, our precision-focused approach achieves 72% precision for blue whale ABZ calls and 80% for fin whale burst pulse calls, though downsweep detection lags at 18% precision. After applying a temporal head for call-specific identification as per the BioDCASE challenge requirements, ABZ call precision drops to 65% and burst pulse calls to 4%, while downsweep calls improve to 29%. Acknowledging the difficulties in call-specific identification, this work highlights the feasibility and potential of edge-optimised architectures for baleen whale detection in real-world monitoring scenarios where computational resources and power consumption are severely constrained, while addressing common challenges and next steps to improve the results.**

***Index Terms*—baleen call detection, signal processing, temporal attention, edge-computing**

1. INTRODUCTION

Our marine environment face significant threats [1], [2], requiring scalable bioacoustics monitoring solutions to assess biodiversity and support conservation. Passive Acoustic Monitoring (PAM) offers a promising non-invasive approach for underwater monitoring but transitioning deep learning (DL) based PAM from research prototypes to operational systems remains challenging, often requiring performance trade-offs through quantization or pruning [3]. Furthermore, current DL approaches often report high accuracies based on biased evaluation protocols that do not consistently account for temporal correlations in acoustic data [4], [5]. By following the guidelines of the BioDCASE 2025 Challenge [6] for the "Supervised Detection of Strongly-Labelled Whale Calls" we aim to reduce this validation bias in real-world deployments and encourage further development.

2. BACKGROUND

Although recent deep learning approaches have shown promising results for automated baleen whale call detection [7]–[11], and traditional spectrogram-based detection methods can achieve low false alarm rates [12], [13], these methods require significant computational resources making them unsuitable for autonomous real-time deployment on memory-constrained devices. Hybrid CNN-LSTM approaches for fin whale detection show that temporal context modelling can improve performance [14], but remain to be benchmarked on edge devices. While specialised systems like DMON/LFDCS [15] have demonstrated successful deployment on autonomous platforms for baleen whale detection, this approach has key limitations - most notably reliance on human expert analysts for final classification decisions and the constraints of data transmission. In contrast compact CNNs can process audio on-board for fully autonomous analysis, allowing for data transmission of just several bytes containing detection information like time, location, and species, rather than large volumes of pitch track data requiring human review.

This work focuses on convolutional neural networks (CNNs) due to their proven success in audio signal processing for animal sounds on edge compute [16]–[18], effectiveness with limited compute and minimal memory footprint [19], and widespread deployment in marine bioacoustic tasks including cetacean detection [7], [9]–[11], [20]–[23]. While recurrent architectures (RNNs/CRNNs) have shown promise in bioacoustics [23], they require more computational resources for training and inference, with [23] observing that temporal CNNs often match or exceed RNN performance while being faster to train. Similarly, although transformer architectures show emerging potential with [24], [25] demonstrating early bioacoustic applications, they remained computationally prohibitive for edge deployment until very recently [26], with advances like EasyViT [27] only now making them feasible for resource-constrained devices.

For this initial edge-optimised implementation targeting immediate deployment on devices like the Raspberry Pi Zero 2 W, CNNs provide the optimal balance of proven performance, computational efficiency, and implementation maturity, with exploration of emerging architectures reserved for future iterations once the baseline system is established. This report presents a development pathway for addressing common challenges: an edge-optimised temporal attention network designed specifically for deployment on tiny low-cost hardware platforms. This network uses an attention mechanism to assign importance to key features in an audio signal, helping the network focus on relevant sounds for call-specific detection, while achieving a balance between detection performance and computational efficiency and laying out clear next steps for development.

3. METHODOLOGY

3.1. Dataset and Preprocessing

The BioDCASE benchmark builds on the baleen whale call dataset containing 1,880 hours of recordings with expert annotations for Antarctic blue whale ABZ calls (bmabz), fin whale burst pulses (bp), and fin/blue whale downsweep calls (d) [28]. Following the challenge protocol, we maintained strict temporal separation between training and validation sets to prevent data leakage. Applying a sliding windowing approach of 11.8s windows (theorising the necessity of longer windows for transient whale sounds and burst calls) with 50% overlap, and undersampling of the heavily over-represented background down to 60% in the training set, we obtained 473,620 training windows and 351,374 validation windows. The overall class distribution is presented in Table 1. We designed the model as a multi-class labelling problem with a confidence score for the presence of each class per window.

