JASA ARTICLE

Beluga whale acoustic signal classification using deep learning neural network models

Ming Zhong,[1] Manuel Castellote,[2,a)] Rahul Dodhia,[1] Juan Lavista Ferres,[1] Mandy Keogh,[3] and Arial Brewer[2]
[1]*AI for Good Research Lab, Microsoft, Redmond, Washington 98052, USA*
[2]*Alaska Fisheries Science Center—National Oceanic and Atmospheric Administration (NOAA) Fisheries and Joint Institute for the Study of the Atmosphere and Ocean (JISAO), University of Washington, Seattle, Washington 98195, USA*
[3]*Alaska Department of Fish and Game, Juneau, Alaska 99802, USA*

ABSTRACT:
Over a decade after the Cook Inlet beluga (*Delphinapterus leucas*) was listed as endangered in 2008, the population has shown no sign of recovery. Lack of ecological knowledge limits the understanding of, and ability to manage, potential threats impeding recovery of this declining population. National Oceanic and Atmospheric Administration Fisheries, in partnership with the Alaska Department of Fish and Game, initiated a passive acoustics monitoring program in 2017 to investigate beluga seasonal occurrence by deploying a series of passive acoustic moorings. Data have been processed with semi-automated tonal detectors followed by time intensive manual validation. To reduce this labor intensive and time-consuming process, in addition to increasing the accuracy of classification results, the authors constructed an ensembled deep learning convolutional neural network model to classify beluga detections as true or false. Using a 0.5 threshold, the final model achieves 96.57% precision and 92.26% recall on testing dataset. This methodology proves to be successful at classifying beluga signals, and the framework can be easily generalized to other acoustic classification problems. https://doi.org/10.1121/10.0000921
(Received 29 October 2019; revised 24 February 2020; accepted 27 February 2020; published online 23 March 2020)
[Editor: James F. Lynch] Pages: 1834–1841

I. INTRODUCTION

A. Background

The beluga whale (*Delphinapterus leucas*), also known as the white whale, is a highly social species that lives in large groups. Like other odontocetes, this species relies heavily on sound. They produce impulsive signals to find and capture prey to feed (Au, 1993), and also produce tonal, impulsive, and combined signals to communicate (Sjare and Smith, 1986). In the U.S., there are five separate populations of beluga whales, all within Alaska. Of those five, the Cook Inlet population is the smallest and has declined by about 75% since 1979. Subsistence hunting contributed to the initial population decrease, but this practice was regulated in 1999, with the last hunt occurring in 2005 (National Marine Fisheries Service, 2008). Still, the Cook Inlet beluga whale population has yet to recover. In 2008, this population was listed as endangered under the Endangered Species Act with hopes that the population would begin to recover. Now, more than a decade later they continue to decline with a current population estimate of 279 whales (Shelden and Wade, 2019).

In order to better assess the seasonal distribution of this population for management purposes, as well as increase our understanding of the effects of noise on beluga whales, there is critical need for accurate and reliable tools that allow the description of beluga whales' presence. Historically, this information has been collected from annual aerial surveys and vessel and shore-based photo identification surveys that use unique markings to identify individuals (e.g., Shelden *et al.*, 2015). More recently, scientists have begun using passive acoustic monitoring to listen for beluga whale signals and describe their long-term spatial and temporal occurrence and habitat use by maintaining a network of underwater acoustic recording moorings (Castellote *et al.*, 2020). However, the acoustic detector performance is degraded when background noise is high due to strong current conditions (including self-noise such as flow noise), commercial shipping, and other diverse anthropogenic noise sources (Castellote *et al.*, 2018). These variables along with often high levels of background noise trigger false detections in the standard semi-automated processing of acoustic recordings, incurring tedious manual validation of automated classifications, which is labor and time intensive.

B. Related work

The classification of biological sounds using machine learning techniques, especially with deep learning models, has recently begun to attract the attention of researchers in the field of bioacoustics.

For example, a random forest model was implemented to improve the analytical efficiency in monitoring and detection of passerine nocturnal flight calls, in which spectro-temporal features are generated (Ross and Allen, 2013). A more popular approach is to conduct image classification

[a)]Electronic mail: Manuel.Castellote@noaa.gov.

https://doi.org/10.1121/10.0000921

TABLE I. PAMGuard whistle and moan detector parameter settings used to detect beluga whale calls in sound recordings from passive acoustic moorings in Cook Inlet, AK, following Keating *et al.* (2015).

FFT parameters	FFT length = 1024
	FFT hop = 512
	Window type = Hann
	Sample rate = 24 kHz
	Time resolution = 42.67 ms
	Frequency resolution= 23.44 Hz
	Time step size = 21.33 ms
Noise and thresholding	Median Filter length = 61
	Average Subtraction = 0.02
	Gaussian Kernel Smoothing = on
	Threshold = 8.0 dB
Detection connections	Minimum Hz = 1000
	Maximum Hz = 12000
	Type = connect 8 (sides and diagonals)
	Minimum length = 5 time slices
	Minimum total size = 50 pixels
	Crossing and joining = Re-link across joins
	Max cross length = 5 time slices

based on spectrograms generated from raw audio recordings, where convolutional neural networks show promising results, such as studies on bat presence (Aodha *et al.*, 2018), or bird call species identification (Sprengel *et al.*, 2016; Adavanne *et al.*, 2017; Grill and Schluter, 2017).

