AUTOMATIC PARAMETER ESTIMATION AND DETECTION OF RINGED SEAL KNOCKING VOCALIZATIONS

Adrià Solana[1], Mairi Young[2], Climent Nadeu[1], Mervi Kunnasranta[2, 3], Ludwig Houégnigan[1]

1. Department of Signal Theory and Communications, Polytechnic University of Catalonia, Barcelona, Spain
2. Department of Environmental and Biological Sciences, University of Eastern Finland, Joensuu, Finland
3. Natural Resources Institute Finland, Joensuu, Finland

Correspondence

Adrià Solana,

Department of Signal Theory and Communications,

Polytechnic University of Catalonia,

Barcelona 08034,

Spain

Email: adria.solana@estudiantat.upc.edu

ABSTRACT

Passive acoustic monitoring is a valuable tool for studying elusive marine mammals, but analyzing large datasets is typically labor-intensive and costly. In this study, we piloted an automatic approach for sound analysis on extensive datasets of acoustic underwater recordings from freshwater Lake Saimaa over a total of 12 months. Our focus was on "knocking" vocalizations, the most commonly found call type of the endangered Saimaa ringed seal (*Pusa saimensis*). The annotated datasets of knock sounds (n = 13,179) were used to train and test binary classification systems to detect this sound type. In addition, the fundamental frequencies of the vocalizations were automatically estimated by an ensemble of methods and corroborated by recent literature. The best classifier was a spectrogram-based convolutional neural network that achieved a minimum F1-score of 97.76% on unseen samples from each dataset, demonstrating its ability to detect knockings amongst noise and other events. Moreover, the estimated fundamental frequencies are comparable to the ones manually computed for the same datasets. These automated approaches can significantly reduce labor and costs associated with manual analysis, making long-term species monitoring more feasible and efficient.

KEYWORDS

convolutional neural networks, fundamental frequency, passive acoustic monitoring, ringed seals

1 | INTRODUCTION

Passive acoustic monitoring (PAM) provides a non-invasive, cost-effective approach for continuously monitoring wildlife across time and space. The process of deploying autonomous recording devices over extended periods generates vast amounts of data, which can provide insights into animal behavior, distribution, abundance, and environmental interactions of various taxa (Bolgan et al., 2023; Gibb et al., 2019; Sugai et al., 2019; Teixeira et al., 2024). This method is particularly advantageous in habitats where visual observation is challenging, and thus is commonly used for monitoring marine mammals, which utilize underwater acoustics in various contexts, such as orientation, feeding, mating, nursing, and other social communications (Fleishman et al., 2023). For elusive pinnipeds, PAM provides a valuable tool to study the presence, abundance, distribution, and spatial/temporal behavior patterns of the species (Åsvestad et al., 2024; de Vincenzi et al., 2019; Marcoux et al., 2017; Muñoz-Duque et al., 2024; Prawirasasra et al., 2021; Roca et al., 2023; Wilmer et al., 2025; Young et al., 2025). PAM can, therefore, play a pivotal role in conservation efforts by enabling long-term systematic monitoring of vulnerable species and populations.

The Saimaa ringed seal (*Pusa saimensis)* is an endangered species endemic to the freshwater Lake Saimaa in Finland (Löytynoja et al., 2025). Conservation efforts over the last decades have brought the population from a hundred individuals in 1984 to approximately 500 individuals (Kunnasranta et al., 2021; Metsähallitus, 2025). Nevertheless, this small and fragmented land-locked population continues to face several threats, including by-catch mortality, human disturbance, and climate change. To understand the impact these threats have, as well as to instate effective mitigation, the population must be carefully monitored (Kunnasranta et al., 2021). Among other population monitoring methods, such as PhotoID (Koivuniemi et al., 2016), lair censuses (Kunnasranta et al., 2021), and genetic monitoring (Valtonen et al., 2015), PAM methods can be applied to population research and conservation as a complement to the current approaches.

Especially in aquatic-mating species, such as ringed seals, underwater sounds play an important role in male competition and mate attraction. Previous studies suggest that, among varied sound types, ringed seals produce a specific call type, referred to as “knocking” (Kunnasranta et al., 1996; Mizuguchi et al., 2016; Rautio et al., 2009; Young et al., 2025). Characterized by a series of rapid knocking sounds, this call is the most common during the breeding and mating seasons. This vocal behavior provides an opportunity to study the species’ presence, behavior, and seasonal activity patterns non-invasively, offering valuable insights into population dynamics and habitat use, especially during critical periods such as breeding and mating. However, current approaches to analysing PAM data are a manual process, requiring trained analysts to visually and audibly inspect the data to identify target seal calls. This intensive process is not only labor intensive but may also introduce human error. As the amount of available acoustic data increases, the need for automatic detection of these vocalizations becomes necessary.

Recent advances in machine learning, especially deep learning, offer potential methods for efficiently processing large acoustic datasets of marine mammals (Allen et al., 2021; Caruso et al., 2020; Kirsebom et al., 2020; Miller et al., 2023; White et al., 2022; Ziegenhorn et al., 2022). Convolutional neural networks (CNNs) have shown success in automatically identifying sounds of other seal species (Escobar-Amado et al., 2022; Licciardi & Carbone, 2024) and could be applied to ringed seals. This study aims to advance our understanding of the fundamental characteristics of the Saimaa ringed seal's knocking calls, allowing for the development of an automatic detection system. By integrating deep learning and signal processing techniques, the proposed system will efficiently identify and analyze these vocalizations in extensive acoustic recordings, reducing manual annotation efforts and ultimately supporting more effective conservation strategies for this vulnerable species. Moreover, the underlying framework can be readily adapted to other seal species sharing similar call characteristics, further broadening its potential impact in marine mammal conservation.

2 | METHODS

2.1 | Acoustic data collection

Data used in this study were collected in Lake Saimaa (61° 05′ to 62° 36′ N, 27° 15′ to 30° 00′ E), Eastern Finland, and were recorded using four acoustic hydrophones (SoundTrap 600 STD, Ocean Instruments, New Zealand) deployed at various dates between 2022 and 2024 (Figure 1 and Table 1). Recording periods included the breeding season (February to April) of the Saimaa ringed seal. The minimum distance between two hydrophones was 13.8 km, increasing the likelihood that vocalizations were obtained from multiple individuals. Each hydrophone had a unique system sensitivity ranging from 175.2 to 176.9 dB re 1 V μPa^{-1} and maintained a flat frequency response across the recording frequency (20 Hz - 60 kHz ± 3 dB). Hydrophones recorded continuously (60/60 duty cycle) with a sampling rate of 96 kHz. The instruments were suspended 2m from the water surface using a subsurface buoy and moored to the lake floor with 20 kg weights (for details see Young et al., 2025).

