frontiers | Frontiers in Artificial Intelligence

TYPE Brief Research Report
PUBLISHED 26 January 2023
DOI 10.3389/frai.2023.1099022

OPEN ACCESS

EDITED BY
Konstantin Markov,
University of Aizu, Japan

REVIEWED BY
Mehul S. Raval,
Ahmedabad University, India
Evaldas Vaiciukynas,
Kaunas University of Technology, Lithuania

*CORRESPONDENCE
Alberto Testolin
✉ alberto.testolin@unipd.it

SPECIALTY SECTION
This article was submitted to Pattern Recognition, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 15 November 2022
ACCEPTED 10 January 2023
PUBLISHED 26 January 2023

CITATION
Nur Korkmaz B, Diamant R, Danino G and Testolin A (2023) Automated detection of dolphin whistles with convolutional networks and transfer learning.
Front. Artif. Intell. 6:1099022.
doi: 10.3389/frai.2023.1099022

Automated detection of dolphin whistles with convolutional networks and transfer learning

Burla Nur Korkmaz[1], Roee Diamant[2], Gil Danino[2] and Alberto Testolin[3,4]*

[1]Department of Information Engineering, University of Padua, Padua, Italy, [2]Hatter Department of Marine Technologies, University of Haifa, Haifa, Israel, [3]Department of General Psychology, University of Padua, Padua, Italy, [4]Department of Mathematics, University of Padua, Padua, Italy

Effective conservation of maritime environments and wildlife management of endangered species require the implementation of efficient, accurate and scalable solutions for environmental monitoring. Ecoacoustics offers the advantages of non-invasive, long-duration sampling of environmental sounds and has the potential to become the reference tool for biodiversity surveying. However, the analysis and interpretation of acoustic data is a time-consuming process that often requires a great amount of human supervision. This issue might be tackled by exploiting modern techniques for automatic audio signal analysis, which have recently achieved impressive performance thanks to the advances in deep learning research. In this paper we show that convolutional neural networks can indeed significantly outperform traditional automatic methods in a challenging detection task: identification of dolphin whistles from underwater audio recordings. The proposed system can detect signals even in the presence of ambient noise, at the same time consistently reducing the likelihood of producing false positives and false negatives. Our results further support the adoption of artificial intelligence technology to improve the automatic monitoring of marine ecosystems.

KEYWORDS

underwater acoustic detection, passive acoustic monitoring, marine biology, environmental monitoring, spectrogram analysis, deep learning, VGG, PamGuard

1. Introduction

Systematic monitoring of marine ecosystems is a key objective to promote sustainability and guarantee natural preservation. Developing and testing innovative monitoring systems is thus rapidly becoming a priority in research agendas, and modern technologies have already shown great potential to advance our understanding of marine communities and their habitat (Danovaro et al., 2016).

Acoustic approaches are widely used to investigate underwater activity thanks to their ability to detect and classify sensitive targets even in low visibility conditions; moreover, passive acoustic technologies (e.g., hydrophones) allow to perform non-invasive continuous monitoring without interfering with biological processes (Sousa-Lima et al., 2013). Notably, most species of marine mammals are acoustic specialists that rely on sounds for communication, reproduction, foraging and navigational purposes. Here we focus on the task of detecting whistles generated by bottlenose dolphins (*Tursiops truncatus*), which can produce a remarkable variety of sound calls for communication purposes (for a review, see Janik and Sayigh, 2013).

Traditional bioacoustic tools to detect odontocete vocalizations typically rely on template matching or algorithmic analysis of audio spectrograms. For example, in the reference approach pursued by Gillespie et al. (2013) three noise removal algorithms are first applied to the spectrogram of sound data, and then a connected region search is conducted to link together

sections of the spectrogram which are above a pre-determined threshold and close in time and frequency. A similar technique exploits a probabilistic Hough transform algorithm to detect ridges similar to thick line segments, which are then adjusted to the geometry of the potential whistles in the image *via* an active contour algorithm (Serra et al., 2020). Other algorithmic methods aim at quantifying the variation in complexity (randomness) occurring in the acoustic time series containing the vocalization, for example by measuring signal entropy (Siddagangaiah et al., 2020).