The application of this methodology to marine mammal detection has just recently begun to be explored, but some studies have already shown the potential on passive acoustic data, such as the use of random forest multivariate analysis to classify southern right whales (*Eubalaena australis*) in New Zealand (Webster *et al.*, 2016), or convolutional neural networks (CNN) to classify humpback whale (*Megaptera novaeangliae*) sounds against background noise (Dorian *et al.*, 2017). Most recently, several related studies have demonstrated success of applying deep learning methods on detecting and classifying whale acoustics, including Bergler *et al.* (2019), Bermant *et al.* (2019), and Shiu *et al.* (2020).

In this paper, we address how machine learning techniques can efficiently enhance one of the several steps of the analysis process, which is the validation of semi-automated classifications as true or false beluga signals. We compare the results from manually validated detections of Cook Inlet beluga whale acoustic signals with the results from four different architectures of CNNs to obtain a binary classification (true or false beluga whale detection), aiming to replace labor-intensive, and time-consuming manual validation processes in the near-future. How machine learning techniques can aid in other steps of the analysis process, such as the detection of beluga signals, or the classification of different cetacean species within Cook Inlet beluga habitat will be treated in future work.

II. APPROACHES

A. Data sources and standard analysis

The original raw data were collected with underwater sound recorders (DSG-ST, Loggerhead Instruments) placed in permanent moorings within the Cook Inlet beluga whale critical habitat, Alaska, and serviced twice per year by the National Marine Fisheries Service (NMFS) and the Alaska Department of Fish and Game (ADF&G) (see Lammers *et al.*, 2013; Castellote *et al.*, 2016). Semi-automated detection of beluga whale social signals (i.e., calls and whistles) was obtained with a whistle and moan detector implemented in PAMGuard software version 2.00.14 beta (https://www.pamguard.org/), following the configuration detailed in Table I. Recording datasets included in this study correspond to 5 to 7 months mooring deployments for the ice-free water season (May to September) or winter season (October to April) in 2017–2018 (Table II) in seven locations (see Fig. 1 for their locations) within Cook Inlet beluga critical habitat. Recorder configuration followed a duty cycle of 5 min recordings every 15 min for the entire deployment duration, and a sampling rate of 25 kHz. In total, 13 000 h of audio files were processed for this study.

The whistle and moan detector implemented in PAMGuard was set up with a relatively low classification threshold (whistle and moan detector v2.00.14 Beta, threshold = 8 dB, see Table I) to reduce the risk of missing beluga whale signals at the expense of triggering false detections, due to self-noise or anthropogenic noise. PAMGuard detection results were manually validated through visual and aural inspection of spectrograms in PAMGuard Viewer Mode, and labeled as either a true detection (i.e., with beluga whale calls) or a false detection (i.e., without beluga

TABLE II. Eight datasets processed for model training, validation and testing. These datasets are collected from seven locations in Alaska, covering both summer and winter seasons.

Mooring location	Recording Period	Area Description
Cairn Point	5/2018–9/2018	Knik Arm, highly disturbed zone, port turning basin, continuous dredging operations, nearby military base
Ship Creek	5/2018–9/2018	Knik Arm, highly disturbed zone, main zone of influence by Port of Alaska (Anchorage and MacKenzie)
Point Woronzof	5/2018–9/2018	Moderate disturbance, access area to Knik Arm
Little Susitna River	5/2018–9/2018	Moderate disturbance, important foraging grounds
Susitna River	10/2017-4/2018 & 5/2018-9/2018	Low disturbance, main foraging grounds
Chuitna River	5/2018–9/2018	Highly disturbed zone, oil and gas operation area.
Kalgin Island	10/2017–4/2018	Moderate disturbance, exposed to shipping traffic, winter concentration area

