BIODCASE 2025 CHALLENGE COMBO: SUPERVISED WHALE CALLS ON TINY HARDWARE

Technical Report

Astrid van Toor

blueOasis
Rua do Norte 16
2655-321 Ericeira, Portugal
avtoor@blueoasis.pt

ABSTRACT

This technical report presents an edge-optimised approach to baleen whale call detection for the BioDCASE 2025 challenge - Task 2. Taking inspiration from Task 3, it focuses on deployment constraints of resource-limited hardware. Where common models range in parameter starting from 4 million training parameters [1] with architectures often unsuitable for real-time edge deployment, our model contains just 35,571 training parameters (159KB) and operates efficiently on a 64-bit ARM Cortex-A53 with 512MB RAM. On a detection window basis of 11.8 second frames, the model performs well on two of the three classes; applying a precision-focused approach we detect blue whale ABZ calls at 72% precision and fin whale burst pulse calls at 80% precision, while downsweep predictions lack behind at 18% precision. Applying our temporal head designed for compression into TFLite, we maintain reasonable precision for ABZ calls at 65%, while downsweep calls rise to 29% precision and burst pulse calls drop significantly to 4%. Acknowledging the difficulties in call-specific identification, this report 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 [2, 3], 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 [4]. Furthermore, current DL approaches often report high accuracies based on biased evaluation protocols that do not consistently account for temporal correlations in acoustic data [5, 6]. By following the guidelines of the BioDCASE 2025 Challenge for the "Supervised Detection of Strongly-Labelled Whale Calls" we aim to reduce this validation bias in real-world deployments and encourage further development.

Thanks to blueOasis Portugal for supporting this research.

Although recent deep learning approaches have shown promising results for whale call detection [7], the deployment of these models on resource-constrained edge devices remains challenging. This technical report presents a development pathway for addressing these challenges: an edge-optimised temporal attention network designed specifically for deployment on tiny hardware platforms, achieving a balance between detection performance and computational efficiency and laying out clear next steps for development.

2. METHODOLOGY

2.1. 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, fin whale burst pulses, and fin/blue whale downsweep calls [8]. 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.

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

Table 1: Class Distribution

Our preprocessing pipeline extracts three-channel acoustic features 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. Channel-wise normalisation parameters were computed from the training set to ensure consistent scaling across diverse recording

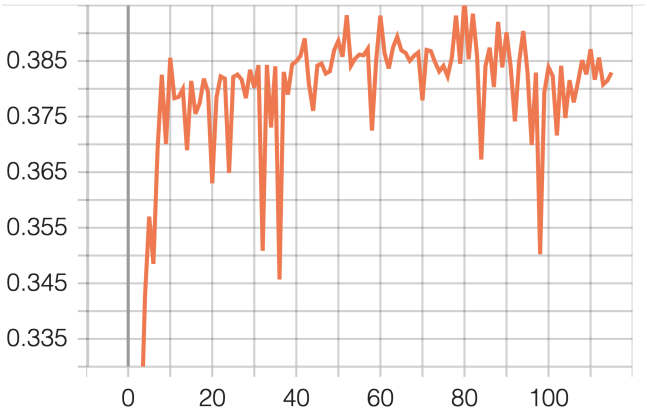

Figure 1: Weighted F1-Macro learning curve on the classification head across the detection of background noise, blue whale ABZ, fin whale pulse calls, and fin/blue whale downsweep calls

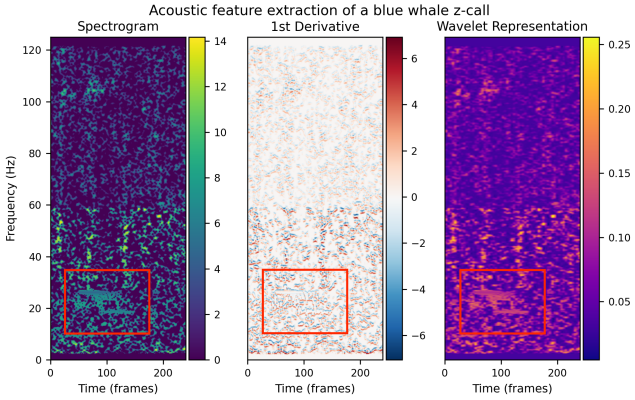

Figure 2: CNN Feature stack for blue and fin whale calls on a 48-second audio snippet. Example shows a blue whale ABZ call [8]. In production these are 11.8-second windows.

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.

2.2. Model Architecture

The proposed temporal attention network employs a lightweight CNN backbone with six convolutional layers, utilising separable convolutions to reduce computational complexity. 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

A detailed architecture diagram is presented in Figure 3. 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 [1], enabling deployment on remote memory-constrained devices.

2.3. 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. 4. Considering the multi-class problem, the background class is only activated when no other classes reach the threshold.

3. 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%.

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),

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

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.

3.1. Post Processing of the Temporal Head

The temporal head processes attention weights to extract precise call boundaries within 11.8-second windows. 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,

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	$(F \times T \times 3)$

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.

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

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.

3.2. Computational Performance

Training was done on 4 NVIDIA A100 GPUs with distributed data-parallel processing and XLA, completing in 10 hours for the 473K 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 0.15 MB. 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.

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

3.3. 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.

Starting with per-window analysis, Fig. 5, 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 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

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

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

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 in real-time after one another, 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 detection accuracy.

We aimed to develop a lightweight solution to detect whale presence in real-time. Considering passive acoustic monitoring 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, with the temporal head attempting to locate exact call boundaries when needed.

