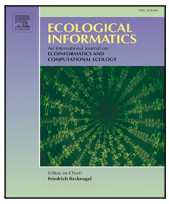
Ecological Informatics 89 (2025) 103130

Contents lists available at ScienceDirect

Ecological Informatics

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



Robust real-time detection of right whale upcalls using neural networks on the edge

Matthew D. Hyer [a,*], Austin T. Anderson [b], David A. Mann [b], T. Aran Mooney [c], Nadège Aoki [c,d], Frants H. Jensen [a,c,e]

[a] Department of Ecoscience, Aarhus University, Frederiksborgvej 399, 4000 Roskilde, Denmark
[b] Loggerhead Instruments, 6576 Palmer Park Circle, Sarasota, FL, 34238, USA
[c] Biology Department, Woods Hole Oceanographic Institution, 266 Woods Hole Road, Woods Hole, MA, 02543, USA
[d] Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 55-101, Cambridge, MA, 02139, USA
[e] Department of Biology, Syracuse University, 107 College Pl, Syracuse, NY, 13244, USA

ARTICLE INFO

Code link: https://doi.org/10.5281/zenodo.13851976

Keywords:
Data augmentation
Deep learning
North Atlantic right whale
Real-time acoustic detection
Sustainable marine development
Wildlife monitoring

ABSTRACT

Animals worldwide are facing ecological pressures from global climate change and increasing anthropogenic activities. To transition to a renewable energy future, extensive offshore wind development is planned globally. In the North Atlantic, future development sites overlap with the migratory range of critically endangered North Atlantic right whales (NARW) and will lead to increased risk of ship strikes, pile driving impacts, and other population risks. New methods to accurately detect cetaceans and provide real-time feedback for mitigation will be increasingly important to enact sustainable management actions to facilitate the recovery of the NARW. Recent developments in acoustic event detection made possible by deep learning have shown significantly improved detection performance across many different taxa, but such models tend to be too computationally expensive to run on existing wildlife monitoring platforms. Here, we use model compression techniques combined with an autonomous acoustic recording platform integrating an ESP32 microcontroller to bring real-time detection with deep learning to the edge. We test if edge-based inference using a compressed network running on a microprocessor entails significant performance loss and find that this loss is negligible. We leverage large, open-source datasets of noise from the NOAA SanctSound project for generating semi-synthetic training datasets that encourage model generalization to novel noise conditions. Our compressed model achieves improved performance across all tested recording sites in the Western North Atlantic Ocean, demonstrating that deep learning powered wildlife monitoring solutions can provide reliable real-time data for mitigation of human impacts and help ensure a sustainable green energy transition.

1. Introduction

Passive acoustic monitoring (PAM) has emerged as a pivotal tool in wildlife conservation, offering non-invasive means to monitor threatened species across diverse habitats (Gibb et al., 2019; Mellinger et al., 2007). Acoustic monitoring leverages sound produced by animals themselves to gather crucial data on their presence, behavior, population density, and dynamics; this is particularly true for species that are elusive or inhabit remote areas (Blumstein et al., 2011; Gillespie et al., 2020; Hutschenreiter et al., 2024; Van Parijs et al., 2009). PAM is especially applicable to the conservation and management of marine species that rely on sound for navigation, socialization, and foraging, and has become a standard component of marine population monitoring (Fleishman et al., 2023) based on a wide range of acoustic recording systems (reviewed in Sousa-Lima et al., 2013). As human activities increasingly encroach on natural habitats, PAM provides a method to assess the impacts of anthropogenic noise, habitat changes, and other disturbances on wildlife, thereby informing conservation strategies and policy decisions. When combined with real-time acoustic detection and transmission, PAM offers a unique opportunity to actively mitigate species loss via targeted interventions that inform dynamic changes in human activities like the cessation of offshore wind construction or temporary changes to dynamic vessel speed limits (Van Parijs et al., 2009). Such improved tools for monitoring species distribution and abundance, and aiding conservation and management actions are in high demand in light of the current biodiversity crisis (Singh, 2002; Keck et al., 2025).

* Corresponding author.
E-mail address: mhyer.eco@gmail.com (M.D. Hyer).

https://doi.org/10.1016/j.ecoinf.2025.103130
Received 4 October 2024; Received in revised form 1 April 2025; Accepted 1 April 2025
Available online 17 April 2025

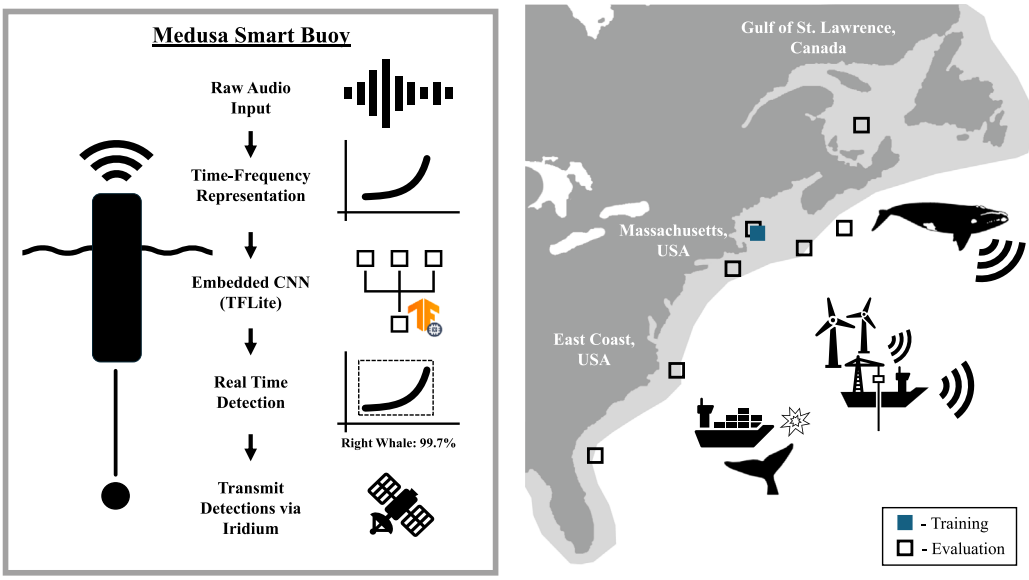


Fig. 1. Visualization of data sources and Medusa smart buoy. Left: Conceptual diagram of the Medusa buoy's real-time signal processing chain. Right: Map of the East Coast of the United States and Canada. The blue square represents Stellwagen Bank National Marine Sanctuary where calls for training originated. White squares denote manually annotated evaluation sites. The light gray area off the coast represents the approximate NARW migration range adapted from Fig. 1 in Hunt et al. (2015). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Among marine species, the North Atlantic right whale (NARW - *Eubalaena glacialis*) exemplifies the importance of acoustic monitoring. The NARW is a critically endangered baleen whale with around 372 remaining individuals (Pettis and Hamilton, 2024). Ship strikes and fishing gear entanglements remain the primary anthropogenic contributions to NARW mortality (Knowlton and Kraus, 2001; Silber and Bettridge, 2012; Pettis and Hamilton, 2024) despite conservation efforts like vessel speed limits in NARW calving and foraging areas (NOAA and NMFS, 2008). Furthermore, ongoing and planned offshore wind developments along the East Coast of the United States and Canada overlap with the NARW migratory range (shown in light grey in Fig. 1) and will increase vessel traffic and potentially harmful pile driving exposures (Madsen et al., 2006), underscoring the need for effective tools to aid conservation efforts.

NARWs produce a variety of sounds including impulsive gunshot calls (primarily vocalized in surface active groups) and long-range upcalls that seem to function as contact signals (Matthews and Parks, 2021). Upcalls are produced by both socializing and isolated whales and are the standard vocalization for detecting NARW populations (Parks et al., 2011; Urazghildiiev and Clark, 2007; Van Parijs et al., 2009; Wade et al., 2006). Typical approaches to upcall detection leverage lightweight feature extraction algorithms operating on spectrogram representations of upcalls (Baumgartner and Mussoline, 2011; Gillespie, 2004). Such algorithms have been widely used for real-time detection on both moored listening systems and gliders (Baumgartner et al., 2013; Palmer et al., 2022).