Nevertheless, automatic environmental monitoring can nowadays be made much more efficient thanks to the deployment of surveying techniques based on Artificial Intelligence. Indeed, deep neural networks have demonstrated great potential in sound detection (Müller et al., 2021) and underwater acoustic monitoring (Testolin and Diamant, 2020; Testolin et al., 2020), and recent work has shown that deep learning can identify signals in large data sets with greater consistency than human analysts, leading to significant advantages in terms of accuracy, efficiency and cost of marine monitoring (Ditria et al., 2022). In particular, Convolutional Neural Networks (CNN) (LeCun et al., 1995) have been applied to detection of whales vocalizations, producing false-positive rates that are orders of magnitude lower than traditional algorithms, while substantially increasing the ability to detect calls (Jiang et al., 2019; Shiu et al., 2020). Deep learning has also been used to automatically classify dolphin whistles into specific categories (Li et al., 2021) and to extract whistle contours by exploiting peak tracking algorithms (Li et al., 2020) or by training CNN-based semantic segmentation models (Jin et al., 2022).

Here we further demonstrate the advantage of deep learning models over alternative algorithmic approaches by testing the detection capability of convolutional neural networks on a large-scale dataset of recordings, collected in a series of sea experiments and carefully tagged by human experts. We show that the performance of deep learning models dramatically exceeds that of traditional algorithms, and we further show that transfer learning (Pan and Yang, 2009) from pre-trained models is a promising way to further improve detection accuracy. The complete dataset of dolphin recordings collected for this study is stored on a cloud server and made publicly available to download (Dataset, 2022).

2. Methods

2.1. Dataset

We created a large-scale database of sound recordings by exploiting a self-made acoustic recorder that comprised a Raspberry Pi-Nano, a sound card sampling at 96 kHz@3B, a pre-amplifier, a battery set, two Geospectrum M18 hydrophones, and a custom made housing. The recorder was anchored by scuba divers at depth of 50 m roughly 200 m from the dolphin's reef in Eilat, Israel. Using floats, the hydrophones were set to hang 1.5 m above the seabed. A picture from the deployment is shown in Figure 1. The recorder was made to continuously log *flac* files for 27 days during the Summer period of year 2021: once recovered, the data passed a quality assurance (QA) procedure to remove sporadic cut-offs and extensive noise periods. The QA involved canceling of noise transients by wavelet denoising, and identifying and discarding cut-off events by thresholding and bias removal.

FIGURE 1
The deployed acoustic recorder with dolphins inspecting the operation. Picture taken from the Eilat deployment site at depth of 50 m.

2.2. Data pre-processing and data tagging

The data passed through a bandpass filter of range 5–20 kHz to fit most dolphins' whistle vocalizations, and through a whitening filter designed to correct for ripples in the hydrophone's open circuit voltage response and the sound card's sensitivity. Recorded audio files consisted of 2 channels, which were averaged before creating the spectrograms in order to reduce noise (see example in Figure 2). Our pre-processing pipeline also removed signal outliers based on their length, using the quartiles-based Tukey method (Tukey, 1949). This resulted in discarding signals longer than 0.78 s and shorter than 0.14 s.

Spectrograms of dolphin whistles were then created by calculating the short-time fast Fourier transform of the signal using MATLAB's *spectrogram* function from the digital signal processing toolbox, using a Blackman function window with 2,048 points, periodic sampling and a hop size obtained by multiplying the window length by 0.8. Subsequent spectrograms were calculated by shifting the signal window by 0.4 s. Spectrogram images were finally produced by applying a gray-scale colormap, converting the frequency to kHz and the power spectrum density to dB and limiting the y-axis between 3 and 20 kHz to focus on the most relevant (dominant) frequency range (Jones et al., 2020).