Table 1: Class Distribution

Class	Train	Val
Background	362,944	312,730
Blue Whale ABZ (bmabz)	64,846	28,931
Fin Whale Pulse (bp)	26,677	6,301
Fin/Blue Downsweep (d)	31,615	5,581
Total	**473,620**	**351,374**

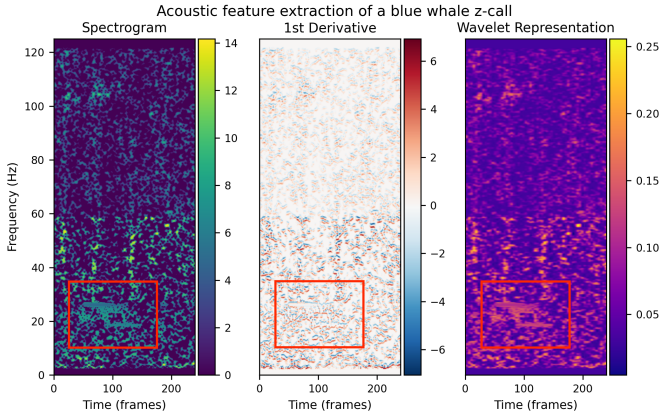

Fig. 1: CNN Feature stack for blue and fin whale calls on a 48-second audio snippet. Example shows a blue whale ABZ call [28]. In production these are 11.8-second windows.

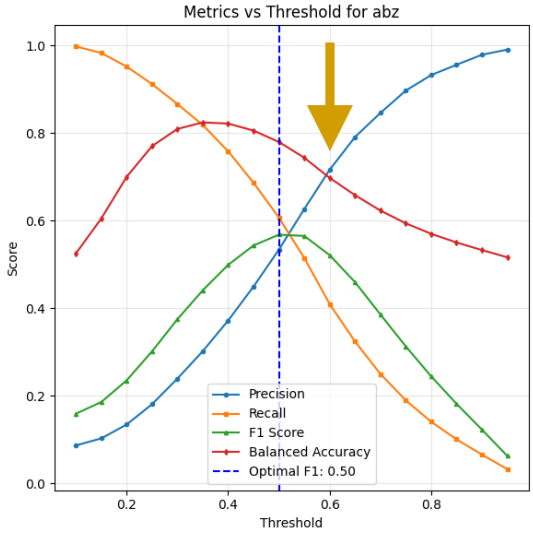

Fig. 2: Threshold analysis approach demonstrated in the blue whale ABZ call class on the evaluation set. "Optimal F1" is maximised F1, the arrow highlights our selected threshold of 0.6 for optimised precision.

Our preprocessing pipeline extracts three-channel acoustic features (Fig. 1) optimised for low-frequency whale vocalisations (5-125Hz range): log-power spectrograms with 250Hz sampling rate, first-order derivatives, and computationally efficient wavelet coefficients processing only approximation and first-level detail coefficients. This adapted wavelet feature was inspired by works that previously presented the effectiveness of wavelet features in identifying low frequency whale calls [8], [29]. Channel-wise normalisation parameters were computed from the training set to ensure consistent scaling across diverse recording conditions. Table 2 details the complete processing pipeline to transform the raw acoustic recordings into standardised sensors suitable for deep learning inference on edge hardware.

3.2. Model Architecture

The proposed network employs a lightweight CNN backbone with six convolutional layers, utilising separable convolutions to reduce computational complexity. A detailed diagram of the network architecture is available from https://doi.org/10.5281/zenodo.17122226.

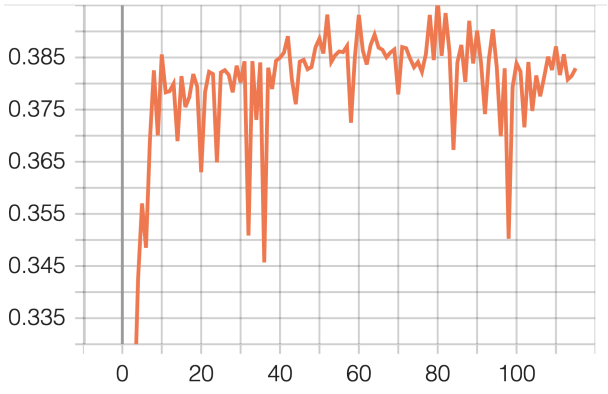

Fig. 3: Weighted F1-Macro learning curve on the classification head across the detection of background noise, blue whale ABZ, fin whale pulse calls, and fin/blue whale downsweep calls

