JASA ARTICLE

A multimodel deep learning algorithm to detect North Atlantic right whale up-calls

Ali K Ibrahim,[1,a)] Hanqi Zhuang,[2] Laurent M. Chérubin,[1] Nurgun Erdol,[2] Gregory O'Corry-Crowe,[1] and Ali Muhamed Ali[1]

[1]*Harbor Branch Oceanographic Institute, Florida Atlantic University, 5600 US1 North, Fort Pierce, Florida 34946, USA*

[2]*Department Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Road, Boca Raton, Florida 33431, USA*

ABSTRACT:

We present a new method of detecting North Atlantic Right Whale (NARW) upcalls using a Multimodel Deep Learning (MMDL) algorithm. A MMDL detector is a classifier that embodies Convolutional Neural Networks (CNNs) and Stacked Auto Encoders (SAEs) and a fusion classifier to evaluate their output for a final decision. The MMDL detector aims for diversity in the choice of the classifier so that its architecture is learned to fit the data. Spectrograms and scalograms of signals from passive acoustic sensors are used to train the MMDL detector. Guided by previous applications, we trained CNNs with spectrograms and SAEs with scalograms. Outputs from individual models were evaluated by the fusion classifier. The results obtained from the MMDL algorithm were compared to those obtained from conventional machine learning algorithms trained with handcrafted features. It showed the superiority of the MMDL algorithm in terms of the upcall detection rate, non-upcall detection rate, and false alarm rate. The autonomy of the MMDL detector has immediate application to the effective monitoring and protection of one of the most endangered species in the world where accurate call detection of a low-density species is critical, especially in environments of high acoustic-masking. © *2021 Acoustical Society of America*.

https://doi.org/10.1121/10.0005898

(Received 22 October 2020; revised 13 July 2021; accepted 29 July 2021; published online 18 August 2021)

[Editor: James F. Lynch] Pages: 1264–1272

I. INTRODUCTION

The North Atlantic Right Whale (*Eubalaena glacialis*, NARW) is one of the most endangered whale species in the world. The current population estimate for NARWs off the east coast of North America is 451 (Hayes *et al.*, 2017; Reeves, 2003), and a decreasing trend and low reproduction rates (Cooke, 2018), combined with high levels of human activities, such as shipping and fisheries, underscore their precarious situation. Efficient tracking of their numbers, migration paths, and habitat use is vital in lowering the number of preventable injuries and deaths and promoting their recovery. Passive acoustics is frequently used for the purpose as a reliable, safe, and effective technology to monitor the NARW by detecting their signature up-calls. Up-calls are narrowband vocalizations with frequency swings in the range of 50–440 Hz (Clark, 1982). Time-frequency representations have, in the past, provided the domain for detecting NARW up-calls with edge detection (Gillespie, 2004), and pattern detection *via* convolutional methods (Mellinger and Clark, 1993). However, these methods have led to high levels of false positives (Urazghildiiev *et al.*, 2009). Feature engineering and machine learning (Mellinger, 2004; Urazghildiiev and Clark, 2006) have reduced false-positive rates and increased detection rates to more than 80%. Gillespie (2004) was able to determine whale types by classifying edge data extracted from spectrograms. Urazghildiiev and Clark (Urazghildiiev and Clark, 2006) applied a generalized likelihood ratio test (GLRT) detector of polynomial-phase signals with unknown amplitude to deal with locally stationary Gaussian noise. In Esfahanian *et al.* (2015), classifying Linear Binary Patterns (LBP) extracted from up-call spectrograms resulted in 93% up-call detection accuracy. In Ibrahim *et al.* (2016), they were able to reduce the false positive rate to 1.48% using linear support vector machines (LSVM) to classify Mel Frequency Cepstral Coefficients (MFCC) obtained from two complementary discrete wavelet transform (DWT) subspaces.

Our overarching research objective is to develop an effective and autonomous set of computational tools for the passive acoustic monitoring of fishes and marine animals such as NARWs. Recent studies suggest that sophisticated preprocessing and handcrafted feature extraction procedures may not be needed for deep learning based detectors and classifiers (Ibrahim *et al.*, 2018b). Deep learning algorithms such as autoencoders, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), can act as feature extractors and classifiers (Ibrahim *et al.*, 2016). Notably, CNNs are excellent choices for identifying spatial patterns in images, and RNNs for the extraction of discriminative patterns from time series or signals (Chérubin *et al.*, 2020; Ibrahim *et al.*, 2018b). However, the vanishing

[a)]Electronic mail: aibrahim2014@fau.edu

https://doi.org/10.1121/10.0005898

gradients phenomenon, involving the decay of feedback information over time, prevents most RNNs from memorizing long-term dependency of an input time sequence. Long Short-Term Memory (LSTM) networks solve this problem through the introduction of additional gate parameters in each neuron, which select the attributes to memorize or forget. A Multi model Deep Learning detector (Ibrahim, 2019) is an ensemble of classifiers consisting of CNNs and Stacked Auto Encoders (SAEs) random in the number and size of such hyperparameters as neuron count and kernel size. The outputs of these classifiers are used to train a fusion algorithm, which is itself a classifier, to determine the class of an input. We show in the study herein that the MMDL model outperforms representative transfer learning methods based on pretrained models such as ResNet101 and VGG19. The remainder of the paper is organized as follows. Section II reviews the vocalization types of NARWs. The description of the proposed MMDL detector and its application to the NARW detection are presented in Secs. III and IV, respectively. Section V presents the detection results from a database of NARW up-calls. Concluding remarks are given in Sec. VI.

