QUANTITY OVER QUALITY? INVESTIGATING THE EFFECTS OF VOLUME AND STRENGTH OF TRAINING DATA IN MARINE BIOACOUSTICS

Andrea Napoli[*], *Paul R. White, Thomas Blumensath*

Institute of Sound and Vibration Research
University of Southampton, UK
{an1g18, P.R.White, Thomas.Blumensath}@soton.ac.uk

ABSTRACT

The trade-off between the quality and quantity of training data is considered for the detection of minke whale (*Balaenoptera acutorostrata*) vocalisations. The performance of two different detectors is measured across a range of label strengths using training sets of different sizes. A detector based on spectrogram correlation and a convolutional neural network (CNN) are considered. The results show that increasing label strength does not benefit either detector past a certain point, corresponding here to a label density of 60 to 70%. Performance is found to be good even when labels are extremely weak (4% label density). Additionally, it is noted that performance of the spectrogram correlation plateaus beyond the use of 5 training calls, whereas the CNN's performance continues to increase up to the maximum training set size tested. Finally, interaction effects are observed between label strength and quantity, indicating that larger training sets are more robust to weaker labels. Overall, these findings suggest that there is indeed a benefit to collecting more, lower quality data when training a CNN, but that for a correlation-based detector this is not the case.

Index Terms— Weak labels, marine bioacoustics, CNN, spectrogram correlation, sound event detection

1. INTRODUCTION

Passive acoustic monitoring (PAM) forms a major part of marine mammal conservation. Acoustic surveys are an effective and non-invasive means to further our understanding of species-wise geographic distributions, migration patterns and feeding grounds, monitor ecosystem health, and help to mitigate the impacts of human activity. Automated analysis of survey data can improve our ability to achieve these goals, whilst substantially reducing the manual effort required [1].

It is generally accepted that larger training sets allow deep neural networks to build richer representations and improve performance across many classification tasks [2]–[4]. However, labelling large-scale datasets is expensive and time-consuming, so weak labels are commonly used to allow more training data to be collected [4]–[7]. This has resulted in a "quantity over quality" mantra which is at risk of being rashly applied without giving due consideration to the data distributions, labelling techniques, and algorithms pertinent to a particular application. So, work is needed to empirically determine whether "quantity over quality" retains its relevance in marine bioacoustics.

Weak data in PAM can take many forms [1], but this study will consider the effects of label noise – when some training instances do not represent the label they are assigned. This is common when audio is labelled without exact temporal localisation of the signal of interest, which is the scenario presented here. The longer the label, the more irrelevant or even confounding information is likely to be present. The question is then how long the labels need to be in order to make best use of the analysts' time – accepting some label noise allows the analyst to work faster and thus collect more samples.

Label noise can also be introduced through computer-assisted labelling. In PAM, manual labelling is often combined with generic energy detection and unsupervised clustering to reduce annotation effort, at the cost of increased label noise [8]–[10]. Here, the relevant parameter is the sensitivity threshold for the energy detectors. A lower threshold yields more training samples, but also results in more erroneous labels.

Regardless of the origin of the label noise, the recurring theme is that a direct trade-off exists between the strength and quantity of the training data created. Determining the best labelling strategy therefore requires a quantification of exactly how classification performance is affected by these two opposing variables. The impact of label noise has been already been investigated for a range of audio tasks [11]–[13]. Additionally, the effect of training set size is often reported when new models are developed [4]. However, varying both strength and quantity concurrently, and in a controlled manner, has not been previously considered. Conducting a two-way study is important, since the impacts of the two are not necessarily independent (termed *factor interaction*). Thus, the description of a method to determine the presence of any interaction effects has broader relevance to other audio domains as well.

Once the training data has been collected, several weak learning techniques can be applied. Strong label assumption training (SLAT) is a basic option that splits the audio into frames and then assigns the same parent label to each [4]. This creates more, smaller training instances, both of which benefit smaller datasets, but a portion of these may be incorrectly labelled. In this case, prior estimates of the size of this portion can be used to improve performance [14]. Other options include multiple-instance learning [15] and scalable variants [16], or the use of attention or recurrence mechanisms [17].

For many years, template matching techniques such as matched filtering [18], [19] and spectrogram correlation [19], [20] have been the PAM algorithms of choice for detecting call types with limited variation (known as stereotyped vocalisations). Following widespread adoption in many other disciplines, CNNs have also recently found success in this field [1], [10], [21], [22]. Since

[*] Thanks to BAE Systems and the Engineering and Physical Sciences Research Council for funding.

these methods differ significantly, but are both extensively used, an implementation of each will be tested. We stress that the objective of this is not to directly compare the performance of the two detection algorithms, but to identify any differences in how gradient-based learning and traditional signal processing methods are affected by the quality and quantity of data used.

