Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics & Computer Science Programmes

Marine mammal sounds classification

Signal Processing final project report

Author:
Khrystyna Mysak

15 December 2024

Contents

Abstract

1 Introduction

1.1 Background

Marine mammals rely on vocalizations for navigation, prey detection, predator avoidance, and communication. Due to the low attenuation of sound in water, these vocalizations can be detected over vast distances, making them a valuable resource for studying marine mammals. Passive acoustic monitoring (PAM) plays a critical role in tracking and assessing populations of marine mammals, particularly in threatened environments.

The large volume of audio data generated by passive acoustic arrays presents a significant challenge. Automating species identification in these recordings is crucial for effective monitoring of marine mammal populations.

1.2 Research Objectives

The primary objective of this project is to develop a deep learning classification model for marine mammal sounds. This model will utilize features such as MFCCs, spectrograms, chromagrams, and Teager-Kaiser energy to accurately identify different species. With the help of these features, the model aims to automate the species identification process and enhance the efficiency of passive acoustic monitoring systems.

1.3 Literature Review

Bioacoustics is a well-established field that combines biology, zoology, physics, and signal processing. There are already several studies that introduced methods and models to classify marine mammals vocalizations.

1.3.1 WhaleNet[1]

This study, conducted by Alessandro Licciardi and Davide Carbone, focuses on data preparation and preprocessing techniques before classifying vocalizations. The authors then examine the application of Wavelet Scattering Transform (WST) and Mel spectrogram as preprocessing tools for feature extraction. This paper introduces WhaleNet (Wavelet Highly Adaptive Learning Ensemble Network), an advanced deep ensemble architecture to classify vocalizations of marine mammals. Using both WST and Mel spectrograms, WhaleNet enhances feature discrimination. As a result, this approach establishes a robust framework for classifying marine mammal vocalizations. Incorporating insights from the WST and Mel spectrograms, the authors achieved an outstanding 97.61% in classification accuracy.

1.3.2 Residual Learning[2]

This research by Daniel Murphy, Elias Ioup, Tamjidul Hoque, and Mahdi Abdelguerfi presents a method for classifying marine mammal vocalizations using residual learning networks (ResNets). The researchers optimized spectrogram generation from acoustic

recordings, testing various configurations of window functions, pre-processing, and multi-channel inputs. The best configuration, using 512×256 spectrograms with Hann window of 1024 and horizontal roll, trained both single- and multichannel networks. The single channel network achieved the highest performance, with an F1 score of 86.7% and an AUC of 92.81% for classifying 32 species.

1.3.3 Detection of Dolphins and Porpoises[3]

The aim of the study conducted by Quentin Hamard, Minh-Tan Pham, Dorian Cazau, and Karine Heerah was to evaluate the ability of a deep learning model to detect, localize, and classify marine mammal sounds from underwater recordings. Using a broadband hydrophone deployed at the Fécamp OWF (Normandy, France), 15-second spectrograms were generated, with dolphin and porpoise sounds manually annotated by species and sound type. The spectrograms were split for five-fold cross-validation, and a R-CNN model was trained to detect and classify the sounds. Three configurations were tested: overall detection of marine mammals, species classification (dolphin vs. porpoise), and sound type classification into five categories. The model achieved detection in 15.4% of the spectrograms with minimal false negatives for the simplest configuration, and high precision for species and sound type classification, with mean Average Precision (mAP) scores of 92.3% and 84.3%, and AUC values of 95.7% and 94.9%, respectively.

1.3.4 Detection of Whales species[4]

This work focused on classifying two species of whales, Blue Whales and Fin Whales, using deep learning models to monitor their populations. The author compared a baseline CNN model (LeNet) with the UPC's Double Multi-Head Attention (DMHA) model to investigate the impact of attention mechanisms on classification performance. Audio files were converted into Mel spectrograms, which were used as input for the models. Both LeNet and DMHA models were trained to classify whale sounds. The DMHA model, which incorporated attention mechanisms, outperformed the baseline model, achieving 91% accuracy compared to LeNet's 88%.

2 Methodology

2.1 Dataset – Watkins Marine Mammal Sound Database [5]

The William A. Watkins Collection of Marine Mammal Sound Recordings features an extensive archive of sounds from various marine mammal species, collected over seven decades in various geographic regions. The recordings were made by Watkins, Schevill, and other contributors, including G.C. Ray, D. Wartzok, D. and M. Caldwell, K. Norris, and T. Poulter.