II. NARW SOUNDS

Right whales vocalize a variety of low-frequency sounds, and the calling repertoires of the three species are similar (Parks and Tyack, 2005). So called moans, groans, belches, and pulses have most of their acoustic energy below 500 Hz. Occasionally, a vocalization will have spectral content up to 4 kHz. One typical right whale vocalization used to communicate with other right whales is the so-called "up-call." It is a short chirp, or a "whoop" sound that rises from about 50 to 440 Hz and lasts about 2 s. Up-calls are often described as "contact" calls as they appear to function as signals that bring whales together (Parks and Tyack, 2005). The Cornell University dataset (Clark *et al.*, 2002; Parks *et al.*, 2009) used in this study includes both NARW up-calls (Fig. 1) and background noise with other sounds (Fig. 2) in 2-s clips sampled at 2000 Hz. Figure 1 shows the spectrogram of commonly encountered types of up-calls. Some up-calls possess more than one chirp [Figs. 1(c)–1(e)] and recordings are typically very noisy as evidenced in Fig. 1(b)

III. MULTIMODEL DEEP LEARNING

Data classification is an iterative process involving problem formulation, data analysis, feature extraction, feature selection, classifier selection, and model validation. There are several common reasons why classification models fail. These reasons include insufficient data preprocessing, lack of model validation, overfitting during the training stage, and the unsuitability of the model for the data. Ibrahim (2019) proposed a multimodel approach based on deep learning for data classification and event forecasting. The MMDL algorithm fuses results from different types of classifiers which collectively cancel the shortcomings of individual classifiers. The proposed MMDL model for NARW upcall detection consists of two types of classifiers: CNNs and SAEs. CNNs are chosen because of their capability of extracting both low-level and high-level features from images. SAEs are selected for their performance in extracting pertinent information-bearing features for data compression.

CNNs have proven to be one of the most effective deep learning algorithms for image classification and identification (Gu *et al.*, 2018). CNN networks became popular after the exceptional performance of AlexNet in the 2012 ImageNet competition (Krizhevsky *et al.*, 2012). The three main components of a CNN network are convolution, pooling, and activation. A convolutional layer convolves input data with a set of $(d \times n)$ kernels or filter impulse responses to produce feature maps. A pooling layer operates on each feature map independently to reduce its size. The non-linear max pooling operation is one of the most frequently used. An activation layer consists of a non-linear operation that, like signal conditioning, controls the range of its input. Rectified Linear Unit (ReLU) is a commonly used activation function. The convolve- pool-activate process is repeated until a set of sufficiently discriminative and concise features is obtained. The feature vector is fed to a fully connected layer with an activation function, mostly either Sigmoid or SoftMax, for decision making.

The SAE is a popular algorithm which consists of multiple layers of unsupervised autoencoders, followed by a fully-connected layer with either SoftMax or Sigmoid as an activation function. SAE training is a two-step process consisting of unsupervised learning followed by supervised learning. Unlabeled samples are input to the SAE's first layer for unsupervised training. The Auto-Encoder (AE) layers are stacked so that the resulting parameter vector of layer *k-1* is used as an input to train the *k*th AE layer. Once the AEs are trained, labelled data are fed to the fully-connected layer to train its parameters. The structure of such a SAE is shown in Fig. 3.

The flow chart of the MMDL model used in this study is now shown in Fig. 4. The output of each CNN and SAE are piped into a fusion block for decision making. The fusion block inspects the outputs from individual models in search of locally consistent, discriminative, and representative patterns. The types of features used in the MMDL components are chosen based on the results of an early study by Moreno-Seco *et al.* (2006) that tested the efficacy of such fusion mechanisms as Majority Voting, Unweighted Average, and PatternNet. The study revealed that PatternNet consistently outperformed other methods; spectrograms (a visual representation of sound based on the time evolution of its Fourier transforms) worked better with CNNs, and scalograms (a visual representation of sound based on the time evolution of its wavelet transforms) worked better with SAEs. Hence, we paired CNNs with spectrograms and SAEs with scalograms, as shown in Fig. 4. That is, the CNNs in the MMDL detector are trained with spectrogram images and the SAEs are trained with scalogram images. For each of the SAEs and CNNs, we defined a range in which its hyperparameters are randomly generated. Hyperparameters,