Recently, Shiu et al. (2020) demonstrated that convolutional neural networks (CNNs) are capable of outperforming traditional feature-based techniques for detecting NARW upcalls by a large margin. Palmer et al. (2022) leverages this neural network in a two-stage detection algorithm which first detects vocalizations using an edge-based detector (Gillespie, 2004), then transmits the signal to a shore-based computer running the neural network which verifies the detection before notifying a user. This configuration results in significantly more accurate detections, but recovers fewer low-amplitude signals than a pure CNN approach. Padovese et al. (2021) demonstrated that data augmentation techniques can further improve the performance of a neural network, particularly in cases of data scarcity which is common when working with endangered species. Data augmentation also enables researchers to incorporate acoustic information from a wide variety of contexts, potentially improving detector performance in novel environments (Stowell, 2022).

Here, we introduce the first CNN for endangered marine species detection capable of running in real-time on a microcontroller by leveraging compressed neural networks. We propose a new data augmentation framework simulating the presence of target signals in highly diverse acoustic soundscapes to improve detection performance across a wide geographic range. We show that model compression has negligible impact on detection performance, and our final model demonstrates significantly improved real-time detection performance across the NARW migratory range.

2. Methods

We develop a robust, real-time NARW detection algorithm suitable for deployment on a CNN-capable acoustic recording unit. We train this network using a mix of real right whale upcalls and semi-synthetic clips including known right whale upcalls inserted into diverse soundscapes to improve model performance in novel noise conditions. Using full-integer quantization, we compress the resulting network to enable real-time operation onboard a microcontroller. We perform a robust performance evaluation of our network to understand detection performance and long-term stability in novel noise contexts. Lastly, we characterize detection performance as a function of signal-to-noise ratio (SNR) to estimate the active space as a function of masking noise level, and compare this to other established NARW detection algorithms.

2.1. Training models for deployment on hardware

We design our feature extraction and processing pipeline for deployment on a Medusa smart buoy, an acoustic monitoring solution currently being developed by Loggerhead Instruments (Sarasota, FL, USA) (Mann et al., 2024). The Medusa utilizes an ESP32-S3-WROOM-1 (hereafter termed ESP) microcontroller developed by Espressif Systems (Shanghai, China) to facilitate distributed real-time detection algorithms running onboard multiple buoys. All models are trained in a desktop environment using full precision floating point (FP-32) operations and data. Trained models are compressed to 8-bit signed

integer (Int-8) format using post-training quantization in TensorFlow Lite Micro (TFLite) before being loaded onto the Medusa microcontroller (Abadi et al., 2015). To facilitate clear downstream comparisons with the different configurations of our approach, we refer to the uncompressed FP-32 models as 'NARWnet' and the compressed Int-8 models as 'NARWnet-Lite'.

2.2. Data sources

We utilize several commonly used datasets for network training and evaluation (see Supplemental Section S1 for a summarizing table). The 2013 Detection, Classification, Localization, and Density Estimation (DCLDE) workshop focused on detecting NARW upcalls and gunshots and provided seven days of continuous recordings from Marine Autonomous Recording Units (MARUs) deployed at the Stellwagen Bank National Marine Sanctuary (SBNMS) off the coast of Massachusetts in 2009 (blue square in Fig. 1) (Gillespie, 2019). Recordings were manually annotated for individual upcalls. Following workshop procedure, the first four days containing 6916 upcalls are used for training and the next three days containing 2766 upcalls are used for evaluation.

To supplement the raw DCLDE upcall recordings, we incorporate additional upcalls recorded at SBNMS for data augmentation. The Marinexplore and Cornell University Whale Detection Kaggle Challenge (hereafter termed Kaggle dataset) contains two-second clips of manually verified upcalls and noise generated from an automated upcall detection system (Karpištšenko et al., 2013). We select a subset of 600 sufficiently high SNR upcalls for generating a semi-synthetic training dataset by visually browsing spectrograms of upcalls for distinct and clear signals.

We utilize a large dataset of background noise recordings during data augmentation to expose the network to diverse acoustic contexts during training. The National Oceanic and Atmospheric Administration (NOAA) Sanctuary Soundscape Monitoring Project (SanctSound) deployed passive acoustic recorders at eight marine sites around the United States in 2018 and has made the recordings freely available (NOAA Office of National Marine Sanctuaries and US Navy, 2020). From each location, we extract one 10-s recording every 30 min for one day a month throughout 2019, except for one site in Hawaii where recordings were only available from December to May at the time of access. These randomly subsampled SanctSound clips are used as noise data for generating semi-synthetic training datasets under diverse background noise conditions.

To evaluate model generalizability, we utilize two additional hand-labeled datasets. Recently, NOAA, along with the Northeast Fisheries Science Center (NEFSC), released a combined dataset totaling eight days of continuous recordings from six sites throughout the Western North Atlantic (white squares in Fig. 1, not including Gulf of St. Lawrence, Canada) (Pacific Marine Environmental Laboratory et al., 2023). Annotators labeled individual upcalls at two levels of confidence to include cases where there are possible right whale upcalls that could not be confidently assigned a species. To directly compare model performance to the human annotator, we evaluate the number of recovered upcalls using the high confidence labels, and false detections against the lower confidence labels to ensure that we do not assign a false detection to a signal that resembles an upcall, but is impossible for a human to distinguish. We supplement this data with the B* test dataset produced by Kirsebom et al. (2020) which contains fifty 30 min recordings annotated by hand for individual upcalls from the Gulf of St. Lawrence, Canada between 2015 and 2017, amounting to 1157 additional upcalls (white square in the Gulf of St. Lawrence, Canada in Fig. 1) (Simard et al., 2020). Together, these datasets amount to nine days of expert annotated, unseen recordings throughout the migratory range of the NARW, providing an openly available benchmark dataset with which to evaluate our network.

In addition to continuous detection performance, we are also interested in our network's ability to detect upcalls as a function of SNR to help quantify the detection range based on background noise conditions. We also utilize the B and C upcall clips and SNR values from Kirsebom et al. (2020) comprised of 3309 and 3000 clips respectively. The B dataset contains expert verified three-second clips generated by a non-deep learning detector recorded during 2018 from the same deployment location as B*, see Kirsebom et al. (2020) and Simard et al. (2020) for more information about the detector and dataset. The C dataset contains a subset of extracted clips from the DCLDE dataset described above. Kirsebom et al. (2020) also computed SNR values for all clips which we use as ground truth values to estimate detection performance across SNR. We also select a subset of 150 high SNR upcalls from the B dataset to use as an unseen set of positive samples during model evaluation using the same SNR criterion outlined above.

2.3. Spectrogram generation

Following previous works, we treat the task of detecting NARW upcalls as an image classification problem and convert acoustic signals to two-dimensional, single channel spectrogram 'images' (Goëau et al., 2016; Stowell, 2022). After resampling audio to 1 kHz using the Resampy package (McFee, 2016) in Python, we calculate a Fast Fourier Transform (FFT) in Tensorflow with a 256 ms Hann window, a 42 ms step length, and a window length of 3 s. We then crop out frequency bins 13 through 77 to construct a tight spectrogram of 64 × 64 pixels representing around 50–300 Hz (app. 4 Hz spectral resolution) and from 0 to 3 s (app. 50 ms temporal resolution) (Fig. 2). To generate an equivalent spectrogram on the ESP microcontroller, we process a 3-s audio buffer with an 8 kHz sampling rate and utilize a 2048-point FFT (equivalent to 256 ms) with a 336-point step length (equivalent to 42 ms).

2.4. Data augmentation framework

To improve performance in novel environments, we design a data augmentation pipeline that represents NARW upcalls in new contexts. First, we select a subset of 600 high SNR upcalls from the two-second Cornell Kaggle clips and track the fundamental frequency contour in 0.2 s steps using custom Matlab software. We use a time-frequency filtering algorithm described in Madsen et al. (2012) to isolate the energy from the fundamental frequency resulting in 'clean' upcall waveforms. We randomly implement time stretching and pitch shifting (Wei et al., 2020; Xu et al., 2018) on de-noised signals of up to 20% and 10% respectively. These represent conservative values based on the natural variation found in upcall production (Matthews and Parks, 2021). Ultimately, signals are injected into NOAA SanctSound recordings at random SNR between −12 and 12 decibels (dB) measured in the 50–225 Hz band. A random offset was introduced to shift signals within the three-second window while still displaying the full contour.

2.5. Training configurations and compressed model architecture

To demonstrate the efficacy of our data augmentation strategy, we train models with multiple data augmentation configurations. As a direct comparison with Shiu et al. (2020) which uses the DCLDE dataset for training, we first train a model with no augmentation. Three-second windows are extracted around all 6916 labeled upcalls in the DCLDE dataset. As above, upcalls are shifted randomly in time so that the full upcall contour is present in the three-second window, and an additional 6916 noise segments are randomly sampled from the periods of time between labeled upcalls.