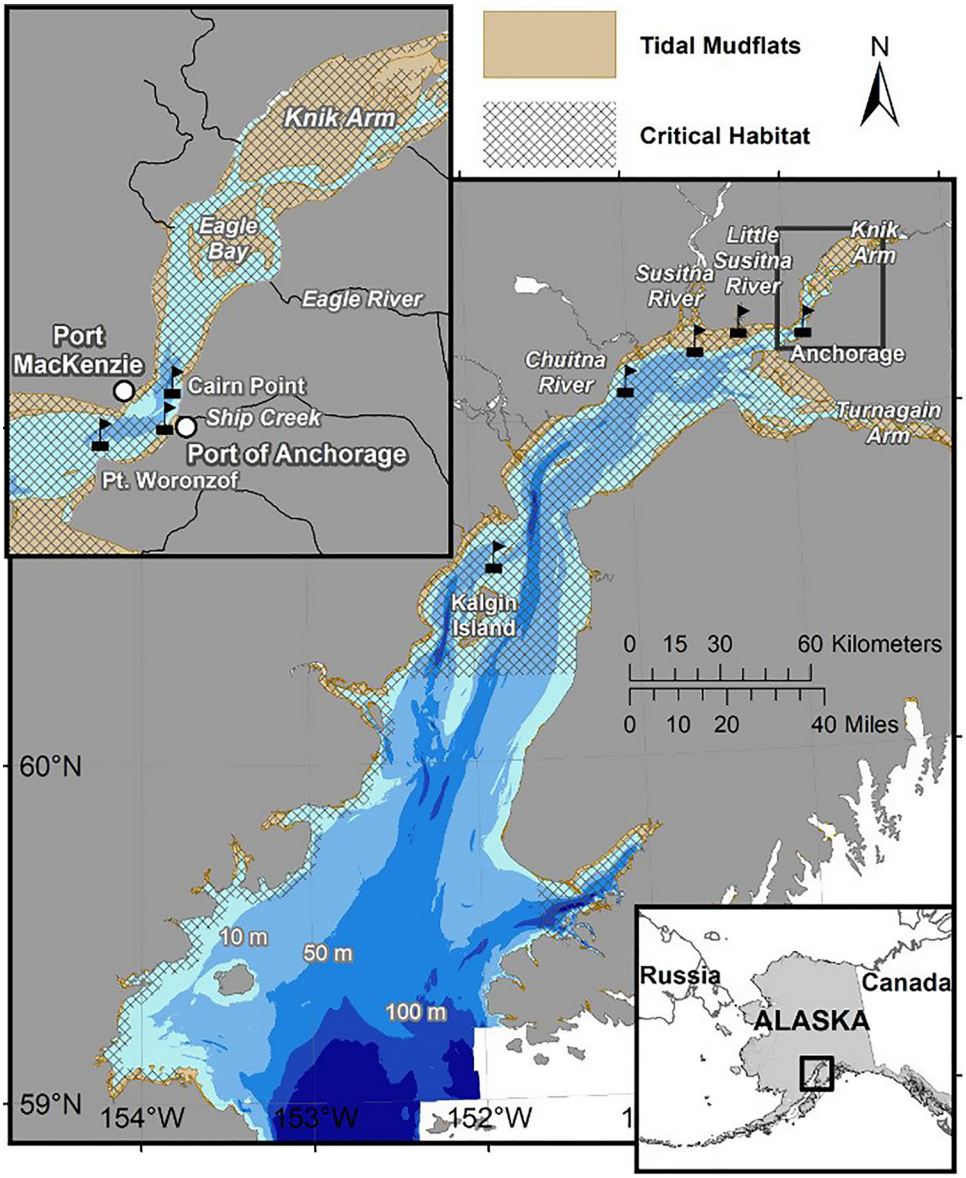

FIG. 1. (Color online) Study area in Cook Inlet, Alaska. Coastal and bathymetric map showing the tidal mudflats exposed at low tide, the designated Cook Inlet beluga whale critical habitat, and the mooring locations where data were sampled from 10/2017 to 9/2018 for this study.

whale calls). In relatively quiet areas (i.e., areas with low levels of anthropogenic noise disturbance and low currents), the whistle and moan detector's precision is close to 95%, while in areas of elevated anthropogenic disturbance and/or high currents (i.e., self-noise), the precision can be as low as 20%-60%.

B. Classification models using convolutional neural networks (CNN)

Using custom written scripts in Python 3.6, spectrograms were extracted from audio files (with NFFT = 256, non overlapping Hanning window). Each spectrogram, resized as 300 pixels by 300 pixels, was generated from a 2-s audio segment from the start time of each of the whistle and moan detections (see Fig. 2, for example). Since the intent of the machine learning model was to replicate the human validation process of the semi-automated detectors' results described in Sec. II A, only audio segments corresponding to PAMGuard detections were processed. The manually validated labeling of detections for each dataset was used as the ground truth to train and evaluate the binary classification model.

A total of eight datasets processed with the PAMGuard whistle and moan detector were used in this study (Table II), and a series of 2-s spectrograms have been generated from the raw audios. Based on the manually validated detections, we generated 89 000 spectrograms from true detections, and 146 000 spectrograms from false detections. Among all the available labeled data from true and false detections, they were split into training and testing portions, according to a 70/30 ratio; and then further split the training data into training and validation using the same ratio. As a result, there are 49% of the data for training, 21% for

https://doi.org/10.1121/10.0000921

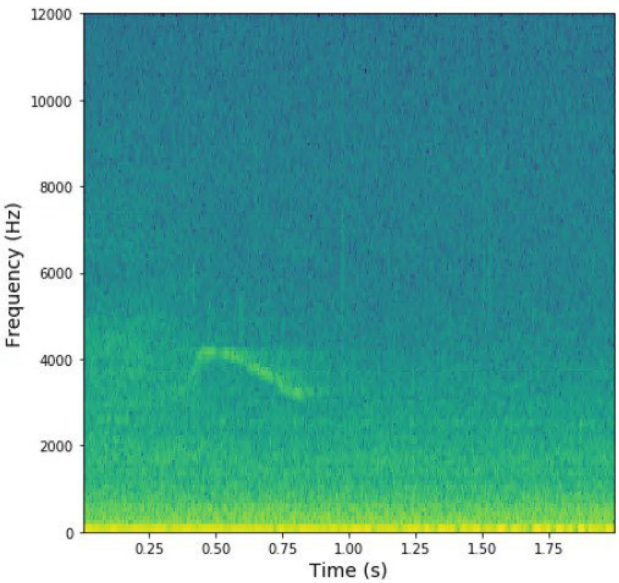

FIG. 2. (Color online) Example of a 2-s spectrogram of audio data containing a low signal-to-noise ratio beluga whale vocalization starting at second 0.40 (FFT = 256, Hanning window).

