Applied Acoustics 254 (2026) 111412

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Lightweight beluga call classification under habitat variability and marine traffic noise

Emmanuel Fernandez[a,*], Irene Roca[a], Emilie Frizzle[a], Robert Michaud[b], Marie-Ana Mikus[c], Valeria Vergara[c], Jaclyn Aubin[c], Clement Chion[a]

[a] Department of Natural Sciences, Université du Québec en Outaouais, Gatineau, Quebec, Canada
[b] Groupe de Recherche et d'Éducation sur les Mammifères Marins, Tadoussac, Quebec, Canada
[c] Raincoast Conservation Foundation, Victoria, British Columbia, Canada

HIGHLIGHTS

- State-of-the-art performance for beluga call classification (0.85 F1 score).
- Multi-label model accurately classifies overlapping beluga vocalizations.
- 320 kilobyte model suitable for resource-constrained edge deployment.
- Quantifies performance degradation due to anthropogenic boat noise.
- Demonstrates strong generalization to unseen recording sites.

ARTICLE INFO

Keywords:
Passive acoustic monitoring
Deep learning
Anthropogenic noise
Bioacoustics
Marine mammals
Edge computing

ABSTRACT

Passive Acoustic Monitoring has become a cornerstone of marine mammal conservation, enabling continuous, non-invasive monitoring through the detection and analysis of vocalizations. As the focus shifts from retrospective analyses to real-time monitoring, classification systems must operate within the strict computational constraints of ultra-low-power embedded devices, generalize across diverse acoustic environments, remain robust to anthropogenic noise from vessel traffic, and classify distinct beluga sound types to enable advanced ecological analyses such as abundance estimation and behavioral inference.

We address these challenges by developing a classification model for the endangered beluga whale population of the St. Lawrence Estuary. Our lightweight, quantized convolutional neural network classifies beluga vocalizations under natural field conditions, including overlapping calls and variable background noise. The model achieves a macro-averaged F1 score of 0.85 with a memory footprint of under 512 kilobytes, satisfying the constraints of edge-deployable microcontrollers. Using labeled examples with and without vessel noise, we directly quantified the impact of anthropogenic interference, revealing a drop to an F1 score of 0.75 in the presence of boat noise. Training on data from four distinct sites, we demonstrate that a single multi-site model generalizes effectively across all locations. In a leave-one-site-out evaluation, the full-precision model maintains strong performance on unseen sites with an average F1 drop of only 0.05, while the quantized model shows a larger gap of 0.10, highlighting the increased sensitivity of compressed models to distribution shifts.

1. Introduction

The beluga whale (*Delphinapterus leucas*) population in the St. Lawrence Estuary (SLEB) remains classified as endangered, as it is continuously threatened by marine traffic, noise pollution, contaminants, and a decline in prey abundance [19,33]. Although these factors are identified, their precise impacts on beluga habitat use and population dynamics largely remain elusive. Enhanced monitoring of SLEB's spatio-temporal dynamics, potentially in real-time, could significantly improve our understanding of these effects, and enhance our capacity to react and mitigate their impacts.

* Corresponding author.
Email address: emmanuelfernandez00@gmail.com (E. Fernandez).

https://doi.org/10.1016/j.apacoust.2026.111412
Received 30 August 2025; Received in revised form 12 March 2026; Accepted 31 March 2026
Available online 6 June 2026

To this end, Passive Acoustic Monitoring (PAM) has emerged as a vital, non-invasive approach for studying the habitat use and behavior of marine mammals, particularly in remote or logistically challenging environments [12,39,44,49]. PAM enables continuous data collection across broad spatial scales, irrespective of weather conditions, time of day, or depth. Its effectiveness, however, depends on the consistent vocal activity and acoustic detectability of the target species. Belugas, often called "canaries of the sea" due to their rich and varied vocalizations, are particularly well-suited for PAM as their calls play crucial roles in communication, navigation, foraging, and social interaction [13,41,51].

Traditionally, PAM has been employed to passively record underwater soundscapes over periods of weeks to months, after which the accumulated data are retrieved, processed, and analyzed retrospectively to address ecological and conservation questions regarding the target species. This approach typically generates extensive volumes of acoustic data that are impractical to process manually. Consequently, the state of the art has involved annotating representative subsets of the data to train deep learning models, which can then be applied to classify and analyze the entire dataset automatically and efficiently [8,14,28,56].

As PAM technology and machine learning capabilities continue to advance, there is growing interest in transitioning from retrospective data analysis to real-time processing. A notable example is the NOAA Right Whale Listening Network, which focuses on the critically endangered North Atlantic right whale *Eubalaena glacialis*. This system employs PAM devices equipped with automated detection algorithms to identify right whale calls [27]. Upon detection, the system transmits alerts to nearby vessels, enabling them to reduce speed or alter course to mitigate the risk of collisions and minimize acoustic disturbance. Similar initiatives include the WhaleSafe project, which employs a moored PAM buoy to detect blue, humpback, and fin whale vocalizations in near real time off the west coast of the United States [4]. In the Mediterranean, the BOMBYX project focuses on sperm whales, fin whales, and pilot whales [42], further demonstrating the growing potential of real-time PAM systems to support conservation efforts across multiple species.

Belugas use a rich vocal repertoire that includes high-frequency pulsed calls and echolocation clicks, broadband signals such as contact calls, and narrow-band tonal signals, hereinafter referred to as whistles [18,29,47]. Additionally, they produce mixed calls that combine high or broadband frequency pulsed components and low frequency whistles [52]. While a general call detector capable of identifying any of these call categories may suffice to infer beluga presence, fine-grained classification offers significant advantages. Specifically, it could enable more complex downstream analyses, such as estimating whale abundance [46], identifying predominant behaviors within a herd (e.g., feeding, social interaction) [12], inferring group composition [54], and potentially recognizing individual or pod-specific acoustic signatures of interest [41,52].

Belugas exhibit complex social behavior and frequently vocalize in groups, creating acoustically dense environments where multiple individuals may produce calls in close succession or simultaneously [47]. Deep learning methods for detecting or classifying animal vocalizations typically operate on fixed-length audio windows ranging from a few seconds to several minutes [48]. Within this framework, multiple or overlapping calls are often captured within the same analysis window, complicating the task of accurate call classification. For models to be reliable they must therefore be robust to signal overlap and capable of detecting multiple call categories concurrently.

To address the challenge of classifying multiple calls within a single time window, object detection models such as YOLO have been adapted to cetacean bioacoustics by predicting bounding boxes in the time–frequency domain of spectrograms [23]. However their performance on overlapping calls remains largely untested. An alternative approach is source separation, which aims to isolate individual calls prior to classification. Models like BioCPPNet [9] have shown promising results with up to two overlapping sources, and separation has been found to enhance classification accuracy in avian datasets [15]. However, these methods require clean, single-source recordings to generate synthetic mixtures for training, a condition rarely met in beluga datasets due to frequent overlapping calls. They also remain untested in complex, real-world settings with such dense vocal activity. Despite efforts to reduce model size, source separation still incurs notable computational overhead, which may be impractical when classification is the primary goal. A simpler approach is multi-label classification: rather than predicting precise onset and offset times or attempting to separate individual calls, this approach simply identifies the presence or absence of all call categories within a fixed-length audio window. It has demonstrated effectiveness in noisy and overlapping conditions, as shown by Cañas et al. [11] in their classification of overlapping anuran calls in natural soundscapes.

The computational efficiency of multi-label classification makes it particularly suitable for the stringent energy and processing constraints of real-time deployment. Autonomous PAM systems are typically deployed in remote marine environments, where all data must be processed onboard and transmitted via satellite links [4,27]. Due to strict energy budgets, these platforms often rely on low-power microcontrollers with severely limited memory, as little as 512 KB of RAM, as seen in the Medusa smart buoys for real-time North Atlantic right whale detection [27], or in biologgers used on sea turtles for soundscape classification [38]. To meet these memory constraints, compact and compressed models must be used [38].

Once a model is trained and optimized for acoustic detection and/or classification in a specific area, a major challenge lies in ensuring model generalizability across the target species' range, which may comprise a wide diversity of habitat structures and conditions [30,55]. The summer habitat of SLEB ranges to the west up to near L'Isle-aux-Coudres, and to the east as far as Rivière Portneuf and Rimouski [24]. It is centered on the confluence of the St. Lawrence and Saguenay rivers with different areas showing higher abundances, individual richness and diverse patterns in individual spatial mixing [10]. As a result, substantial variation is expected in the acoustic environment, driven both by site-specific background noise and by the localized presence of beluga communities with distinct vocal behaviors [3]. Therefore, automated systems should not only achieve high accuracy at specific sites but also generalize effectively across SLEB's essential habitat.