Next, we apply our data augmentation strategy and mix clean upcalls from the Kaggle dataset with randomly selected background noise segments from SanctSound. Here, we use a 30x upscaling factor to produce 18,000 augmented upcalls, as well as an additional 18,000 noise segments. We trained models using (a) only synthetic data, and

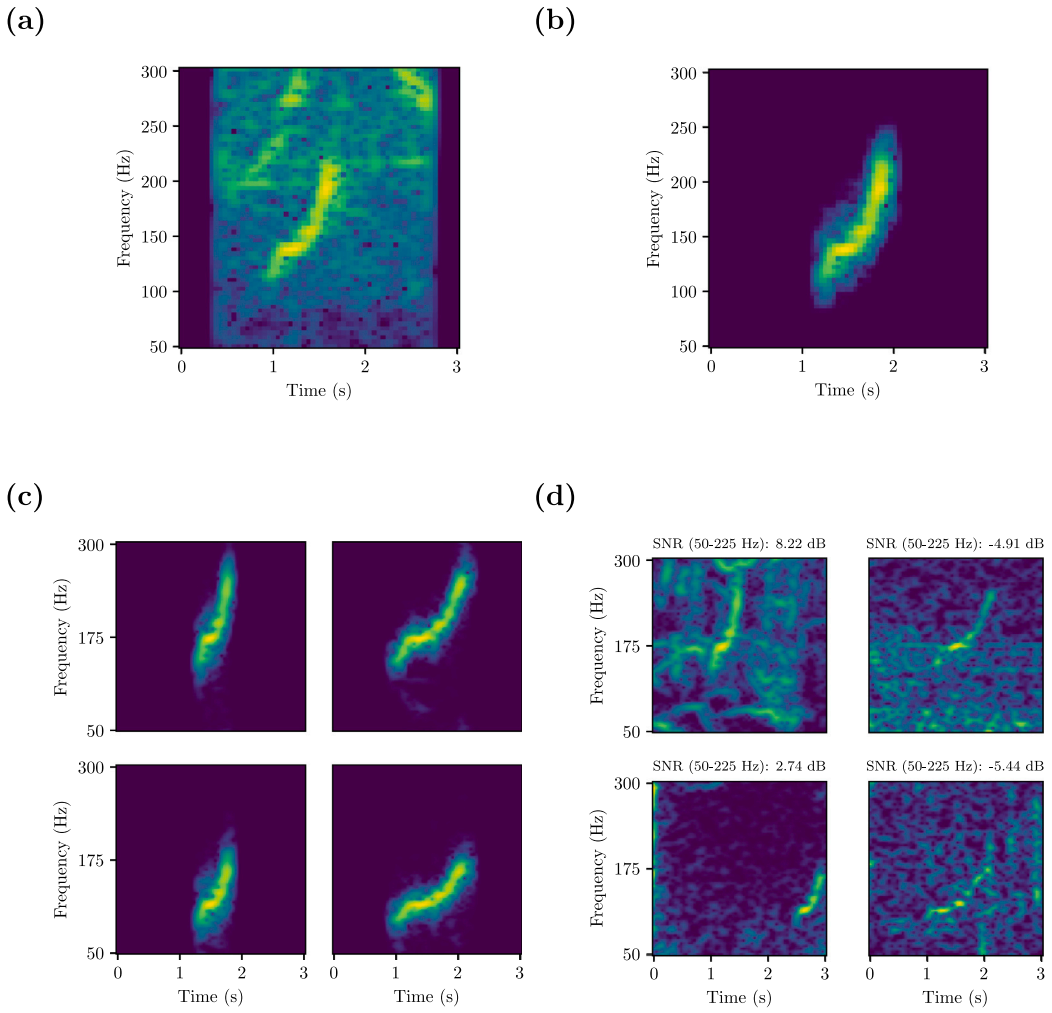


Fig. 2. Data extraction and augmentation. The full data augmentation pipeline showing (a) raw two-second clips from the Kaggle dataset, (b) the de-noised upcall contour, (c) the same contour with different degrees of pitch shifting and time stretching, and (d) time shifting and noise mixing with upcalls embedded in NOAA SanctSound clips.

(b) a combination of the original and synthetic datasets totaling 49,832 three-second clips.

Models were trained using either an NVIDIA A6000 Desktop or an NVIDIA RTX 3070 Laptop. Model training was done in Tensorflow 2.10.1 for 50 epochs using the Adam optimizer with a learning rate of 0.001, β_1 of 0.9, β_2 of 0.999, ϵ of 1e−7, and a batch size of 64 (Kingma and Ba, 2014). Following He et al. (2016) we do not train with dropout. To monitor progress across all data configurations throughout training, we use the same test dataset made up of 2766 upcalls and an equal number of noise segments from the last three days of DCLDE recordings. As a result, the performance on the DCLDE test dataset may be inflated due to observer bias when selecting a model, even if models were never trained directly with the test data. However, we consider the standard DCLDE 2013 train/test splits to be auto-correlated as they are sequential in nature. As such, we denote the 2013 test dataset a 'familiar' noise environment.

We implement a Resnet style CNN based on the Tensorflow implementation of Resnet-50, but with significantly fewer parameters (He et al., 2016). We first learn a batch normalization on the spectrogram input to discourage very small or very large weights throughout inference to aid in downstream quantization (Ioffe and Szegedy, 2015; LeCun et al., 2002). The rest of the model architecture closely resembles (He et al., 2016), but with smaller convolution kernels and fewer layers throughout (Supplemental Info Section S3). For single class detection, we convert the 256 convolution output features to a distribution between 0 and 1 using a densely connected prediction layer with a sigmoid activation. In total, our NARWnet and NARWnet-Lite networks are made up of around 375,000 trainable parameters.

For model compression, we utilize full integer post-training quantization using TFLite and convert spectrogram inputs and model weights to 8-bit signed integer format. We treat each model's training data as its representative dataset during compression for simplicity. Lastly, the resulting TFLite model is converted to a C source file for compilation into the larger Medusa firmware, along with its corresponding scale and shift factors for operation on the edge.

2.6. Performance assessment using continuous data

To process continuous recordings with uncompressed NARWnet models, we inference every 0.1 s to produce a near continuous time-series of model outputs following Shiu et al. (2020). To mirror real world operating conditions, we evaluate compressed NARWnet-Lite networks every 0.5 s, generating a coarser time-series output. In both cases, we compare the network output with a predefined detection threshold t to determine if an upcall is present or not. The start and end points of a detection correspond to the midpoints of three-second detection windows at which the network output crosses t. Similar to Kirsebom et al. (2020), we apply a moving average to the output time-series to account for transient spikes or misses in the network output. We determine the moving average for each model empirically by testing values between 50%–85% of the window length. See Supplemental Section S2 for a detailed figure showing the continuous data processing pipeline.

To evaluate model performance, we calculate precision, recall, and the number of false positives using the scoring tool provided in the DCLDE 2015 Workshop (Roch, 2015). Precision and recall are common metrics for model evaluation; precision quantifies the proportion of correct predictions, while recall quantifies the proportion of calls that were detected, whereas false positives are incorrect predictions (see Roch, 2015 for detailed explanations). Since continuous data is primarily made up of ambient noise, these metrics do not account for background noise 'detections' which are a common component in model evaluations. Instead, the rate at which a model incorrectly classifies background noise as an upcall indicates its robustness to variable background noise conditions. Following previous works, we report overall model performance by calculating the Area under the Precision–Recall Curve (AUC) using Scikit-Learn (Pedregosa et al., 2011) in Python. AUC curves are generated by computing continuous detections at fifty detection threshold points between 0.1 and 1.0. Shiu et al. (2020) suggest that 20 FP/H is the maximum false detection rate that can be verified by a human analyst, so we also report the recall intercept at 20 and 5 false positives per hour (FP/H) calculated by dividing the number of false positives by the total recording time to offer insight into different operating modes for our model; a model configured to run at 20 FP/H will require more user supervision compared to a model at 5 FP/H.