Key architectural features include:

- A multi-class label classification head for per-window predictions
- Asymmetric max-pooling to reduce the frequency dimension while preserving full temporal resolution for precise event localisation
- Multi-scale temporal processing via parallel dilated convolutions to capture patterns across different time scales
- A lightweight temporal attention mechanism to weigh and aggregate features across 60 time steps
- Batch normalisation throughout for training stability
- Focal loss optimisation addressing severe class imbalance
- Cyclical learning rate scheduling for improved convergence

The model contains only 35,571 training parameters and is just 159KB in size, representing a 95-99% size reduction compared to standard and edge-optimised CNNs [30], enabling deployment on remote memory-constrained devices using Tensorflow Lite (TFLite).

3.3. Thresholding

Performance evaluation followed a precision-focused approach, recognising that in ocean sustainability and conservation false positives often incur higher costs than false negatives. We therefore implemented class-specific detection thresholds optimised for precision on the classification head. An example of this approach is presented in Fig. 2. Considering the multi-class problem, the background class is only activated when no other classes reach the threshold.

3.4. Weighted monitoring metric

For the early stopping mechanism, we employed a weighted F1 macro metric for a weighted precision-recall rating favouring precision at 70% over 30% (Fig. 3).

The weighted F1 score for each class i is computed as shown in (1),

$$F_{1,w}^{(i)} = \frac{(w_p + w_r) \cdot P^{(i)} \cdot R^{(i)}}{w_p \cdot R^{(i)} + w_r \cdot P^{(i)} + \epsilon}, \tag{1}$$

where $P^{(i)}$ and $R^{(i)}$ are the precision and recall for class i, respectively, w_p is the precision weight, $w_r = 1 - w_p$ is the recall weight, and $\epsilon = 10^{-8}$ is a small constant to prevent division by zero. The precision and recall are calculated using the standard definitions in (2),

$$P^{(i)} = \frac{TP^{(i)}}{TP^{(i)} + FP^{(i)} + \epsilon}, \quad R^{(i)} = \frac{TP^{(i)}}{TP^{(i)} + FN^{(i)} + \epsilon}, \tag{2}$$

Table 2: Feature Extraction Pipeline for Blue and Fin Whale Detection. F = frequency bins (513), T = time frames per window (60), Output tensor dimensions: $(513 \times 60 \times 3)$ per analysis window. Processing optimised for 5-125Hz whale vocalisations on edge hardware with under 40k training parameters CNN models.

Processing Stage	Parameter	Implementation Details	Output
Signal Preprocessing	Resampling	Target sampling rate: 250Hz using librosa resample	250Hz audio
	Bandpass Filter	4th-order Butterworth filter, 5-125Hz cutoff frequencies	Filtered signal
Spectrogram Generation	Window Function	2.0s Hann window (500 samples at 250Hz)	STFT frames
	Hop Length	0.2s hop (50 samples, 90% overlap)	Time resolution
	FFT Size	1024-point FFT providing 0.244Hz frequency resolution	513 freq bins
	Power Conversion	Log-power: $S_{dB} = 10\log_{10}(\|STFT\|^2) - \max(S_{dB})$	dB spectrogram
Spectral Enhancement	Background Subtraction	Time-averaged profile with 15-sample Gaussian smoothing	Enhanced signal
	Contrast Enhancement	5-120Hz whale frequency band	Band-enhanced
Normalisation	Percentile Scaling	Map 5th-95th percentile to [-20, +20] dB range	Standardised
Multi-Channel Features	Channel 1	Log-power spectrogram	$(F \times T)$
	Channel 2	First-order frequency derivatives with safety checks	$(F \times T)$
	Channel 3	Daubechies-4 wavelet (3 levels, approx + detail coeffs)	$(F \times T)$
Temporal Windowing	Window Size	60 frames (11.8s at 0.2s hop)	Feature windows
	Overlap	30 frames (50% overlap between windows)	Sliding windows
Final Standardisation	Channel-wise Norm	Training-set calculated mean/std applied to all data splits	$(F \times T \times 3)$

where $TP^{(i)}$, $FP^{(i)}$, and $FN^{(i)}$ represent the true positives, false positives, and false negatives for class i, respectively. The final monitoring metric is the macro-averaged weighted F1 score across all classes as given in (3),

$$F_{1,weighted-macro} = \frac{1}{C}\sum_{i=1}^{C} F_{1,w}^{(i)}, \tag{3}$$

where C is the total number of classes. In our experiments, we used a precision weight of $w_p = 0.7$ to emphasise precision over recall in the early stopping criterion.