This collection holds significant historical and scientific value, providing a unique resource for studying long-term changes in vocal behavior linked to ambient noise and serving as a vital reference for 55 marine mammal species. However, for this project, only the 'best-of' recordings, representing 32 species, were used, as the rest lacked sufficient quality and quantity for classification purposes.

2.2 Exploratory Data Analysis (EDA)

Upon conducting some exploratory data analysis, key insights into the dataset's sampling rates, durations, and file counts were revealed. First of all, sampling rates range from 320 Hz to 192000 Hz with an average of 63056 Hz. The duration of the audio signals also has a wide range from 0.02 to 1455.93 seconds, but overall stays the same with an average of 7.04 seconds.

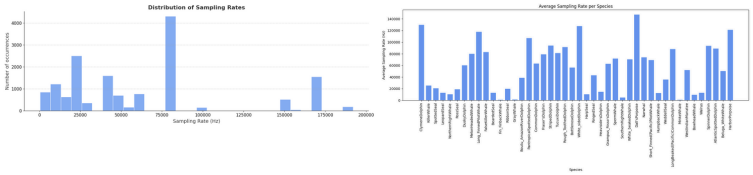

Figure 1: Distribution of the sampling rate

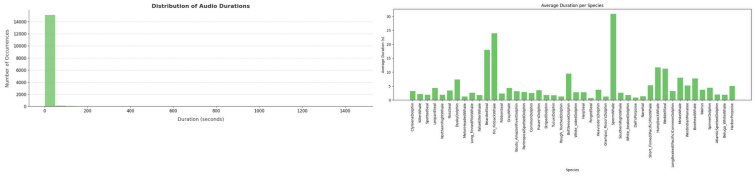

Figure 2: Distribution of the durations

Regarding the number of files for each species, the data set exhibits a significant imbalance in the number of files per species, and many species have a noticeably lower number of audio samples compared to others.

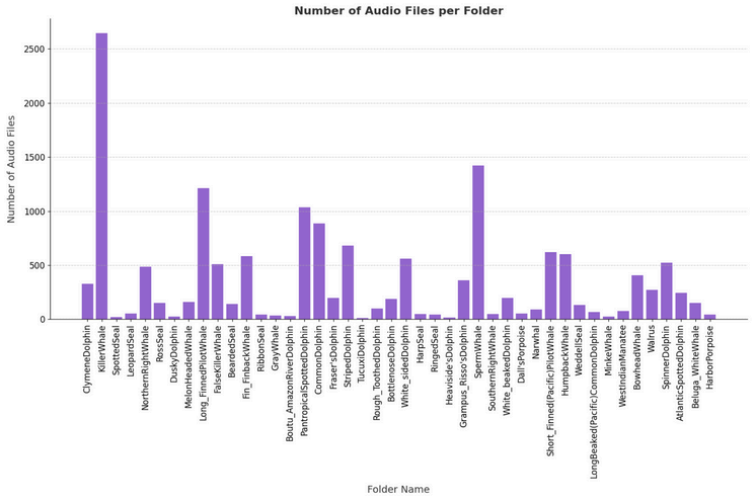

Figure 3: Number of files for each species

The spectrogram and amplitude spectrum plots for each species provided valuable insights into the frequencies of sounds produced by marine mammals. For instance, dolphins typically produce high-frequency clicks and whistles, while whales and seals generate low to mid-frequency sounds. Additionally, each species exhibits a distinct spectrogram pattern, as shown in the image below.

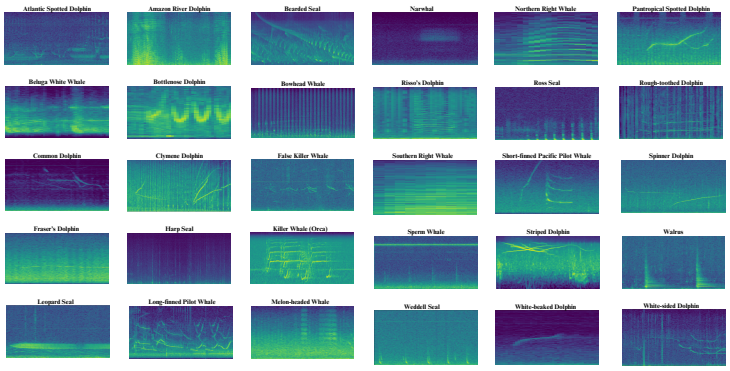

Figure 4: Spectrogram plots for various species

2.3 Chosen approach and justification

After thoroughly reviewing and analyzing existing approaches, the decision was made to use a Convolutional Neural Network (CNN) for sound classification, a common method across the studies, although with different architectures. However, to simplify the model, a less complex CNN was chosen. This decision aimed to prioritize the extraction of features from the audio signals and to leverage signal processing techniques.