While AUC scores offer descriptive results that encapsulate both recall and precision, real-time operation necessitates a single, fixed detection threshold. Thus, quantifying a model's 'fixed' performance by using a single threshold in different contexts is crucial to facilitate consistent and comparable performance across use cases. We leverage the detection threshold values corresponding to an average of 5 and 20 FP/H on the Kirsebom B* dataset to understand how each model behaves 'out of the box' in novel acoustic environments. The Kirsebom B* dataset provides a unique perspective into the long-term variability of a detection algorithm in different contexts. The fifty 30 min recordings represent unique seasonal and diel contexts between 2015 and 2017. To visualize performance, we plot the precision and recall for each of the fifty sites along with the kernel density function computed using the $gaussian\_kde$ function with four equipotential lines from the SciPy python package (Virtanen et al., 2020). Furthermore, the NEFSC dataset comprises unseen recordings in mostly novel contexts throughout the NARW migratory corridor. We test if a single detection configuration is sufficient to enable consistent, distributed detection performance across multiple sites.

2.7. Detection probability vs. SNR

To assess model performance in different background noise conditions, we report the Recall at 20 FP/H across SNR using our data augmentation pipeline. In order to investigate performance of the continuous detection framework, we leverage our data augmentation strategy to create semi-synthetic 30-s clips with a single, centered upcall. Using the unseen subset of 150 high SNR calls taken from the B dataset (Kirsebom et al., 2020), we first extract 'clean' upcall contours using the same procedure from the data augmentation step. Next, we mix clean upcalls with 30-s background noise clips from periods between upcalls in the B* dataset (Kirsebom et al., 2020). We generate four, 30-s clips per clean upcall at 20 SNR levels from −20 to 20 dB for a total of 12,000 semi-synthetic clips, each with a single upcall at a specified SNR value.

By leveraging the results of the theoretical detection curve, we estimate the detection range of our network using the noise propagation model and ambient noise measurements off the coast of Maryland, USA described by Bailey et al. (2018). The model is defined as $RL = SL - 16.1log_{10}(R)$ where R is the distance (m) from source to receiver, RL is received level, and SL is the source level. Here, RL, SL, and noise level (NL) measurements are measured between 70.8–224 Hz and are reported as root-mean-square (rms) values. If a call is detected when the RL exceeds the background NL by a critical SNR corresponding to 50% detection probability, we can calculate detection range using $R = 10^{\frac{SL-NL-SNR}{16.1}}$. We use noise measurements taken around the Maryland Wind Energy Area from an array of 12 hydrophones over a 3-year survey period (Bailey et al., 2018). Some of the recording sites were nearby a shipping channel, thus noise levels represent realistic noise conditions that could occur during a deployment. The 50th percentile noise level from this study, measured in the 70.8–224 Hz band, ranged between 100–105 dB re. 1 μPa (rms) across recording sites. Similarly, the lowest 10th percentile noise levels ranged between 90–97 dB re. 1 μPa, and the highest 90th percentile noise levels ranged between 109–117 dB re. 1 μPa (Fig. 4.7.1k in Bailey et al., 2018). We also assume an upcall SL of 155 dB re. 1 μPa (rms) based on Trygonis et al. (2013).

2.8. Baseline configurations

We utilize the model and pretrained weights in Shiu et al. (2020) released along with the Deep learning in PAMGuard tutorial as a baseline deep learning model (hereafter termed Baseline) (Macaulay, 2021; Shiu et al., 2020). As the full detection pipeline was not released with the model weights, we follow all implementation details from the original paper for spectrogram generation and normalization with two-second sound clips. Spectrogram values are normalized by dividing each value by the sum of all squared spectrogram values. However, the non-maximum suppression classification step was not released, so we instead use the same continuous classification method detailed in this work operating every 100 ms.

As a real-time baseline, we run the PAMGuard (Gillespie et al., 2008) implementation of the Right Whale Edge Detector (hereafter abbreviated RWED) which is used to detect upcalls on the Cornell autobuoy and CABOW systems (Gillespie, 2004; Palmer et al., 2022; Spaulding et al., 2009). Following the implementation details from Gillespie (2004) and the corresponding PAMGuard documentation, we compute a FFT using a hann window of 256 frames with an overlap of 131 frames. We evaluate minimum detection thresholds 3–6 with a 6 dB SNR sound threshold between 0 and 1000 Hz. With a detection threshold of 4, the RWED can recover about 89% of upcalls above 9 dB SNR (Shiu et al., 2020).

3. Results

3.1. Real-time inference

When running on an ESP microcontroller, inference for the compressed NARWnet-Lite model takes about 75 ms and FFT calculation and spectrogram generation take about 44 ms and 81 ms respectively, for a total step-time of 200 ms. Thus, our NARWnet-Lite configuration, which operates every 500 ms, is capable of running in real-time on an ESP microcontroller without a significant performance drop relative to the uncompressed NARWnet configuration despite a 10x reduction in the resulting model weights file size (Fig. 3 and Table 1). Real-time audio processing, including spectrogram generation and model inference, results in a power draw of approximately 263 mW (71 mA at 3.7 V), while the full system draws approximately 666 mW (180 mA) with GPS and Iridium enabled. In a real-world test, a Medusa ran for 36 h on one 3.7 V 5Ah LiPo battery with satellite transmissions every 10 min and GPS on continuously. However, a 9 W solar panel is sufficient to power the Medusa during the Summer.

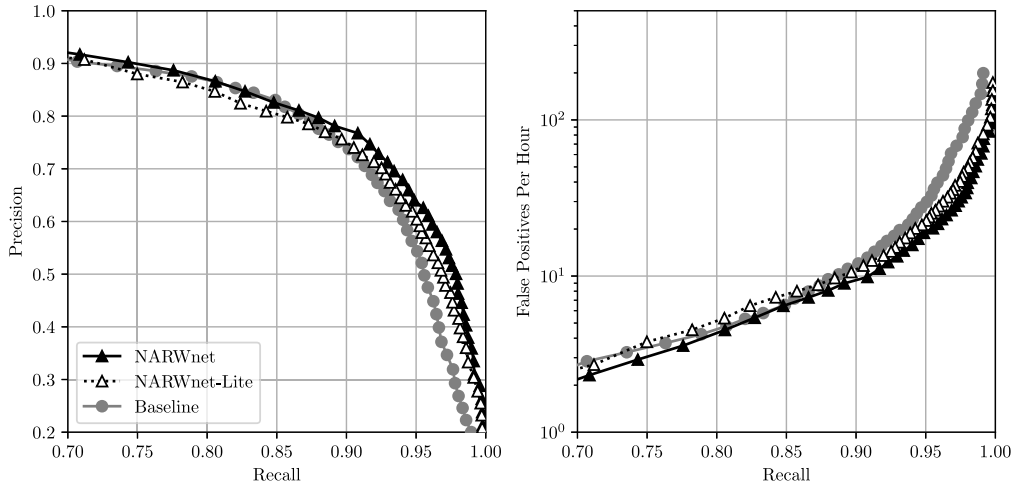


Fig. 3. Continuous performance in familiar noise environments. Left: Precision vs. Recall and Right: FP/H vs. Recall curves on three sequential days of DCLDE test data. The NARWnet and Baseline models operate with a step of 0.1 s, whereas the NARWnet-Lite model operates with a step of 0.5 s.

Table 1
Comparison of network configurations in familiar noise environments. Results for AUC and Recall at 5 and 20 FP/H on three days of familiar DCLDE test data. All models, including the compressed NARWnet-Lite model, are evaluated at 0.1 and 0.5 s step size to quantify the performance loss due to reduced inference speeds.

Model	AUC	Step	Recall	
			5 FP/H	20 FP/H
NARWnet	0.915	0.1 s	0.82	0.95
	0.904	0.5 s	0.79	0.94
NARWnet-Lite	0.914	0.1 s	0.82	0.95
	0.904	0.5 s	0.79	0.94
Baseline (Shiu et al., 2020)	0.898	0.1 s	0.81	0.93
	0.874	0.5 s	0.79	0.91

Table 2
Comparison of models in novel noise environments. Results for AUC and Recall at 5 and 20 FP/H on fifty 30 min recordings from the Gulf of St. Lawrence (B* dataset). We report the threshold required to operate at an average of 5 and 20 FP/H across the entire dataset. Also shown are results for the RWED with thresholds 4 and 5 which represent the high recall and high precision configurations which correspond to approximately 16.4 and 3.6 FP/H respectively. We do not compute AUC for the RWED as we only evaluate four different configurations.