3.5. Post Processing of the Temporal Head

The temporal head processes attention weights generated by the Temporal Attention Layer (see also the architecture diagram in https://doi.org/10.5281/zenodo.17122226). This layer produces time-step-specific attention scores across the 60 temporal frames (11.8-second window), where each score indicates the model's focus on potential whale call events at that time point. These attention weights form a temporal attention map that highlight regions of acoustic significance, which we then process to extract precise call boundaries. Class-specific processing parameters were derived from statistical analysis of the training set (Table 3).

The temporal processing pipeline applies class-specific attention weight smoothing using median filters (kernel sizes: bmabz=5, bp=1, d=1), followed by adaptive thresholding based on attention statistics. High-attention regions exceeding $\mu + \alpha\sigma$ (where $\alpha \in [0.05, 0.08, 0.1]$ for 'bp', 'd', 'mbabz') are identified as potential call boundaries. Duration constraints derived from 5th and 95th percentiles filter detected events: bmabz (4.69-13.62s), bp (0.92-1.93s), and d (0.83-4.40s).

Post-processing applies class-specific merging of nearby events (max gaps: bmabz=8.0s, bp=1.0s, d=1.5s), overlap-based deduplication (thresholds: 0.2-0.3), and confidence filtering to produce final temporal boundaries. Events are ranked by a composite score combining classification confidence (40%), peak attention (40%), and duration bonus (20%) when resolving overlapping detections.

3.6. Computational Performance

Training was done on 4 NVIDIA A100 GPUs with distributed data-parallel processing and XLA, completing in 10 hours for the 474K training windows. Processing occurred with sharded TFRecord files across 16 parallel workers. Implementation includes deterministic data shuffling, fixed random seeds, and a batch size of 128.

To evaluate the feasibility of real-time, on-device deployment, the model's computational performance was benchmarked on a resource-constrained edge device: a Raspberry Pi Zero 2 W which features a 64-bit ARM Cortex-A53 CPU and 512 MB of RAM. The performance was measured across five 60-second audio files with varied whale call patterns, with multiple trials ensuring stable results. The system demonstrates exceptional efficiency, achieving a mean real-time factor (RTF) of 24.5×, indicating it can process audio over 24 times faster than it is recorded. The complete end-to-end processing of a 60-second audio clip takes, on average, just 2.45 seconds. The compressed TFLite model has a minimal disk footprint of only 159KB. The inference latency for a single 11.8-second analysis window is 187.42 ± 1.07 ms. The temporal post-processing adds negligible overhead. The peak memory footprint of the application during runtime was 213.1 MB, well within the device's operational limits. These results confirm the model's suitability for long-term, low-power, and real-time acoustic monitoring applications. Key performance metrics are summarised in Table 4.

4. RESULTS AND DISCUSSION

Bearing in mind the computational constraints and potential of the edge-optimised architecture, we present both the per-window analysis and the evaluation following the BioDCASE benchmark.

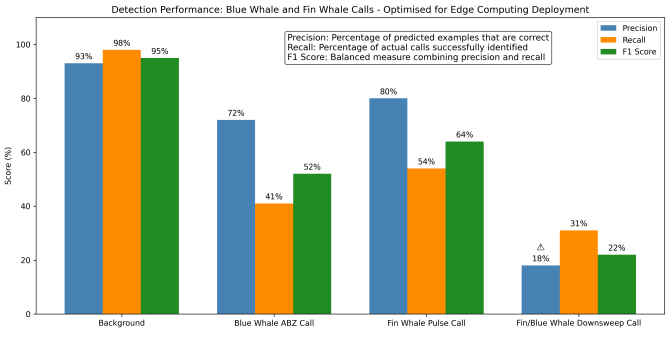

Fig. 4: Performance metrics per class in the classification head.

Starting with per-window analysis from the classification head, Fig. 4, we achieve a precision of 72% for the blue whale ABZ call and 80% for the fin whale burst calls, but poor precision of 18% on downsweep - reflective of previous reports on the challenges of

Table 3: Training set duration statistics (seconds) for merged annotation classes used to derive class-specific temporal processing parameters.