Spectrograms were then manually tagged by one human expert in two phases: (1) marking tagging and (2) validation tagging. The former involved accurate annotation of 5 s spectrograms over 10 days of data collection, in order to train a preliminary version of a deep learning classifier that was then used to select new portions of recordings containing putative dolphin sounds. This allowed to more efficiently tag the remaining data during the validation tagging phase, which only involved the verification of positive samples detected by the preliminary deep learning classifier. Although the accuracy of the preliminary classifier was not as high as that reported here for the final classifiers, it nevertheless allowed to significantly speed-up the labeling process by automatically selecting the portions or recordings that most likely contained dolphin whistles.

The human expert was instructed to identify dolphin's whistles as curving lines in the time-frequency domain and to ignore contour lines produced by shipping radiated noise. When discrimination

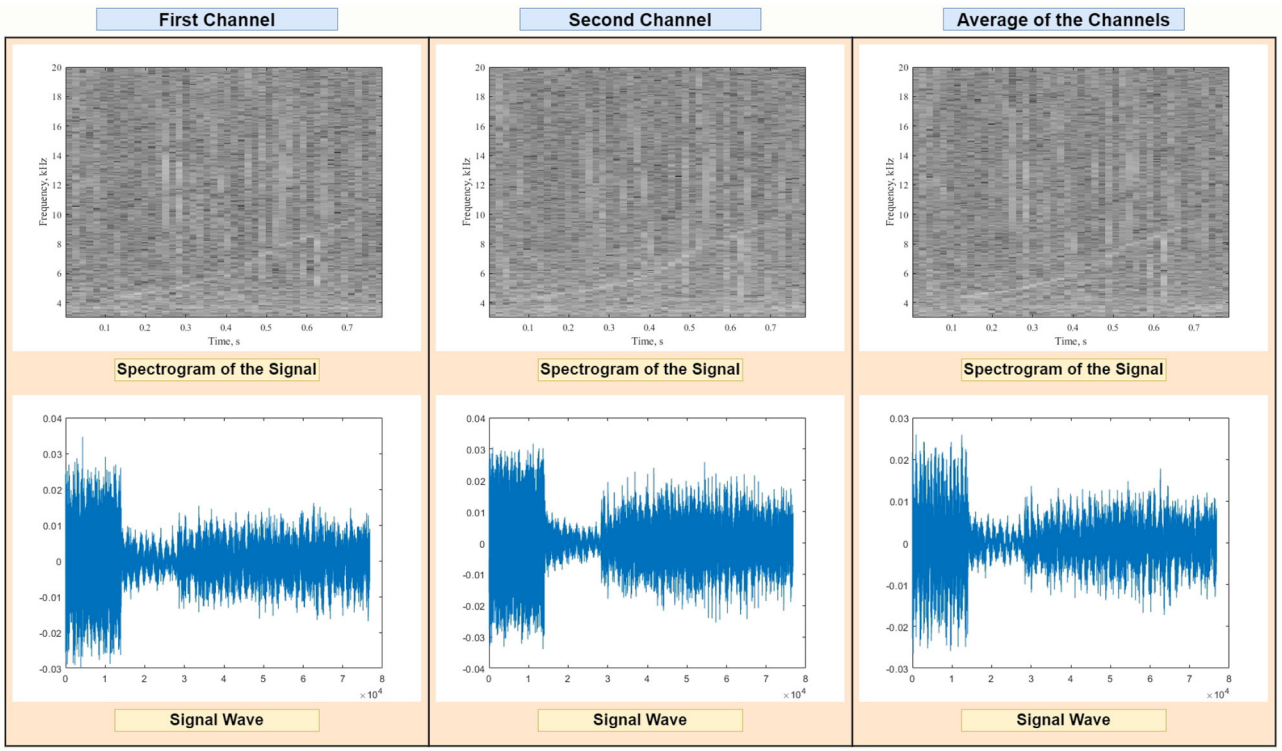

FIGURE 2
Visualization of the spectrogram **(Top)** and raw audio data **(Bottom)** of a representative sample containing a dolphin whistle (curved line in the time-frequency plots). Our detection system receives as input the average of the two recording channels.

was challenging, the expert directly listened to the recorded audio track to identify whistle-like sounds. The tagging resulted in a binary classification (whistle vs. noise) and a contour line marking the time-frequency characteristic of the identified whistle. The latter was used to explore the quality of the manual tagging by checking that the bandwidth of the identified whistle met expected thresholds for a dolphin's whistle, namely between 3 and 20 kHz. A second quality assessment was made by measuring the variance of the acoustic intensity of the identified whistle along the time-frequency contour, where we expect the acoustic intensity of a valid whistle to be stable.