In summary, this paper contributes the following:

- Design of two detectors for minke whale vocalisations, one based on spectrogram correlation, and the other a CNN.
- A quantification of how label strength affects the performance of these detectors. This is the first such study in marine bioacoustics, and also tests more datapoints than similar studies in other audio domains.
- The addition of a second dimension to the problem space, so as to explore the impact of both quantity and quality of labels simultaneously. Comments are also provided regarding interaction effects between the two factors.

2. DATA

The scenario considered is the detection of minke whales in towed-array data from the 2017 Hawaiian Islands Cetacean and Ecosystem Assessment Survey [23]. The dataset comprises over 23,000 60-second audio files, sampled at 500 kHz on 6 channels. The channels have varying signal-to-noise ratios (SNRs) based on the distances between the hydrophones and the ship. Like other marine mammals, minke whales are under threat from human activities including vessel strikes, fishing gear entanglement, noise pollution, and ingestion of debris, so collecting regular and accurate abundance estimates is important [24].

The minke whales in this area produce stereotyped “boing” calls. A boing comprises a brief pulse followed by a longer, frequency and amplitude modulated component, and has a peak frequency of 1.4 kHz, harmonics up to 9 kHz, and source levels around 150 dB re 1 µPa·m [25], Figure 1. Note that this image is thresholded to improve clarity for illustrative purposes, but the spectrograms used for the experiments are unthresholded to maximise detectability and for better generalisability of the results.

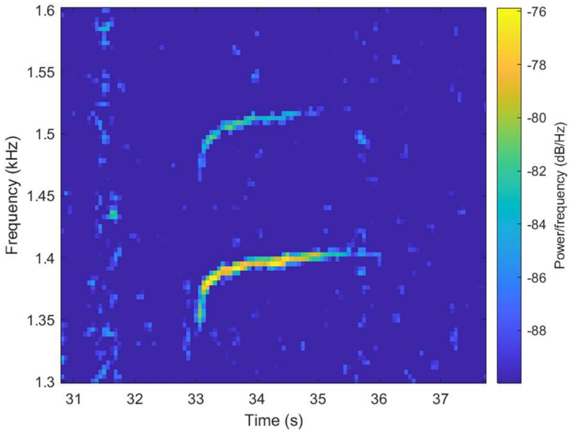

Figure 1: A typical minke whale boing.

For this experiment, 40 calls are identified, and selected such that no audio file contains more than one call. These are strongly labelled by hand, with the start and end times determined to 0.1 s precision. Call durations range from 1.4 to 4.3 s, with a mean of 2.7 s. For the non-target class, 40 files are manually verified to contain only ambient noise. All audio is taken from the same survey day to minimise data leakage. The files are split into 4 validation folds, each containing 30 training calls and 10 test calls. Hereinafter, audio will be referred to as *positive* if it contains a call, and *negative* otherwise.

In addition to the strong labels, 7 sets of weaker labels are generated by increasing the length of the label. Thus, weakly-labelled *positive* audio will contain varying amounts of ambient noise in addition to a call, depending on the label length. For the first 3 sets, the label length is variable, and is equal to the call length plus a fixed time quantity of 0.5, 1 or 1.5 s. Since no call is longer than 5 s, for weaker sets, the label length is fixed, and varies from 5 to 60 s. The extra time added is split randomly between the start and end of the strong label, so the calls can occur at any point in the audio. The strength of each set can be quantitively measured by considering the call duration as a percentage of the overall label length, referred to as the *label density* [11]. Table 1 shows the label density for each training case.

Table 1: Average label density for each of the label strengths tested.

Training Set	*Label Density*
Strong	100%
Strong +0.5 s	84%
Strong +1 s	72%
Strong +1.5 s	63%
5 s	52%
10 s	26%
20 s	13%
60 s	4%

Spectrogram representations of the audio are generated using 250 ms Hamming windows with 75% overlap. The spectrograms are cropped to a narrow band between 1.3 and 1.6 kHz, containing only the peak frequency and one additional harmonic, as per Figure 1. The spectrograms are then divided into 1 s frames with 50% overlap, discarding any excess beyond the length of the label. The resulting frames measure 76 by 16 pixels. Each audio channel is treated independently, resulting in 6 times as many samples. Since the channels have varying SNRs, this acts as a form of data augmentation and helps to regularise the data. SLAT is then applied, so all the frames are assigned the label from their parent audio segment. Note that:

- The “Strong” set contains only samples where the calls entirely fill the frames.
- The “Strong +0.5 s” and “Strong +1 s” sets also include samples that only partially contain calls.
- The weaker sets include partial samples as well as samples that do not contain calls at all (but are still labelled *positive*).