validation, and the remaining 30% for out-of-sample testing. In order to train the model to better differentiate between the true beluga whale calls from various other types of acoustic signals (such as human voices, sonar signals, waves, and rain), we also generated 25 000 spectrograms from randomly chosen clips when there were no detections, and included them into the training and validation datasets. The spectrograms from true detections were assigned with label 1, and those from false detections and without detections were assigned with label 0. All the generated spectrograms along with their assigned labels were used as input for building a binary image classification model. The model is initially fit on the training dataset, and the validation dataset provides an unbiased evaluation of the model which tunes the model's hyperparameters to avoid overfitting; finally, the testing dataset is used to evaluate how the model generalizes to an unseen dataset. With this data, we trained four different CNNs to implement binary classification of beluga whale calls, and the final prediction is an ensemble of the weighted average of each individual model.

Among these models, we trained one CNN model from scratch, while the remaining three models were built upon pre-trained model weights known as transfer learning (Oquab *et al.*, 2014). Transfer learning is a machine learning technique where a model trained on one task (or domain) is re-purposed on a second related task (or domain). For example, the model weights were initially trained on ImageNet (Deng *et al.*, 2009) database with 1000 classes of objects, but their pre-trained weights can be leveraged by a different task or domain. This approach is effective because the images were trained on a large corpus of photographs and require the model to make predictions on a relatively large number of classes, in turn, requiring that the model efficiently learn to extract features from photographs in order to perform well on the problem.

1. Build CNN using AlexNet architecture

AlexNet (Krizhevsky *et al.*, 2012) is a deep neural network which learns visual knowledge from images to classify a given image. Based on the original architecture of the AlexNet model, we made minor modifications by adding batch normalization layer after each max-pooling layer to solve the internal covariate shift problem (this refers to the change in the input distribution to each layer of the deep neural network, where even small changes might get amplified down the network) as it forces the input of every layer to have approximately the same distribution, and by adding dropout layer (with dropout rate = 0.5) after each fully connected layer to reduce interdependent learning amongst the neurons and thus to prevent over-fitting.

2. Fine-tuning with pre-trained VGG16 model

VGG16 is a convolutional neural network model proposed by (Simonyan and Zisserman, 2015). It makes improvements over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another.

Instead of training the entire model from scratch, we used transfer learning and fine-tuning hyperparameters. With transfer learning, the weights of the first 13 layers from pre-trained VGG16 model were retained, and we fine-tuned the hyperparameters of the last three layers, and added a fully connected layer, a dropout layer, and output layer in the end.

3. Fine-tuning with pre-trained ResNet model

Deep residual networks (He *et al.*, 2016) were arguably the most groundbreaking work in the computer vision/deep learning community in the last few years. ResNet makes it possible to train up to hundreds or even thousands of layers and still achieves compelling performance. Because of the compelling results, ResNet quickly became one of the most popular architectures in various computer vision tasks. Without training the entire model from scratch, we used the pre-trained weights of ResNet, and fine-tuned the parameters by adding a fully connected layer, dropout layer and an output layer.

4. Fine-tuning with pre-trained DenseNet model

DenseNet (Densely Connected Convolutional Networks) (Huang *et al.*, 2017) is one of the latest neural networks for visual object recognition. It is quite similar to ResNet but has some fundamental differences. In DenseNet, a layer receives all the outputs of previous layers and concatenates them in the depth dimension; while in ResNet, a layer only receives outputs from the previous second or third layers. Using a similar approach described in Secs. II and III, we used the

pre-trained weights of DenseNet, and fine-tuned the parameters by adding a fully connected layer, dropout layer and an output layer.

5. Ensemble modeling

Ensemble modeling is a process where multiple diverse models are created to predict an outcome in order to improve the overall performance, either by using many different modeling algorithms or using different training datasets. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and independent, ensemble methods help to minimize the prediction error and improve the stability of the predictions. To ensemble, we take the weighted average of each individual model's predictions while their optimal weights are derived towards the optimization of the overall accuracy.

The developed model code for this study is open source and freely available at https://github.com/microsoft/belugasounds.

III. RESULTS

To evaluate and compare the performance of different models on the same testing dataset, we report four key metrics: precision, recall, accuracy and area under a curve (AUC). Precision measures the proportion of positive identifications that were actually correct [i.e., TP/(TP+FP)]; recall measures the proportion of actual positives that were identified correctly [i.e., TP/(TP+FN)]; and accuracy measures the overall percentage of correct classifications [i.e., (TP + TN)/(TP+FP+TN+FN)]. While precision, recall and accuracy are dependent on the choice of threshold score, average precision provides an aggregate measure of performance across all possible classification thresholds, and it is not affected by the class imbalance.