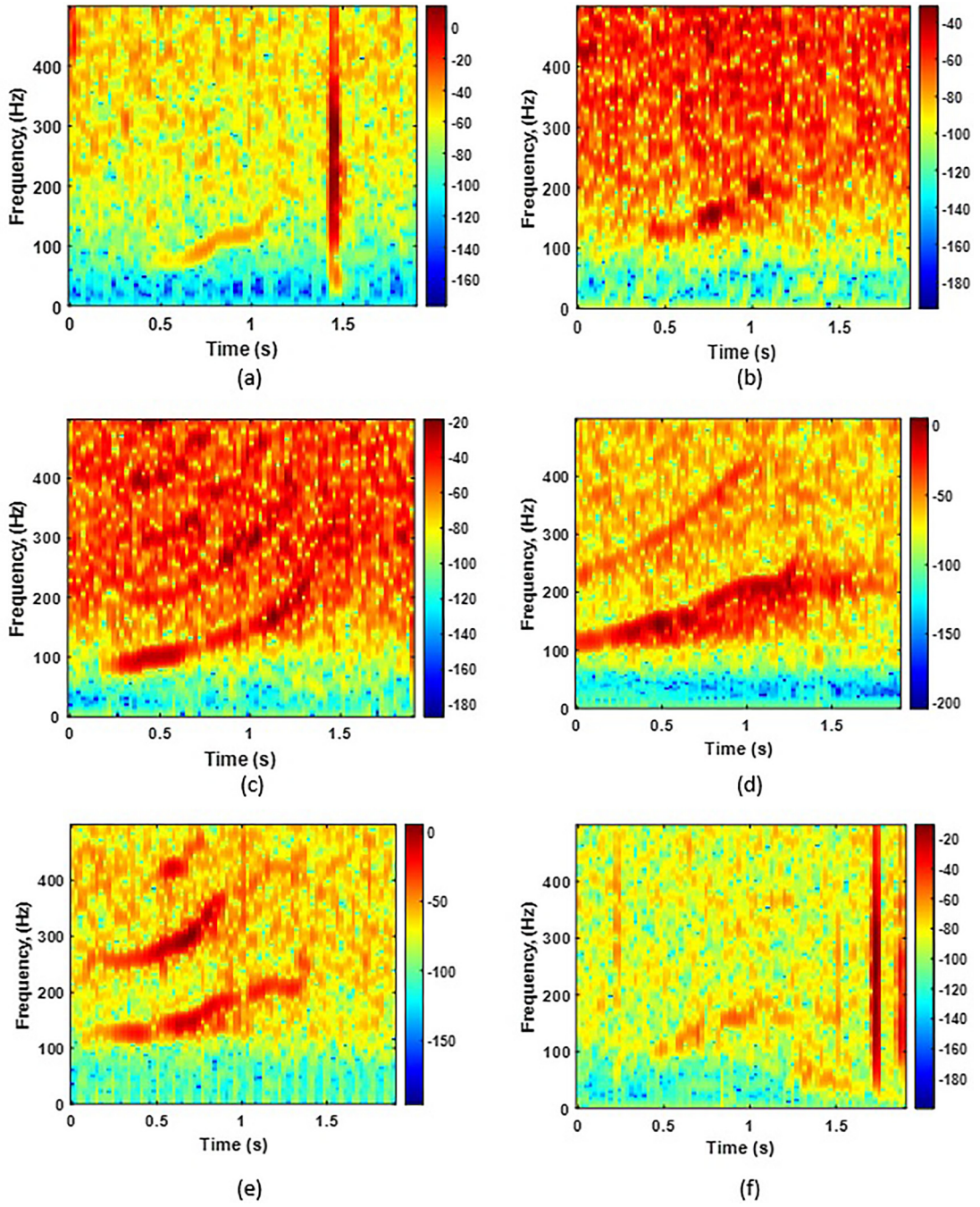

FIG. 1. (Color online) NARW up-call sound spectrograms with various types of background noise.

described in Sec. IV C, can be chosen to control the learning curves. They can make the network architecture both computationally inexpensive and structurally simpler than a deep CNN, and they can be chosen to adapt to data size, thus reducing the probability of overfitting.

IV. NARW UP-CALL CLASSIFICATION

A. Data preparation

The NARW sound dataset was collected with an array of 19 Marine Autonomous Recording Units (Urazghildiiev and Clark, 2006) during 22 deployment periods from June 26, 2007, to May 8, 2013. The dataset consisted of 2-s audio clips, sampled at 2 kHz, that were transformed to both a spectrogram and a scalogram, then resized for a resolution of 100×100 pixels. Spectrogram images were generated by organizing the sound signal into 80-ms frames with 50% overlap. A 1024-point discrete Fourier transform (DFT) of each frame, multiplied by a Hamming window to reduce sidelobe leakage, was computed. The resulting spectrograms were saved as pseudo RGB images to be processed by CNNs.

A scalogram displays an approximation of the magnitude of the continuous wavelet transform (CWT) of a signal

https://doi.org/10.1121/10.0005898

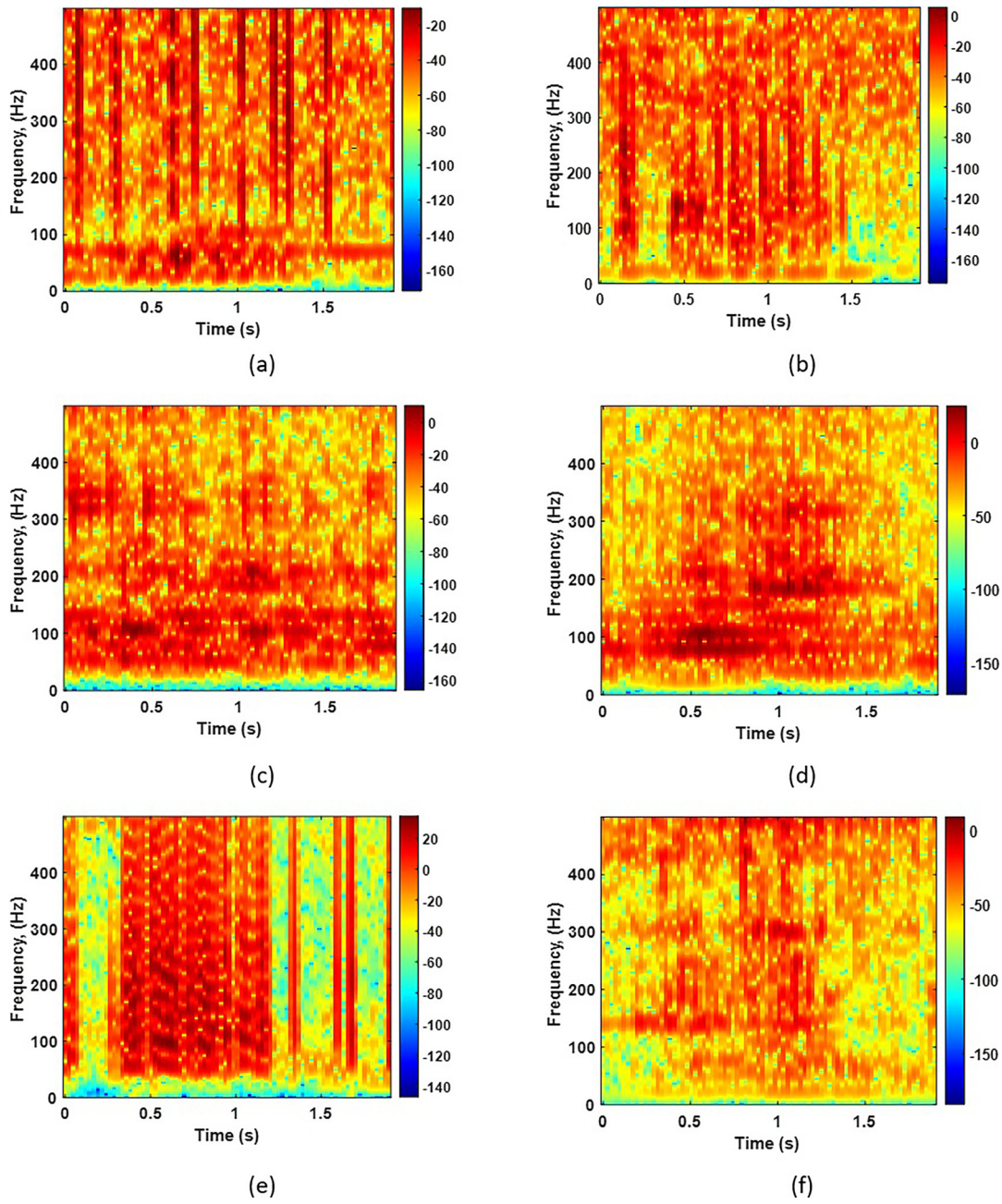

FIG. 2. (Color online) Background noise and other sound spectrograms.