For each label strength, the number of calls in the training sets is varied from 30 down to 1. Training on a single channel of one call is also considered. The test data is the same for every training case in the fold, and is always strongly labelled.

In every case, the *negative* audio contains only ambient sounds, which is not unrealistic for PAM data. Thus, the label noise is entirely one-sided, effectively rendering this a positive-unlabelled (PU) learning problem [26]. The classes are kept balanced by randomly undersampling from the *negative* audio.

Table 2: Average detection accuracies and standard deviations for different sized training sets and label strengths.

№ Calls	*Strong*	*Strong +0.5 s*	*Strong +1 s*	*Strong +1.5 s*	*5 s*	*10 s*	*20 s*	*60 s*
	Spectrogram Correlation							
30	99.0 (0.8)	98.6 (0.7)	99.0 (0.8)	98.4 (1.2)	98.0 (1.2)	97.5 (1.4)	97.6 (1.7)	95.9 (2.1)
20	99.0 (0.8)	98.8 (0.9)	99.0 (0.8)	98.4 (1.1)	98.1 (1.2)	97.5 (1.5)	97.3 (1.7)	96.1 (2.1)
10	98.9 (0.9)	99.1 (0.7)	98.8 (0.7)	98.5 (0.8)	98.7 (1.4)	97.4 (1.5)	97.2 (1.4)	95.7 (2.0)
5	99.1 (0.7)	98.9 (1.0)	98.9 (0.5)	98.5 (0.8)	98.2 (1.5)	97.1 (1.5)	97.0 (1.6)	94.9 (1.7)
1	96.7 (1.7)	96.5 (2.1)	96.8 (1.6)	96.6 (2.2)	96.7 (2.2)	95.4 (2.9)	93.7 (3.7)	90.9 (4.6)
1/6	96.4 (2.2)	96.6 (2.1)	96.6 (2.1)	97.3 (2.2)	96.8 (2.1)	96.3 (2.6)	96.2 (2.3)	93.1 (3.8)
	CNN							
30	98.4 (1.5)	98.0 (2.0)	98.1 (1.1)	96.0 (1.0)	94.1 (0.6)	86.9 (1.2)	82.0 (8.2)	83.5 (3.9)
20	97.5 (2.3)	98.0 (2.0)	97.4 (1.7)	95.7 (1.7)	93.2 (2.0)	83.8 (6.4)	83.1 (5.3)	79.4 (5.7)
10	94.4 (5.7)	96.6 (3.0)	96.7 (2.2)	93.7 (1.8)	92.1 (1.8)	81.6 (1.4)	80.4 (4.0)	78.4 (8.6)
5	94.8 (5.8)	95.9 (3.8)	96.1 (2.7)	93.6 (1.9)	88.9 (1.9)	82.5 (3.5)	79.5 (1.7)	76.0 (6.4)
1	85.6 (2.8)	88.4 (6.1)	85.1 (2.5)	90.9 (5.1)	81.9 (3.7)	75.2 (1.9)	73.2 (5.0)	61.9 (11.3)
1/6	76.2 (8.9)	62.6 (15.0)	66.2 (11.9)	61.4 (20.5)	60.9 (15.4)	54.5 (10.9)	51.7 (10.8)	47.8 (5.6)

3. DETECTORS

Design of the spectrogram correlation detector follows Mellinger and Clark [20]. For each test sample, a 2D correlation is performed with each *positive* training sample and the highest correlation value is taken as the recognition score. The *negative* training samples are unused. The decision threshold is then set as the median recognition score across the test samples. Thus, this implementation implicitly assumes that half of the test samples are *positive*. However, this can easily be modified to use a fixed threshold or a threshold-moving algorithm for imbalanced data [27].

A simple CNN is designed with three convolutional layers and one dense layer. The convolutional layers have 3 by 3 kernels, [2, 2] stride, 8, 16 and 32 filters, and are followed by batch normalisation [28] and RELU activations. The network has 7,170 trainable parameters in total. Spectrogram values are rescaled to the range [0, 1] before input. Training is performed using the Adam optimiser [29] with an initial learning rate of 0.003, a batch size of 50, and early stopping.