Model	AUC	Recall		Threshold	
		5 FP/H	20 FP/H	5 FP/H	20 FP/H
NARWnet	0.957	0.91	0.97	0.76	0.46
NARWnet-Lite	0.948	0.90	0.97	0.83	0.50
Baseline (Shiu et al., 2020)	0.853	0.74	0.86	0.82	0.68
RWED (Gillespie, 2004)	–	0.43	0.65	5	4

3.2. Performance in familiar noise environments

Our NARWnet and NARWnet-Lite models achieve an AUC of 0.915 and 0.904 respectively, exceeding or matching the current best models on the three days of continuous DCLDE Test data despite seeing 5x fewer samples in the compressed NARWnet-Lite case. Our configuration of the baseline model achieves an AUC of 0.898, a decrease of 0.005 from the published results (Shiu et al., 2020) (Fig. 3 and Table 1). Results for models trained with limited data configurations can be found in Supplemental Info Section S4.

3.3. Performance in novel noise environments

On the 25-h continuous B* Kirsebom dataset from the Gulf of St. Lawrence, our NARWnet and NARWnet-Lite models achieve an AUC of 0.957 and 0.948 respectively, compared to 0.853 for the baseline (Fig. 4 and Table 2). From Shiu et al. (2020) and Supplemental Info Section S4, we see a slight decrease in performance for networks trained with DCLDE data only when evaluating on the B* Kirsebom dataset compared to a significant increase when utilizing our data augmentation strategy during training. Thus, our data augmentation strategy that mixes known right whale upcalls into highly diverse background noise conditions improves performance when deploying in novel noise environments. As the performance decrease due to single site training far outweighs the performance decrease caused by differences between our configuration of the baseline and the published version, we consider our results an accurate comparison between the models. See Supplemental Info Section S4 for results from other training configurations.

Furthermore, by inspecting the per-recording recall using a fixed threshold across two years of recordings for the compressed NARWnet-Lite model compared to the baseline configurations, we observe a significant mean recall boost at the same mean false detection rate (Fig. 5). See Supplemental Info Section S5 for results of the pairwise statistical tests and detailed recall and FP/H values.

Of the six sites processed from the NEFSC data, our two model configurations perform comparably to, or better than the baseline in all locations when using a fixed threshold. The models perform most similarly at SBNMS, where the training data was recorded, and at NE-Offshore, which is relatively close by. However, in novel contexts the performance gain is more noticeable (Fig. 6, Table 3, and Supplemental Info Section S6).

3.4. Performance at varying SNR

Both NARWnet and NARWnet-Lite versions of our model demonstrate strong detection ability across SNR compared to the baseline algorithms. While the models have comparable detection performance at very low SNR (<10 dB theoretical), our data augmentation strategy appears to improve detection consistency with better signal clarity regardless of the context (Fig. 7 and Supplemental Info Section S7). For our real-time NARWnet-Lite configuration, 50% detection probability occurs at around −6.5 dB SNR, compared to approximately −5.5 dB SNR and −1 dB SNR for the baseline CNN and RWED. As evidenced by Fig. 7(b), these values correspond to improved theoretical real-time detection ranges across noise levels. Combined with our simple detection range approximation, $R = 10^{\frac{SL-NL-SNR}{16.1}}$, an improvement of 5.5 dB SNR in the critical detection threshold effectively doubles the

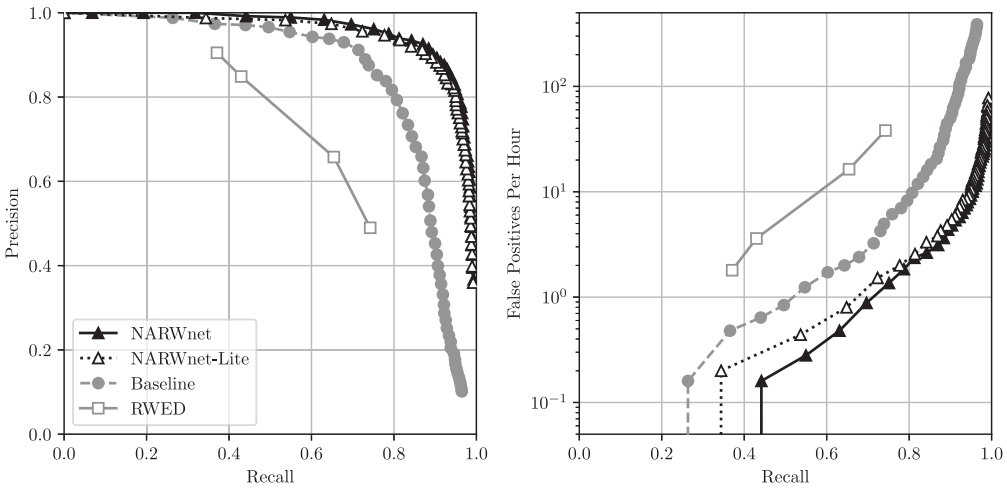


Fig. 4. Continuous performance in novel noise environments. Left: Precision vs. Recall and Right: FP/H vs. Recall curves on fifty 30 min recordings from the Gulf of St. Lawrence (B* dataset). Curves represent NARWnet and NARWnet-Lite configurations of our model running at 0.1 and 0.5 s step size respectively compared with the baseline model running at 0.1 s step size. Results for the RWED are shown across thresholds 3–6.

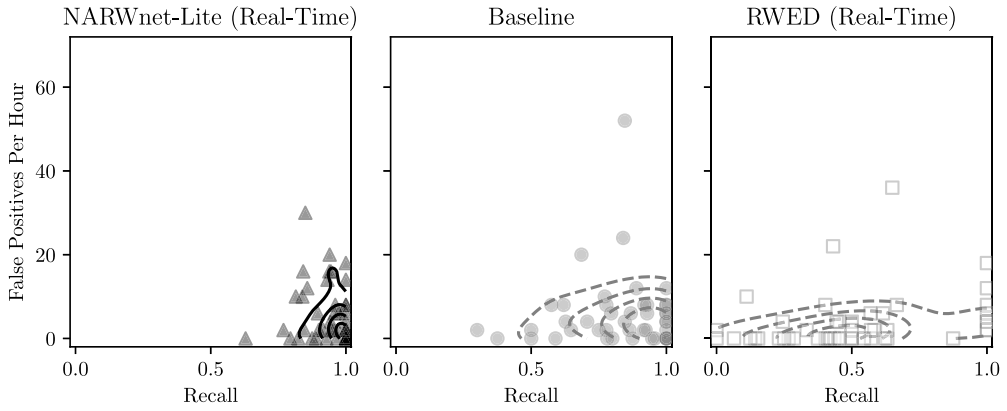


Fig. 5. Variation in detector performance across time with a fixed threshold. Points represent false positives per hour vs. recall values calculated for fifty 30-m recordings sampled intermittently between 2015 and 2017 from the Gulf of St. Lawrence (B* dataset). Contours represent four equipotential lines on the kernel density approximation for each detector. Data is shown for three different detectors operating with a fixed threshold tuned to 5 FP/H, two of which are capable of running in real-time on a microprocessor. Results for 20 FP/H can be found in Supplemental Info Section S5.

Table 3
Generalized performance in the NARW migratory corridor with a fixed 5 FP/H threshold. Recall and false positive rate when operating at a fixed 5 FP/H threshold using eight days of manually labeled continuous recordings throughout the Western North Atlantic Ocean from the NOAA NEFSC NARW Annotations. Corresponding results for 20 FP/H can be found in Supplemental Info Section S6.

Model	Recall/False positives per hour					
	Georgia N = 88	N. Carolina N = 94	NE-Offshore N = 21	Georges Bank N = 96	MA-RI N = 218	SBNMS N = 129
NARWnet	0.98/5.42	0.99/2.08	1.00/0.12	0.83/0.04	0.92/11.50	0.78/0.38
NARWnet-Lite	0.97/4.38	0.98/2.29	1.00/0.25	0.83/0.03	0.93/13.50	0.87/1.08
Baseline Shiu et al. (2020)	0.69/3.63	0.92/1.88	1.00/0.53	0.69/0.03	0.87/10.08	0.64/0.08
RWED Gillespie (2004)	0.60/46.88	0.67/0.67	0.81/1.72	0.20/0.92	0.28/4.79	0.28/0.71

detection range of a given system. We estimate that the NARWnet-Lite model should be able to detect upcalls at about 5 km on average, with ranges extending up to 25 km in low-noise conditions. Our model demonstrates a modest improvement over the neural network baseline which also significantly outperforms the RWED across SNRs. In periods of high noise, the active space of all networks is significantly reduced to less than 1 km. These values reflect a limitation of acoustic detection in general, especially in the presence of vessels. See Supplemental Info Section S7 for results with the precomputed clips released in Kirsebom et al. (2020).