With a default neutral threshold score of 0.5 used for classification, the results show that each individual CNN model has relatively similar performance, while leveraging the pre-trained model with transfer learning yields better results, especially in recall (Table III). The ensembled model has a 1.5% increase in precision when compared to best individual model results, as well as the highest accuracy and AUC among all models. Testing on the out of sample dataset, the CNN models were able to satisfactorily classify most detections and achieve high accuracy, while the output probabilities also indicate the level of confidence by the model predictions.

While the ensemble model has high confidence scoring for most observations (that is, with probability score >=0.9 for true positives, and <=0.1 for true negatives, see Fig. 3), in the real implementation, the classification threshold score can be tuned with flexibility, for example, from 0.5 to 0.8 or 0.9, to include fewer false positives and thus to get higher precision, or vice versa.

The amount of time required for an expert analyst to manually validate the detections triggered by the PAMGuard whistle and moan detector from the eight datasets included in this study, ranged from 56 to 112 h per dataset (on average 70 h). As a comparison, after the ensemble model has been trained (it is a one-time task, taking about 30 h for training four individual models and their ensembles), on average it only takes 4.5 h for the ensemble model to score each dataset with the same detections using Azure Deep Learning Virtual Machine NC24 (with 224 GB memory and 4 x K80 GPU).

IV. DISCUSSION

The main objective of this study was to test the performance of a machine learning approach to support the manual validation process of semi-automated detections of beluga whale acoustic signals. Compared to an expert analyst manually validating the detections used in this study, we calculate that scoring by our model could save 93% of that time. The CNN ensemble model presented here has high accuracy in differentiating Cook Inlet beluga whale sounds from various types of noises. This methodology could replace the labor-intensive manual validation process. In some cases, faint beluga whale calls were missed by the PAMguard detector or wrongly labeled by manual validation process suggesting the detection process could be improved; on the other hand, there are some cases where various types of signals that look similar to beluga whale calls may be wrongly classified by the model (see Fig. 4). To further improve the ability of the model to distinguish such false signals from true beluga whale calls, a simple solution is to collect more comprehensive data to be included in the training dataset. A manual inspection of these events revealed most are associated with tonal elements of distant ship noise, outboard engine noise, or ice noise from mechanical stress.

Our model was built based on eight datasets from seven locations within the Cook Inlet beluga whale critical habitat, and the majority of the datasets were collected during the ice-free season (May to September), while only one location

TABLE III. Classification results (precision, recall, accuracy, AUC, and average precision) for each individual CNN model and the ensemble model. Bold values denote highest results for each class.

Model	Precision	Recall	Accuracy	AUC	Average Precision
CNN with AlexNet Architecture	0.9504	0.8550	0.9362	0.9783	0.9681
Fine-tuning with pre-trained VGG16	0.9383	0.8894	0.9437	0.9798	0.9700
Fine-tuning with pre-trained ResNet50	0.9422	0.9195	0.9544	0.9868	0.9796
Fine-tuning with pre-trained DenseNet	0.95.03%	**0.9248**	0.9588	0.9898	0.9842
Ensemble	**0.9656**	0.9226	**0.9633**	**0.9906**	**0.9859**