4. RESULTS

The average detection accuracies and standard deviations across the 4 validation folds are shown in Table 2. The results are also given graphically, also called interaction plots, in Figure 2.

As expected, the overall trend is that the detectors perform better when trained with stronger labels, and higher quantities of data. However, the results show that both methods can tolerate some label noise (corresponding to a label density of 60 to 70%) without a meaningful reduction in performance. This has an important corollary: increasing the strength of the labels does not improve performance beyond a certain point.

Since the analysis frames are 1 s long, “Strong +1.5 s” is the first dataset for which samples can be labelled *positive* but not contain even a partial call, and this coincides with the accuracy beginning to drop. This suggests a possible interaction between the length of the analysis frames and the strength of the labels, especially for the CNN. However, further work is needed to establish to what extent, if at all, this is the case, and determine whether frame length should also be considered when choosing a label strength.

The spectrogram correlation responds similarly to increasing label quantity, with performance plateauing beyond the use of 5 training calls. Performance of the CNN, however, continues to increase up to the maximum training set size tested. Thus, the CNN is shown to scale better with the quantity of training data available. On the other hand, the spectrogram correlation is more robust to fewer training samples, with the accuracy dropping at most only a few percent between the highest and lowest values tested.

It is observed that using all 6 audio channels instead of only one actually makes the spectrogram correlation perform worse. This is because channels for the same call are highly correlated, so comparing additional channels provides insufficient new information to compensate for the reduced SNR, which only serves to confuse the detector. On the other hand, the varying SNR is shown to be an effective regulariser for the CNN, improving accuracy by an average of 20%. When only a single channel is available, it is likely that similar gains can be achieved by varying the SNR artificially (i.e., standard data augmentation).

Both methods perform well even when labels are extremely weak. In the weakest case, only 4% of the *positive* audio distinguishes it from the non-target class. The spectrogram correlation has far better robustness to label noise, with its accuracy dropping by at most only 6%. This is likely because samples of ambient noise correlate poorly with each other, as well as with samples of minke whale call. The CNN, on the other hand, does not hold this bias, making it susceptible to overfitting, and learning to erroneously discriminate the two classes based on spurious information in the background noise. With only 7,000 parameters, the extremely small capacity of the model may have helped to avoid this.