The initial plan was to classify vocalizations using only spectrograms. However, this approach would have been computationally expensive and restricted the number of experiments that could be performed on the audio signals. Therefore, an alternative approach was adopted: creating a dataset with various extracted features from the audio signals and classifying the vocalizations based on those features.

2.4 Pipeline of the implementation

As a result, the following implementation pipeline can be outlined to summarize the project's methods and steps.

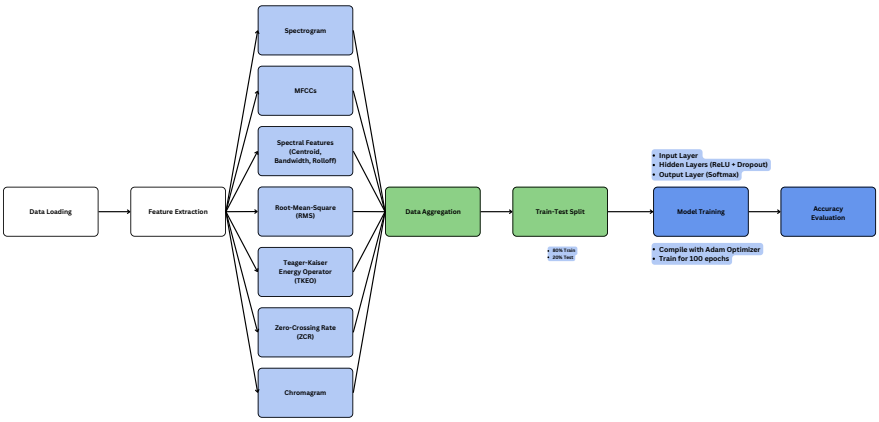

Figure 5: The implementation pipeline

2.4.1 Discrete-time Short-Time Fourier Transform (STFT) and Spectrogram

The Discrete-time Short-Time Fourier Transform (STFT) provides a localized time-frequency analysis of non-stationary signals. Unlike the standard Fourier Transform,

which assumes stationarity, the STFT uses a window function $w[n]$ to divide the signal into overlapping segments, sliding across the signal with an offset k. This approach captures the evolution of frequency content over time, making it ideal for analyzing dynamic signals such as marine mammal vocalizations.

$$c[m,k] = \sum_{n=0}^{N-1} x[n]w[n-k]e^{-\frac{2\pi j}{N}mn}$$

where:

- $w[n]$ is the window function,
- k is the parameter representing the offset of the window function in time.

Hann window function is one of the most commonly used to reduce spectral leakage. Moreover, it has been shown to deliver optimal performance, as demonstrated in [2].

$$w[n] = \begin{cases} \cos^2\left(\pi\frac{n}{N}\right) = \frac{1}{2}\left(1-\cos\left(2\pi\frac{n}{N}\right)\right), & -\frac{N}{2} \leq n \leq \frac{N}{2} \\ 0, & \text{in all other cases} \end{cases}$$

A spectrogram is a representation of the power distribution of spectral components, which reflects the change in the spectrum over time. It is especially useful for analyzing non-stationary signals like animal vocalizations.

$$C_w[m,k] = \left|\sum_{n=0}^{N-1} x[n]w[n-k]e^{-\frac{2\pi j}{N}mn}\right|^2 = |c[m,k]|^2$$

2.4.2 Mel-frequency cepstral coefficients (MFCCs) [6]

Mel-frequency Cepstral Coefficients (MFCCs) are a feature extraction technique used in sound processing. The underlying idea behind MFCCs is that the human auditory system does not perceive all frequencies with the same sensitivity—especially at higher frequencies. The Mel scale is designed to model this characteristic, where the ear is more sensitive to changes in frequency at lower frequencies and less sensitive at higher frequencies. The Mel-frequency cepstrum (MFC) represents the short-term power spectrum of a sound, and the MFCCs are the coefficients that make up this representation, capturing the most relevant features of the sound's spectral content.