https://doi.org/10.1121/10.0000921

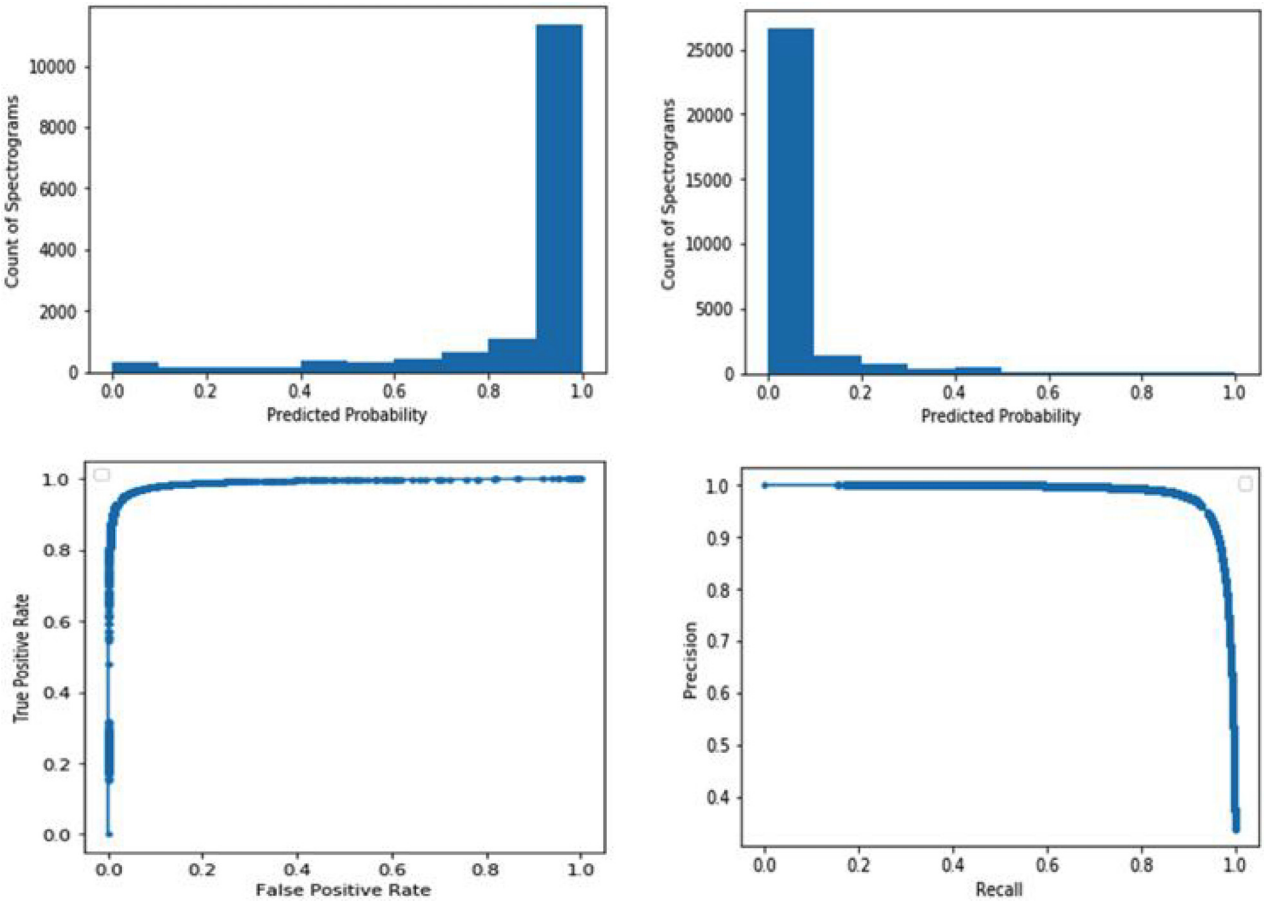

FIG. 3. (Color online) Top left: histogram of predicted probabilities for true positive (i.e., manually validated as "with beluga whale call") testing dataset consisting of 15 000 observations of manually validated detections with beluga whale signals. Top right: histogram of predicted probabilities for true negative (i.e., manually validated as "false beluga whale call") testing dataset consisting of 30 000 observations of manually validated periods false beluga whale signals. Bottom left: receiver operating characteristic (ROC) curve. Bottom right: precision-recall curve.

included data for both ice-free and winter seasons. Considering the underwater acoustic environment may vary seasonally as well as spatially, it would be helpful to expand this study for both seasons in all selected locations. Refreshing the model with a more comprehensive and larger size training datasets will likely yield improved accuracy.

In building each CNN model, we used the default viridis color scales when generating spectrograms, although some previous studies have shown that the choice of color scales may impact model performance (Amiriparian *et al.*, 2017). Even though the average frequency range of beluga whale calls in this dataset is from 2.0 to 5.9 kHz, we did not apply this restriction when plotting spectrograms to make our approach more generic. On the other hand, it will yield some false detections due to convolutional neural models' translation invariant property (that is, the model cannot distinguish noises that have similar frequency modulation but are outside the normal frequency range of beluga whale calls).

Building deep neural network models typically requires a large amount of annotated data due to the multitude of parameters that need to be optimized. However, the application of transfer learning with pre-trained models makes it a viable approach when only small datasets are available for training purposes. While transfer learning leverages the learning from one task, which is generally trained on a large dataset, it does not require learning from scratch for the new task, which directly addresses the smart parameter initialization point for training neural networks. So, for situations where a large amount of data are being collected or archived, but manually validated subsets are small (e.g., only a subset of data are processed, typical of large underwater acoustic efforts), this method is still applicable. While each individual model performs well already, ensembling them with their respective optimal weights will further improve the prediction accuracy and make the final outcome more robust.

The manual validation of detections as false or true beluga whale acoustic signals is the most time-consuming, expensive, and subjective process of all the currently applied acoustic data analysis methods in this long-term study. Because the ground truth data are predefined by what the PAMGuard whistle and moan detector was able to identify as beluga signal, the annotated data used in this study may not be completely representative of the diversity of beluga calls and whistles present in the datasets. However, this approach ensures that, even if the methods to validate semi-automated detections are improved, the detection performance is equivalent and comparable to previous acoustic