(Halberstadt, 2020). This representation of the localized wavelet transform is well suited for the analysis of nonstationary phenomena by revealing the frequency content of the signal for each frame, while tracking evolving phenomena in both time and scale. Unlike the spectrogram, which decomposes an input signal into sinusoids of infinite duration, CWT decomposes the signal into wavelets.

To create a scalogram image, we processed each audio clip with a CWT filter bank and formed an image of the magnitude of the CWT coefficients. Wavelet filters are logarithmically spaced bandpass filters. We recorded the center frequency of the filters on the ordinate of the scalogram. Figure 5 shows a spectrogram and a scalogram of both an up-call and background noise sample taken from the data set. We used a data augmentation procedure to increase the number and diversity of our training data and to improve the performance of the MMDL detector. The data were augmented by adding Gaussian-distributed noise with zero mean and 0.2 variance. Other augmentation schemes for images include rotation ($\pm 15°$), scaling (0.6–1), reflection around x axis, and shearing (0°–30°). After applying the augmentation procedure, the training data set contained approximately 5000 images per class (up-calls and non-calls) as compared to 2000 prior to data augmentation. The proposed classifier was trained solely with the augmented training data set. A testing dataset with additional 80 665 audio files was used solely for model validation.

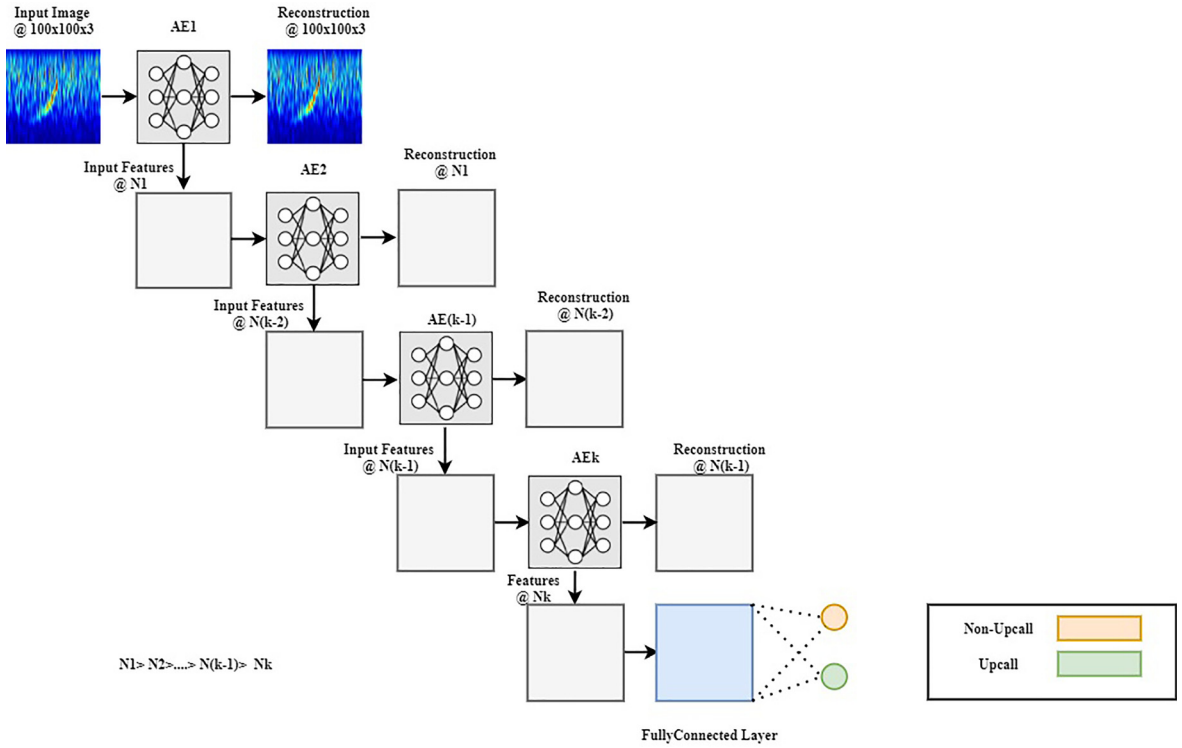

FIG. 3. (Color online) SAE network for up-call detection.

B. Model set-up and implementation

Deep neural network algorithms require significant fine tuning to work with specific data sets, thus posing the question of finding suitable structures and architectures as an important research challenge (He *et al.*, 2021). The proposed MMDL method automates the architecture construction process by introducing diversity in the classifiers and by randomizing a wide range of hyperparameters to allow the system to learn a suitable network architecture for a given dataset. The fusion mechanism ensures that the system takes