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). Alternatively, we would want to look at hydrophone calibration to enhance performance for production.

For the evaluation set submission, we supplied a prediction set of a model trained on the training data only ($\sim$80 epochs with early stop), as evaluated on the validation set. We also trained a version that includes both the training and validation set for 120 epochs (13.5hrs) for pure evaluation on the unseen evaluation set.

4. 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 suitable for edge compute. 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 and cross-hydrophone generalisation to improve deployment robustness.

5. ACKNOWLEDGMENT

A.v.T. thanks blueOasis for supporting this research and [8] for providing such a comprehensive public dataset to aid advancements in the field of marine mammal detection.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] 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.

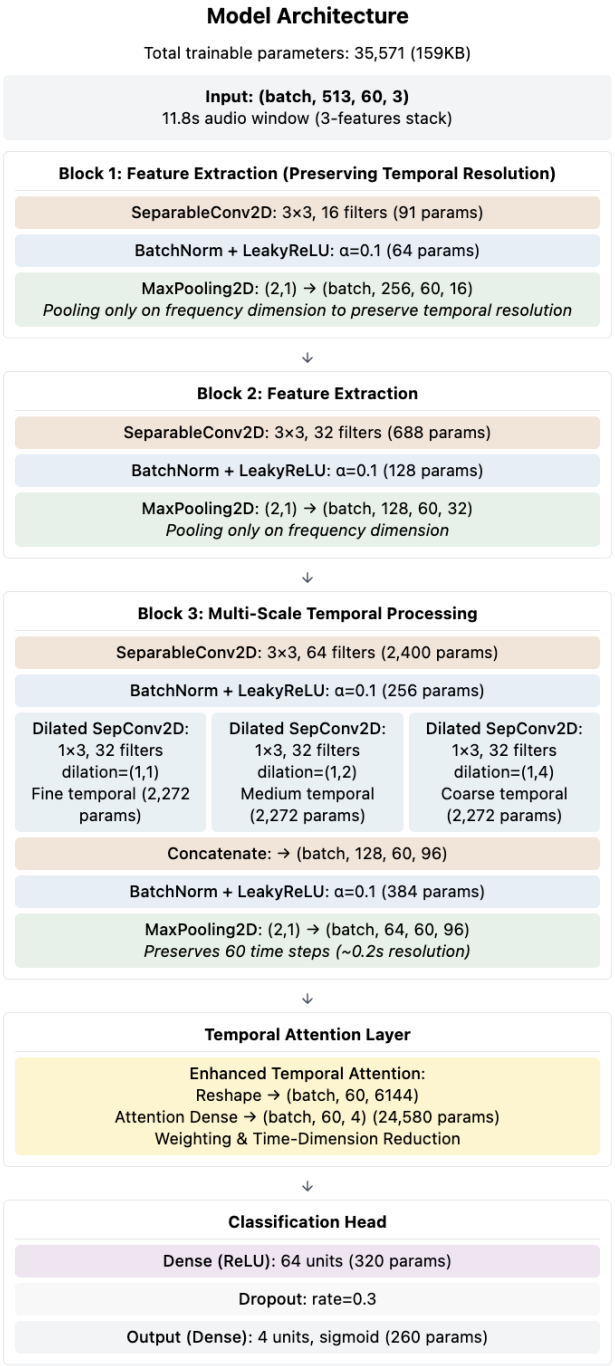

Figure 3: Architecture diagram. Includes: Adam optimizer (lr=1e-4, clipnorm=1.0) and Focal Loss ($\gamma = 2.0$, class $\alpha_A = [0.2, 0.6, 0.6, 0.9]$). Metrics: Accuracy, Precision, Recall, F1-macro, F1-micro, Weighted-F1-macro (70% precision, 30% recall).

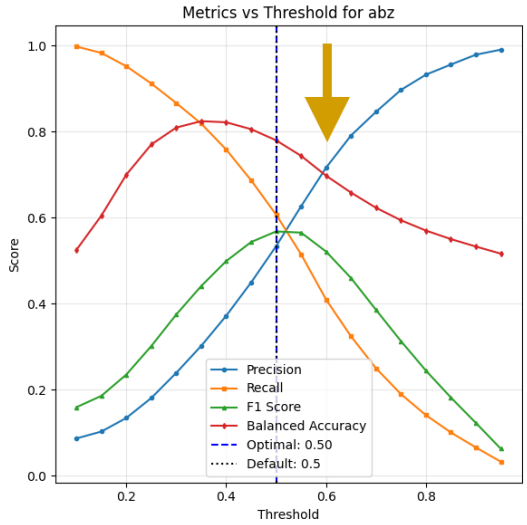

Figure 4: Threshold analysis approach demonstrated in the blue whale ABZ call class on the evaluation set. "Optimal" is maximised F1, the arrow highlights our selected threshold of 0.6 for optimised precision.

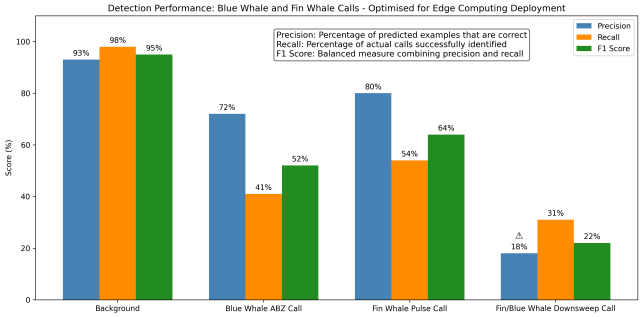

Figure 5: Performance metrics per class in the classification head.