2.3. Baseline detection method

As a benchmark detection method we used PamGuard (Gillespie et al., 2013), which is a popular software specifically developed to automatically identify vocalizations of marine mammals. The working parameters of PamGuard were set as follows:

- The "Sound Acquisition" module from the "Sound Processing" section was added to handle a data acquisition device and transmit its data to other modules;
- The "FFT (spectrogram) Engine" module from the "Sound Processing" section was added to compute spectrograms;
- The "Whistle and Moan Detector" module from the "Detectors" section was added to capture dolphin whistles;
- The "Binary Storage" module from the "Utilities" section was added to store information from various modules.
- The "User Display" module from the "Displays" section was added to create a new spectrogram display.

Input spectrograms were created using the FFT analysis described above, using the same parameters: FFT window length was set to 2,048 points, and the hop size was set to the length multiplied by 0.8 using the Blackman window from the "FFT (spectrogram) Engine" module in the software settings. The frequency range was set between 3 and 20 kHz, and "FFT (spectrogram) Engine Noise free FTT data" was selected as source of FFT data from the "Whistle and Moan Detector" module in settings. While creating a new spectrogram display, the number of panels was set to 2 to visualize both channels. The PamGuard output was considered as a true positive detection if the signal window identified by the software overlapped with at least 5% of the ground truth signal interval. Although this might seem a permissive criterion, it allowed to consider many PamGuard detections that otherwise would have been discarded.

2.4. Deep learning detection methods

We explored two different deep neural network architectures: a vanilla CNN and a pre-trained CNN based on the VGG16 architecture used in object recognition (Simonyan and Zisserman, 2014). Note that the spectrogram images were resized to 224 x 224 and converted into 3D tensors in order to match the number of input channels required by VGG. This was simply achieved by replicating

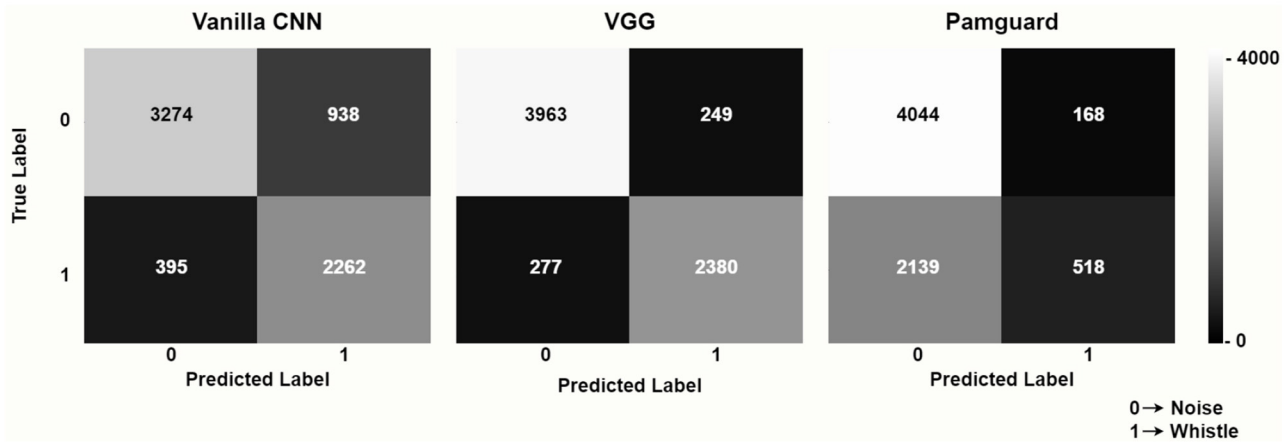

FIGURE 3
Confusion matrices for the deep learning models (vanilla CNN and VGG with transfer learning) and for the PamGuard baseline.

the same image array across the 3 dimensions. Image pixels were normalized by dividing each gray-scale intensity value by 255.