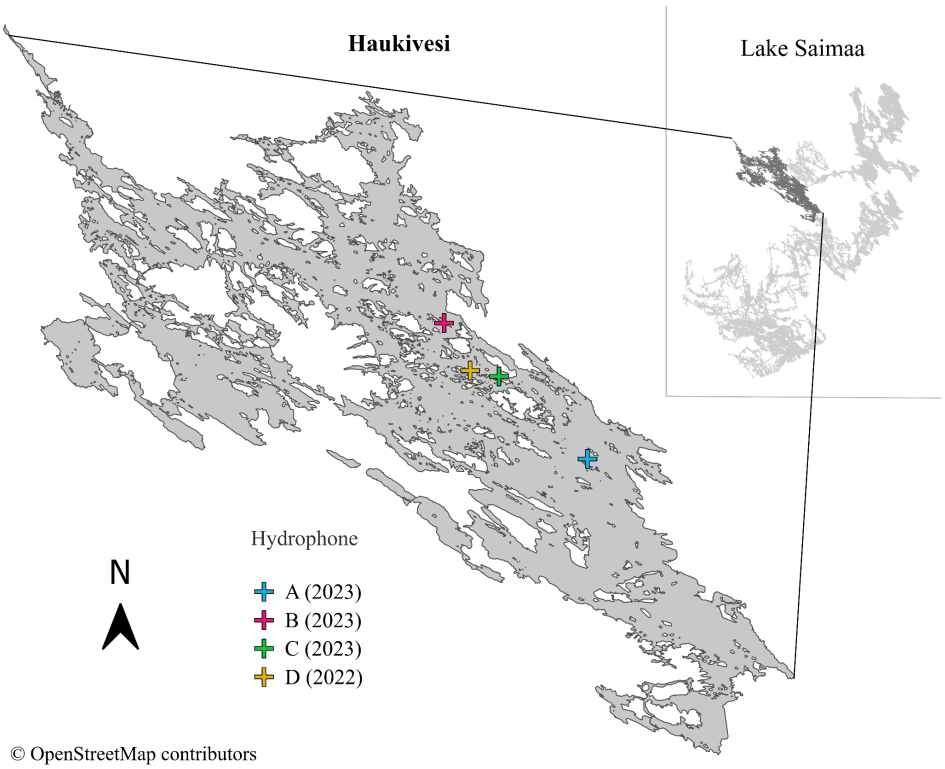

Figure 1 Map of Lake Saimaa with the hydrophone locations in the Haukivesi basin.

Table 1 Hydrophone deployment and recording parameters for recording underwater vocalization of Saimaa ringed seals.

Hydrophone ID	Depth (m)	Recording dates
A	8.3	06 Jan 2023 – 01 Jun 2023
B	6.8	05 Jan 2023 – 01 Jun 2023
C	7.5	20 Dec 2022 – 18 May 2023
D	10.2	28 Jan 2022 – 01 Jul 2022

2.2 | Manual Analysis

A total of 14,309 hours of audio data were manually analyzed. Spectrograms were generated using a Hann window (sample length 512, 50% overlap) using Raven Pro software (version 1.6.4). The data were visually and audibly inspected to identify and annotate significant acoustic events. In total, 13,353 distinct audio events were annotated. The majority of these events (98.7%) were identified as seal vocalizations, while a smaller proportion (0.7%) consisted of scraping sounds produced by seals maintaining breathing holes in the ice. The remaining events (0.5%) were of human or unknown origin.

All seal vocalizations recorded in this study consisted of a single call type, referred to as knocking (Figure 2). This vocalization is characterized by a series of short, repeated pulses with a consistent fundamental frequency (Rautio et al., 2009; Young et al., 2025). Among other sound types, two variations of this call type have been described in Lake Saimaa, with a slower version associated with mother-pup communication (Rautio et al., 2009), and the faster is suggested to play a role in adult mating behavior (Young et al., 2025). During this study, only the second, faster variation of the call was recorded.

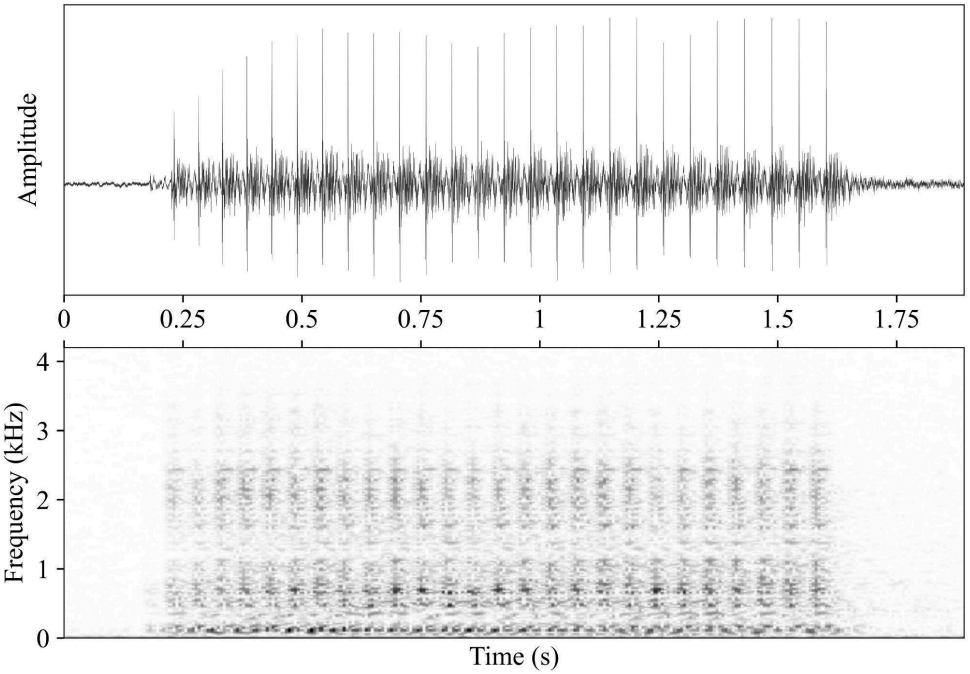

Figure 2 Waveform and spectrogram of a Saimaa ringed seal knocking vocalization. The spectrogram was computed using a 1024-point Hann window with 87.5% overlap and a 1024-point FFT.

2.3 | Data preprocessing

All the sound files were recorded with a sampling frequency of 96 kHz. However, after analyzing the bandwidth of several knockings, the files were downsampled to 19.2 kHz to safely reduce the overall computational cost.