4. Discussion

We describe the first neural network for NARW upcall detection capable of running on a microprocessor, combining the flexibility and high performance of deep neural networks with the conservation benefits of real-time acoustic detection systems. By minimizing the hardware requirements and audio processing time for a device, we open the door to more accurate large scale population monitoring of an endangered species. In contrast to existing approaches for real-time detection and

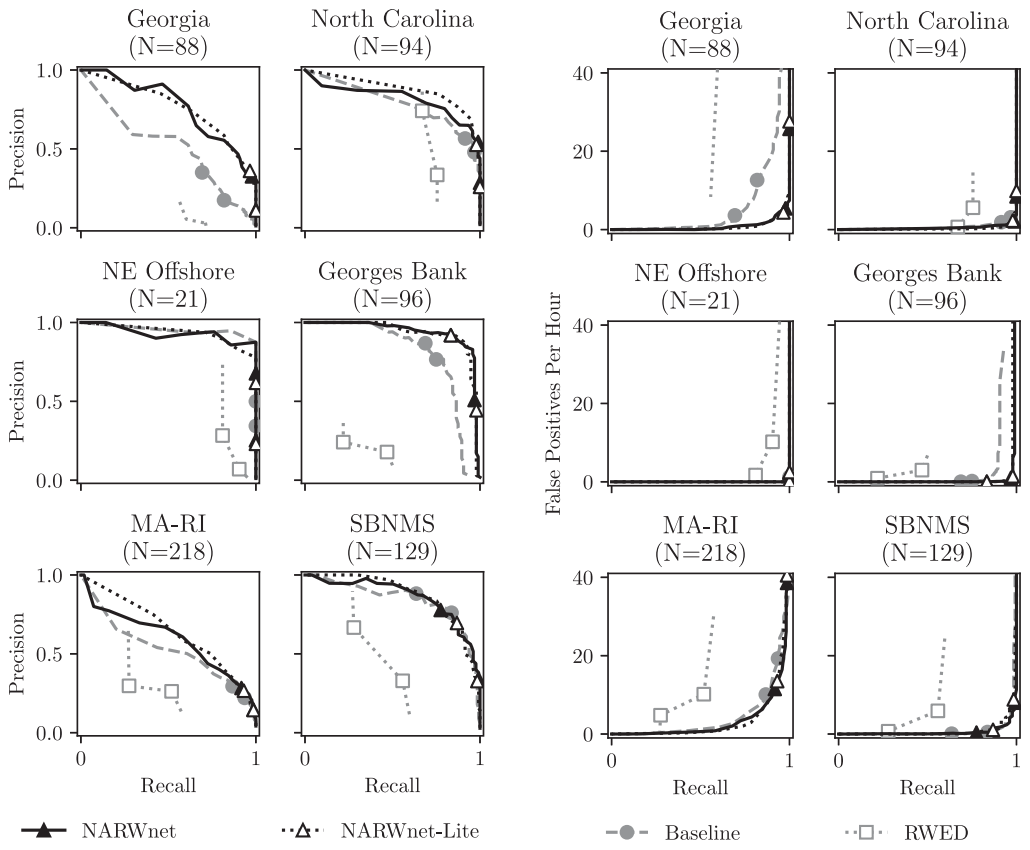


Fig. 6. Detection performance throughout the NARW migratory range. Left: Precision vs. Recall and Right: FP/H vs. Recall curves on eight days of NEFSC data from the Western North Atlantic Ocean. Curves represent NARWnet and NARWnet-Lite configurations of our model running at 0.1 and 0.5 s step size respectively compared with the baseline model running at 0.1 s step size. Points correspond to model performance at fixed 5 and 20 FP/H thresholds. For the recordings from Georgia, the fixed results for the RWED surpass the 40 FP/H limit of the figure; corresponding values can be found in Table 3 and Supplemental Info Section S6.

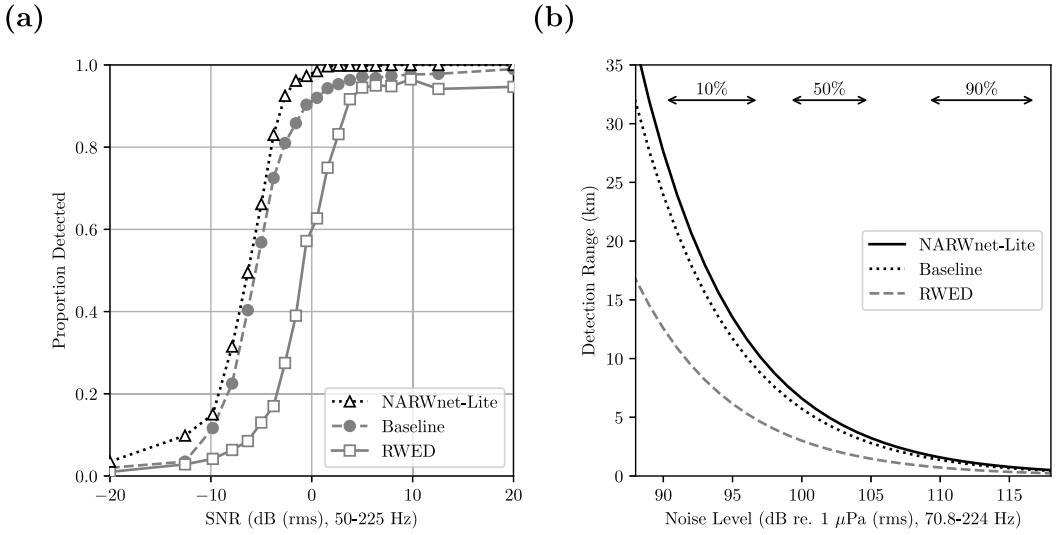


Fig. 7. Theoretical performance across SNR levels. Proportion of augmented upcalls recovered by various detectors across SNR while operating at a fixed 20 FP/H configuration. (a) Theoretical performance using 30-s augmented clips extracted from periods of noise in the continuous B* dataset between −20 and 20 dB. (b) Estimated detection range at 50% recall across noise levels with a fixed source level of 155 dB for the compressed NARWnet-Lite model running at 20 FP/H compared with the baseline neural network (Shiu et al., 2020) and RWED (Gillespie, 2004). Arrows show 10th, 50th, and 90th percentile noise levels from 12 long-term listening stations recording in and around the Maryland wind energy area (Bailey et al., 2018).

localization which utilize feature-based methods for signal processing (Baumgartner and Mussoline, 2011; Gillespie, 2004), our approach can be configured to operate at a lower false detection rate while still achieving significantly higher recall rates with less reliance on a secondary classification step or human-in-the-loop validation. This directly reduces the operating costs while simultaneously expanding the detection range of such systems. Additionally, our solution allows for further downstream classification improvements by utilizing the temporal pattern of detections which has been shown to further reduce false detections for NARW and other species (Madhusudhana

et al., 2021). Furthermore, current regulations require manual verification of automated acoustic detections to implement temporary slow zones. We perform a robust and transparent performance evaluation of our network operating in realistic deployment conditions using openly available datasets. We hope such transparency will help instill confidence in the capabilities of real-time detection systems to automatically implement time-sensitive conservation actions like slow zones.

Our NARWnet-Lite network demonstrates equal or superior performance when running in a real-time configuration compared to the baseline CNN in all scenarios. In contrast to the NARWnet configuration and the baseline CNN, the NARWnet-Lite configuration operates on 5x fewer frames and is limited to Int-8 precision, enabling lightweight real-time operation on an ESP microcontroller without a significant drop in accuracy, an ability that has previously been unattainable. The performance improvement is most pronounced in novel contexts like the Gulf of St. Lawrence and Georgia, whereas the baseline model is more competitive at sites like NE-Offshore and SBNMS that were closer in proximity to the original training data, indicating that the effect of our data augmentation strategy is most pronounced in new contexts. Furthermore, our model recovers more known upcalls, detecting 17% more upcalls at 5 FP/H in the Gulf of St. Lawrence. Additionally, Fig. 7(a) indicates the model is more accurate above −5 dB SNR.

When training a model with generalization in mind, the composition of training data is paramount to achieve good performance. As shown in Supplemental Info Section S4, it is possible to achieve very good performance using data from a single deployment with limited variation in background noise when evaluating data recorded at the same location. However, performance decreases on data from a different location (Gulf of St. Lawrence). While the baseline still outperforms feature-based methods in novel environments (Shiu et al., 2020), generalized results suggest our approach, which combines real DCLDE data with synthetic data from a variety of contexts, significantly improves performance when extrapolating to novel acoustic environments. The DCLDE data includes consistent system noise, so models trained with only DCLDE data are likely specialized to detect upcalls when this background noise is present. Consequently, models trained using only semi-synthetic data based on our data augmentation framework maintain high performance on the Kirsebom dataset from the Gulf of St. Lawrence, but show slightly lower DCLDE performance (Supplemental Info Section S4).