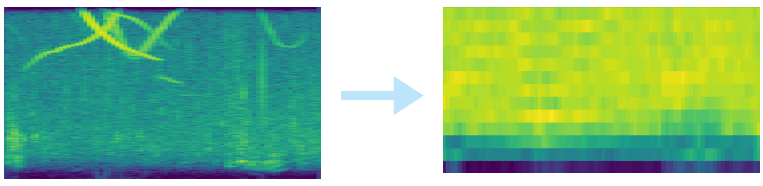

Figure 6: MFCCs extracted from Striped Dolphin vocalizations

The MFCC feature extraction process can be divided into the following steps:

- Framing and Windowing

$$y[n] = x[n] \cdot w[n]$$

- Calculation of the Discrete Fourier Transform on each frame

$$c_w[m] = \sum_{n=0}^{N-1} y[n] e^{-\frac{2\pi j}{N} mn}$$

- Computing the power spectrum

$$P[m] = \frac{|c_w[m]|^2}{N}$$

- Apply Mel Filterbanks

$$S[m] = \ln \left[\sum_{k=0}^{N-1} P[m] H_m[k] \right], \quad 0 \leq m < M$$

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

- Apply Discrete Cosine Transform (DCT)

$$\text{MFCC}(n) = \sum_{m=0}^{M-1} \log |S(m)| \cdot \cos \left[\frac{\pi n (m + 1/2)}{M} \right], \quad 0 \leq n < M$$

2.4.3 Chromagram [7]

Chroma features reduce sensitivity to the tonal qualities of instruments or voices (timbre) by grouping pitches across octaves, emphasizing harmonic and pitch-related content instead. Each pitch has two components: tone height (octave) and chroma (pitch class). The 12 chroma values (0–11) represent pitch classes cyclically. A chromagram aggregates spectral energy for each chroma by summing coefficients of all pitches sharing the same chroma.

$$C(n, c) := \sum_{\{p \,:\, p \bmod 12 = c\}} Y_{\text{LF}}(n, p)$$

where $Y_{\text{LF}}(n, p)$ — Log-frequency spectrogram coefficient for frame n and pitch p.

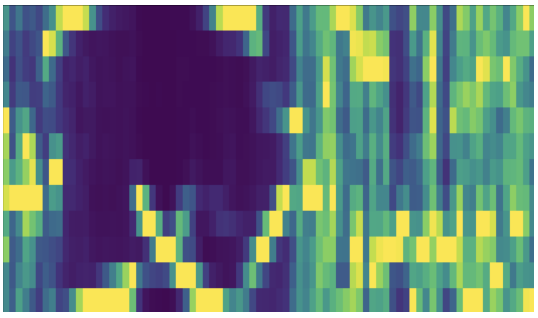

Figure 7: Chromagram for the Striped Dolphin vocalizations

2.4.4 Root Mean Square (RMS) [8]

RMS is the square root value of the mean of the sum of squares of signal. It is generally used to evaluate the signal amplitude and signal energy in time domain.

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x[n]^2}$$

2.4.5 Spectral Centroid, Bandwidth and Rolloff [8]

- Spectral rolloff is the frequency below which a specified percentage of the total spectral energy lies.
- Spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located.

  $$f_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)}$$

  where:

  - $S(k)$ is the spectral magnitude at frequency bin k,
  - $f(k)$ is the frequency at bin k.
- Spectral bandwidth quantifies the spread of frequencies around the spectral centroid. It gives an indication of how "wide" or "narrow" a spectrum is.

  $$\left(\sum_k S(k) \left(f(k) - f_c\right)^p\right)^{\frac{1}{p}}$$

  where:

  - $S(k)$ is the spectral magnitude at frequency bin k,
  - $f(k)$ is the frequency at bin k,
  - f_c is the spectral centroid.

2.4.6 Zero-crossing rate [8]

Zero-crossing rate – is the rate at which the signal changes sign, that is, the number of times the signal crosses the zero amplitude threshold. It is a simple, yet effective feature for characterizing the noisiness or "percussiveness" of a sound.

$$\text{ZCR} = \frac{1}{N-1}\sum_{n=1}^{N-1} \mathbf{1}\left(x[n] \cdot x[n-1] < 0\right)$$

where:

- $x[n]$ represents the signal at time step n,
- $\mathbf{1}(\cdot)$ is the indicator function, which is 1 if the condition inside the parentheses is true (i.e., if the signal crosses zero between $x[n-1]$ and $x[n]$), and 0 otherwise,
- N is the total number of samples in the signal.