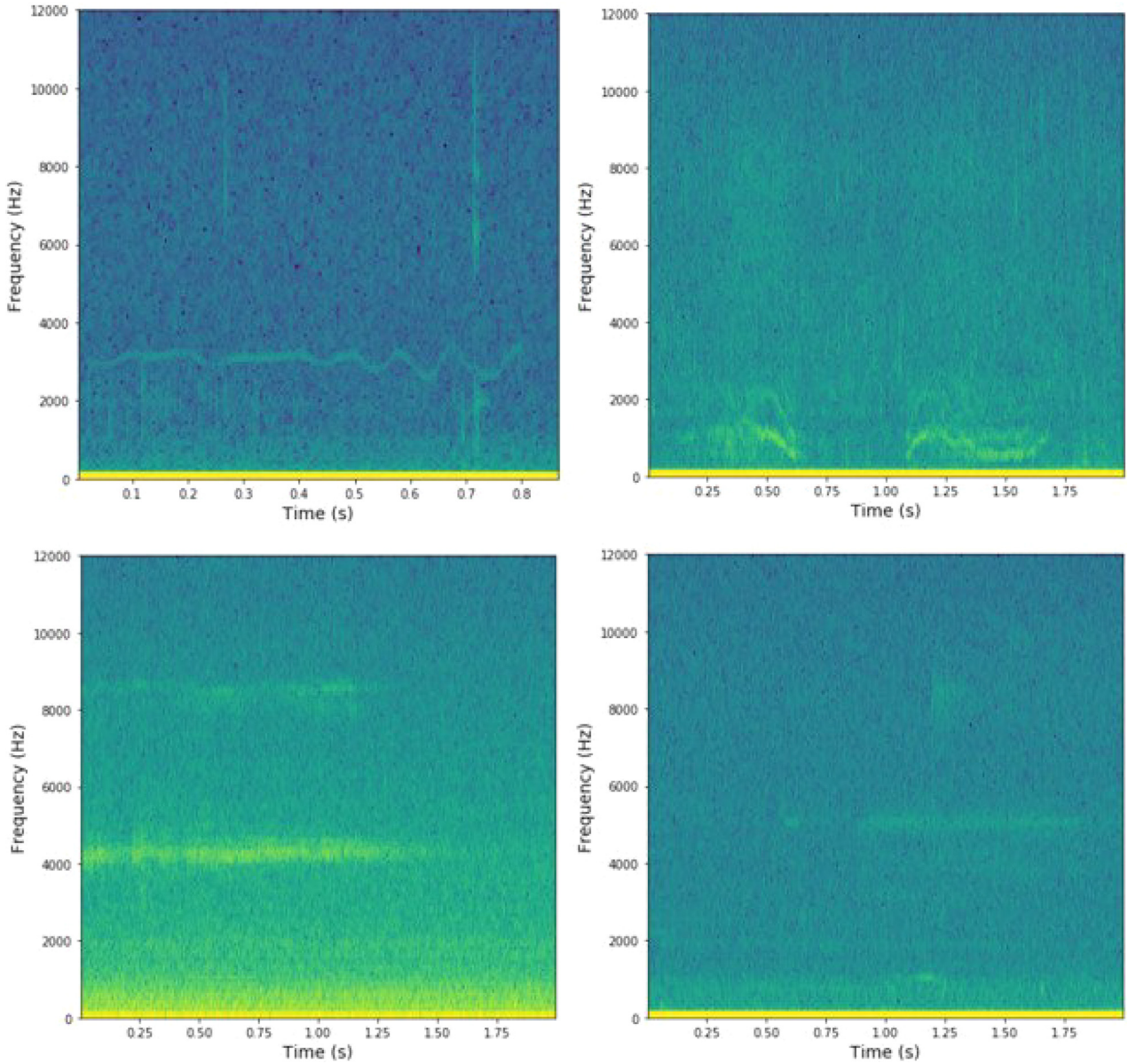

FIG. 4. (Color online) Examples of spectrograms which have been validated as "False beluga whale calls" but were wrongly predicted as beluga whale signals by the model. Upper left—tonal component of noise generated by ice mechanical stress, upper right—outboard motor noise, lower left and right—large size ship noise.

monitoring efforts. Once the manual validation process is replaced by this initial machine learning framework, the next step is to add a detection process routine before the CNN ensemble-based classification, and conduct further comparative testing between this new procedure and the standard analysis procedure based on manually validated semi-automated detections. By doing this, an easy end-to-end processing methodology fully based on deep learning CNN models would be achieved, providing a more comprehensive, efficient, objective, and accurate approach to new efforts in the description of Cook Inlet beluga whale spatial and temporal distribution within the critical habitat.

Finally, the methodology and implementation framework presented in this study can be easily adopted by other similar bioacoustics applications, where target signals require manual validation. This study describes initial steps for placing deep learning CNN analysis, based on weighted output from a model ensemble, as the natural evolution of analysis methods for passive acoustic monitoring data.

ACKNOWLEDGMENTS

The authors would like to thank those who participated in this project for their expertise and support: Jennifer Keating McCullough, for providing insight and assistance with the PAMGuard whistle and moan detector; Benjamin X. Hou, for his support in the integration of machine learning into the IT system of the Marine Mammal Laboratory; the Marine Mammal Program of the Alaska Department of Fish and Game, for their partnership in the

passive acoustic monitoring efforts in Cook Inlet; the NMFS Alaska Regional Office, for their logistics support in Anchorage. This work was supported by AI for Earth grants at Microsoft. Our appreciation goes to Dan Morris for connecting different parties for fruitful discussions and helpful online materials. Passive acoustic data collection was funded by the NOAA Fisheries Species Recovery Grants to States Program.

Adavanne, S., Drossos, K., Cakir, E., and Virtanen, T. (**2017**). "Stacked convolutional and recurrent neural networks for bird audio detection," in *25th European Signal Processing Conference (EUSIPCO)*, pp. 1729–1733.