Class	Count	Mean	Std	Median	Min	Max	P5	P10	P25	P75	P90	P95	IQR	CV
bmabz	9463	7.95	2.76	7.36	1.29	36.62	4.69	5.23	6.14	9.09	11.60	13.62	2.95	0.35
bp	5308	1.39	0.31	1.38	0.46	2.82	0.92	1.01	1.20	1.60	1.82	1.93	0.40	0.22
d	2856	2.44	1.11	2.41	0.37	7.36	0.83	0.98	1.56	3.19	3.93	4.40	1.63	0.46

Table 4: Detailed performance breakdown of the whale call detection system on Raspberry Pi Zero 2 W.

Component	Metric	Value
Model	Size (MB)	0.15
	Input shape	[1, 513, 60, 3]
Inference	Latency per window (ms)	187.42 ± 1.07
	95th percentile (ms)	189.21
	Real-time factor (RTF)	24.5×
Temporal Processing	Latency per window (ms)	0.79 ± 0.17
	Overhead vs inference	0.4%
Resource Usage	Peak memory (MB)	213.1
	Peak CPU (%)	155.6

Table 5: Detection performance results as per the BioDCASE evaluation across datasets for different classes (bmabz, bp, d).

Dataset	Method	TP	FP	FN	Recall	Precision
casey2017	bmabz	676	446	1742	0.280	0.602
	bp	0	13	292	0.000	0.000
	d	77	339	476	0.139	0.185
kerguelen2014	bmabz	473	154	3824	0.110	0.754
	bp	4	56	3742	0.001	0.067
	d	70	138	709	0.090	0.337
kerguelen2015	bmabz	329	215	2419	0.120	0.605
	bp	1	39	1269	0.001	0.025
	d	126	181	1398	0.083	0.410
Final Results	bmabz	1478	815	7985	0.156	0.645
	bp	5	108	5303	0.001	0.044
	d	273	658	2583	0.096	0.293

labelling and predicting this class [7]. Recall rates vary between 31%, 41% and 54% for downsweep, ABZ, and burst pulse calls respectively.

The results of ‘evaluation.py‘ provided by the BioDCASE benchmark are presented in Table 5. There is a noticeable drop in performance, which was somewhat expected given the minimal temporal head restriced to edge limitations. Nevertheless, 65% precision for ‘bmbabz‘ is promising for a model with just 35K training parameters. The temporal processing pipeline may benefit from further tuning to enhance both precision and recall for this class.

The significant drop in precision for the burst pulse calls was unanticipated given its high performance in the classification head. It seems that the 11.8-second window approach is fundamentally mismatched to the 1.39-second average calls. Despite custom class tuning, the temporal attention mechanism lacks the resolution needed for short events within the 11.8-second windows. Downsweep calls being longer interestingly achieved higher precision than both temporal burst pulse calls and the ‘d‘ classification head at 29%, but recall is low.

Given the considerable characteristic differences between the targeted events both in the time and frequency domain and the extremely efficient processing pipeline, future endeavours might seek to build custom per-class inference models with feature extraction tailored specifically for each individual class. Three models could run sequentially in real-time, and still run smoothly on tiny microcontrollers such as the one used in this experiment (Raspberry Pi Zero 2 W, 512MB RAM). Such class-specific pipelines could significantly enhance temporal call-specific detection accuracy.

We aimed to develop a lightweight solution to detect whale presence in real-time on edge. Considering PAM of baleen whales for real-time applications such as adapting shipping routes based on mammal presence, we argue that the windowing approach in the classification head would be sufficient and perhaps preferable over individual call detection when deployed on edge devices. The impact of inconsistent labelling is reduced since the classification simply determines whether a class is present or absent within a given window, which is sufficient to support policy decisions. Additionally, once the data is retrieved, high-performance computing (HPC) analysis can be performed on land to refine the predictions. Statistical inference of call presence could be adjusted for known precision and recall errors, though hydrophone hardware and environmental differences must be taken into account. In production, the classification head could serve as an initial detection stage, while the temporal head attempts to locate exact call boundaries when required.