The vanilla CNN model included two convolutional layers interleaved with max pooling layers (pool size = 2) and dropout layers (dropout factor = 0.2). The convolutional layers used 16 and 32 kernels, respectively, with kernel sizes of (7.7) and (5.5), and a stride value of 2. The last convolutional layer was then flattened and fed to 2 fully connected layers containing 32 and 16 nodes, respectively. All layers used a ReLU activation function; only the output layer used a softmax activation. The model was trained using the Adam optimizer with an initial learning rate of 0.0001.

To implement the transfer learning architecture, the top layers of a pre-trained VGG16 were replaced by 2 new fully connected layers with size 50 and 20, respectively, and the *trainable parameter* was set to "True." This allowed the optimizer to jointly train all layers of the VGG model, in order to also adjust low-level features to the new data domain. A ReLU activation function was used in both fully connected layers, while the output layer used a softmax activation. The model was trained using the Adam optimizer with an initial learning rate of 0.00001.

In both cases, binary cross-entropy was used as a loss function and overfitting was monitored by using an early stop criteria (with patience parameter of 15 epochs) applied to a separate validation set. Deep learning models were implemented using Tensorflow (Abadi et al., 2016). All model hyperparameters were automatically optimized using the Optuna framework (https://optuna.org/), considering the following ranges:

- Number of convolutional layers of Vanilla CNN (min 2 to max 4).
- Number of dense layers of VGG and Vanilla CNN (min 1 to max 3).
- Learning rate (min 0.00001 to max 0.01).
- Number of units for dense layers (min 5 to max 100).
- Number of filters of convolutional layers for Vanilla CNN (min 8 to max 256).
- Filter size of convolutional layers for Vanilla CNN (min 3 × 3 to max 7 × 7).
- Patience parameter for early stopping (min 5 to max 20.

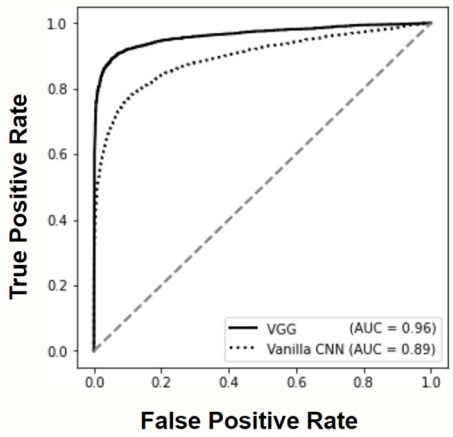

FIGURE 4
Receiver Operating Characteristic (ROC) curves and corresponding Area Under the Curve (AUC) values for both deep learning models.

- Dropout rate, if dropout added (min 0.1 to max 0.3).

2.5. Evaluation procedure

To guarantee a robust assessment of our detection method, the dataset was split into separate training and test sets. The training set only contained spectrograms obtained from audio files recorded between July 24th and July 30th, while the test set only contained spectrograms of audio files recorded between July 13th and July 15th. This allowed to test the generalization performance of our models using a completely different set of recordings, thus evaluating the detection accuracy with variable sea conditions. Overall, the training set contained 108,317 spectrograms, of which 49,807 were tagged as noise and 58,510 as dolphin whistles. The test set contained 6,869 spectrograms, of which 4,212 were tagged as noise and 2,657 were tagged as dolphin whistles. The training set was then randomly

shuffled and further split into training and validation sets, using 5-fold cross-validation. Cross-validation was implemented using the "StratifiedKFold" function from the scikit-learn library in order to make sure that each validation set contained a balanced amount of data from both classes. After cross-validation, all training data was used to produce the final model.