2.4 | Fundamental frequency estimation

The fundamental frequency (f_0) is defined as the lowest frequency component of a periodic signal, or the frequency with which the periods of the signal repeat (Gerhard, 2003). In the domain of speech processing, the fundamental frequency is responsible for the pitch of the human voice. In the context of ringed seal vocalizations, it has been observed that their knocking calls are characterized by periodic waveforms that repeat at a constant fundamental frequency (Figure 2). Therefore, it is

reasonable to hypothesize that the methods employed for the estimation of pitch in human speech can be utilized to estimate the fundamental frequency of knocking calls.

Knocking calls are commonly characterized by manually measuring the number of periodic pulses in a call, its duration, and the duration of each pulse (Mizuguchi et al., 2016; Rautio et al., 2009; Young et al., 2025). An estimation of the fundamental frequency can thus be derived from these measurements, either by computing the inverse of the duration of a pulse or by dividing the number of pulses by the duration of the call. Hence, exploring a method to automatically estimate the fundamental frequency of knockings might result in a tool that could help researchers in their tasks.

To estimate the fundamental frequency in an automated and unsupervised manner, several widely used time-domain-based estimators were implemented. The results of each method were aggregated to obtain more robust estimates. These estimators were the autocorrelation function, the average magnitude difference function (AMDF), and the YIN fundamental frequency estimator (de Cheveigné et al., 2002; Tan & Karnjanadecha, 2003). These tools transform audio signals so that their periodicity can be estimated by identifying the positions of local extrema. They present some differences, namely, the AMDF can be computationally lighter since it avoids the multiplications needed in the autocorrelation function, and YIN is a more sophisticated version of the autocorrelation method that presents lower error rates in pitch detection while incorporating tunable parameters. In addition, the same techniques were used but preceded by a prefiltering step with the non-linear Teager-Kaiser energy operator (TK) (Kaiser, 1990) (Equation 1), in order to achieve more distinctive extrema, especially in noisier or lower amplitude knockings.

$$\Phi[x[n]] = x^2[n] - x[n-1]x[n+1]$$

Equation 1. Teager-Kaiser energy operator.

All of the aforementioned methods require selecting a frequency range in which to find the fundamental frequency. These margins were chosen according to previous studies that included manual characterization of knocking calls (Mizuguchi et al., 2016; Rautio et al., 2009; Young et al., 2025). Based on the reported statistics of the call's characterizations, it was inferred that the fundamental frequencies were usually between 3 and 80 Hz, with an average of roughly 25 Hz and a small standard deviation. To avoid missing uncharacterized calls with higher fundamental frequencies, the range was initially established from 3 to 100 Hz. However, after conducting an initial set of experiments, the results showed that the fundamental frequencies were rarely over 60 Hz. Therefore, the frequency range was ultimately shortened to [3, 70] Hz to improve the performance of the algorithms. In addition, only the most central samples from each vocalization were utilized, discarding 25% of the samples from the beginning and the end. This mitigated both noise and potential transition effects within the calls, which would otherwise lead to inaccuracies. Finally, YIN was implemented using the code from Guyot (2018), and a harmonic threshold of 0.6 was set based on experimental tuning.

2.5 | Classification data

To train the classifiers, fixed-length frames were extracted from the knocking calls and non-annotated parts of the signal (hereafter referred to as "noise"). The frames belonging to knocking calls were labeled as positive samples, and the rest as negative. The length of the samples was determined with the results of the fundamental frequency estimation and the manual analysis in Young et al. (2025). In consequence, the frame length was set to 166 ms, which can capture at least one period in a signal with a minimum fundamental frequency of 6 Hz. The samples were then grouped into datasets, resulting in one dataset for each hydrophone. Finally, the data from each dataset were split into three subsets: 70% of the samples were used for training, 20% were utilized to validate the training process, and 10% of the data were reserved to test the models on unseen samples. Table 2 shows a description of the datasets.

Table 2 Number of train, validation, and test samples with class balance (proportion of positive samples) for each dataset.

Dataset	Class balance	Train samples	Validation samples	Test samples
Hydrophone A (H_A)	50.17%	25,312	7,232	3,616
Hydrophone B (H_B)	50.00%	217	62	31
Hydrophone C (H_C)	50.70%	9,413	2,689	1,345
Hydrophone D (H_D)	57.96%	40,803	11,658	5,829

An initial set of experiments was conducted using Hydrophone D (H_D), as it was the first dataset that was made available. These tests aimed to assess whether there were input features and classifier architectures that outperformed the others. To achieve that, data from H_D were transformed in various ways to experiment with several features and classifiers. Initially, 18 features (Table A1) related to fundamental frequency estimation and voicing detection were extracted with the methods presented in Section 2.4. Other common voicing detection features, such as power and zero-crossing rate, were also added. In addition to the classical features, the neural network-based pitch estimator CREPE (Kim et al., 2018) was used to obtain the network's confidence in the presence of pitch and the average activation along the last layer. Although CREPE does not operate within the estimated range of fundamental frequencies (3 to 70 Hz), these two metrics contributed positively to the separability between the knocking and noise samples. Furthermore, Principal Component Analysis (PCA) (Jolliffe, 2005) was used in some experiments to reduce the feature space into more representative components.

In addition to the features derived from the fundamental frequency estimation, time-frequency representations of the data were also used to train the classifiers. These representations include raw and Teager-Kaiser filtered signals and autocorrelation curves, spectrograms, mel spectrograms, and

Mel Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980). The spectrograms were obtained using a Hann window with a 512-point Discrete Fourier Transform (DFT) and 87.50% overlapping. The mel spectrograms were computed using 20 energy bands. Finally, 20 MFCCS and their first and second-order time derivatives were obtained from the mel spectrograms.

2.6 | Classification models

The architectures used for classification were Support Vector Machines (SVM) (Hearst et al., 1988), Multi Layer Perceptrons (MLP) (Hornik et al., 1989), and Convolutional Neural Networks (CNN) (LeCun & Bengio, 1998). These models operated in the feature, time, and time-frequency domains, respectively. They were trained, evaluated, and tested on H_D, and the list of the used architectures and the type of data that they were fit to can be found in Table 3.

To obtain the best possible results from every classifier, an optimization process was conducted for each tested pair of architecture and data type. The goal was to attain the best results without excessively overfitting the classifiers or creating architectures with a disproportionate number of parameters. To achieve that, both the hyperparameters and dimensions of the models were iteratively adjusted and evaluated. Each experiment was assessed with widely used binary classification metrics, and the number of false negatives was given particular emphasis, as it is of particular importance to avoid missing calls when studying endangered species.