Beyond habitat-driven variability, anthropogenic noise represents an additional and pervasive challenge for PAM systems. Vessel traffic in particular produces sustained broadband noise that can mask biological signals or share spectral properties with certain call categories in overlapping frequency bands, potentially leading to false detections [57]. Accounting for boat noise is therefore an important consideration in the development of robust PAM systems intended for deployment in ship-trafficked waters such as the St. Lawrence Estuary.

The goal of this study is to advance the development of scalable, accurate, and field-deployable PAM systems for beluga conservation. We build upon the pipeline introduced by Cotillard et al. [14], which demonstrated the feasibility of using deep learning to detect and classify the full range of beluga vocalizations in the St. Lawrence Estuary, establishing a strong foundation for automated monitoring at a single high-residency site. In this work, we extend their methodology toward real-world deployment by addressing overlapping calls and vessel noise, evaluating cross-site generalization across SLEB's broader habitat, and optimizing the model architecture for resource-constrained edge devices.

2. Methods

We trained and evaluated convolutional neural networks to perform multi-label classification, predicting all beluga call categories present within fixed 1-second analysis windows. We used ResNet-18 as a baseline architecture and explored lightweight MobileNet models as efficient alternatives, further compressing them using 8-bit quantization to reduce the model footprint. Because our dataset contains labeled examples both with and without vessel noise, we were able to directly assess model

robustness to anthropogenic interference. We also evaluated cross-site generalization across four distinct recording locations.

2.1. Data

We collected acoustic recordings using SoundTrap hydrophones (Ocean Instruments, NZ) at four sites across the St. Lawrence estuary (Fig. 1): Baie Sainte-Marguerite (BSM), Cacouna (CAC), Rivière-du-Loup (RDL), and Kamouraska (KAM) (see deployment details in Table A.1 in the appendix).

We selected these sites based on current knowledge of areas highly frequented by SLEBs. A large part of the SLEB population frequents the Saguenay River up to BSM during summer [10] and BSM constitutes an essential habitat for SLEB for calf rearing, feeding, and socializing [31]. CAC, RDL and KAM recording sites were situated off the southern shore of the St. Lawrence River within SLEB critical summer habitat, close to high residency areas where belugas are known to spend time foraging and socializing (e.g., [32]). Marine traffic (both commercial vessels and pleasure craft) is high along both the St. Lawrence and Saguenay rivers [36], and its radiated noise is a regular presence in the background noise at all sites.

2.1.1. Acoustic categories studied

This study aims to comprehensively cover the full range of beluga vocalizations by classifying them into four broad acoustic categories: echolocation clicks and three types of calls spanning high-frequency calls and low-frequency whistles (Fig. 2).

- **Echolocation Clicks (ECHO)**: ECHOs range from approximately 30 to 120 kHz or higher and are used primarily for navigation and prey localization [41,47]. In this study, ECHOs refer to echolocation events comprising multiple pulses, rather than the detection and classification of individual clicks.
- **High-Frequency Pulsed Calls (HFPC)**: These are pulsed bursts of ultrasonic broadband sounds (~ >20 kHz). HFPCs can be distinguished from burst pulses in echolocation buzz trains by their high and constant pulse repetition rate throughout the signal [53].
- **Broadband Pulsed Calls (BBPC)**: As their name implies, BBPCs are broadband signals (200 Hz–144 kHz) with substantial acoustic energy. The most studied subcategory is the Contact Call, a long-duration (typically >1 s) pulsed signal used for group cohesion, such as mother–calf interactions [52]. Shorter BBPCs, known as short broadband burst-pulse signals, have been studied less extensively [53].
- **Whistles**: Whistles are low-frequency narrow-band tonal calls that can be modulated or pulsed and may include harmonics. They typically range from 500 Hz to 20 kHz, and are commonly associated with social communication [54].

Belugas are capable of biphonation, meaning they can produce two different sounds simultaneously [2,50]. This is most often observed in combination with HFPC and Contact Calls, resulting in so-called complex calls. In complex Contact Calls, the lower frequency component is suspected to encode individual identity or group membership [52]. Although such mixed calls are included in the analysis, this study does not attempt to distinguish whether they originate from a single individual producing both vocalizations simultaneously or from two animals vocalizing at the same time.

2.1.2. Annotations

We conducted stratified random sampling across the dataset to identify recording periods with SLEB acoustic activity, ensuring a balanced representation across acoustic categories and, as far as possible, capturing the heterogeneity of each acoustic category of the SLEB repertoire. For labeling, we cut raw audio into fixed one-second windows and applied a weak multi-label scheme, annotating the presence or absence of each acoustic category in every window. Expert bioacousticians annotated the data through visual and aural analysis of spectrograms created

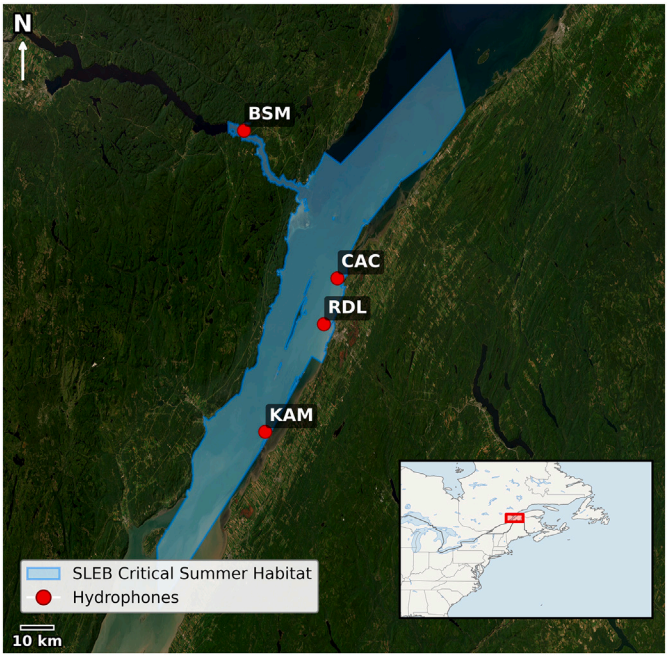

Fig. 1. Locations of hydrophone deployments within the St. Lawrence Estuary (Québec, Canada) beluga critical summer habitat. Basemap: Esri World Imagery.

with Raven (Raven Pro 2025). If multiple calls of the same category occurred within a single window, we counted them only once, with the label reflecting presence rather than a count. Calls extending across two one-second windows were labeled in both windows. To ensure the model could reliably distinguish between the presence and absence of vocalizations, we retained windows without any beluga acoustic activity in the dataset.

To improve robustness to boat noise and analyze its impact on detection, we added samples overlapping with boat passages. We first identified boat passages across the four sites, then extracted two-minute clips from the start, middle, and end of each passage. We also annotated these clips for the presence or absence of the studied beluga calls following the same methodology. We also aimed for a balanced and representative inclusion of the acoustic categories in the presence of boat noise. However, this task proved to be challenging, since the relative prevalence of the different acoustic categories within the SLEB acoustic repertoire varies under vessel noise conditions (e.g., [22,34,53]), and some calls are difficult to detect even for human annotators.

The resulting dataset comprises 6854 one-second windows with beluga vocal activity and 10,287 without (see Table 1). Among acoustic categories, ECHOs are the most prevalent with 5384 occurrences, while HFPCs and BBPCs are less represented with 806 and 1059 occurrences respectively. Although we did not explicitly annotate temporal overlap between categories, co-occurrence was frequent, particularly among the higher-frequency categories (ECHO, HFPC, and BBPC) and between BBPCs and Whistles, which share overlapping frequency ranges. Counting windows where multiple categories with overlapping frequency ranges co-occur yields 1377 windows, approximately 20% of all vocally active windows, providing an estimate of temporal overlap in the dataset. Regarding background conditions, 5978 windows overlap with boat noise passages (see Table 1b); however, the representation of acoustic categories under these conditions is limited, with fewer than 400 occurrences for all categories except ECHO.