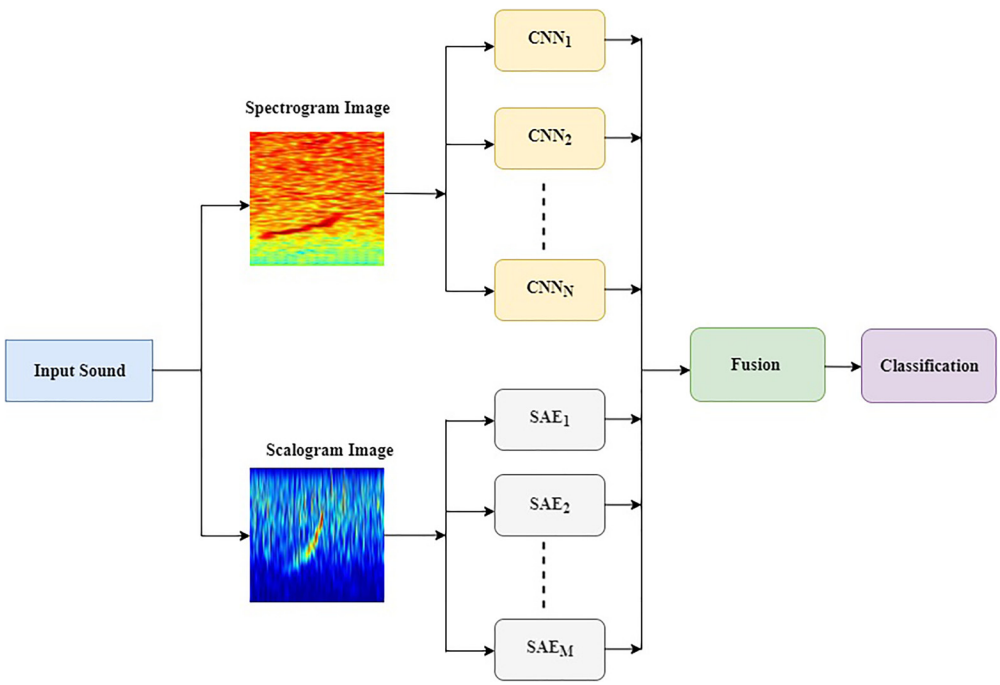

FIG. 4. (Color online) The proposed MMDL model for NARW up-call detection.

https://doi.org/10.1121/10.0005898

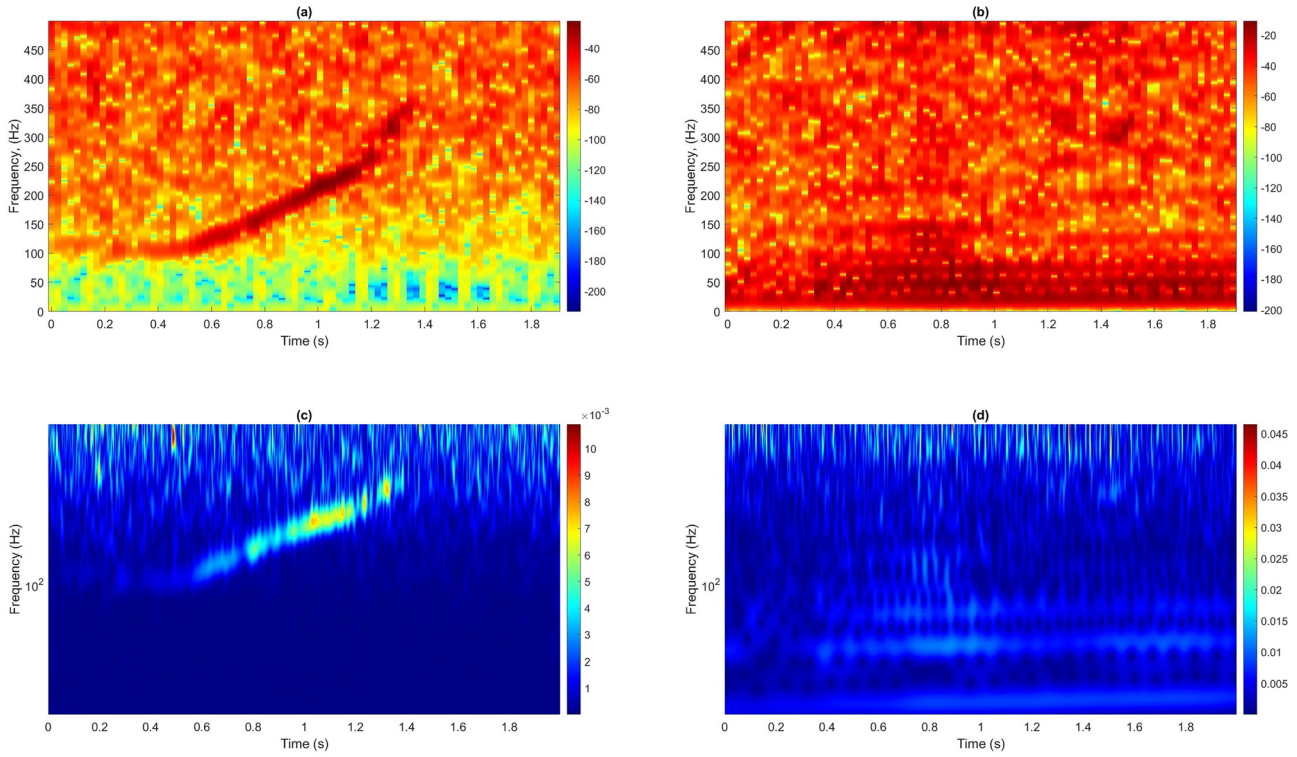

FIG. 5. (Color online) (a) Spectrogram of a NARW up-call, (b) spectrogram of background noise, (c) scalogram of NARW up-call in (a). (d) Scalogram of background noise in (b). The image resolution is 100×100 pixels. The x and y axes represent time and frequency, respectively.

advantage of the strengths of both "good classifiers" and exclude the "outliers." The algorithmic steps for training the MMDL detector are as follows:

(1) Prepare a training dataset by converting sound files to spectrograms and scalograms. If necessary, perform data augmentation.
(2) Define a range for the hyperparameter values for each CNN and SAE.
(3) Generate n_1 CNNs with hyperparameter values randomly assigned from the preset ranges.
(4) Train each one of the n_1 CNNs with the spectrograms.
(5) Generate n_2 SAEs with hyperparameter values randomly assigned from the preset ranges.
(6) Train each one of the n_2 SAEs with the scalograms.

In the test phase, the following steps are used:

(1) Compute a spectrogram and a scalogram for each input sound.
(2) Input each spectrogram simultaneously to all the CNNs.
(3) Input each scalogram simultaneously to all the SAEs.
(4) Extract the predicted label and the predicted probability vector of each sound from each CNN. The predicted probability vector is the SoftMax layer output corresponding to the input spectrogram.
(5) Extract the predicted label and the predicted probability vector of each sound from each SAE.
(6) Pipe the labels and the predicted probability vectors obtained in Steps 4 and 5 into the fusion classifier to make a final decision on the label of the input.