Model performance on the separate test set was assessed by computing mean detection accuracy and by visualizing confusion matrices. True Positive rate and False Positive rate were also computed in order to calculate Precision/Recall, produce Receiver Operating Characteristic (ROC) curves and measure the corresponding Area Under the Curve (AUC) (Davis and Goadrich, 2006):

$$\begin{aligned} &Precision = \frac{TP}{TP+FP};\\ &Recall = \frac{TP}{TP+FN};\\ &True\ Positive\ Rate = \frac{TP}{TP+FN};\\ &False\ Positive\ Rate = \frac{FP}{FP+TN} \end{aligned} \tag{1}$$

Where *TP* indicates True Positives, *TN* True Negatives, *FP* False Positives and *FN* False Negatives.

3. Results

The vanilla CNN model achieved a remarkable mean detection accuracy of 80.6%, significantly outperforming the PamGuard baseline, which achieved 66.4%. Most notably, the performance of the VGG model implementing the transfer learning approach was even more impressive, achieving a mean detection accuracy of 92.3%.

The advantage of deep learning models is even more striking when considering the confusion matrices: as shown in Figure 3, although the amount of True Negatives (Label = 0) was comparable across different methods, the number of True Positives was remarkably higher for deep learning models, especially for the VGG architecture. The low sensitivity of PamGuard was highlighted by the very high number of False Negatives (n = 2,139), suggesting that this method is not very effective in identifying dolphin whistles when the level of signal to noise ratio makes detection particularly challenging. The VGG model achieved much higher performance also in terms of Precision (VGG = 90.5%; Vanilla CNN = 70.7%; PamGuard = 75.5%) and Recall (VGG = 89.6%; Vanilla CNN = 85.1%; PamGuard = 19.5%).

The ROC curves and AUC scores reported in Figure 4 allow to further compare the performance of deep learning models. The superior accuracy of the VGG architecture is evident also in this case, approaching the performance of the ideal classifier.

4. Discussion

With the large increase in human marine activity, our seas have become populated with boats and ships projecting acoustic emissions of extremely high power that often affect areas of more than 20 km^2. The underwater radiated noise level from large ships can exceed 100 PSI with a clear disturbance impact on the hearing, self-navigation and foraging capabilities of marine mammals and especially coastal dolphins (Ketten, 2008; Erbe et al., 2019). Monitoring the marine ecosystem and the sea life is thus a crucial task to promote environment preservation.

Nevertheless, traditional monitoring technologies rely on sub-optimal detection methods, which limit the possibility of conducting long-term and large-scale surveys. Automatic detection methods can greatly improve our surveying capability, however algorithmic solutions do not achieve satisfactory performance in the presence of high levels of background noise. In this paper we demonstrated that modern deep learning approaches can detect dolphin whistles with an impressive accuracy, and are thus well-suited to become the new standard for the automatic processing of underwater acoustic signals. Although further research is needed to validate these methods in different marine environments and with different animal species, we believe that deep learning will finally enable the creation and deployment of cost-effective monitoring platforms.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the animal study because our study only exploited passive acoustic monitoring equipment, which does not interfere with maritime fauna.

Author contributions

RD and AT contributed to conception and design of the study. RD performed the sea experiments and provided the recordings database. BN and AT designed the deep learning models. BN implemented the models and performed the analyses. GD performed data tagging. All authors contributed to manuscript writing, revision, read, and approved the submitted version.

Funding

The research was funded in part by a grant from the University of Haifa's Data Science Research Center.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

(2022). *Link to the Publicly Available Repository Containing All Our Acoustic Recordings*. Available online at: https://csms-acoustic.haifa.ac.il/index.php/s/2UmUoK80Izt0Roe (accessed January 11, 2022).

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation* (Savannah, GA: USENIX), 265–283.

Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., et al. (2016). Implementing and innovating marine monitoring approaches for assessing marine environmental status. *Front. Mar. Sci.* 3, 213. doi: 10.3389/fmars.2016.00213

Davis, J., and Goadrich, M. (2006). "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine learning* (Pittsburgh, PA: Association for Computing Machinery), 233–240.