Understanding performance as a function of masking noise is important for determining thresholds for successful detection, and ultimately estimating the active space, or detection range, for real-time detection buoys deployed for mitigating human impacts. Using synthetic data with known signal level and known noise level within the frequency band of interest, we infer that our real-time configuration can effectively double the detection range of an existing system running the RWED. Our algorithm not only improves the accuracy of a system, thereby reducing the time and effort required to verify real-time detections, it can directly increase the active space of an existing listening system, decreasing the chance of missed whales. While these values represent simulated detection ranges, it is clear that an improved real-time network will lead to more successful and efficient management interventions. However, these values do not originate from a Medusa smart buoy and represent a theoretical active space. All of the recordings used in this work were collected with bottom-moored recorders which offer significantly reduced background noise levels compared to drifting recorders. While our analysis offers a fair comparison between detection algorithms, a full characterization of in-situ performance and detection range of a drifting recorder is under development.

Our network is configured to run in real-time by leveraging information from a single three-second frame to make a prediction. This results in an efficient implementation that can be easily compressed and incorporated into a larger detection framework. Consequently, our approach misses out on key contextual information that can help discern between a right whale upcall and upcalls present in the vocalizations of other baleen species. All the examined models here show an increased false detection rate in MA-RI and Georgia that seems to be a result of acoustically flexible humpback whales that also use upcalls in the same frequency range (Fig. 6 and Table 3). Other systems like the LFDCS use engineered features which extract contextual information from a longer window and can distinguish NARW upcalls from humpback whale song (Baumgartner and Mussoline, 2011). Furthermore, the LFDCS algorithm transmits pitch tracks of detections for an analyst to verify, thus improving the detection accuracy. However, every vocalization matters when listening for critically endangered right whales; deep learning approaches demonstrate a significant recall boost even with a human-in-the-loop. This work focuses primarily on refining existing deep learning approaches for detecting right whale upcalls to run in real-time on a microprocessor. As such, we do not investigate significant architectural changes like multi-class or multi-species detection, or the use of more recent deep learning models. While it would be possible to modify our current approach to handle additional biological classes like right whale gunshots without significantly increasing power and processing requirements, the absence of high quality benchmark datasets precludes a similarly robust performance evaluation. Moreover, alternative network architectures like transformers can handle much longer context windows and may be better suited for differentiating between baleen species (Vaswani et al., 2017). Transformers typically require more training data and have less support for compression due to architectural and memory constraints, but are a promising option for future generalized baleen whale detectors.

5. Conclusions

Viable solutions to the impending biodiversity crisis must be simple, low-cost, and extensible. This study leverages openly available data sources and code bases to further existing methods for detecting vocalizations from critically endangered NARW using passive listening devices in real-time. Our approach exemplifies the efficacy of compressed, lightweight deep neural networks for remote sensing and furthers the notion that robust data augmentation techniques can improve performance in novel environments. By conducting a theoretical detection range approximation, we demonstrate that our real-time algorithm can effectively double the active space of an existing system running an automated upcall detector, resulting in clear conservation benefits. Additionally, our data augmentation approach offers a potential solution to generating suitable training datasets for developing models to detect other endangered species. By performing an extensive model evaluation, we demonstrate robust performance throughout the NARW migratory range, but also highlight the need for user validation in the presence of vocally complex baleen species that are known to incorporate similar upcalls in their acoustic repertoire. We hope that these improved real-time detection systems can contribute to future conservation and ensure sustainable management actions for critically endangered species.

CRediT authorship contribution statement

Matthew D. Hyer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Austin T. Anderson:** Writing – review & editing, Software, Methodology. **David A. Mann:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **T. Aran Mooney:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Nadège Aoki:** Writing – review & editing, Conceptualization. **Frants H. Jensen:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

Funding was provided by an Aarhus University Research Foundation Recruiting Grant. Development of the Medusa was supported by National Science Foundation (NSF) Ocean Technology and Interdisciplinary Coordination (OTIC) award #1736359 to DAM/Loggerhead Instruments and #2024077 to TAM/WHOI. Performance evaluation and field validation were supported by the U. S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Wind Energy Technologies Office (WETO) Award Number DE-EE0010287 (Wildlife and Offshore Wind Project) and with co-funding from Woods Hole Oceanographic Institution and Aarhus University.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David A. Mann is the Founder and President of Loggerhead Instruments, the company that designs and manufactures the Medusa Buoys used in this study. He was not involved in the analysis or interpretation of the data. Austin Anderson is an employee of Loggerhead Instruments. The remaining 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.

Acknowledgments

We are grateful for productive discussions with Drs. Genevieve Davis and Sofie Van Parijs. All support is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103130.

Data availability

Clips of upcalls to recreate Fig. 7(a) are available upon request. All other data used in this work is openly available and cited in the main text. Code to reproduce results can be found here: https://doi.org/10.5281/zenodo.13851976. See README.md for detailed instructions.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/.

Bailey, H., Rice, A., Wingfield, J., Hodge, K., Estabrook, B., Hawthorne, D., Garrod, A., Fouda, L., McDonald, E., 2018. Determining Habitat Use by Marine Mammals and Ambient Noise Levels Using Passive Acoustic Monitoring Offshore of Maryland. Technical Report, Sterling (VA): US Department of the Interior, Bureau of Ocean Energy Management. OCS Study BOEM 2019-018, p. 232.

Baumgartner, M.F., Fratantoni, D.M., Hurst, T.P., Brown, M.W., Cole, T.V.N., Van Parijs, S.M., Johnson, M., 2013. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. J. Acoust. Soc. Am. 134 (3), 1814–1823. http://dx.doi.org/10.1121/1.4816406.

Baumgartner, M.F., Mussoline, S.E., 2011. A generalized baleen whale call detection and classification system. J. Acoust. Soc. Am. 129 (5), 2889–2902. http://dx.doi.org/10.1121/1.3562166.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A.M., Kirschel, A.N.G., 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus: Acoustic monitoring. J. Appl. Ecol. 48 (3), 758–767. http://dx.doi.org/10.1111/j.1365-2664.2011.01993.x.

Fleishman, E., Cholewiak, D., Gillespie, D., Helble, T., Klinck, H., Nosal, E.-M., Roch, M.A., 2023. Ecological inferences about marine mammals from passive acoustic data. Biol. Rev. 98 (5), 1633–1647. http://dx.doi.org/10.1111/brv.12969.

Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 10 (2), 169–185. http://dx.doi.org/10.1111/2041-210X.13101.

Gillespie, D., 2004. Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram. Can. Acoust. 32, 39–47.

Gillespie, D.M., 2019. DCLDE 2013 Workshop dataset. http://dx.doi.org/10.17630/62C3EEBC-5574-4EC0-BFEF-367AD839FE1A.

Gillespie, D.M., Gordon, J., McHugh, R., Mclaren, D., Mellinger, D., Redmond, P., Thode, A., Trinder, P., Deng, X.Y., 2008. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans.

Gillespie, D., Palmer, L., Macaulay, J., Sparling, C., Hastie, G., 2020. Passive acoustic methods for tracking the 3D movements of small cetaceans around marine structures. PLoS One 15 (5), e0229058. http://dx.doi.org/10.1371/journal.pone.0229058.

Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R., Joly, A., 2016. LifeCLEF bird identification task 2016: The arrival of deep learning. In: CLEF: Conference and Labs of the Evaluation Forum. vol. CEUR Workshop Proceedings, Évora, Portugal, pp. 440–449, URL: https://hal.science/hal-01373779. Issue: 1609.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Las Vegas, NV, USA, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hunt, K.E., Rolland, R.M., Kraus, S.D., 2015. Conservation physiology of an uncatchable animal: The North Atlantic right whale (*Eubalaena glacialis*). Integr. Comp. Biol. 55 (4), 577–586. http://dx.doi.org/10.1093/icb/icv071.

Hutschenreiter, A., Andresen, E., Briseño-Jaramillo, M., Torres-Araneda, A., Pinel-Ramos, E., Baier, J., Aureli, F., 2024. How to count bird calls? Vocal activity indices may provide different insights into bird abundance and behaviour depending on species traits. Methods Ecol. Evol. 15 (6), 1071–1083. http://dx.doi.org/10.1111/2041-210X.14333.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 448–456.