C. Network architecture

To evaluate and characterize the proposed MMDL architecture, we tested three different detectors comprised of 5, 10, and 15 CNNs and SAEs, respectively. Each CNN block has a convolutional layer with randomly-picked hyperparameters (i.e., neuron count and kernel size), a batch normalization layer, a max-pooling layer, and a ReLu activation layer. The hyperparameters are shown in Tables I and II. The number of AEs in each SAE and the number of neurons in each hidden layer are the two randomly-generated hyperparameters. Each randomized SAE is designed as follows: L random numbers, representing the number of hidden neurons for a single AE in a stack, are generated. The AEs are stacked together, in decreasing order of their numbers, to form an SAE. The sorting ensures that each layer reduces the dimension of the captured features. The generated CNN and SAE structures are trained individually. As mentioned earlier, SAEs are trained using scalograms, and CNNs are trained using spectrograms. The outputs of the SAEs and CNNs are subsequently fused to

TABLE I. Range of randomly generated hyperparameters for each CNN.

Hyperparameter	Range
Number of convolutional layers block	1–5
Kernel size	(1–3–5–7) odd numbers
Number of kernels	(8–64)
Mini-batch size	16–256
Optimizer	Adam, SGDM, SGD, RMSProp

TABLE II. Range of randomly generated hyperparameters for each SAE.

Hyperparameter	Range
Number of AEs	1–5
Number of hidden neurons in each AE	150–800
Sparsity regularization	16–4
l2 weight regularization	0.01–0.05

determine the signal's class. The fusion classifier is described in the next subsection.

D. Fusion strategies

We tested three types of fusion strategies: Majority Vote, Average, and PatternNet. Average and PatternNet fuse predicted probabilities provided by the Softmax layers of individual models. Majority Voting fuses the predicted class labels of individual models.

The Weighted Average strategy takes the vote and confidence of a model into consideration by assigning weights according to the uncertainty of each model. Ju *et al.* (2018) showed that a weighted average is a good strategy when models have similar performances. Following Abidalkareem *et al.* (2020) and Wang *et al.* (2018a), activations of individual models were concatenated into a vector that was then piped into a PatternNet, essentially forming an optimized feed forward network for pattern recognition (Wang *et al.*, 2018b). In our approach, we used the predicted probability vectors from individual shallow models (in general weak classifiers) to train the PatternNet in order to create a strong classifier. The PatterNet here acts as a mixer of the predicted probability vectors. In order to optimize its performance, we adopted the cross-entropy measure as the loss function and the scaled conjugate gradient method as the training procedure for the PatternNet. Our experiments showed that this approach was appropriate for fusing predicted probabilities because the MMDL architecture allows each model to contribute according to its strength.

TABLE III. Results of different input-model combinations (each has five CNN models or five SAE models).

Models type	Upcall detection	Non-upcall detection	False alarm
Spectrogram + CNNs	90.1%	99.%	0.9%
Scalogram + CNNs	88.6%	98.9%	1.08%
Spectrogram + SAEs	86.25%	98.97%	1.02%
Scalogram + SAEs	89.34%	99.01%	0.99%

V. APPLICATION TO THE NAWR UPCALL DETECTION

The data used in this study was initially labeled by using an edge detector, resulting in a substantial number of false labels. We generated spectrograms of all the audio clips and relabeled them based on visual and audio inspection of the samples. We trained the models with the correctly labeled images.

To evaluate our model performance at detecting NARW up-calls, we compared its detection rate to those of conventional machine learning algorithms trained with handcrafted features as done in Ibrahim *et al.* (2018a), Ibrahim *et al.* (2018b), and Ibrahim *et al.* (2018c,d). The handcrafted features were derived from combinations of MFCCs, Gammatone Filter Cepstral Coefficients (GFCC), and wavelet subspace projections of the recordings. We showed that coupling MFCC or GFCC features with two-stage Daubechies wavelet (db4) projections significantly improved the performance and provided the best results with 92.27% up-call detections and 1.48% false alarm rates (Fig. 6).

For the MMDL testing, we first confirmed the best type of input for each type of models. Four random combinations of either CNNs or SAEs with scalogram or spectrogram were tested. Five CNNs or SAEs were used per model type. The number of layers for each of the five CCNs or SAEs models was also randomly assigned between 3 and 5. The results are shown in Table III for CNNs and SAEs. They confirm that the best detection rate is obtained when

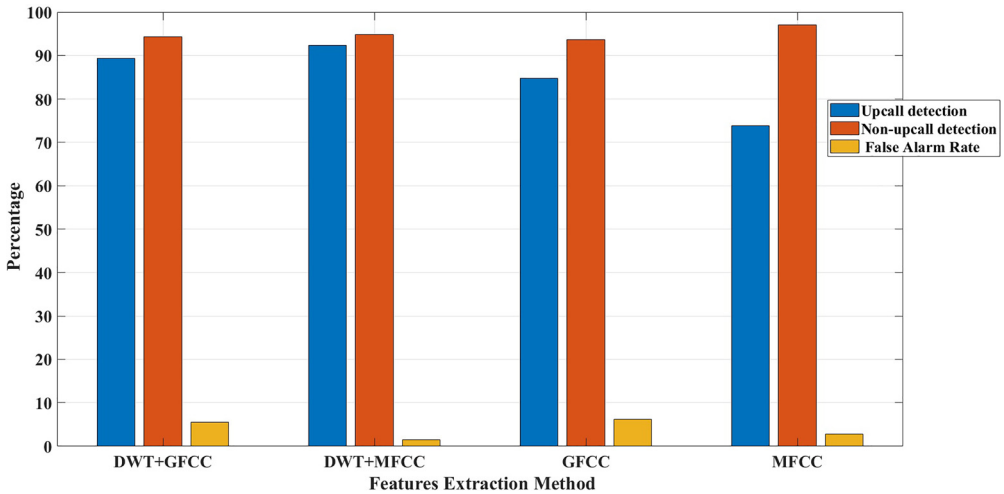

FIG. 6. (Color online) NARW upcall, non-upcall detection, and false alarm rate results using conventional machine learning methods with handcraft features. Features tested were combinations indicated along the abscissa. DWT is the 2-stage DWT with the Daubechies-4 wavelet.

https://doi.org/10.1121/10.0005898

TABLE IV. Results of different ensemble methods (each has five CNN models and five SAE models).