Ditria, E. M., Buelow, C. A., Gonzalez-Rivero, M., and Connolly, R. M. (2022). Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: a perspective. *Front. Mar. Sci.* 9, 918104. doi: 10.3389/fmars.2022.918104

Erbe, C., Marley, S. A., Schoeman, R. P., Smith, J. N., Trigg, L. E., and Embling, C. B. (2019). The effects of ship noise on marine mammals-a review. *Front. Mar. Sci.* 6, 606. doi: 10.3389/fmars.2019.00606

Gillespie, D., Caillat, M., Gordon, J., and White, P. (2013). Automatic detection and classification of odontocete whistles. *J. Acoust. Soc. Am.* 134, 2427–2437. doi: 10.1121/1.4816555

Janik, V. M., and Sayigh, L. S. (2013). Communication in bottlenose dolphins: 50 years of signature whistle research. *J. Comp. Physiol. A* 199, 479–489. doi: 10.1007/s00359-013-0817-7

Jiang, J.-j., Bu, L.-r., Duan, F.-j., Wang, X.-q., Liu, W., Sun, Z.-b., et al. (2019). Whistle detection and classification for whales based on convolutional neural networks. *Appl. Acoust.* 150, 169–178. doi: 10.1016/j.apacoust.2019.02.007

Jin, C., Kim, M., Jang, S., and Paeng, D.-G. (2022). Semantic segmentation-based whistle extraction of indo-pacific bottlenose dolphin residing at the coast of jeju island. *Ecol. Indicat.* 137, 108792. doi: 10.1016/j.ecolind.2022.108792

Jones, B., Zapetis, M., Samuelson, M. M., and Ridgway, S. (2020). Sounds produced by bottlenose dolphins (tursiops): a review of the defining characteristics and acoustic criteria of the dolphin vocal repertoire. *Bioacoustics* 29, 399–440. doi: 10.1080/09524622.2019.1613265

Ketten, D. R. (2008). Underwater ears and the physiology of impacts: comparative liability for hearing loss in sea turtles, birds, and mammals. *Bioacoustics* 17, 312–315. doi: 10.1080/09524622.2008.9753860

LeCun, Y., and Bengio, Y. (1998). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT Press).

Li, L., Qiao, G., Liu, S., Qing, X., Zhang, H., Mazhar, S., et al. (2021). Automated classification of tursiops aduncus whistles based on a depth-wise separable convolutional neural network and data augmentation. *J. Acoust. Soc. Am.* 150, 3861–3873. doi: 10.1121/10.0007291

Li, P., Liu, X., Palmer, K., Fleishman, E., Gillespie, D., Nosal, E.-M., et al. (2020). "Learning deep models from synthetic data for extracting dolphin whistle contours," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow: IEEE), 1–10.

Müller, R., Illium., S., Ritz., F., and Schmid., K. (2021). "Analysis of feature representations for anomalous sound detection," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, INSTICC* (SciTePress), 97–106.

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Serra, O., Martins, F., and Padovese, L. R. (2020). Active contour-based detection of estuarine dolphin whistles in spectrogram images. *Ecol. Inform.* 55, 101036. doi: 10.1016/j.ecoinf.2019.101036

Shiu, Y., Palmer, K., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., et al. (2020). Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-57549-y

Siddagangaiah, S., Chen, C.-F., Hu, W.-C., Akamatsu, T., McElligott, M., Lammers, M. O., et al. (2020). Automatic detection of dolphin whistles and clicks based on entropy approach. *Ecol.Indicat.* 117, 106559. doi: 10.1016/j.ecolind.2020.106559

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Testolin, A., and Diamant, R. (2020). Combining denoising autoencoders and dynamic programming for acoustic detection and tracking of underwater moving targets. *Sensors* 20, 2945. doi: 10.3390/s20102945

Testolin, A., Kipnis, D., and Diamant, R. (2020). Detecting submerged objects using active acoustics and deep neural networks: A test case for pelagic fish. *IEEE Trans. Mobile Comput.* 21, 2776–2788. doi: 10.1109/TMC.2020.3044397

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*. 5, 99–114. doi: 10.2307/3001913