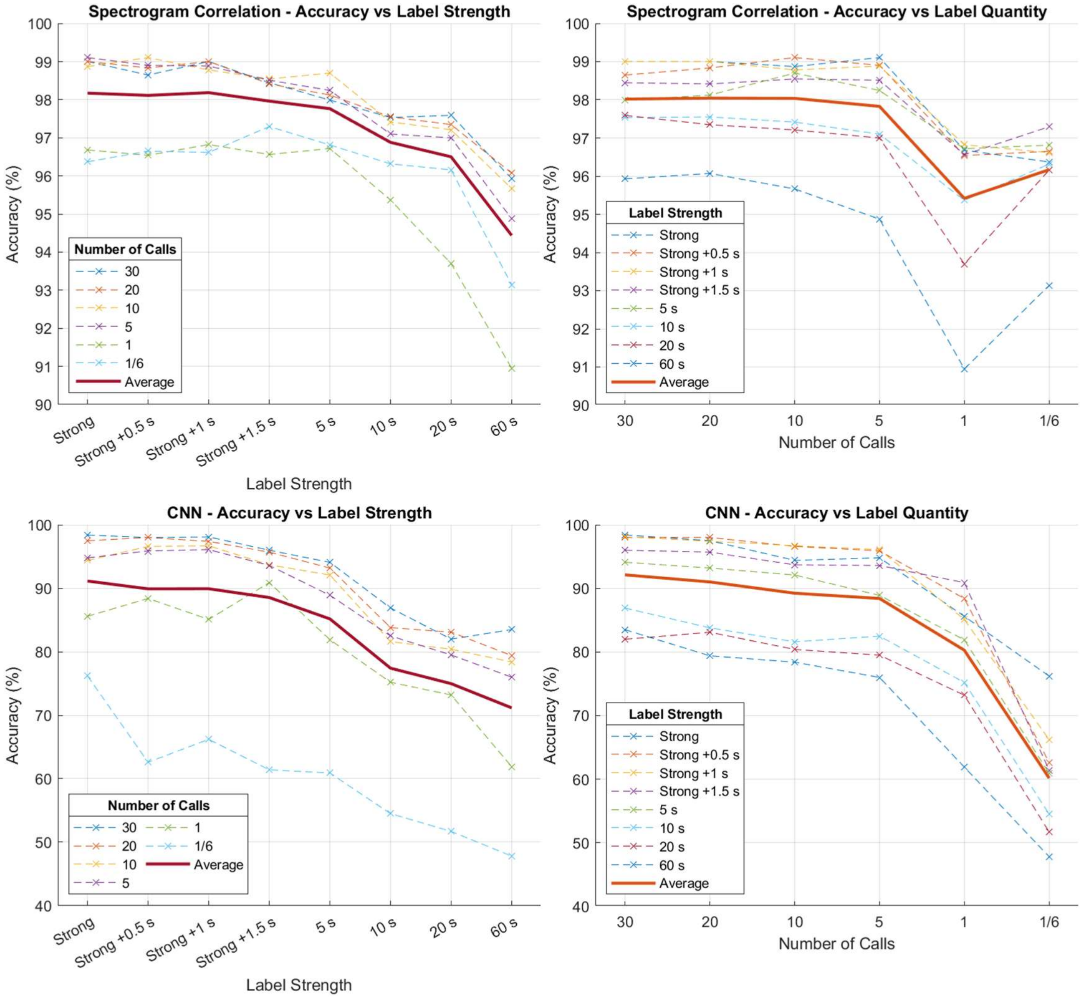

Figure 2: Interaction plots for label quantity and strength, for the spectrogram correlation (top) and CNN (bottom).

The presence of factor interaction is indicated by the lines in an interaction plot being nonparallel. In general, the lines in each plot of Figure 2 can be seen to spread out from left to right, showing that some interaction between label strength and quantity is occurring. Specifically, the results indicate that stronger labels are more robust to smaller training sets, and larger training sets are more robust to weaker labels. The presence of interaction also demonstrates the importance of conducting multi-factor studies.

5. CONCLUSION

This paper studied the effects of volume and strength of training data on the performance of two detectors for minke whale calls. The aim was to determine the relevance to marine bioacoustics of the oft-quoted principle of "quantity over quality". Given the CNN's good scalability to larger training sets, and the performance saturation that occurs when enhancing label strength, the study concludes that "quantity over quality" does indeed hold for CNNs for the call detection scenario presented. Consistency in other machine learning domains suggests that this conclusion is likely to be valid for other gradient-based models as well. However, this is not the case for the spectrogram correlation, which is found to be incapable of exploiting additional training data. On the other hand, the spectrogram correlation demonstrates greater robustness to noisy labels and smaller training sets, making it more appropriate for few-shot learning scenarios.

Future work includes investigating the effects of frame length, mixing labels of different strengths, and using larger training sets to find the saturation points of the CNN's performance. Extensions can also be made to include more advanced weak learning methods, and classification of multiple marine mammal species. Finally, statistical significance tests such as two-way analysis of variance can be used to provide a more objective measure of factor interaction.

6. REFERENCES

[1] P. Nguyen, "Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context," Doctoral Thesis, Sorbonne University, 2020.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.48550/arxiv.1512.03385.

[3] H. Wei, L. Tao, R. Xie, and B. An, "Open-set Label Noise Can Improve Robustness Against Inherent Label Noise," *Adv. Neural Inf. Process. Syst.*, vol. 10, pp. 7978–7992, Jun. 2021, doi: 10.48550/arxiv.2106.10891.

[4] S. Hershey *et al.*, "CNN Architectures for Large-Scale Audio Classification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 131–135, Sep. 2016, doi: 10.1109/ICASSP.2017.7952132.

[5] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 776–780, Jun. 2017, doi: 10.1109/ICASSP.2017.7952261.

[6] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," *MM 2015 - Proc. 2015 ACM Multimed. Conf.*, pp. 1015–1018, Oct. 2015, doi: 10.1145/2733373.2806390.

[7] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, Oct. 2020, doi: 10.48550/arxiv.2010.00475.

[8] K. E. Frasier, "A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets," *PLoS Comput. Biol.*, vol. 17, no. 12, pp. 1–26, 2021, doi: 10.1371/JOURNAL.PCBI.1009613.

[9] K. E. Frasier, M. A. Roch, M. S. Soldevilla, S. M. Wiggins, L. P. Garrison, and J. A. Hildebrand, "Automated classification of dolphin echolocation click types from the Gulf of Mexico," *PLOS Comput. Biol.*, vol. 13, no. 12, p. e1005823, Dec. 2017, doi: 10.1371/JOURNAL.PCBI.1005823.