Ensemble method	Up-call detection	Non-up-call detection	False alarm
Majority voting	96%	98.7%	0.41%
Unweighted average	98.1%	99.1%	0.18%
PatternNet	100.0%	100.0%	0.07%

TABLE V. MMDL model classification accuracy for three combinations of subnetwork numbers in each model.

Number of Models	Up-call detection	Non-up-call detection	False alarm
5 CNNs, 5 SAEs	99.3%	99.9%	0.07%
10 CNNs, 10 SAEs	99.8%	100.0%	0.02%
15 CNNs, 15 SAEs	100.0%	100.0%	0%

spectrograms (scalograms) are used as input to CNNs (SAEs).

Three different fusion methods with varying relative numbers of CNN and SAE subnetworks were tested. These networks were tested using 80 665 test files compared to only 1500 files in the previous studies by Ibrahim *et al.* (2018c,d). All results were compiled through a fivefold cross-validation test. Using 15 CNN subnetworks on the spectrogram images, and 15 SAE subnetworks on the scalogram images, the most accurate fusion method is shown to be PatternNet (Table IV). When the relative number of subnetworks was randomly varied between 5, 10, and 15 and fused with PatternNet, increasing the number of subnetworks increased the classification accuracy (Table V).

The performance of the model composed of five CNNs + five SAEs was compared to the following standard deep CNN models: ResNet101, VGG19, MobileNetv2, and EfficentNet. We adopted a transfer learning approach to transfer knowledge of the pretrained models to detect upcalls. The concept of transfer learning is to use an existing deep learning model trained in one domain (usually with a large dataset) to perform a classification task in another domain (usually with a smaller dataset). To use a pretrained model, a fine-tuning procedure needs to be applied. In the procedure, the outer layers of the pretrained model are replaced with additional layers whose weights are trained using the dataset in the new domain. The number of layers added to a pretrained model affects the performance of the model. Another parameter to be tuned is the learning rate, which is also application-dependent. In our application, for each pretrained model, we removed the output layers and replaced them with one (for ResNet101, VGG19) or three dense layers (for MobileNetv2 and EfficentNet) and a softmax activation layer. We froze the other layers in the pretrained model and only train the weights of the newly added layers. We also used the following learning rates: 0.001, 0.004, 0.005 for RestNet101, VGG19, MobileNetv2 and 0.0001 for EfficentNet. Table VI shows that the MMDL model led to more upcall detections than these pretrained Deep Learning models. In addition, the MMDL model is computationally less expensive and structurally simpler than VGG19 and ResNet101, but more complex than MobileNetv2 and EfficentNet.

VI. CONCLUSION

In this study, a new approach for NARW upcall detection was proposed. The NARW sound dataset was collected with multiple recording units equipped with passive acoustic sensors over a period of many years. These recorded signals were converted to spectrograms and scalograms and classified by our proposed MMDL detector. Our algorithm is composed of a number of CNNs and SAEs with randomly chosen design parameters. The detector does not require sophisticated preprocessing and it automates its architecture construction. MMDL combines the advantages of diversity offered by CNNs, extracting discriminative features at both local and global levels, and SAEs, which are designed for data abstraction and reproduction. We showed that the randomness of the model structure and the distinct characteristics of CNNs and SAEs render the integrated MMDL detector robust against data variability. The effectiveness of the proposed MMDL model for NARW upcall detection was verified with Cornell University's (Clark *et al.*, 2002) dataset after relabeling the entire set by visual and aural inspection of the audio files and their spectrograms. Our labels are available on Github (https://github.com/Aliklawat/North-Atlantic-Right-Whales-Data_Corrected-labels) for use by the research community. Our experimental study demonstrated that the MMDL detector outperformed conventional machine learning methods as well as representative deep CNN models which we focused our analysis on, in terms of

TABLE VI. Performance comparison with standard deep CNN models.

Model	Up-call detection	Non-up-call detection	#Parameters	FLOPS
5 CNNs, 5 SAEs	99.3%	99.9%	35M	1.2B
ResNet101 + Spectrogram	96%	99.4%	44M	7.6B
ResNet101 + Scalogram	89%	98%	44M	7.6B
VGG19 + Spectrogram	92%	99%	138M	19.6B
VGG19 + Scalogram	87.4%	98.8%	138M	19.6B
MobileNetv2 + Spectrogram	85.9%	97.5%	6.9M	585M
MobilenetV2 + Scalogram	85.2%	97.4%	6.9M	585M
EfficentNet + Spectrogram	87.35%	97.6%	5.3M	0.39B
EfficentNet + Scalogram	86.1%	97.53%	5.3M	0.39B

upcall detection rate, non-upcall detection rate, and false alarm rate; though it is computationally less efficient than MobileNet and EfficientNet. Since the attributes of the MMDL system are not signal specific, we conjecture that it can be used as a classifier for all applications in which multiple classes are involved. As such, the deep learning algorithm is a significant advancement on conventional machine learning methods. The near zero false-positive, false-negative and false alarm rates indicate that this new MMDL detector could be a powerful tool in the detection and monitoring of the low density, endangered NARW, especially in environments with high acoustic-masking.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Protect Florida Whales Specialty License Plate provided through the Harbor Branch Oceanographic Institute Foundation and continuing support from Harbor Branch Oceanographic Institute and Florida Atlantic University. The authors also gratefully acknowledge the acquisition of the acoustic data made available by Cornell University. The authors also acknowledge the NSF MRI Grant No. 1828181, which provided us with the necessary computing equipment. Finally, the authors express their sincere gratitude towards the reviewers, whose comments and suggestions were crucial in improving the quality of the paper.

Abidalkareem, A. J., Abd, M. A., Ibrahim, A. K., Zhuang, H., Altaher, A. S., and Ali, A. M. (**2020**). "Diabetic retinopathy (DR) severity level classification using multimodel convolutional neural networks," in *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, July 20–24, Montreal, Canada, pp. 1404–1407.