Support Vector Machines were fit and optimized to classify features derived from fundamental frequency estimation. A grid-search optimization was performed using 80% of the data to fit the models, and the remaining 20% was used to test them on unseen data, instead of the default split defined in section 2.5. The SVMs used a Radial Basis Function (RBF) kernel with different penalty parameter (C) values and several kernel coefficients (ɣ). The tested parameters were in a logarithmically spaced range between 0.001 and 100.

The neural network-based architectures were optimized hierarchically by adjusting the models' depth and width and tuning the training hyperparameters, namely, the batch size, the number of epochs, and the learning rate. The Multi-Layer Perceptrons were fit to extensive 1D data, such as an audio frame or its autocorrelation, while the Convolutional Neural Networks were trained with 2D data, including spectrograms and MFCCs.

Table 3 Tested combinations of architectures and data types.

Architecture	Data type
SVM	f_0 features
	PCA f_0 features
MLP	Raw signal
	TK-filtered signal
	Autocorrelation curve
	TK-filtered autocorrelation curve
CNN	Spectrogram
	Mel spectrogram
	MFCCs
	MFCCs + 1st time derivative
	MFCCs + 1st time derivative + 2nd time derivative

3 | RESULTS

3.1 | Fundamental frequency estimation

The fundamental frequency estimation experiments produced relevant outcomes when compared with existing literature. Each available call was analyzed, and its f_0 was automatically estimated by all of the available methods (autocorrelation, AMDF, YIN, TK autocorrelation, TK AMDF, and TK YIN). The histograms of these estimates grouped by method for each dataset can be seen in Figure 3.

To validate these results, the estimates were compared with the manual characterizations of a subset of vocalizations from the same datasets in Young et al. (2025). First, the means and standard deviations of the estimates for each dataset were computed, excluding the YIN estimates that were equal to 0 Hz (meaning the call was perceived as noise by the estimator). While Young et al. (2025) do not present fundamental frequency estimates, Table 3 in that study shows the statistical description of three characteristics from which the f_0 can be derived: the duration of a call, the number of knock pulses, and the inter-pulse interval. Two separate approximations of the f_0 can be computed from these characteristics: a pulse-rate-derived approximation obtained by dividing the mean number of pulses in a call by the mean duration of the call, and an inter-pulse-duration-derived approximation obtained by computing the inverse of the mean inter-pulse intervals. Table 4 in this study presents the results of these calculations. It can be observed how the means of the estimated fundamental frequencies are close to the ones derived from manually analyzed calls, with a mean absolute error of 1.98 Hz in the case of the pulse-rate-derived approximations.

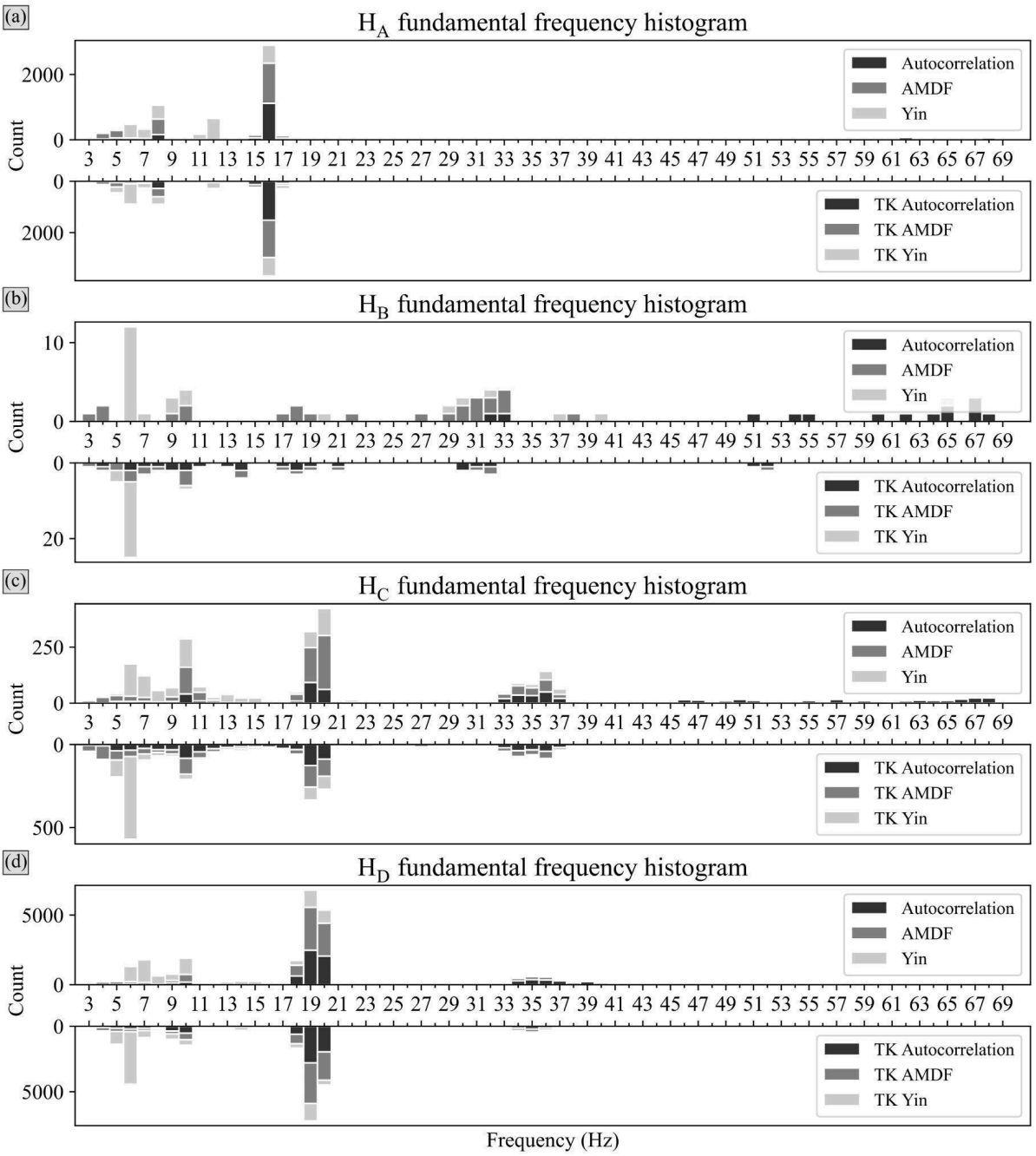

Figure 3 Fundamental frequency estimation histogram for H_A (a), H_B (b), H_C (c), and H_D (d).