2.2. Spectrogram preprocessing

We use mel spectrograms as a unified input representation for all acoustic data. The mel scale, although originally designed to reflect the nonlinear frequency sensitivity of human hearing, has proven

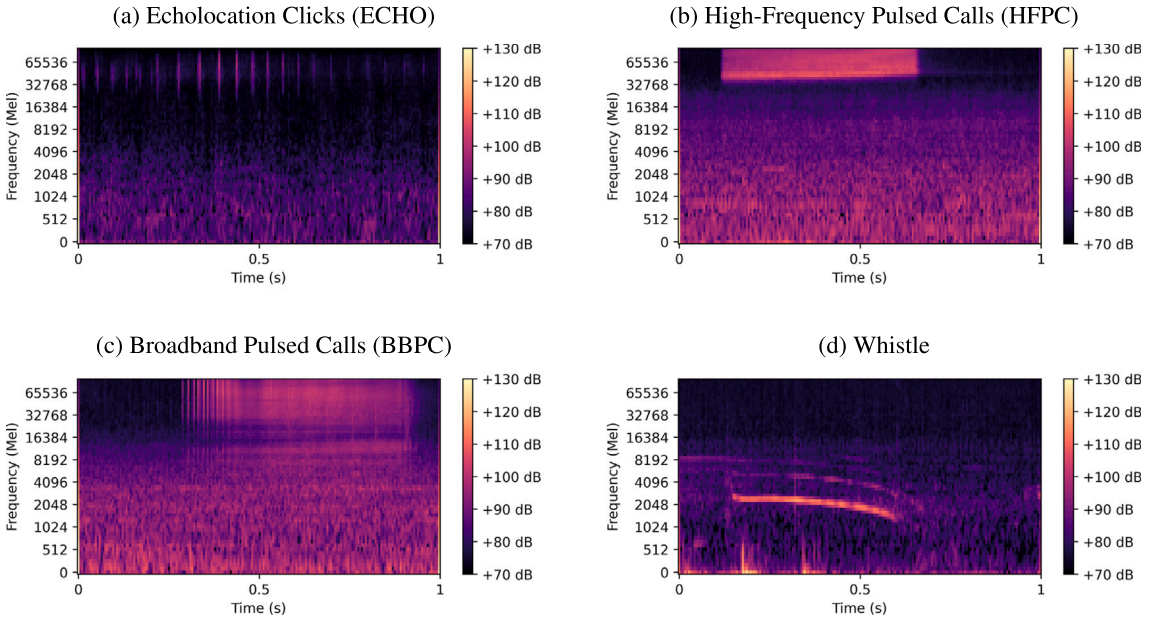

Fig. 2. Representative mel spectrograms of the four beluga acoustic categories considered in this study: (a) Echolocation clicks (ECHO), (b) High-Frequency Pulsed Calls (HFPC), (c) Broadband Pulsed Calls (BBPC), and (d) Whistles.

Table 1
Summary of annotated dataset: number of windows with and without beluga vocal activity, and total occurrences per acoustic category. A single window may contain multiple categories; therefore, per-category counts can exceed the number of vocally active windows. (a) Distribution across recording sites. (b) Distribution across background acoustic conditions.

	Vocal Activity		Per-Category Counts			
Site	Present	Absent	HFPC	BBPC	ECHO	Whistle
(a) Dataset summary by recording site.						
BSM	1198	7161	293	193	1021	338
CAC	2159	1608	179	318	1321	1630
KAM	2983	1082	177	327	2819	646
RDL	514	436	157	221	223	333
Total	**6854**	**10,287**	**806**	**1059**	**5384**	**2947**
(b) Dataset summary by background condition.						
	Vocal Activity		Per-Category Counts			
Background	Present	Absent	HFPC	BBPC	ECHO	Whistle
Ambient	3833	7330	577	818	2628	2547
Boat noise	3021	2957	229	241	2756	400
Total	**6854**	**10,287**	**806**	**1059**	**5384**	**2947**

effective in classifying a wide range of cetacean vocalizations [1,20]. Its logarithmic scaling provides sufficient frequency resolution across the full bandwidth of interest, enabling the simultaneous capture of both low-frequency whistles and high-frequency calls. This choice allows for a streamlined preprocessing pipeline without the need to generate multiple spectrogram types tailored to specific frequency bands, as was done in previous approaches (e.g., [14,23]).

We first downsampled all audio recordings to 192 kHz to match the lowest sampling rate among the hydrophones used. We then generated spectrograms using the `librosa` library with an FFT window size of 2048, 64 mel frequency bins, and a 50% overlap between successive frames.

We calibrated amplitude values using the sensitivity profile of each hydrophone, allowing spectrogram amplitudes to be expressed in decibels relative to1 μPa (dB re 1 μPa), in accordance with marine bioacoustic standards. This calibration ensures that amplitude information is physically interpretable and comparable across devices.

Before being passed to the neural network, we normalized all spectrograms using min–max scaling and resized them from their original resolution of 64×188 pixels to 224×224 pixels to match the expected input dimensions of the pretrained models.

2.3. Models

We employed convolutional neural networks (CNNs) for multi-label classification of beluga acoustic categories. Each architecture consists of a convolutional feature extractor followed by a fully connected classification head that outputs one logit per category. A sigmoid activation is applied independently to each output to obtain per-class scores. In this multi-label framework, each output head effectively performs binary detection for its corresponding acoustic category, and the collection of these binary decisions constitutes the overall classification. For this reason, when referring to our own model, we use the terms ‘detection’ and ’classification’ interchangeably throughout the paper.

We selected ResNet-18 as our baseline architecture based on Bergler et al. [8], which demonstrated comparable performance between the compact ResNet-18 and larger variants for orca call classification. However, ResNet-18’s 50 MB model size remains prohibitive for real-time deployment on resource-constrained edge devices, where computational efficiency and power consumption are critical limitations [38].

To address these constraints, we selected MobileNetV3 Small [25] as an alternative, motivated by the successful deployment of MobileNet-based soundscape classifiers in [38]. Although the referenced study employed MobileNetV1 [26], we selected the more recent MobileNetV3 Small to leverage improved architectural optimizations and efficiency gains. The original MobileNetV3 Small architecture contains 12 convolutional layers of increasing size. To further reduce computational requirements while maintaining classification performance, we systematically investigated truncated versions of this architecture by removing later convolutional layers. Hereafter, we refer to MobileNetV3 Small as “MobileNet” for brevity.

We also explored 8-bit quantization for MobileNet using quantization-aware training (QAT) to further reduce model size and enable deployment on resource-constrained hardware. We employed PyTorch’s default QAT configuration for x86 architectures with a staged freezing protocol: we froze Batch Normalization statistics

and disabled weight and activation observers once the average F1 score on the validation set exceeded 0.80. This threshold ensured that we finalized quantization parameters only after the model had begun to converge. We did not apply quantization to ResNet-18, as its post-quantization footprint remained significantly larger than that of MobileNet in full precision.

We adapted both architectures to accept single-channel spectrogram inputs in place of the default three-channel RGB format and modified their classification heads to match the dimensions of the final convolutional feature maps. We initialized all models with ImageNet-pretrained weights provided by PyTorch.

2.4. Training

We implemented all models in PyTorch and trained them on an NVIDIA GeForce RTX 2060 GPU with 8 GB of memory. We followed the training protocol of Cotillard et al. [14], using the Adam optimizer with an initial learning rate of 10^{-3} and a learning rate scheduler that halved the rate after three consecutive epochs without improvement in validation loss. We stopped training early if validation performance did not improve for five consecutive epochs. We optimized models using binary cross-entropy loss with logits (`BCEWithLogitsLoss`), the standard loss function for multi-label classification, which allows each label to be learned independently.

2.5. Metrics

We evaluated model performance using class-wise and macro-averaged precision, recall, and F1 score, with F1 serving as the primary metric due to its balanced treatment of false positives and false negatives. For comparability with prior work, we also included accuracy and balanced accuracy. We additionally reported the false alarm rate to provide an indication of false positive frequency.

2.6. Experimental setup

The first set of experiments focused on optimizing model architecture and computational efficiency. We compared the baseline ResNet-18 with MobileNet and its truncated versions, as well as their respective 8-bit quantized variants. These experiments allowed us to explore trade-offs between model size and classification accuracy in the context of deployment constraints. For each configuration, we systematically evaluated performance on samples with and without boat noise to assess robustness to anthropogenic interference.

The second set of experiments focused on cross-site generalization. We first evaluated how well the best-performing model from the previous experiment generalized across sites by assessing its performance on each individual site within the test set. We then investigated whether training site-specific models using data from only one location could improve performance on that site compared to models trained on all sites.

To evaluate generalization to unseen environments, we conducted a leave-one-site-out experiment. In each round, we trained the model on data from three of the four sites and evaluated it on the held-out site. We repeated this procedure across all site combinations, providing a robust measure of the model's ability to adapt to previously unseen acoustic conditions.