[10] M. Ferrari, H. Glotin, R. Marxer, and M. Asch, "DOCC10: Open access dataset of marine mammal transient studies and end-to-end CNN classification," *Proc. Int. Jt. Conf. Neural Networks*, Jul. 2020, doi: 10.1109/IJCNN48605.2020.9207085.

[11] A. Shah, A. Kumar, A. G. Hauptmann, and B. R. Fellow, "A Closer Look at Weak Label Learning for Audio Events," *CoRR*, Apr. 2018, doi: 10.48550/arxiv.1804.09288.

[12] N. Turpault, R. Serizel, and E. Vincent, "Limitations of Weak Labels for Embedding and Tagging," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 131–135, May 2020, doi: 10.1109/ICASSP40776.2020.9053160.

[13] S. Hershey *et al.*, "The Benefit Of Temporally-Strong Labels In Audio Event Classification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 366–370, May 2021, doi: 10.48550/arxiv.2105.07031.

[14] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating Labels from Label Proportions," *J. Mach. Learn. Res.*, pp. 2349–2374, 2009, Accessed: Jun. 09, 2022. [Online]. Available: https://www.jmlr.org/papers/v10/quadrianto09a.html

[15] A. Kumar and B. Raj, "Audio Event Detection using Weakly Labeled Data," *MM 2016 - Proc. 2016 ACM Multimed. Conf.*, pp. 1038–1047, May 2016, doi: 10.1145/2964284.2964310.

[16] A. Kumar and B. Raj, "Weakly Supervised Scalable Audio Content Analysis," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2016-August, Jun. 2016, doi: 10.48550/arxiv.1606.03664.

[17] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and Localization based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-August, pp. 3083–3087, Mar. 2017, doi: 10.48550/arxiv.1703.06052.

[18] G. L. Turin, "An introduction to matched filters," *IRE Trans. Inf. Theory*, vol. 6, no. 3, pp. 311–329, 1960, doi: 10.1109/TIT.1960.1057571.

[19] J. R. Potter, D. K. Mellinger, and C. W. Clark, "Marine mammal call discrimination using artificial neural networks," *J. Acoust. Soc. Am.*, vol. 96, p. 2636, 1994, doi: 10.1121/1.410274.

[20] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *J. Acoust. Soc. Am.*, vol. 107, p. 3518, 2000, doi: 10.1121/1.429434.

[21] T. Lu, B. Han, and F. Yu, "Detection and classification of marine mammal sounds using AlexNet with transfer learning," *Ecol. Inform.*, vol. 62, p. 101277, May 2021, doi: 10.1016/J.ECOINF.2021.101277.

[22] Y. Shiu *et al.*, "Deep neural networks for automated detection of marine mammal species," *Sci. Reports 2020 101*, vol. 10, no. 1, pp. 1–12, Jan. 2020, doi: 10.1038/s41598-020-57549-y.

[23] NOAA Pacific Islands Fisheries Science Center, "Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) towed array data. Edited and annotated for DCLDE 2022," *NOAA National Centers for Environmental Information*. 2022. doi: https://doi.org/10.25921/e12p-gj65.

[24] D. Risch, T. Norris, M. Curnock, and A. Friedlaender, "Common and Antarctic Minke Whales: Conservation Status and Future Research Directions," *Front. Mar. Sci.*, vol. 0, p. 247, May 2019, doi: 10.3389/FMARS.2019.00247.

[25] J. N. Oswald, W. W. L. Au, and F. Duennebier, "Minke whale (Balaenoptera acutorostrata) boings detected at the Station ALOHA Cabled Observatory," *J. Acoust. Soc. Am.*, vol. 129, no. 5, p. 3353, May 2011, doi: 10.1121/1.3575555.

[26] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, "Classification with Asymmetric Label Noise: Consistency and Maximal Denoising," *Electron. J. Stat.*, vol. 10, no. 2, pp. 2780–2824, Mar. 2013, doi: 10.48550/arxiv.1303.1208.

[27] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning," *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2623–2640, Jun. 2021, doi: 10.1021/ACS.JCIM.1C00160/ASSET/IMAGES/LARGE/CI1C00160_0015.JPEG.

[28] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, Feb. 2015, doi: 10.48550/arxiv.1502.03167.

[29] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, Dec. 2014, doi: 10.48550/arxiv.1412.6980.