Table 4 Comparison of the automatically estimated fundamental frequency statistics and the manual characterization of a subset of calls in Young et al. (2025).

Hydrophone	Total analyzed calls	Estimated f_0 (Hz) mean + SD	Pulse rate-derived f_0 (Hz)	Inter-pulse-duration-derived f_0 (Hz)

A	2,623	15.71 ± 12.39	17.64	20.00
B	24	21.21 ± 19.71	20.14	25.00
C	953	19.79 ± 15.44	22.90	25.00
D	8,995	17.73 ± 10.43	19.56	25.00

A manual review of a limited set of samples indicated that the estimators failed more often with low-SNR calls. This phenomenon was common in all of the employed methods. While the Teager-Kaiser energy operator enhanced the periods in some cases (Figure 4), this improvement was not consistent across all of the samples. Overall, the estimates of unprocessed calls and the ones from TK-filtered samples showed similar statistics.

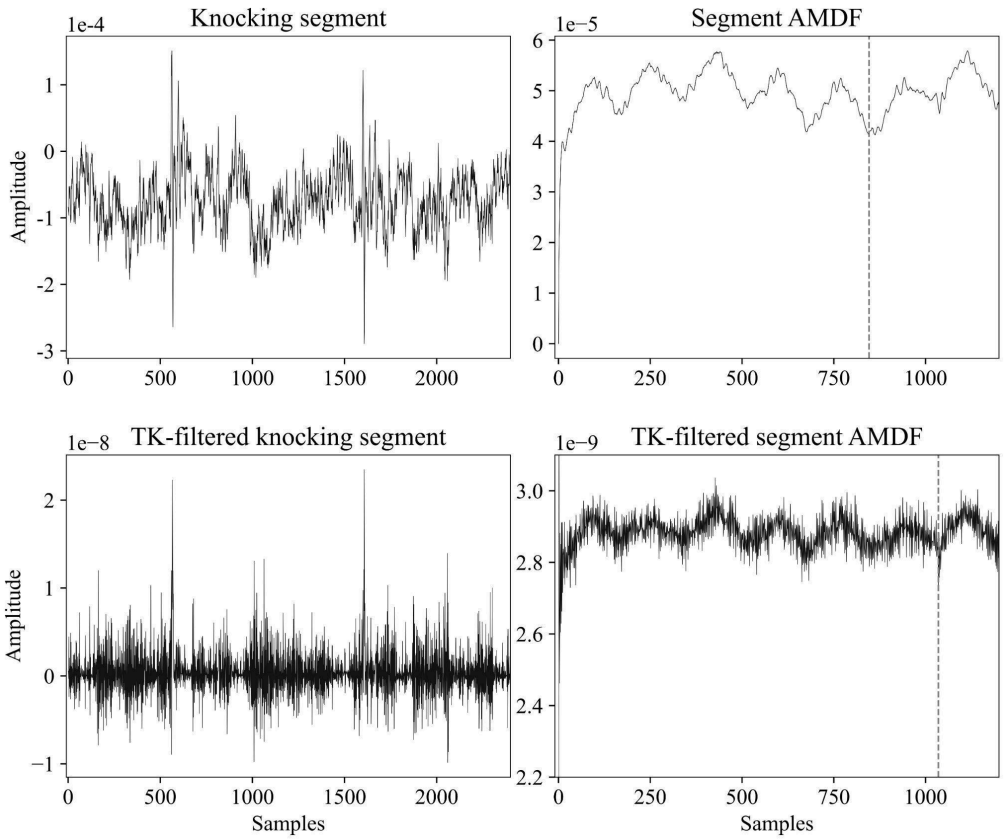

Figure 4 Periodic segment of a ringed seal knocking vocalization (top left) and its AMDF (top right). Below, the effects of the Teager-Kaiser energy operator can be seen with the filtered signal (bottom left) and its AMDF (bottom right). The dashed lines show the period length in samples estimated by the AMDF. A period of roughly 1000 samples can be observed in the left panels. In this example, the periodicity is estimated more accurately when the signal is prefiltered with Teager-Kaiser.

3.2 | Classification

The assessment of the classifiers on unseen test data from H_D can be found in Table 5. This assessment collects the models that yielded the most noteworthy results, in terms of evaluation metrics, features used, and model dimensions altogether. It is worth noting that more complex classifiers, namely CNNs with larger architectures, attained slightly better scores. In particular, the results showed that the CNN models that tripled the number of parameters listed in Table 5 generally outperformed their analogous, smaller versions by less than 0.5% F1-score. However, it was determined that the marginal performance improvement did not justify the significant increment in size.

Table 5 Assessment of the best classifiers using unseen data from H_D. The Precision, Recall, F1-Score, and FN ratio columns are given in percentages.

Classifier	Input features	Num. parameters	Precision	Recall	F1-score	FN ratio
SVM	f_0 features	-	97.55	86.66	93.44	10.34
CNN	Spectrograms	221,057	99.17	96.38	97.76	3.62
CNN	Mel spectrograms	1,053,633	99.49	94.87	97.12	5.13
CNN	MFCCs	46,273	99.23	94.91	97.02	5.09
CNN	MFCCs & D1	128,193	99.39	94.44	96.85	5.56

Note: The abbreviations correspond to: "f_0 features" - fundamental frequency estimation-related features; "MFCC & D1" - mel frequency cepstral coefficients with their first-order time derivative; "Num. parameters" - number of trainable parameters; "FN ratio" - false negative ratio.

As shown in Table 5, the most promising classification approach was using a small CNN to predict spectrograms. This method achieved the highest F1-score and the lowest false negative ratio while having a relatively low number of parameters. Thus, the same architecture and time-frequency representation were evaluated on the remaining datasets. The architecture was trained independently with spectrograms from each dataset, and the results of deploying each model on unseen test samples are shown in Table 6.

Table 6 Assessment of the best model and features on unseen samples from every dataset.

Dataset	Precision (%)	Recall (%)	F1-score (%)	FN ratio (%)
Hydrophone A	99.78	98.35	99.06	1.65
Hydrophone B	100.00	100.00	100.00	0.00
Hydrophone C	98.69	99.56	99.12	0.44
Hydrophone D	99.17	96.38	97.76	3.62

4 | DISCUSSION

The knocking vocalizations of Saimaa ringed seals exhibited several characteristics that made them easily distinguishable from background ambient noise and other events. Their most relevant feature was the presence of clear periods, or pulses, that suggested qualitative similarities to the time-frequency features of human speech. Thus, widely used pitch estimation methods proved to be effective for estimating the fundamental frequency of knocking calls.