Given the relatively small size of our dataset compared to those typically used in deep learning, we adopted a modified version of the nested K-fold cross-validation procedure described in [43] to ensure statistically robust and reproducible results. We first partitioned the data into five folds at the recording file level, with each file containing at most one boat passage, to prevent data leakage across splits. We held out one fold as the test set and used the remaining four for training. To tune hyperparameters, we randomly selected 25% of the training data as a validation set, instead of performing a full internal two-fold split as in the original procedure. This modification preserved a larger portion of the data for training while still supporting early stopping and learning rate scheduling. We repeated the procedure five times so that each fold served once as the test set. In each round, approximately 60% of the data went to training, 20% to validation, and 20% to testing. We computed final results by averaging performance metrics across all five folds and reported standard deviations to assess the stability of the results across splits.

We applied this cross-validation procedure to all experiments except the leave-one-site-out evaluation, where the site-based partitioning already provides a natural train-test split. For the site-specific experiments, we filtered training and validation sets to include only the sites relevant to each experimental condition. For boat noise evaluation, we applied the partitioning to the test set only.

3. Results

3.1. General performance and optimization

The ResNet-18 model achieved strong baseline performance in detecting and classifying SLEB vocalizations, even in the presence of overlapping calls, reaching an average F1 score of 0.85 (see Table 2). However, this came with a substantial memory footprint of 45.30 MB. MobileNet matched this performance while being more than ten times smaller. Further truncating the architecture to 8 layers resulted in no performance loss, while degradation became evident below 6 layers, with average F1 scores dropping to 0.76 for the 4-layer variants, highlighting the limited capacity of overly compact networks.

Quantization consistently reduced model size by over 70% without affecting classification performance for models with at least 6 layers. The 8-layer quantized MobileNet achieved an average F1 score of 0.86 with only 0.35 MB, representing a 99.2% size reduction compared to ResNet-18.

Differences in average F1 score between models, whether comparing ResNet-18 to MobileNet or full-precision to quantized variants, typically fell within 0.02. Given that the standard deviation across folds ranged from 0.01 to 0.02 for all models with more than four layers, these variations fall within expected noise and reflect rounding sensitivity and minor fluctuations inherent to cross-validation, rather than systematic performance differences.

Although the 8-layer and 6-layer models performed similarly, the quantized 8-layer model had already surpassed the 512 KB memory constraint of the most limited edge devices. We therefore selected this model for subsequent experiments, evaluating both its full-precision and quantized variants. See Table A.2 in the appendix for a detailed breakdown of all metrics per acoustic category for the 8-layer quantized model.

When tested on the subset of windows with a high likelihood of temporal overlap, as described in Section 2.1.2, the quantized 8-layer model shows no performance degradation, demonstrating robust classification even during periods of high vocal activity with co-occurring sounds. The model achieves an overall false alarm rate of 2% on the full dataset, with variation across acoustic categories: HFPCs and BBPCs exhibit the lowest rates at under 1% (see Table A.2). When evaluated on continuously annotated segments with no beluga activity totaling over 30min, false alarm rates drop further, with HFPCs producing zero false positives (see Table A.3).

3.2. Robustness to boat noise

Table 3 reports the performance of the quantized 8-layer MobileNet under ambient and vessel noise conditions. We observed no significant difference between the full-precision and quantized variants, so we report only the quantized results.

Partitioning the test set by noise condition reveals that the model performs notably better in ambient conditions, achieving an average F1 score of 0.88. In the presence of boat noise, performance drops by 0.13 points. ECHOs prove resilient, maintaining a high F1, while HFPC, BBPC, and Whistle all fall to around 0.68-0.70, indicating substantially reduced detectability. The standard deviation for these categories

Table 2
Mean F1 scores for overall and per-category classification across model configurations. Results are reported for both full-precision and quantized variants, along with corresponding model sizes.

Model	Model Variant	Size (MB)	Avg F1	ECHO F1	HFPC F1	BBPC F1	Whistle F1
ResNet 18	*Full Precision*	45.28	0.85	0.92	0.82	0.80	0.86
MobileNet (12 layers)	*Full Precision*	4.96	0.85	0.91	0.84	0.82	0.86
	Quantized	1.36	0.84	0.91	0.81	0.81	0.87
MobileNet (10 layers)	*Full Precision*	2.96	0.85	0.92	0.82	0.81	0.87
	Quantized	0.83	0.85	0.91	0.82	0.80	0.87
MobileNet (8 layers)	*Full Precision*	1.22	0.86	0.92	0.84	0.81	0.88
	Quantized	0.35	0.85	0.92	0.84	0.80	0.86
MobileNet (6 layers)	*Full Precision*	0.89	0.84	0.92	0.84	0.77	0.86
	Quantized	0.26	0.84	0.92	0.84	0.77	0.86
MobileNet (4 layers)	*Full Precision*	0.31	0.76	0.89	0.68	0.66	0.80
	Quantized	0.09	0.75	0.89	0.67	0.66	0.80

Table 3
Mean F1 scores (± standard deviation) across cross-validation folds for the quantized 8-layer MobileNet model, evaluating overall and per-category classification performance under ambient and boat-noise conditions.

Background	Avg F1	ECHO F1	HFPC F1	BBPC F1	Whistle F1
Ambient	0.88 ± 0.01	0.90 ± 0.04	0.88 ± 0.04	0.85 ± 0.02	0.90 ± 0.03
Boat Noise	0.75 ± 0.08	0.94 ± 0.03	0.75 ± 0.12	0.64 ± 0.20	0.68 ± 0.07

also rises sharply, exceeding 0.12 for HFPC and Whistle and 0.15 for BBPC, compared to 0.02-0.04 under ambient conditions. This instability across folds suggests that performance is highly sensitive to which boat passages appear in training versus test splits.

As shown in Table A.4 in the appendix, the false alarm rate remained stable at 2% regardless of acoustic conditions, indicating that the model successfully learned not to confuse vessel-generated signals with beluga calls. However, this robustness to false positives comes at the cost of recall, which drops significantly in the presence of boat noise as the model fails to detect calls masked by vessel interference.

3.3. Cross-site generalization

To evaluate cross-site generalization, we conducted three sets of experiments: training on all sites, site-specific training, and leave-one-site-out evaluation (see Table 4).

When trained on all sites, the quantized MobileNet achieved consistently high performance across all test locations, with scores closely matching the overall dataset average. CAC exhibited the lowest performance at 0.82 F1 score, while the remaining sites achieved between 0.84 and 0.87.

In the site-specific experiments, each model performed well on its respective training site, with the exception of RDL, which had the fewest samples in the dataset (see Table 1b). Across all sites, training on the combined dataset consistently outperformed site-specific training, confirming the benefit of data diversity.

Since training on multiple sites yields superior overall performance, the leave-one-site-out experiments provide a more meaningful assessment of the model's generalization to unseen locations. In this setting, the full-precision model remained robust for CAC, KAM, and RDL, with an average drop of only 0.03 F1 score relative to the all sites topline. BSM, however, exhibited a more notable decline of 0.12. The quantized model showed greater sensitivity to unseen sites, with the generalization gap widening considerably: CAC increased from 0.02 to 0.06, KAM from 0.05 to 0.12, and BSM from 0.12 to 0.16. These results highlight that quantization amplifies the generalization gap.

4. Discussion

We successfully trained and optimized a classification model that achieved strong performance in detecting and classifying SLEB sounds. The quantized 8-layer MobileNet reached a macro-averaged F1-score of 0.85 across all acoustic categories. Among these, ECHOs and Whistles achieved the highest classification performance, 0.92 and 0.86 respectively, likely due to their higher number of examples in the training dataset.

Our results compare favorably with those of Cotillard et al. [14], though direct comparisons are limited by methodological differences. Their approach relied on separate models for high-frequency sound classification and whistle detection, reporting 0.86 and 0.88 balanced accuracy respectively. Using a single unified model, we achieve 0.91 balanced accuracy for high-frequency sounds and 0.94 for whistles (see

Table 4
Site generalization performance of the 8-layer MobileNet in full precision and quantized formats. Values are reported as full precision/quantized. Bold values indicate topline performance (generally the All-Sites full-precision model). The generalization gap shows the performance drop relative to the corresponding topline result for each site. To reduce clutter, only the diagonal entries are shown for both the site-specific and leave-one-site-out experiments.

		Test Sites				
Experiment	Train Sites	BSM	CAC	KAM	RDL	vs Topline
All Sites	All	**0.86**/0.84	**0.82/0.82**	**0.85/0.85**	**0.87**/0.86	–
Site Specific	BSM	0.83/0.83	–	–	–	0.03/0.03
	CAC	–	0.80/0.79	–	–	0.02/0.03
	KAM	–	–	0.81/0.80	–	0.04/0.05
	RDL	–	–	–	0.67/0.46	0.20/0.41
Leave Site Out	BSM out	*0.75/0.71*	–	–	–	0.11/0.15
	CAC out	–	*0.80/0.76*	–	–	0.02/0.06
	KAM out	–	–	*0.80/0.73*	–	0.05/0.12
	RDL out	–	–	–	*0.84/0.83*	0.03/0.04

Table A.2 in the Appendix for all comparative metrics), while also supporting multi-label predictions that capture the full acoustic content of each segment. By contrast, Cotillard et al. [14] used single-label annotations constrained to the dominant sound type, which also likely introduced label noise during training. Additionally, their whistle detector relied on a substantially heavier transformer-based architecture (AVES) [21], whereas our model achieves superior performance at a fraction of the computational cost.