2.4.7 Teager-Kaiser Energy Operator [9]

The Teager–Kaiser energy operator (TKEO) is a nonlinear operator that tracks the energy of a data stream. An important property of the TK energy operator in is that it is nearly instantaneous given that only three samples are required in the energy computation at each time instant: $x(n-1)$, $x(n)$, and $x(n+1)$.

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

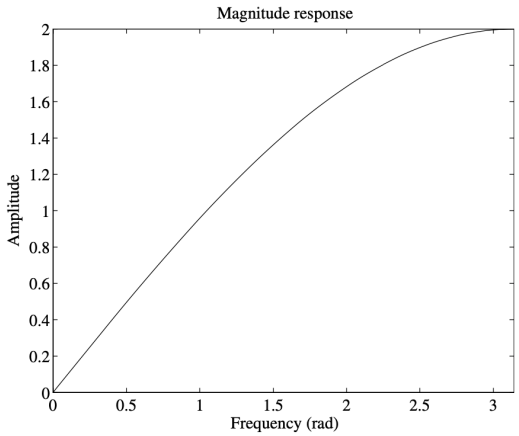

Figure 8: Magnitude of the frequency response of the TK-energy operator filter [10]

The Teager-Kaiser Energy Operator (TKEO) is noted for its success in detecting high-frequency sounds, such as clicks from bats or Sperm Whales[10], or odontocete biosonar. While Mel-frequency cepstral coefficients (MFCCs) capture features where the ear is more sensitive to frequency changes at lower frequencies, the TKEO feature was used to complement this by emphasizing higher frequencies. As illustrated in the plot, TKEO acts as a high-pass filter, resulting in an output that mostly contains the high-frequency components of the input signal.

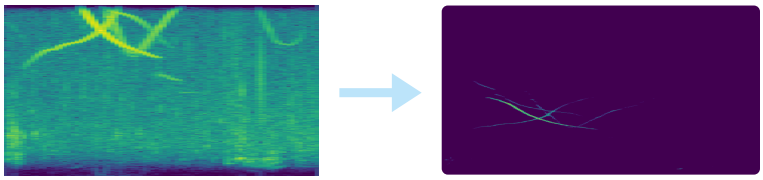

Figure 9: TKEO applied on Striped Dolphin vocalizations

2.5 Model

The model is a simple Convolutional Neural Network model. For the CNN model, all hidden layers use a ReLU activation function, the output layer a Softmax function and a Dropout is used to avoid overfitting. Adam optimizer is used to train the model over 100 epochs. This choice was made because it allows us to obtain better results. The loss is calculated with the sparse-categorical-crossentropy function.

2.6 Evaluation

The accuracy of the model was measured using the following metrics:

- Accuracy measures the overall correctness of the model.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. Precision shows how many of the positively classified instances were actually positive, while recall measures how many of the actual positives were correctly identified by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  where:

  - TP – True Positive: Correctly classified as positive.
  - FP – False Positive: Incorrectly classified as positive.
  - FN – False Negative: Incorrectly classified as negative.

- F1-Score is the harmonic mean of precision and recall, providing a balance between the two.