The results of the automatic fundamental frequency estimation of knocking calls were reasonably coherent with those derived from Young et al. (2025). Although studying other datasets, the results were also comparable to those of Kunnasranta et al. (1996), Mizuguchi et al. (2016), and Rautio et al. (2009), who also manually analyzed the same type of calls. However, as the available data did not include means to identify the emitters of the vocalizations, the estimates might not have been representative of the plurality of the seals. Despite the fact that the means of both the estimates and the manually analyzed calls were close, the estimates presented a high standard deviation that could have been the result of different factors. First, it is common for the estimators to incorrectly detect a frequency that is *n* times smaller or larger than the actual frequency. Figure 3 (a) depicts this phenomenon, showing that most calls have a f_0 of 16 Hz, while a notable subset has a f_0 of 8 Hz. Second, many seal species are known to produce individually specific calls (Parson et al., 2025), and therefore some individuals in this study may have higher or lower characteristic fundamental frequencies. Figure 3 (c) shows this, as there appears to be a high occurrence of calls with a frequency of 36 Hz, although most calls have a fundamental frequency of 19-20 Hz. Finally, estimation methods failed more often with low SNR samples. In these cases, the f_0 was estimated to be close to the lowest possible frequency, increasing the deviation towards lower values.

A visual examination of a small set of vocalizations revealed that the periods of low-power calls were less distinct than those of higher-power instances. When filtering out a percentile of lower-power vocalizations, the estimated f_0 means generally approached those of the manually analyzed calls, and the standard deviation decreased significantly. This effect was accentuated by increasing the percentile of filtered low-power samples. While this method is not ideal for accurately characterizing all knocking vocalizations, it may be useful for portraying tendencies in the areas captured by each hydrophone.

To conclude the fundamental frequency estimation assessment, there did not seem to be a method that notably outperformed the rest. Every estimator had its advantages and inconveniences; for instance,

the autocorrelation method tended to assign frequencies to the higher end of the admitted range, while YIN did the opposite (Figure 3). The Teager-Kaiser filter was useful in some instances (Figure 4), but not systematically. Altogether, the results suggest that the most reliable method to automatically estimate the fundamental frequencies of knocking calls would be ensembling all of the presented techniques.

Regarding the classification experiments, the separability inherent to the data allowed for a wide variety of architectures to yield satisfactory results. Models with a relatively reduced number of parameters proved to be good alternatives, and increasing the complexity of the model was often not beneficial. Nevertheless, maintaining a low false negative ratio constituted a great challenge for all the tested systems. Amongst the tested architectures and input features, the ones that produced the worst results were the classifiers working in the time domain, namely MLPs that analyzed raw samples or signals derived from the autocorrelation function. Overall, the knockings with lower power seemed to be harder to identify, which can be related to their lack of clear periods. However, the energy distribution along their bandwidth proved to be a good descriptor. The usage of spectrograms and convolutional architectures is coherent with the recent advancements in deep-learning based acoustic detection of seals and other marine mammals (e.g., Escobar-Amado et al., 2022; 2023; Licciardi & Carbone, 2024; Tyshko et al., 2023).

Moreover, the generalization capabilities of the classifiers were studied by deploying the model that achieved the best results on H_D (Table 5, row 2) on H_A, H_B, and H_c without training it with new data. The performance was inferior to the one attained by the same model trained and evaluated separately on each dataset (Table 6), achieving F1-scores of 90.80% for H_A, 70.78% for H_B, 83.33% for H_c and 97.52% for H_D. Albeit the decrease in performance, these results suggest that the best model is able to mostly capture the variability of the data recorded by the employed hydrophones, exhibiting a degree of generalization capabilities.

Finally, this study could be expanded to improve the results or include other types of data. Additional fundamental frequency estimators could include unsupervised wavelet transforms (Kadambe & Boudreaux-Bartels, 1992), while feasible supervised methods could be derived from the CREPE pitch tracker (Kim et al., 2018), provided that at least a subset of the data was manually labeled beforehand. In regards to knocking detection, a number of alternative architectures could be suitable for the task. For example, Recurrent Neural Networks (RNNs) could enhance classification accuracy by incorporating contextual information into the decision-making process. Another promising approach would be to implement CNN-based object detection models, as explored by Hamard et al. (2024). A more advanced approach could involve utilizing a transformer-based architecture to harness the power of attention mechanisms to detect knocking calls, as outlined by Moritz et al. (2020). To improve both the f_0 estimation and the knocking detection, commonly used noise reduction techniques could be implemented, including spectral subtraction and frequency filtering (Xie et al., 2021). In a broader scale, other potential experiments could apply these architectures to additional species of seal that produce similar types of calls, such as other ringed seal populations (Kunnasranta et al., 1996, Mizuguchi et al., 2016) or other pinniped species like the Weddell seal (*Leptonychotes weddelli)* (Thomas et al., 1982), the gray seal (*Halichoerus grypus*) (Asselin et al., 1993), or the Pacific walrus (*Odobenus rosmarus divergens*) (Schusterman et al., 2008).

5 | CONCLUSIONS

To help monitor Saimaa ringed seals, this study aims to provide a system capable of automatically analyzing and detecting instances of knocking vocalizations amongst noise segments from underwater sound recordings. The proposed solution uses the time-frequency representations of audio segments to detect possible knocking vocalizations by classifying such frames in a deep learning framework. This proved to be the best alternative among using time sequences and features derived from fundamental frequency estimation, achieving F1-scores as high as 97.76% when utilizing spectrograms. To automatically characterize the events, the fundamental frequency was estimated with an assembly of time-domain techniques and compared with manual analysis from existing literature. The means of

the estimated frequencies were quantitatively compared to those derived from the manual analyses of a subset of the data, and they presented a mean absolute error of 1.98 Hz. In a real-world setting, the joined classifier and analyzer could be easily expanded to provide the onsets, offsets, and fundamental frequencies of the detected knocking calls in a continuous audio stream.

ACKNOWLEDGMENTS

The authors wish to thank Metsähallitus and Miina Auttila, Marja Niemi, and Riikka Alakoski, who contributed to the deployment and collection of the hydrophones, as well as our collaborators at Turku University of Applied Sciences. MY is funded by the Finnish Cultural Foundation [55241449]. This work is also supported by WWF Finland [2421/02.07.02/2010] and Tampereen Särkänniemi Ltd. [351/02.01/2022]. This study is part of SealHabitat cross-border project that is funded by the EU-programme Interreg Aurora.

REFERENCES

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., ... & Oleson, E. M. (2021). A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Frontiers in Marine Science*, 8, 607321.