These results are consistent with recent work in related species. Hamard et al. [23] reported F1 scores of 0.91 for dolphin whistle detection and 0.88 for click trains using an object detection-based approach. Our work demonstrates that comparable performance can be achieved for beluga call classification using a lightweight multi-label framework suitable for edge deployment.

To meet the constraints of edge deployment, we focused on reducing model size and achieved a 99% reduction, from the 45 MB ResNet-18 baseline to just 350 KB using an 8-layer quantized MobileNet, with no loss in classification or detection performance. This places the model well below the 512 KB threshold commonly associated with ultra-low-power microcontrollers, as used in prior work [27,38]. These results suggest that high-performance beluga sound classification is feasible even on highly constrained embedded hardware, paving the way for real-time, on-board inference in autonomous and remote monitoring systems, though performance in the presence of anthropogenic boat noise remains a challenge.

Anthropogenic noise affecting PAM detections is a widely known issue, leading to both false detections and missed calls due to masking. Researchers have commonly addressed this by training models on datasets with varying background noise levels (e.g., [8,23,55]). However, to the best of our knowledge, none have performed a direct comparison using a clearly partitioned dataset containing samples both with and without boat noise alongside target species calls, allowing for the computation of both false alarm rate and recall.

Our partitioned dataset enables precisely this analysis. On ambient noise samples, model performance rises to an F1 of 0.88, with HFPC, BBPC, and Whistle scores improving significantly. In the presence of boat noise, however, performance drops to 0.75, a decrease of 13 points. ECHO remains largely unaffected, while BBPC and Whistle drop most substantially, to 0.64 and 0.68 respectively. Two factors explain this pattern. First, BBPC and Whistle occupy frequency ranges that overlap most with boat noise, whereas higher-frequency calls like ECHO and HFPC are less affected. However, HFPC performance also drops significantly despite minimal spectral overlap with boat noise. This suggests that a second factor is more influential: the limited number of training examples containing both boat noise and calls other than ECHO (fewer than 400 samples per category). This scarcity prevents the model from learning to detect calls reliably in noisy conditions, demonstrating that even calls like HFPC, which theoretically do not overlap with boat noise, require substantial examples for the model to generalize to that scenario. The increased standard deviation across folds when evaluating on boat noise further reflects this difficulty: depending on cross validation fold, some held-out passages prove harder than others, leading to considerable variation in performance.

Although we systematically sought boat passages containing beluga vocalizations, certain acoustic categories, such as HFPC and BBPC, occur less frequently in the natural acoustic repertoire. In addition, belugas have been shown to alter their vocal behavior in the presence of vessels [53], further limiting the availability of representative samples under boat-noise conditions. Together, these factors make it challenging to construct a more balanced dataset for this scenario.

We explored several approaches to improve performance under boat-noise conditions. Noise reduction techniques were tested but proved fragile: performance depended strongly on hyperparameters that were difficult to tune across diverse acoustic environments, and the processing often removed faint whistles together with the noise. Optimizing detection thresholds to balance recall against the false positive rate was also considered. However, the high variability in performance across folds indicates that such tuning would not yield a robust solution with the current dataset size.

With the data currently available, our analysis therefore primarily quantifies the extent to which boat noise degrades detection performance. This also highlights the inherent difficulty of the task: even human annotators frequently need to adjust spectrogram settings and spend additional time inspecting recordings to distinguish calls from vessel noise (see Fig. 3(b). While informative, this level of robustness would not yet be sufficient for fully reliable PAM deployment in heavily trafficked environments.

We instead identify several promising directions for future work. First, a dedicated labeling effort focused specifically on boat-noise samples, or data augmentation through synthetically adding boat noise to ambient recordings, could substantially increase the number of training examples [45]. Second, characterizing boat passages by vessel type and estimated loudness could provide a proxy for signal-to-noise ratio, enabling a finer-grained analysis of performance degradation.

Our results show that training on data from all sites yields robust performance across geographic locations within the essential summer habitat of the SLEB. The multi-site model consistently outperformed site-specific models, as expected given its exposure to broader acoustic variability [40]. This confirms that training separate models for each location offers no benefit; incorporating data from multiple sites into a single model is both more effective and efficient.

The leave-one-site-out setup provides the clearest assessment of generalization to unseen environments, closely simulating real-world deployment in novel locations. Under this evaluation, we observed the first notable difference between full-precision and quantized models. The full-precision model showed strong generalizability, with an average F1 drop of only 0.05 compared to training on all sites. The quantized model, however, showed a more substantial drop of 0.10. This likely reflects increased sensitivity to distribution shifts: quantization parameters such as clipping ranges for weights and activations are learned during training and may not adapt well to unseen input distributions [17].

These results suggest that our full-precision model generalizes well and could be applied to new sites within the St. Lawrence Estuary. The quantized model, by contrast, shows limited generalizability. For live deployment, even with the full-precision model, we recommend collecting data from the new site and evaluating performance before operational use. If results are unsatisfactory, retraining with data from the new site would be advisable. As the number of monitored sites increases, periodically repeating this leave-one-site-out evaluation would help confirm whether generalization improves with additional data and variability, as expected.

Future work should also examine generalization across distinct beluga populations, such as those in the Arctic, to evaluate the model's applicability to a broader range of acoustic environments. Beyond spatial transfer across populations, temporal robustness remains another important consideration for long-term deployments. Our dataset included recordings spanning multiple years for most sites (see Table A.1 in the appendix), and results suggest stable performance over the period covered. Nevertheless, systematic evaluation of temporal generalization should be explicitly addressed in future work to ensure long-term reliability under changing acoustic conditions.

From a practical integration perspective, both for offline analysis and eventual real-time PAM applications, we advocate for a simple strategy that prioritizes computational efficiency and ease of maintenance. In contrast to [14], we do not recommend a preliminary Region of Interest (ROI) detection step. Although originally introduced to reduce computation by skipping silent segments, this step adds unnecessary overhead given the compact size of our model. More importantly, ROI-based methods rely on hardcoded thresholds that are difficult to generalize across sites and often merge overlapping calls into a single segment, limiting classification performance in dense acoustic environments.