While the margin of 24.5× exceeds the strict requirement of real-time operation, this computational headroom ensures robust operation when edge devices must simultaneously handle data processing and transmission across potentially parallel models - realistic scenarios for autonomous ocean monitoring. The presented architecture offers scalability and flexibility in development paths: vertical scaling through deeper networks and expanded channels could improve accuracy at moderate computational costs, while horizontal scaling via class-specific pipeline appears promising given the distinct characteristics of each call type. Initial experiments with GRU layers showed potential for improved temporal modelling but required custom TensorFlow Lite operators for edge compression, this remains a route to explore. Future work will explore lightweight vision transformers as they mature for edge applications, layer normalisation and cosine learning rate schedulers for training stability, and systematic benchmarking to identify optimal accuracy-efficiency trade-offs.

It has to be noted that all validation datasets are recorded with the ”AAD-MAR” hydrophone. As there can be considerable differences in acoustic data recorded with different hardware, performance on hydrophones outside of this domain remains to be addressed. For example, only 12% of the training data was recorded with ”AAD-MAR”, whereas 77% of the training data was recorded with ”AURAL”. Given that the vast majority of training data comes from a single hydrophone type, the model may perform better on hardware within this domain (AURAL). Additionally, hydrophone calibration across hardware and environments must be considered for production.

5. CONCLUSION

This work demonstrates the feasibility of edge-optimised deep learning for baleen whale detection, achieving competitive performance on the classification head with just 35K parameters. The proposed model is suitable for edge compute and real-time detection. While temporal localisation remains challenging, particularly for short burst pulse calls, the classification head provides reliable presence detection suitable for real-time conservation applications. Future work should explore class-specific models, model scaling, and cross-hydrophone generalisation to improve deployment robustness.

6. ACKNOWLEDGMENT

A.v.T. thanks blueOASIS for supporting this research and Miller et al. [28] for providing such a comprehensive public dataset to aid advancements in the field of marine mammal detection.

REFERENCES

[1] K. Dube, "A Comprehensive Review of Climatic Threats and Adaptation of Marine Biodiversity," *Journal of Marine Science and Engineering*, vol. 12, no. 2, p. 344, Feb. 2024.

[2] R. Ahmed and M. T. R. Tamim, "Marine and Coastal Environments: Challenges, Impacts, and Strategies for a Sustainable Future," *International Journal of Science Education and Science*, vol. 2, no. 1, pp. 53–60, Mar. 2025.

[3] M. G. S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, and F. Hussain, "Machine Learning at the Network Edge: A Survey," *ACM Comput. Surv.*, vol. 54, no. 8, pp. 170:1–170:37, Oct. 2021.

[4] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, and K. Sammut, "A Survey of Underwater Acoustic Data Classification Methods Using Deep Learning for Shoreline Surveillance," *Sensors*, vol. 22, no. 6, p. 2181, Jan. 2022.

[5] X. Luo, L. Chen, H. Zhou, and H. Cao, "A Survey of Underwater Acoustic Target Recognition Methods Based on Machine Learning," *Journal of Marine Science and Engineering*, vol. 11, no. 2, p. 384, Feb. 2023.

[6] BioDCASE, "BioDCASE Challenge - BioDCASE," https://dcase.community/challenge2025/index, 2025.

[7] E. Schall, I. I. Kaya, E. Debusschere, P. Devos, and C. Parcerisas, "Deep learning in marine bioacoustics: A benchmark for baleen whale detection," *Remote Sensing in Ecology and Conservation*, vol. 10, no. 5, pp. 642–654, 2024.

[8] F. Sattar, "A New Acoustical Autonomous Method for Identifying Endangered Whale Calls: A Case Study of Blue Whale and Fin Whale," *Sensors*, vol. 23, no. 6, p. 3048, Jan. 2023.

[9] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine Mammal Species Classification Using Convolutional Neural Networks and a Novel Acoustic Representation," in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds. Cham: Springer International Publishing, 2020, pp. 290–305.

[10] J. H. Rasmussen and A. Širović, "Automatic detection and classification of baleen whale social calls using convolutional neural networks," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3635–3644, May 2021, publisher: AIP Publishing. [Online]. Available: https://pubs.aip.org/asa/jasa/article/149/5/3635/607542/Automatic-detection-and-classification-of-baleen

[11] D. Wang, L. Zhang, Z. Lu, and K. Xu, "Large-Scale Whale Call Classification Using Deep Convolutional Neural Network Architectures," in *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Sep. 2018, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8567758

[12] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, pp. 39–47, Jun. 2004. [Online]. Available: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1586

[13] M. F. Baumgartner and S. E. Mussoline, "A generalized baleen whale call detection and classification system," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 2889–2902, May 2011.