Asselin, S., Hammill, M. O., & Barrette, C. (1993). Underwater vocalizations of ice breeding grey seals. *Canadian Journal of Zoology*, 71(11), 2211-2219.

Åsvestad, L., Ahonen, H., Menze, S., Lowther, A., Lindstrøm, U., & Krafft, B. A. (2024). Seasonal acoustic presence of marine mammals at the South Orkney Islands, Scotia Sea. *Royal Society Open Science*, 11(1), 230233.

Bolgan, M., Parmentier, E., Picciulin, M., Hadjioannou, L., & Di Iorio, L. (2023). Use of passive acoustic monitoring to fill knowledge gaps of fish global conservation status. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 33(12), 1580–1589.

Caruso, F., Dong, L., Lin, M., Liu, M., Gong, Z., Xu, W., ... & Li, S. (2020). Monitoring of a nearshore small dolphin species using passive acoustic platforms and supervised machine learning techniques. *Frontiers in Marine Science*, 7, 267.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366.

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917–1930.

de Vincenzi, G., Parisi, I., Torri, M., Papale, E., Mazzola, S., Nuth, C., & Buscaino, G. (2019). Influence of environmental parameters on the use and spatiotemporal distribution of the vocalizations of bearded seals (*Erignathus barbatus*) in Kongsfjorden, Spitsbergen. *Polar Biology*, 42(7), 1241–1254.

Escobar-Amado, C., Badiey, M., & Pecknold, S. (2022). Automatic detection and classification of bearded seal vocalizations in the northeastern Chukchi Sea using convolutional neural networks. *The Journal of the Acoustical Society of America*, 151(1), 299–309.

Escobar-Amado, C., Badiey, M., & Wan, L. (2023). Computer vision for bioacoustics: Detection of bearded seal vocalizations in the Chukchi shelf using YOLOV5. *IEEE Journal of Oceanic Engineering*, 49(1), 133-144.

Fleishman, E., Cholewiak, D., Gillespie, D., Helble, T., Klinck, H., Nosal, E. M., & Roch, M. A. (2023). Ecological inferences about marine mammals from passive acoustic data. *Biological Reviews*, 98(5), 1633–1647.

Gerhard, D. (2003). *Pitch extraction and fundamental frequency: History and current techniques* (pp. 0–22). Regina, SK, Canada: Department of Computer Science, University of Regina.

Gibb, R., Browning, E., Glover‐Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2), 169–185.

Guyot, P. (2018). Fast Python Implementation of the Yin Algorithm. Version 1.1.1, Zenodo. https://doi.org/10.5281/zenodo.1220947

Hamard, Q., Pham, M. T., Cazau, D., & Heerah, K. (2024). A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data. Ecological Informatics, 84, 102906.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4), 18–28.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.

Jolliffe, I. (2005). Principal component analysis. In *International Encyclopedia of Statistical Science* (pp. 1094–1096). Springer, Berlin, Heidelberg.

Kadambe, S., & Boudreaux-Bartels, G. F. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 38(2), 917–924.

Kaiser, J. F. (1990, April). On a simple algorithm to calculate the'energy'of a signal. In *International conference on acoustics, speech, and signal processing* (pp. 381-384). IEEE.

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018, April). CREPE: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 161–165). IEEE.

Kirsebom, O. S., Frazao, F., Simard, Y., Roy, N., Matwin, S., & Giard, S. (2020). Performance of a deep neural network at detecting North Atlantic right whale upcalls. *The Journal of the Acoustical Society of America*, 147(4), 2636-2646.

Koivuniemi, M., Auttila, M., Niemi, M., Levänen, R., & Kunnasranta, M. (2016). Photo-ID as a tool for studying and monitoring the endangered Saimaa ringed seal. *Endangered Species Research*, 30, 29–36.

Kunnasranta, M., Hyvärinen, H., & Sorjonen, J. (1996). Underwater vocalizations of Ladoga ringed seals (*Phoca hispida ladogensis Nordq.*) in summertime. *Marine Mammal Science*, 12(4), 611–618.

Kunnasranta, M., Niemi, M., Auttila, M., Valtonen, M., Kammonen, J., & Nyman, T. (2021). Sealed in a lake—Biology and conservation of the endangered Saimaa ringed seal: A review. *Biological Conservation*, 253, 108908.

LeCun, Y., & Bengio, Y. (1998). Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*.

Licciardi, A., & Carbone, D. (2024). Whalenet: A novel deep learning architecture for marine mammals vocalizations on Watkins marine mammal sound database. *IEEE Access*.

Löytynoja, A., Pohjoismäki, J., Valtonen, M., Laakkonen, J., Morita, W., Kunnasranta, M., Väinölä, R., Olsen, M.T., Auvinen, P., & Jernvall, J. (2025). Deep origins, distinct adaptations, and species-level status indicated for a glacial relict seal. *Proceedings of the National Academy of Sciences*, 122(25), e2503368122.

Marcoux, M., Ferguson, S. H., Roy, N., Bedard, J. M., & Simard, Y. (2017). Seasonal marine mammal occurrence detected from passive acoustic monitoring in Scott Inlet, Nunavut, Canada. *Polar Biology*, 40(5), 1127–1138.

Metsähallitus. (2025). *Saimaannorppakannan seuranta* (Monitoring of the Saimaa ringed seal population). https://www.metsa.fi/luonto-ja-kulttuuriperinto/lajien-suojelu/saimaannorppa/norppakannan-seuranta/

Miller, B. S., Madhusudhana, S., Aulich, M. G., & Kelly, N. (2023). Deep learning algorithm outperforms experienced human observer at detection of blue whale D‐calls: a double‐observer analysis. *Remote Sensing in Ecology and Conservation*, 9(1), 104-116.

Mizuguchi, D., Tsunokawa, M., Kawamoto, M., & Kohshima, S. (2016). Underwater vocalizations and associated behavior in captive ringed seals (Pusa hispida). *Polar Biology*, 39(4), 659–669. https://doi.org/10.1007/s00300-015-1821-x

Moritz, N., Wichern, G., Hori, T., & Le Roux, J. (2020). All-in-one transformer: Unifying speech recognition, audio tagging, and event detection. *Interspeech* (pp. 3112-3116).

Muñoz-Duque, S., Vieira, M., Fonseca, P.J., Quintella, B., Charrier, I., Monteiro, J.G. et al. (2024). First assessment of passive acoustics as a tool to monitor the endangered Mediterranean monk seal in the Madeira Archipelago (Portugal). *Aquatic Conservation: Marine and Freshwater Ecosystems*, 34(2), e4100. https://doi.org/10.1002/aqc.4100

Parsons, S. K., Erbe, C., Marley, S. A., & Parsons, M. J. (2025). Pinniped Sounds. *In Marine Mammal Acoustics in a Noisy Ocean* (pp. 351-440). Cham: Springer Nature Switzerland.