Chérubin, L. M., Dalgleish, F., Ibrahim, A. K., Schärer -Umpierre, M., Nemeth, R. S., Matthews, A., and Appeldoorn, R. (**2020**). "Fish spawning aggregations dynamics as inferred from a novel, persistent presence robotic approach," Front. Mar. Sci. **6**, 779.

Clark, C. W. (**1982**). "The acoustic repertoire of the southern right whale, a quantitative analysis," Animal Behav. **30**(4), 1060–1071.

Clark, C. W., Borsani, J., and Notarbartolo-Di-sciara, G. (**2002**). "Vocal activity of fin whales, balaenoptera physalus, in the ligurian sea," Mar. Mammal Sci. **18**(1), 286–295.

Cooke, J. (**2018**). "Eubalaena glacialis. IUCN red list of threatened species 2018," https://dx.doi.org/10.2305/IUCN.UK.2018-1.RLTS.T41712A50380891.en.

Esfahanian, M., Zhuang, H., Erdol, N., and Gerstein, E. (**2015**). "Comparison of two methods for detection of north atlantic right whale upcalls," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, August 31–September 4, Nice, France, pp. 559–563.

Gillespie, D. (**2004**). "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," Can. Acoust. **32**(2), 39–47, available at https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1586.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T. (**2018**). "Recent advances in convolutional neural networks," Pattern Recognit **77**, 354–377.

Halberstadt, A. L. (**2020**). "Automated detection of the head-twitch response using wavelet scalograms and a deep convolutional neural network," Sci. Rep. **10**(1), 8344.

Hayes, S. A., Josephson, E., Maze-Foley, K., Rosel, P. E., Byrd, B., and Cole, T. (**2017**). *US Atlantic and Gulf of Mexico Marine Mammal Stock Assessments—2016* (US Department of Commerce, National Oceanic and Atmospheric Administration, Washington, DC).

He, X., Zhao, K., and Chu, X. (**2021**). "AutoML: A survey of the state-of-the-art," Knowl. Based Syst. **212**, 106622.

Ibrahim, A. K. (**2019**). "Multi-model deep learning for grouper sound classification and seizure prediction," Ph.D. thesis, Florida Atlantic University, Boca Raton, FL.

Ibrahim, A. K., Chérubin, L. M., Zhuang, H., Schärer Umpierre, M. T., Dalgleish, F., Erdol, N., Ouyang, B., and Dalgleish, A. (**2018a**). "An approach for automatic classification of grouper vocalizations with passive acoustic monitoring," J. Acoust. Soc. Am. **143**(2), 666–676.

Ibrahim, A. K., Zhuang, H., Chérubin, L. M., Schärer-Umpierre, M. T., and Erdol, N. (**2018b**). "Automatic classification of grouper species by their sounds using deep neural networks," J. Acoust. Soc. Am. **144**(3), EL196–EL202.

Ibrahim, A. K., Zhuang, H., Erdol, N., and Ali, A. M. (**2016**). "A new approach for north Atlantic right whale upcall detection," in *Proceedings of the 2016 International Symposium on Computer, Consumer and Control (IS3C)*, July 4–6, Xi'an, China, pp. 260–263.

Ibrahim, A. K., Zhuang, H., Erdol, N., and Ali, A. M. (**2018c**). "Detection of north Atlantic right whales with a hybrid system of CNN and dictionary learning," in *Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, December 13–15, Las Vegas, NV, pp. 1210–1213.

Ibrahim, A. K., Zhuang, H., Erdol, N., and Ali, A. M. (**2018d**). "Feature extraction methods for the detection of north Atlantic right whale upcalls," in *Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, December 13–15, Las Vegas, NV, pp. 179–185.

Ju, C., Bibaut, A., and van der Laan, M. (**2018**). "The relative performance of ensemble methods with deep convolutional neural networks for image classification," J. Appl. Stat. **45**(15), 2800–2818.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (**2012**). "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems,* December 3–8, Lake Tahoe, NV, pp. 1097–1105.

Mellinger, D. K. (**2004**). "A comparison of methods for detecting right whale calls," Can. Acoust. **32**(2), 55–65.

Mellinger, D. K., and Clark, C. W. (**1993**). "A method for filtering bioacoustic transients by spectrogram image convolution," in *Proceedings of OCEANS'93. Engineering in Harmony with Ocean*, October 18–21, Victoria, BC, Canada, pp. III122–III127.

Moreno-Seco, F., Inesta, J. M., De Léon, P. J. P., and Micó, L. (**2006**). "Comparison of classifier fusion methods for classification in pattern recognition tasks," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (Springer, New York), pp. 705–713.

Parks, S. E., and Tyack, P. L. (**2005**). "Sound production by north Atlantic right whales (*Eubalaena glacialis*) in surface active groups," J. Acoust. Soc. Am. **117**(5), 3297–3306.

Parks, S. E., Urazghildiiev, I., and Clark, C. W. (**2009**). "Variability in ambient noise levels and call parameters of north Atlantic right whales in three habitat areas," J. Acoust. Soc. Am. **125**(2), 1230–1239.

Pylypenko, K. (**2015**). "Right whale detection using artificial neural network and principal component analysis," in *Proceedings of the 2015 IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO)*, April 21–24, Kyiv, Ukraine, pp. 370–373.

Reeves, R. R. (**2003**). *Dolphins, Whales and Porpoises: 2002–2010 Conservation Action Plan for the World's Cetaceans* (IUCN, Gland, Switzerland).

Urazghildiiev, I. R., and Clark, C. W. (**2006**). "Acoustic detection of north Atlantic right whale contact calls using the generalized likelihood ratio test," J. Acoust. Soc. Am. **120**(4), 1956–1963.

Urazghildiiev, I. R., Clark, C. W., Krein, T. P., and Parks, S. E. (**2009**). "Detection and recognition of north Atlantic right whale contact calls in the presence of ambient noise," IEEE J. Oceanic Eng. **34**(3), 358–368.

Wang, J. L., Li, A. Y., Huang, M., Ibrahim, A. K., Zhuang, H., and Ali, A. M. (**2018**). "Classification of white blood cells with PatternNet-fused ensemble of convolutional neural networks (PECNN)," in *Proceedings of the 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, December 6–8, Louisville, KY, pp. 325–330.