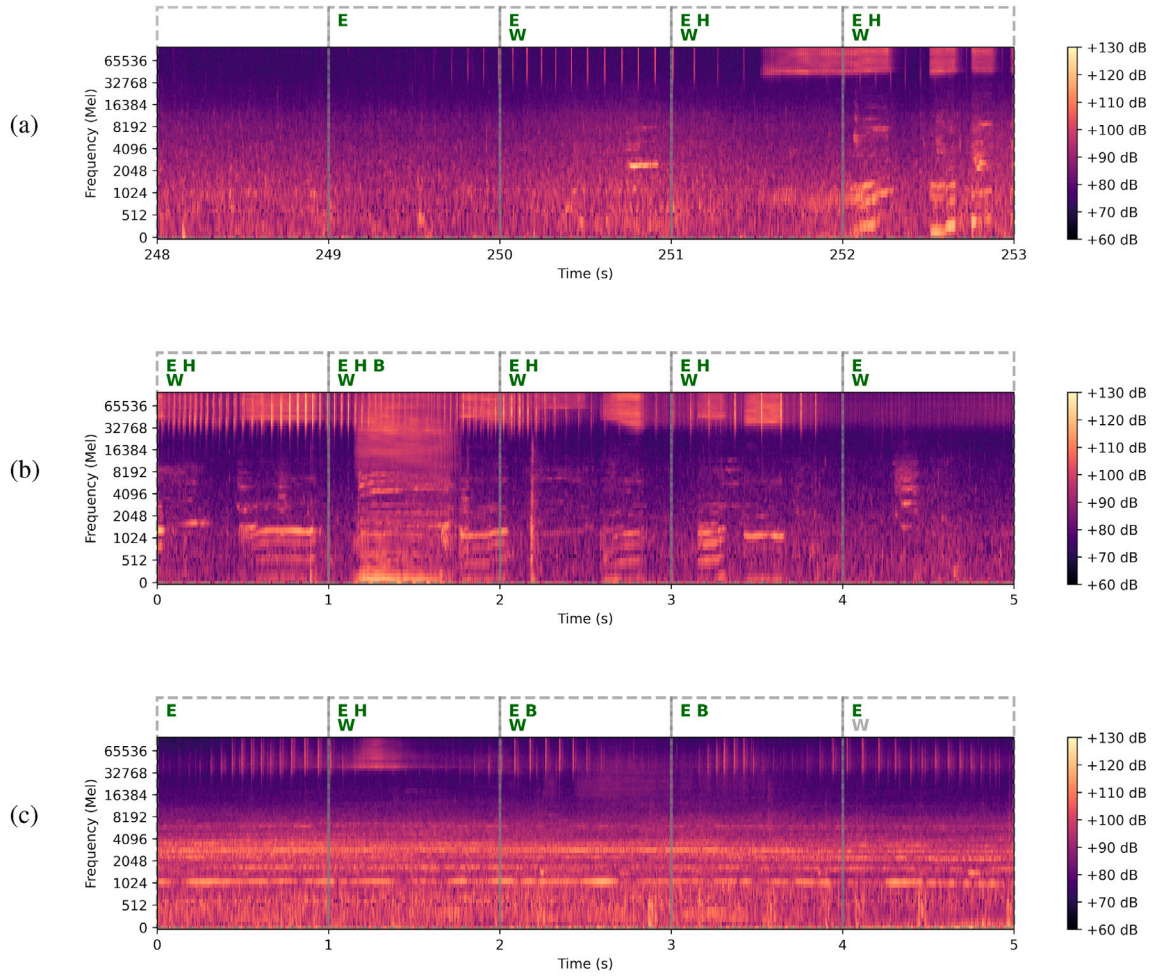

Fig. 3. Example outputs of the 8-layer quantized MobileNet model on unseen recordings, with one prediction per second. Detected acoustic categories are overlaid above the spectrogram: E (ECHO), H (HFPC), B (BBPC), and W (Whistle). Green labels indicate correct predictions and grey labels denote false negatives; no false positives are present in these examples. (a) Illustration of counting limitations due to the fixed-window approach. (b) Predictions under ambient conditions during a high vocal activity period, demonstrating accurate detection of overlapping ECHO and HFPC sounds. (c) Predictions under boat-noise conditions, correctly identifying the first two whistles despite strong spectral masking, but missing the final whistle.

We instead recommend applying the model directly to non-overlapping one-second windows, as illustrated in Fig. 3. Since the model was trained to recognize acoustic categories even when only a short portion is present within a window, overlapping segments are not required. However, it should be noted that due to the fixed window size, this model cannot precisely count individual vocalizations. For example, a long call spanning two windows may be detected twice, while a window containing multiple short HFPCs will still yield a single binary prediction (see Fig. 3). We do not consider this a major limitation. The one-second window is short enough to capture temporal trends, and overcounting calls is expected to be consistent over time. As such, the pipeline's outputs remain a reliable proxy for call abundance and vocal activity, supporting applications such as estimating beluga presence, relative density, or behavior.

Another important consideration for PAM deployments is the false alarm rate [27]. Our model achieves an overall false alarm rate of approximately 2%, with no significant differences across sites or in the presence of anthropogenic vessel noise. At a one-second resolution, this corresponds to roughly 72 false positives per hour. While this may appear substantial for continuous monitoring, several factors mitigate its impact. First, the reported false alarm rate is computed per class. In windows where multiple call types are present, predicting an additional absent class counts as a false alarm for that class, effectively reflecting a form of misclassification. This metric is therefore informative for evaluating classification performance during periods of intense vocal activity, but it does not directly capture the primary concern for PAM applications: predicting presence when the focal species is completely absent. Moreover, during highly vocal periods, a small amount of labeling uncertainty is unavoidable, particularly for faint or partially truncated calls, which can artificially inflate this metric.

To better assess performance under conditions relevant to monitoring, we evaluated the model on extended ambient recordings totaling more than 30min where no beluga activity was detected (Table A.3, Appendix). Under these conditions, ECHOs account for most false alarms at 1.8%, while all other call types remain below 1%, and HFPCs produce no false positives. This indicates that the model remains highly reliable when belugas are absent. Future work should estimate false alarm rates over longer beluga-free periods to obtain more robust estimates. Finally, two practical strategies can further reduce false positives during deployment. Monitoring can prioritize call categories with inherently low false alarm rates, such as HFPCs, depending on the monitoring objective. Alternatively, adjusting detection thresholds per category provides a flexible way to balance precision and recall according to application-specific requirements.

While the model performs well at detecting SLEB vocalizations, its classification scheme operates at a relatively coarse level of acoustic

categorization. Depending on the classification scheme, up to 52 call types have been reported [5–7], with extensive work focusing on whistles, which are often subdivided by contour shape, number of inflection points, frequency range, and other acoustic features [5]. Moreover, while our model jointly detects both whistles and high-frequency calls from a shared spectrogram input, it does not attempt to determine whether simultaneous vocalizations originate from a single beluga producing a biphonation or from multiple individuals vocalizing at the same time. Future work could aim to expand the number of call categories and enable the detection of complex calls, such efforts would require more detailed and fine-grained annotations. This level of granularity may become increasingly important as our understanding of call functionality and beluga communication deepens. For example, the whistle-like component in complex contact calls is believed to encode individual identity or group membership [52], and a classifier with this level of precision could be invaluable for tracking individual belugas or social groups.

The St. Lawrence Estuary and Saguenay Fjord form a highly productive ecosystem where belugas share critical habitats with pinnipeds, baleen whales, and other odontocetes during the summer [35,37]. Although no co-occurrence of beluga and other species' vocalizations was observed in our dataset, harbour porpoises (*Phocoena phocoena*) produce echolocation clicks with spectral characteristics similar to beluga ECHOs, and baleen whale vocalizations may overlap with beluga whistles. A systematic evaluation of model performance in the presence of other species would be valuable to assess whether these overlaps lead to false detections.

Given sufficient data, the simplicity of our architecture allows for straightforward extensions. Output nodes could be added to support multi-species classification, as demonstrated by [1,16]. Such an extension would increase the model's utility for edge deployment, enabling a single compact model to monitor multiple species and support a wider range of conservation applications.

5. Conclusion

This study demonstrates the feasibility of using a lightweight, quantized convolutional neural network to classify beluga calls across multiple acoustic environments in the St. Lawrence Estuary.

Our approach achieves strong performance across all acoustic categories under realistic conditions involving overlapping calls and variable background noise. Notably, the adoption of a truncated, quantized MobileNet architecture reduces the model size to under 512 KB without measurable degradation in performance. This substantial compression makes the model suitable for real-time inference on ultra-compact embedded platforms.

We also evaluated the model's robustness to anthropogenic boat noise and observed a clear degradation in performance under noisy conditions. Future work should address this limitation through targeted labeling efforts or data augmentation strategies designed to increase the representation of calls masked by vessel noise. Additionally, future research should evaluate model robustness in multi-species acoustic scenarios, which may introduce false detections.

By supporting multi-label classification, this work enables finer-grained acoustic analysis in natural field conditions, accurately identifying each category even when multiple sound types co-occur within the same segment. This capability opens the way to downstream applications such as behavioral inference, abundance estimation, and spatio-temporal monitoring of habitat use.

Overall, the methods presented here represent a concrete step toward deploying compact, accurate, and field-ready tools for real-time marine mammal monitoring. Such systems could support conservation not only through improved data collection, but also through direct mitigation strategies, for example by detecting beluga presence in real time and enabling timely vessel alerts to reduce disturbance or collision risk.

CRediT authorship contribution statement

Emmanuel Fernandez: Writing – original draft, Visualization, Software, Methodology, Data curation. **Irene Roca:** Writing – review & editing, Methodology, Data curation. **Emilie Frizzle:** Data curation. **Robert Michaud:** Resources, Investigation, Data curation. **Marie-Ana Mikus:** Data curation. **Valeria Vergara:** Writing – review & editing, Validation, Data curation. **Jaclyn Aubin:** Data curation. **Clement Chion:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Claude (Anthropic) for minor language editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs for providing full funding for this work.

Appendix A

Table A.1
Deployment details of PAM sites. Total effective recording time across all deployments: 8916 h.