$$\text{F-1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

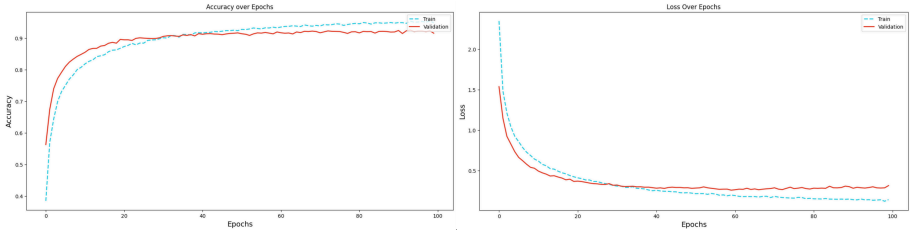

Figure 10: Accuracy and Loss over the training epochs

Accuracy	**Precision**	**Recall**	**F1-Score**
0.9297	0.9302	0.9297	0.9283

Table 1: Evaluation Metrics

3 Results

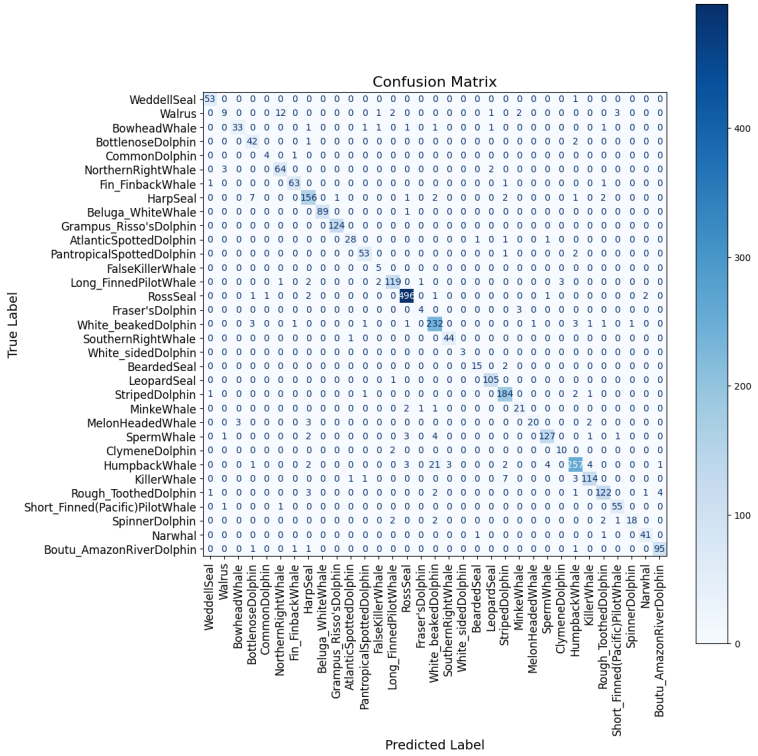

Figure 11: Confusion matrix of the predictions

After thoroughly analyzing the evaluation metrics, we observe that all metrics are approximately 93%, indicating that the model predicts with a high degree of accuracy.

With only minor misclassifications (1-2 errors) for most labels, species with fewer samples in the dataset showed a higher rate of misclassification. For instance, the Humpback Whale and Walrus were misclassified more frequently than other species, likely due to the limited data available for these categories.

All things considered, the model serves as a strong baseline for predictions and offers potential for improvement through feature engineering, such as incorporating the location of the recorded sound etc.

4 Discussion

The resulting model achieved a higher accuracy score than most of the models provided in the literature review. One model, the WhaleNet model, eventually had a greater score of 97.61%, which implies that further usage of Wavelet Scattering Transform might be benefitial in the current approach.

Despite the encouraging results, the study is constrained by limitations in the Watkins Marine Mammal Sound (WMMS) Database. The dataset suffers from a severe class imbalance, with some species, such as Killer Whales, represented by over 2,500 samples, while others have fewer than 100. This imbalance poses challenges for model training, as models may become biased toward species with larger sample sizes, reducing their generalizability across underrepresented species. Addressing this issue will require either the

collection of additional vocalization data for underrepresented species or the implementation of advanced data augmentation techniques tailored to bioacoustics.

Another limitation is the dataset's environmental variability. Acoustic recordings are influenced by factors such as background noise, recording equipment, and oceanographic conditions, which may impact the model's performance in real-world applications. This variability also makes it challenging to filter noise from the signals, as the diverse characteristics of the sounds prevent the creation of a generalized filter that can effectively handle all recordings.

References

[1] *WhaleNet: A Novel Deep Learning Architecture for Marine Mammals Vocalizations on Watkins Marine Mammal Sound Database*, https://ieeexplore.ieee.org/abstract/document/10720021

[2] *Residual Learning for Marine Mammal Classification*, https://ieeexplore.ieee.org/abstract/document/9943551

[3] *A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data*, https://www.sciencedirect.com/science/article/pii/S1574954124004485

[4] *Acoustic classification of marine mammals by deep learning*, https://upcommons.upc.edu/handle/2117/404731

[5] *Watkins Marine Mammal Sound Database*, https://whoicf2.whoi.edu/science/B/whalesounds/index.cfm

[6] *Mel Frequency Cepstral Coefficient (MFCC)*, http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[7] *Fundamentals of Music Processing*, https://books.google.com.ua/books?id=HCI_CgAAQBAJ&printsec=copyright&redir_esc=y#v=onepage&q&f=false

[8] *Music Information Retrieval* https://musicinformationretrieval.com/index.html

[9] *Exploring Animal Behavior Through Sound: Volume 1*, https://library.oapen.org/handle/20.500.12657/60833

[10] *Detection of sperm whale clicks based on the Teager–Kaiser energy operator*, https://www.researchgate.net/publication/222823042_Detection_of_sperm_whale_clicks_based_on_the_Teager-Kaiser_energy_operator