Karpištšenko, A., Cukierski, W., Spaulding, E., 2013. The Marinexplore and Cornell University whale detection challenge. URL: https://kaggle.com/competitions/whale-detection-challenge.

Keck, F., Peller, T., Alther, R., Barouillet, C., Blackman, R., Capo, E., Chonova, T., Couton, M., Fehlinger, L., Kirschner, D., Knüsel, M., Muneret, L., Oester, R., Tapolczai, K., Zhang, H., Altermatt, F., 2025. The global human impact on biodiversity. Nature http://dx.doi.org/10.1038/s41586-025-08752-2.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. http://dx.doi.org/10.48550/ARXIV.1412.6980.

Kirsebom, O.S., Frazao, F., Simard, Y., Roy, N., Matwin, S., Giard, S., 2020. Performance of a deep neural network at detecting North Atlantic right whale upcalls. J. Acoust. Soc. Am. 147 (4), 2636–2646. http://dx.doi.org/10.1121/10.0001132.

Knowlton, A.R., Kraus, S.D., 2001. Mortality and serious injury of northern right whales (*Eubalaena glacialis*) in the western North Atlantic Ocean. J. Cetacean Res. Manag. 193–208. http://dx.doi.org/10.47536/jcrm.vi.288.

LeCun, Y., Bottou, L., Orr, G.B., Müller, K.-R., 2002. Efficient backprop. In: Neural Networks: Tricks of the Trade. Springer, pp. 9–50.

Macaulay, J., 2021. Deep learning in PAMGuard. URL: https://github.com/macster110/PAMGuard_resources/tree/main/deep_learning.

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., Širović, A., Roch, M.A., 2021. Improve automatic detection of animal call sequences with temporal context. J. R. Soc. Interface 18 (180), 20210297. http://dx.doi.org/10.1098/rsif.2021.0297.

Madsen, P.T., Jensen, F.H., Carder, D., Ridgway, S., 2012. Dolphin whistles: a functional misnomer revealed by heliox breathing. Biol. Lett. 8 (2), 211–213. http://dx.doi.org/10.1098/rsbl.2011.0701.

Madsen, P., Wahlberg, M., Tougaard, J., Lucke, K., Tyack, P., 2006. Wind turbine underwater noise and marine mammals: implications of current knowledge and data needs. Mar. Ecol. Prog. Ser. 309, 279–295. http://dx.doi.org/10.3354/meps309279.

Mann, D., Hall, M., Anderson, A., Donner, A., Aoki, N., Formel, N., Jensen, F., Hyer, M., Mooney, T.A., 2024. Medusa: An AI-powered buoy and drifter for large-scale passive acoustic monitoring. In: 2024 Ocean Sciences Meeting. AGU.

Matthews, L.P., Parks, S.E., 2021. An overview of North Atlantic right whale acoustic behavior, hearing capabilities, and responses to sound. Marine Poll. Bull. 173, 113043. http://dx.doi.org/10.1016/j.marpolbul.2021.113043.

McFee, B., 2016. Resampy: efficient sample rate conversion in Python. J. Open Source Softw. 1 (8), 125. http://dx.doi.org/10.21105/joss.00125, Publisher: The Open Journal.

Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P., Matsumoto, H., 2007. An overview of fixed passive acoustic observation methods for cetaceans. Oceanography 20 (4), 36–45, URL: http://www.jstor.org/stable/24860138.

NOAA, NMFS, 2008. Endangered fish and wildlife; Final rule to implement speed restrictions to reduce the threat of ship collisions with North Atlantic right whales. Fed. Regist. 73 (198), 60173–60191.

NOAA Office of National Marine Sanctuaries and US Navy, 2020. SanctSound raw passive acoustic data. http://dx.doi.org/10.25921/SACA-SP25.

Pacific Marine Environmental Laboratory, Oregon State Univeristy, Cornell University, Scripps Research Insitute, 2023. NOAA NEFSC North Atlantic right whale annotations. http://dx.doi.org/10.25921/2C09-NG58.

Padovese, B., Frazao, F., Kirsebom, O.S., Matwin, S., 2021. Data augmentation for the classification of North Atlantic right whales upcalls. J. Acoust. Soc. Am. 149 (4), 2520–2530. http://dx.doi.org/10.1121/10.0004258.

Palmer, K.J., Tabbutt, S., Gillespie, D., Turner, J., King, P., Tollit, D., Thompson, J., Wood, J., 2022. Evaluation of a coastal acoustic buoy for cetacean detections, bearing accuracy and exclusion zone monitoring. Methods Ecol. Evol. 13 (11), 2491–2502. http://dx.doi.org/10.1111/2041-210X.13973.

Parks, S., Searby, A., Célérier, A., Johnson, M., Nowacek, D., Tyack, P., 2011. Sound production behavior of individual North Atlantic right whales: implications for passive acoustic monitoring. Endanger. Species Res. 15 (1), 63–76. http://dx.doi.org/10.3354/esr00368.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pettis, H.M., Hamilton, P.K., 2024. North Atlantic Right Whale Consortium 2023 Report Card. Technical Report, North Atlantic Right Whale Consortium, http://dx.doi.org/10.1575/1912/69694.

Roch, M.A., 2015. Scoring tools for the 2015 DCLDE workshop. URL: https://www.cetus.ucsd.edu/dclde/scoringTool.html.

Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10 (1), 607. http://dx.doi.org/10.1038/s41598-020-57549-y.

Silber, G.K., Bettridge, S., 2012. An assessment of the final rule to implement vessel speed restrictions to reduce the threat of vessel collisions with North Atlantic right whales. U. S. Dept. Commer..

Simard, Y., Kirsebom, O.S., Frazao, F., Roy, N., Matwin, S., Giard, S., 2020. Acoustic recordings of North Atlantic right whale upcalls in the Gulf of St. Lawrence. http://dx.doi.org/10.20383/101.0241.

Singh, J.S., 2002. The biodiversity crisis: A multifaceted review. Current Sci. 82 (6), 638–647, URL: http://www.jstor.org/stable/24106689.

Sousa-Lima, R.S., Norris, T.F., Oswald, J.N., Fernandes, D.P., 2013. A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. Aquatic Mammals 39 (1), 23–53. http://dx.doi.org/10.1578/AM.39.1.2013.23.

Spaulding, E., Robbins, M., Calupca, T., Clark, C.W., Tremblay, C., Waack, A., Warde, A., Kemp, J., Newhall, K., 2009. An Autonomous, Near-Real-Time Buoy System for Automatic Detection of North Atlantic Right Whale Calls. Portland, Oregon, 010001. http://dx.doi.org/10.1121/1.3340128.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152. http://dx.doi.org/10.7717/peerj.13152.

Trygonis, V., Gerstein, E., Moir, J., McCulloch, S., 2013. Vocalization characteristics of North Atlantic right whale surface active groups in the calving habitat, southeastern United States. J. Acoust. Soc. Am. 134 (6), 4518–4531. http://dx.doi.org/10.1121/1.4824682.

Urazghildiiev, I.R., Clark, C.W., 2007. Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics. J. Acoust. Soc. Am. 122 (2), 769–776. http://dx.doi.org/10.1121/1.2747201.

Van Parijs, S., Clark, C., Sousa-Lima, R., Parks, S., Rankin, S., Risch, D., Van Opzeeland, I., 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. Mar. Ecol. Prog. Ser. 395, 21–36. http://dx.doi.org/10.3354/meps08123.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 30, Curran Associates, Inc..

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Wade, P., Heide-Jørgensen, M.P., Shelden, K., Barlow, J., Carretta, J., Durban, J., LeDuc, R., Munger, L., Rankin, S., Sauter, A., Stinchcomb, C., 2006. Acoustic detection and satellite-tracking leads to discovery of rare concentration of endangered North Pacific right whales. Biol. Lett. 2 (3), 417–419. http://dx.doi.org/10.1098/rsbl.2006.0460.

Wei, S., Zou, S., Liao, F., Lang, W., 2020. A comparison on data augmentation methods based on deep learning for audio classification. J. Phys.: Conf. Ser. 1453 (1), 012085. http://dx.doi.org/10.1088/1742-6596/1453/1/012085.

Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., Cai, H., Liu, S., 2018. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (Eds.), Advances in Multimedia Information Processing – PCM 2018. In: Lecture Notes in Computer Science, vol. 11166, Springer International Publishing, Cham, pp. 14–23. http://dx.doi.org/10.1007/978-3-030-00764-5_2.