Site	Year	Coordinates (long, lat)	Deployment Dates	Deployment Duration (days)	Recording Schedule	Effective Recording Time (h)	Depth (m)	Hydrophone Model	Sample Rate (Hz)
BSM	2017	−69.967525, 48.24992	7/24 – 8/18	23	Continuous	552	15	ST300HF	288 000
BSM	2018	−69.967525, 48.24992	7/7 – 8/17	41	Continuous	984	15	ST300HF	288 000
BSM	2021	−69.96673, 48.24913	7/6 – 9/7	64	Continuous	1 536	28.4	ST500HF	288 000
BSM	2022	−69.966483, 48.24913	7/15 – 10/4	64	Continuous	1 536	25	ST600HF	384 000
KAM	2020	−69.868966, 47.626388	7/21 – 10/5	68	12 h/day (08:00–20:00)	816	58	ST500HF	288 000
KAM	2021	−69.86351, 47.62425	7/13 – 9/19	68	Continuous	1 632	38.7	ST600HF	192 000
RDL	2020	−69.599403, 47.849971	7/21 – 10/12	47	12 h/day (08:00–20:00)	564	11	ST300HF	288 000
RDL	2022	−69.59964, 47.84996	7/28 – 9/6	40	Continuous	960	10	ST600HF	384 000
CAC	2021	−69.537467, 47.9449	7/13 – 8/9	28	12 h/day (08:00–20:00)	336	31.5	ST300HF	192 000

Table A.2
Per-category performance of the quantized 8-layer MobileNet averaged across all data.

Class	Precision	Recall	F1	Accuracy	Bal. Acc.	False Alarm Rate
ECHO	0.96	0.89	0.92	0.95	0.94	0.02
HFPC	0.85	0.83	0.83	0.98	0.91	0.01
BBPC	0.83	0.77	0.80	0.98	0.88	0.01
Whistle	0.88	0.87	0.88	0.96	0.92	0.02
Macro Avg	**0.88**	**0.84**	**0.86**	**0.97**	**0.91**	**0.02**

Table A.3
False alarm rates per acoustic category evaluated on extended beluga-free ambient segments (n = 1917 windows, >30min of audio). These segments were selected from periods with no annotated beluga activity, providing a controlled estimate of model behavior in the absence of vocalizations.

	ECHO	HFPC	BBPC	Whistle
False Alarm Rate	0.018	0.000	0.007	0.005
False positives	35	0	13	10

Table A.4
Per-category performance of the quantized 8-layer MobileNet by background condition (Ambient/Boat).

Class	Precision	Recall	F1	Accuracy	Bal. Acc.	False Alarm Rate
ECHO	0.93/0.97	0.84/0.92	0.88/0.94	0.96/0.95	0.92/0.95	0.01/0.03
HFPC	0.86/0.84	0.88/0.73	0.87/0.77	0.99/0.98	0.94/0.86	0.01/0.01
BBPC	0.88/0.74	0.80/0.68	0.84/0.67	0.98/0.97	0.90/0.83	0.01/0.01
Whistle	0.89/0.76	0.90/0.67	0.90/0.70	0.95/0.97	0.94/0.83	0.03/0.01
Macro Avg	**0.89/0.83**	**0.86/0.75**	**0.87/0.77**	**0.97/0.97**	**0.92/0.87**	**0.02/0.02**

Data availability

Data will be made available on request.

References

[1] Allen AN, Harvey M, Harrell L, Wood M, Szesciorka AR, McCullough JLK, Oleson EM. Bryde's whales produce biotwang calls, which occur seasonally in long-term acoustic recordings from the Central and Western North Pacific. Front Mar Sci 2024;1394695. https://doi.org/10.3389/fmars.2024.1394695

[2] Ames AE, Beedholm K, Madsen PT. Lateralized sound production in the beluga whale (delphinapterus leucas). J Exp Biol 2020;jeb226316. https://doi.org/10.1242/jeb.226316

[3] Aubin JA. Vocal communication, social structure, and responses to disturbance in belugas (delphinapterus leucas) [Ph.D. thesis]. University of Windsor (Canada); 2025. https://hdl.handle.net/20.500.14776/12251.

[4] Baumgartner MF, Bonnell J, Van Parijs SM, Corkeron PJ, Hotchkin C, Ball K, Pelletier LP, Partan J, Peters D, Kemp J, et al. Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: system description and evaluation. Methods Ecol Evol 2019;1476–89. https://doi.org/10.1111/2041-210X.13244

[5] Belikov R, Bel'Kovich V. High-pitched tonal signals of beluga whales (delphinapterus leucas) in a summer assemblage off solovetskii island in the white sea. Acoust Phys 2006;125–31. https://doi.org/10.1134/S1063771006020023

[6] Belikov R, Bel'Kovich V. Whistles of beluga whales in the reproductive gathering off solovetskii island in the white sea. Acoust Phys 2007;528–34. https://doi.org/10.1134/S1063771007040148

[7] Belikov R, Bel'kovich V. Communicative pulsed signals of beluga whales in the reproductive gathering off solovetskii island in the white sea. Acoust Phys 2008;115–23. https://doi.org/10.1134/S1063771008010168

[8] Bergler C, Schröter H, Cheng RX, Barth V, Weber M, Nöth E, Hofer H, Maier A. ORCA-SPOT: an automatic killer whale sound detection toolkit using deep learning. Sci Rep 2019;10997. https://doi.org/10.1038/s41598-019-47335-w

[9] Bermant PC. BioCPPNet: automatic bioacoustic source separation with deep neural networks. Sci Rep 2021;23502. https://doi.org/10.1038/s41598-021-02790-2

[10] Bonnell TR, Michaud R, Dupuch A, Lesage V, Chion C. Estimating spatial mixing within the St. Lawrence estuary beluga population by comparing local individual diversity and abundance. Marine Mammal Science 2024;e13162. https://doi.org/10.1111/mms.13162

[11] Cañas JS, Toro-Gómez MP, Sugai LSM, Benítez Restrepo HD, Rudas J, Posso Bautista B, Toledo LF, Dena S, Domingos AHR, de Souza FL, et al. A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. Sci Data 2023;771. https://doi.org/10.1038/s41597-023-02666-2

[12] Castellote M, Small RJ, Lammers MO, Jenniges J, Mondragon J, Garner CD, Atkinson S, Delevaux JM, Graham R, Westerholt D. Seasonal distribution and foraging occurrence of cook inlet beluga whales based on passive acoustic monitoring. Endangered Species Research 2020;225–43. https://doi.org/10.3354/esr01023

[13] Chmelnitsky EG, Ferguson SH. Beluga whale, delphinapterus leucas, vocalizations from the Churchill river, Manitoba, Canada. The Journal of the Acoustical Society of America 2012;4821–35. https://doi.org/10.1121/1.4707501

[14] Cotillard T, Sécheresse X, Aubin J, Mikus MA, Vergara V, Gambs S, Michaud R, Martins CCA, Turgeon S, Chion C, Roca I. Automatic detection and classification of beluga whale calls in the St. Lawrence estuary. The Journal of the Acoustical Society of America 2024;3723–40. https://doi.org/10.1121/10.0030472

[15] Denton T, Wisdom S, Hershey JR. Improving bird classification with unsupervised sound separation. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2022. p. 636–40. https://doi.org/10.48550/arXiv.2110.03209

[16] Duan D, Lü LG, Jiang Y, Liu Z, Yang C, Guo J, Wang X. Real-time identification of marine mammal calls based on convolutional neural networks. Appl Acoust 2022;108755. https://doi.org/10.1016/j.apacoust.2022.108755

[17] Finkelstein A, Almog U, Grobman M. Fighting quantization bias with bias. arXiv, arXiv:1906.03193. 2019.

[18] Fish MP, Mowbray WH. Production of underwater sound by the white whale or beluga, delphinapterus leucas (pallas). Journal of Marine Research 1962;149–62.

[19] Fisheries and Oceans Canada. Recovery strategy for the beluga whale (delphinapterus leucas) St. Lawrence estuary population in Canada. species at risk act recovery strategy series Fisheries and Oceans Canada; 2012.

[20] Frazao F, Joy R, Dowd M. Comparing acoustic representations for deep learning-based classification of underwater acoustic signals: a case study on orca (orcinus orca) vocalizations. Ecol Inform 2025;103297. https://doi.org/10.1016/j.ecoinf.2025.103297

[21] Hagiwara M. Aves: animal vocalization encoder based on self-supervision. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2023. p. 1–5. https://doi.org/10.48550/arXiv.2210.14493

[22] Halliday WD, Scharffenberg K, MacPhee S, Hilliard RC, Mouy X, Whalen D, Loseto LL, Insley SJ. Beluga vocalizations decrease in response to vessel traffic in the Mackenzie river estuary. Arctic 2019;337–46. https://doi.org/10.14430/arctic69294

[23] Hamard Q, Pham MT, Cazau D, Heerah K. A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data. Ecol Inform 2024;102906. https://doi.org/10.1016/j.ecoinf.2024.102906

[24] Harvey V, Mosnier A, St-Pierre AP, Lesage V, Gosselin JF. Seasonal variation in distribution and habitat use of St. Lawrence estuary beluga (*Delphinapterus leucas*) estimated from systematic photographic and visual line-transect aerial surveys. Res. Doc. 2025/075 DFO Can. Sci. Advis. Sec.; 2025, iv + 81.