Aodha, O. M., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., and Jones, K. E. (**2018**). "Bat detective—Deep learning tools for bat acoustic signal detection," Comput. Biol. **14**, 1–19.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., and Schuller, B. (**2017**). "Snore sound classification using image-based deep spectrum features," in *Proceedings of INTERSPEECH*, pp. 3512–3516.

Au, W. (**1993**). *The Sonar of Dolphins* (Springer, New York), 278 pp.

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Hofer, H., and Maier, A. (**2019**). "ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning," Sci. Rep. **9**, 10997.

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (**2019**). "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," Sci. Rep. **9**, 12588.

Castellote, M., Small, R. J., Lammers, M. O., Jenniges, J. J., Mondragon, J., and Atkinson, S. (**2016**). "Dual instrument passive acoustic monitoring of belugas in Cook Inlet, Alaska," J. Acoust. Soc. Am. **139**, 2697–2707.

Castellote, M., Small, R. J., Lammers, M. O., Jenniges, J. J., Mondragon, J., Garner, C., Atkinson, S., Delevaux, J., Graham, R., and Westerholt, D. (**2020**). "Seasonal distribution and foraging occurrence of Cook Inlet beluga whales based on passive acoustic monitoring," Endangered Species Res. **41**, 225–243.

Castellote, M., Thayre, B., Mahoney, M., Mondragon, J., Lammers, M. O., and Small, R. J. (**2018**). "Anthropogenic noise and the endangered Cook Inlet beluga whale, *Delphinapterus leucas*: Acoustic considerations for management," Mar. Fish. Rev. **80**(3), 63–88.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, Fei-Fei. (**2009**). "ImageNet: A large-scale hierarchical image database," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dorian, C., Lefort, R., Bonnel, J., Zarader, J. L., and Adam, O. (**2017**). "Bi-class classification of humpback whale sound units against complex background noise with Deep Convolution Neural Network," arXiv:1703.10887.

Grill, T., and Schluter, J. (**2017**). "Two Convolutional Neural Networks for Bird Detection in Audio Signals," in *Proceedings of EUSIPCO*, pp. 1764–1768.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Keating, J. L., Rankin, S., and Barlow, J. (**2015**). "Burst pulse detector for California Current odontocetes using PAMGUARD," Poster presentation at the *7th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using Passive Acoustics*, 13–15 July 2015, La Jolla, CA.

Krizhevsky, A., Sutskever, I., and Hinton, G. (**2012**). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS)*.

Lammers, M. O., Castellote, M., Small, R. J., Atkinson, S., Jenniges, J. J., Rosinski, A., Oswald, J. N., and Garner, C. (**2013**). "Passive acoustic monitoring of Cook Inlet beluga whales (*Delphinapterus leucas*)," J. Acoust. Soc. Am. **134**(3), 2497–2504.

National Marine Fisheries Service (**2008**). "Conservation plan for the Cook Inlet beluga whale (*Delphinapterus leucas*)," Juneau, Alaska.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (**2014**). "Learning and transferring mid-level image representations using convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724.

Ross, J. C., and Allen, P. E. (**2013**). "Random forest for improved analysis efficiency in passive acoustic monitoring," Ecol. Inf. **21**, 34–39.

Shelden, K. E., Goetz, K. T., Rugh, D. J., Calkins, D. G., Mahoney, B. A., and Hobbs, R. C. (**2015**). "Spatio-temporal changes in Beluga whale, *Delphinapterus leucas*, distribution: Results from aerial surveys (1977–2014), opportunistic sightings (1975–2014), and satellite tagging (1999–2003) in Cook Inlet, Alaska," Mar. Fish. Rev. **77**(2), 1–60.

Shelden, K. E. W., and Wade, P. R. (**2019**). "Aerial surveys, distribution, abundance, and trend of belugas (*Delphinapterus leucas*) in Cook Inlet, Alaska, June 2018," AFSC Processed Report 2019-09, Alaska Fisheries Science Center, Seattle, WA.

Shiu, Y., Palmer, K. J., Roch, M. A., Fleishman, E., Liu, X., Nosal, E-M., Helble, T., Cholewaik, D., Gillespie, D., and Klinck, H. (**2020**). "Deep neural networks for automated detection of marine mammal species," Sci. Rep. **10**, 607.

Simonyan, K., and Zisserman, A. (**2015**). "Very deep convolutional networks for large-scale image recognition," The International Conference on Learning Representations.

Sjare, B. L., and Smith, T. G. (**1986**). "The vocal repertoire of white whales, *Delphinapterus leucas*, summering in Cunningham Inlet, Northwest Territories," Can. J. Zool. **64**(2), 407–415.

Sprengel, E., Jaggi, M., Kilcher, Y., and Hofmann, T. (**2016**). "Audio based bird species identification using deep learning techniques," *CLEF (working notes)*, pp. 547–559.

Webster, T. A., Dawson, S. M., and Rayment, W. J. (**2016**). "Quantitative analysis of the acoustic repertoire of southern right whales in New Zealand," J. Acoust. Soc. Am. **140**(1), 322–333.