Prawirasasra, M. S., Mustonen, M., & Klauson, A. (2021). The underwater soundscape at Gulf of Riga marine-protected areas. *Journal of Marine Science and Engineering*, 9(8), 915.

Rautio, A., Niemi, M., Kunnasranta, M., Holopainen, I. J., & Hyvärinen, H. (2009). Vocal repertoire of the Saimaa ringed seal (Phoca hispida saimensis) during the breeding season. *Marine Mammal Science*, 25(4), 920–930. https://doi.org/10.1111/j.1748-7692.2009.00299.x

Roca, I. T., Kaleschke, L., & Van Opzeeland, I. (2023). Sea‐ice anomalies affect the acoustic presence of Antarctic pinnipeds in breeding areas. *Frontiers in Ecology and the Environment*, 21(5), 227–233.

Schusterman, R. J., & Reichmuth, C. (2008). Novel sound production through contingency learning in the Pacific walrus (Odobenus rosmarus divergens). *Animal Cognition, 11*(2), 319-327.

Spell, B. (2022). torch-yin: A PyTorch implementation of the YIN pitch tracking algorithm. GitHub. Retrieved July 30, 2025, from https://github.com/brentspell/torch-yin

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr, J. W., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: Review and perspectives. *BioScience*, 69(1), 15–25.

Tan, L., & Karnjanadecha, M. (2003, September). Pitch detection algorithm: autocorrelation method and AMDF. In *Proceedings of the 3rd international symposium on communications and information technology* (Vol. 2, pp. 551-556).

Teixeira, D., Roe, P., van Rensburg, B. J., Linke, S., McDonald, P. G., Tucker, D., & Fuller, S. (2024). Effective ecological monitoring using passive acoustic sensors: Recommendations for conservation practitioners. *Conservation Science and Practice*, 6(6), e13132.

Thomas, J. A., & Kuechle, V. B. (1982). Quantitative analysis of Weddell seal (L eptonychotes weddelli) underwater vocalizations at McMurdo Sound, Antarctica. *The Journal of the Acoustical Society of America*, 72(6), 1730-1738.

Tyshko, A. A., Krinitskiy, M. A., Shatravin, A. V., & Belikov, R. A. (2023). Automatic Detection of Acoustic Signals of Beluga Whales and Bottlenose Dolphins. *Moscow University Physics Bulletin,* 78(Suppl 1), S217-S225.

Valtonen, M., Heino, M., Aspi, J., Buuri, H., Kokkonen, T., Kunnasranta, M., ... & Nyman, T. (2015, April). Genetic monitoring of a critically-endangered seal population based on field-collected

placentas. In *Annales Zoologici Fennici* (Vol. 52, No. 1–2, pp. 51–65). Finnish Zoological and Botanical Publishing Board.

White, E. L., White, P. R., Bull, J. M., Risch, D., Beck, S., & Edwards, E. W. (2022). More than a whistle: Automated detection of marine sound sources with a convolutional neural network. *Frontiers in Marine Science*, 9, 879145.

Willmer, J., Watson, M. S., Eriksson, B. K., & Van Opzeeland, I. (2025). Underwater vocalization behavior of harbor seals (Phoca vitulina). Marine Mammal Science, e70056.

Xie, J., Colonna, J. G., & Zhang, J. (2021). Bioacoustic signal denoising: a review. *Artificial Intelligence Review, 54*(5), 3575-3597.

Young, M. , Nykänen, M. , Niemi, M. , Kurkilahti, M. , & Kunnasranta, M. (2025). Aquatic vocalization activity indicates the timing of the mating season of Saimaa ringed seals. *Marine Mammal Science*, 41(3), e13225. https://doi.org/10.1111/mms.13225.

Ziegenhorn, M. A., Frasier, K. E., Hildebrand, J. A., Oleson, E. M., Baird, R. W., Wiggins, S. M., & Baumann-Pickering, S. (2022). Discriminating and classifying odontocete echolocation clicks in the Hawaiian Islands using machine learning methods. *PLOS ONE*, 17(4), e0266424.

APPENDIX

Table A1 Features engineered for classification.

Feature	Description
Power (dB)	Overall energy of the signal in decibels.
Range dimension	Number of discrete values that the amplitude of a signal takes.
ZCR	Zero-Crossing Rate: the rate at which the signal waveform crosses zero.
Autocorrelation f_0 (Hz)	Fundamental frequency estimated with the autocorrelation method.
Autocorrelation rel. amp.	Amplitude of the maximum value of the autocorrelation curve relative to the energy of the signal (relative amplitude).
AMDF f_0 (Hz)	Fundamental frequency estimated with the AMDF method.
AMDF. rel. amp.	Amplitude of the minimum value of the AMDF curve relative to the maximum of the curve (relative amplitude).
YIN custom f_0 (Hz)	Fundamental frequency estimated with a custom implementation of YIN based on (Guyot, 2018).
YIN custom harmonic rate	Confidence score of the fundamental frequency estimated with a custom implementation of YIN based on (Guyot, 2018).
YIN torch f_0 (Hz)	Fundamental frequency estimated with a YIN implementation by (Spell, 2022). This implementation was more accurate than the manual one in some cases.

TK Autocorrelation f_0 (Hz)	Fundamental frequency of a Teager-Kaiser prefiltered signal estimated with the autocorrelation method
TK Autocor. rel. amp.	Relative amplitude of a Teager-Kaiser prefiltered signal estimated with the autocorrelation method
TK AMDF f_0 (Hz)	Fundamental frequency of a Teager-Kaiser prefiltered signal estimated with the AMDF method
TK AMDF rel. amp.	Relative amplitude of a Teager-Kaiser prefiltered signal estimated with the AMDF method
TK YIN man. f_0 (Hz)	Fundamental frequency estimated with a custom implementation of YIN based on (Guyot, 2018) from a Teager-Kaiser pre-filtered signal.
TK YIN man. harmonic rate	Confidence score of the fundamental frequency estimated with a custom implementation of YIN based on (Guyot, 2018) from a Teager-Kaiser pre-filtered signal.
CREPE Confidence	Neural-network-based pitch estimation confidence score from the CREPE model (Kim et al., 2018).
CREPE avg. activation	Average activation across CREPE's output neurons (Kim et al., 2018), quantitatively indicating to what extent one pitch was favored over others.