[25] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, et al. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 1314–24. https://doi.org/10.1109/ICCV.2019.00140

[26] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv, arXiv:1704.04861. 2017.
[27] Hyer MD, Anderson AT, Mann DA, Mooney TA, Aoki N, Jensen FH. Robust real-time detection of right whale upcalls using neural networks on the edge. Ecol Inform 2025;103130. https://doi.org/10.1016/j.ecoinf.2025.103130
[28] Jiang JJ, Bu LR, Duan FJ, Wang XQ, Liu W, Sun ZB, Li CY. Whistle detection and classification for whales based on convolutional neural networks. Appl Acoust 2019;169–78. https://doi.org/10.1016/j.apacoust.2019.02.007
[29] Karlsen J, Bisther A, Lydersen C, Haug T, Kovacs K. Summer vocalisations of adult Male white whales (delphinapterus leucas) in Svalbard, Norway. Polar Biology 2002;808–17. https://doi.org/10.1007/s00300-002-0415-6
[30] Kather V, Seipel F, Berges B, Davis G, Gibson C, Harvey M, Henry LA, Stevenson A, Risch D. Development of a machine learning detector for North Atlantic humpback whale song. The Journal of the Acoustical Society of America 2024;2050–64. https://doi.org/10.1121/10.0025275
[31] Kingsley MC. Status of the belugas of the st Lawrence estuary, Canada. NAMMCO Scientific Publications 2002;239–58. https://doi.org/10.7557/3.2847
[32] Lefebvre SL, Michaud R, Lesage V, Berteaux D. Identifying high residency areas of the threatened St. Lawrence beluga whale from fine-scale movements of individuals and coarse-scale movements of herds. Mar Ecol Prog Ser 2012;243–57. https://doi.org/10.3354/meps09570
[33] Lesage V. The challenges of a small population exposed to multiple anthropogenic stressors and a changing climate: the St. Lawrence estuary beluga. Polar research 2021. https://doi.org/10.33265/polar.v40.5523
[34] Lesage V, Barrette C, Kingsley MCS, Sjare B. The effect of vessel noise on the vocal behavior of belugas in the st. Lawrence river estuary, Canada. Marine Mammal Science 1999;65–84. https://doi.org/10.1111/j.1748-7692.1999.tb00782.x
[35] Lesage V, Hammill MO, Kovacs KM. Marine mammals and the community structure of the estuary and gulf of st Lawrence, Canada: evidence from stable isotope analysis. Mar Ecol Prog Ser 2001;203–21. https://doi.org/10.3354/meps210203
[36] Lesage V, McQuinn IH, Carrier D, Gosselin J, Mosnier A. Exposure of the beluga (delphinapterus leucas) to marine traffic under various scenarios of transit route diversion in the St. Lawrence estuary. Canadian Science Advisory Secretariat; 2014.
[37] Martins C, Turgeon S, Michaud R, Ménard N. Seasonal occurrence and spatial distribution of four species of baleen whales vulnerable to ship strikes in the Saguenay-St. Lawrence marine park (Quebec, Canada). Canadian Science Advisory Secretariat (CSAS); 2022.
[38] Noda T, Koizumi T, Yukitake N, Yamamoto D, Nakaizumi T, Tanaka K, Okuyama J, Ichikawa K, Hara T. Animal-Borne soundscape logger as a system for edge classification of sound sources and data transmission for monitoring near-real-time underwater soundscape. Sci Rep 2024;6394. https://doi.org/10.1038/s41598-024-56439-x
[39] Oedekoven CS, Marques TA, Harris D, Thomas L, Thode AM, Blackwell SB, Conrad AS, Kim KH. A comparison of three methods for estimating call densities of migrating bowhead whales using passive acoustic monitoring. Environmental and Ecological Statistics 2022;101–25. https://doi.org/10.1007/s10651-021-00506-3
[40] Okujeni A, Canters F, Cooper SD, Degerickx J, Heiden U, Hostert P, Priem F, Roberts DA, Somers B, van der Linden S. Generalizing machine learning regression models using multi-site spectral libraries for mapping vegetation-impervious-soil fractions across multiple cities. Remote Sens Environ 2018;482–96. https://doi.org/10.1016/j.rse.2018.07.011
[41] Panova E, Belikov R, Agafonov A, Bel'Kovich V. The relationship between the behavioral activity and the underwater vocalization of the beluga whale (delphinapterus leucas). Oceanology 2012;79–87. https://doi.org/10.1134/S000143701201016X
[42] Poupard M, Ferrari M, Best P, Glotin H. Passive acoustic monitoring of sperm whales and anthropogenic noise using stereophonic recordings in the Mediterranean Sea, North West pelagos sanctuary. Sci Rep 2022;2007. https://doi.org/10.1038/s41598-022-05917-1
[43] Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv, arXiv:1811.12808. 2018.
[44] Roca IT, Kaleschke L, Van Opzeeland I. Sea-ice anomalies affect the acoustic presence of Antarctic pinnipeds in breeding areas. Front Ecol Environ 2023;227–33. https://doi.org/10.1002/fee.2622
[45] Shiu Y, Palmer K, Roch MA, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Klinck H. Deep neural networks for automated detection of marine mammal species. Sci Rep 2020;607. https://doi.org/10.1038/s41598-020-57549-y
[46] Simard Y, Roy N, Giard S, Gervaise C, Conversano M, Ménard N. Estimating whale density from their whistling activity: example with St. Lawrence beluga. Appl Acoust 2010;1081–6. https://doi.org/10.1016/j.apacoust.2010.05.013
[47] Sjare B, Smith T. The vocal repertoire of white whales, delphinapterus leucas, summering in cunningham inlet, Northwest Territories. Canadian Journal of Zoology 1986;407–15. https://doi.org/10.1139/z86-063
[48] Stowell D. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 2022;e13152. https://doi.org/10.7717/peerj.13152
[49] Todd NR, Cronin M, Luck C, Bennison A, Jessopp M, Kavanagh AS. Using passive acoustic monitoring to investigate the occurrence of cetaceans in a protected marine area in northwest Ireland. Estuar Coast Shelf Sci 2020;106509. https://doi.org/10.1016/j.ecss.2019.106509
[50] Vergara V, Barrett-Lennard LG. Vocal development in a beluga calf (delphinapterus leucas). Aquatic Mammals 2008;123–43. https://doi.org/10.1578/AM.34.1.2008.123
[51] Vergara V, Michaud R, Barrett-Lennard L. What can captive whales tell us about their wild counterparts? Identification, usage, and ontogeny of contact calls in belugas (delphinapterus leucas). International Journal of Comparative Psychology 2010. https://doi.org/10.46867/ijcp.2010.23.03.08
[52] Vergara V, Mikus MA. Contact call diversity in natural beluga entrapments in an arctic estuary: preliminary evidence of vocal signatures in wild belugas. Marine Mammal Science 2019;434–65. https://doi.org/10.1111/mms.12538
[53] Vergara V, Mikus MA, Chion C, Lagrois D, Marcoux M, Michaud R. Effects of vessel noise on beluga (delphinapterus leucas) call type use: ultrasonic communication as an adaptation to noisy environments? Biol Open 2025;bio061783. https://doi.org/10.1242/bio.061783
[54] Vergara V, Wood J, Lesage V, Ames A, Mikus MA, Michaud R. Can you hear me? Impacts of underwater noise on communication space of adult, sub-adult and calf contact calls of endangered St. Lawrence belugas (delphinapterus leucas). Polar Research 2021. https://doi.org/10.33265/polar.v40.5521
[55] White EL, Klinck H, Bull JM, White PR, Risch D. One size fits all? Adaptation of trained CNNs to new marine acoustic environments. Ecol Inform 2023;102363. https://doi.org/10.1016/j.ecoinf.2023.102363
[56] Zhong M, Castellote M, Dodhia R, Lavista Ferres J, Keogh M, Brewer A. Beluga whale acoustic signal classification using deep learning neural network models. The Journal of the Acoustical Society of America 2020;1834–41. https://doi.org/10.1121/10.0000921
[57] Zhong M, Castellote M, Dodhia R, Lavista Ferres J, Keogh M, Brewer A. Beluga whale acoustic signal classification using deep learning neural network models. The Journal of the Acoustical Society of America 2020;1834–41. https://doi.org/10.1121/10.0000921