[14] S. Madhusudhana, Y. Shiu, H. Klinck, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, A. Širović, and M. A. Roch, "Improve automatic detection of animal call sequences with temporal context," *Journal of The Royal Society Interface*, vol. 18, no. 180, p. 20210297, Jul. 2021, publisher: Royal Society. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsif.2021.0297

[15] M. F. Baumgartner, J. Bonnell, P. J. Corkeron, S. M. Van Parijs, C. Hotchkin, B. A. Hodges, J. Bort Thornton, B. L. Mensi, and S. M. Bruner, "Slocum Gliders Provide Accurate Near Real-Time Estimates of Baleen Whale Presence From Human-Reviewed Passive Acoustic Detection Information," *Frontiers in Marine Science*, vol. 7, Feb. 2020, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2020.00100/full

[16] T. Mahbub, A. Bhagwagar, P. Chand, I. Zualkernan, J. Judas, and D. Dghaym, "Bat2Web: A Framework for Real-Time Classification of Bat Species Echolocation Signals Using Audio Sensor Data," *Sensors*, vol. 24, no. 9, p. 2899, Jan. 2024, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/24/9/2899

[17] J. Miquel, L. Latorre, and S. Chamaillé-Jammes, "Energy-Efficient Audio Processing at the Edge for Biologging Applications," *Journal of Low Power Electronics and Applications*, vol. 13, no. 2, p. 30, Jun. 2023, publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9268/13/2/30

[18] Z. Huang, A. Tousnakhoff, P. Kozyr, R. Rehausen, F. Bießmann, R. Lachlan, C. Adjih, and E. Baccelli, "TinyChirp: Bird Song Recognition Using TinyML Models on Low-power Wireless Acoustic Sensors," Sep. 2024, arXiv:2407.21453 [cs]. [Online]. Available: http://arxiv.org/abs/2407.21453

[19] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," 2015, pp. 1478–1482. [Online]. Available: https://www.isca-archive.org/interspeech_2015/sainath15b_interspeech.html

[20] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. Maier, "ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning," *Scientific Reports*, vol. 9, no. 1, p. 10997, Jul. 2019, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-019-47335-w

[21] Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Scientific Reports*, vol. 10, no. 1, p. 607, Jan. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-020-57549-y

[22] A. N. Allen, M. Harvey, L. Harrell, A. Jansen, K. P. Merkens, C. C. Wall, J. Cattiau, and E. M. Oleson, "A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset," *Frontiers in Marine Science*, vol. 8, Mar. 2021, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2021.607321/full

[23] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8944344/

[24] D. Elliott, C. E. Otero, S. Wyatt, and E. Martino, "Tiny Transformers for Environmental Sound Classification at the Edge," Mar. 2021, arXiv:2103.12157 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2103.12157

[25] P. Wolters, L. Sizemore, C. Daw, B. Hutchinson, and L. Phillips, "Proposal-based Few-shot Sound Event Detection for Speech and Environmental Sounds with Perceivers," Dec. 2023, arXiv:2107.13616 [eess]. [Online]. Available: http://arxiv.org/abs/2107.13616

[26] S. Saha and L. Xu, "Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies," *Neurocomputing*, vol. 643, p. 130417, Aug. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225010896

[27] D. Wen, G. Liang, T. Li, L. Chen, J. Li, and T. Li, "EasyViT: An Adaptive Collaborative Edge Computing Framework for Vision Transformer," *IEEE Internet of Things Journal*, vol. 12, no. 16, pp. 33 885–33 898, Aug. 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11030760

[28] B. S. Miller, N. Balcazar, S. Nieukirk, E. C. Leroy, M. Aulich, F. W. Shabangu, R. P. Dziak, W. S. Lee, and J. K. Hong, "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, Jan. 2021.

[29] M. W. Rademan, D. J. J. Versfeld, and J. A. du Preez, "Detecting and classifying blue whale calls with wavelet scattering and spectral entropy," *The Journal of the Acoustical Society of America*, vol. 157, no. 2, pp. 1448–1457, Feb. 2025.

[30] K. Team, "Keras documentation: Keras Applications," https://keras.io/api/applications/.