TASK Quarterly 26 (4) 2022

Application of visual classification algorithms for identification of underwater audio signals

Piotr Gnyś
Department of Computer Science ,

Polish-Japanese Academy of Information Technology
Koszykowa 86 Street, 02-008 Warsaw, Poland

Date: 08 January 2024

Gabriela Szczęsna
Department of Computer Science ,

Polish-Japanese Academy of Information Technology
Koszykowa 86 Street, 02-008 Warsaw, Poland

Date: 08 January 2024

Antonio C. Domínguez-Brito
Instituto Universitario SIANI and Departamento de Informática y Sistemas
Universidad de las Palmas de Gran Canaria, 35017 Spain

Date: 08 January 2024

Jorge Cabrera-Gámez
Instituto Universitario SIANI and Departamento de Informática y Sistemas
Universidad de las Palmas de Gran Canaria, 35017 Spain

Date: 08 January 2024

https://doi.org/10.34808/wr60-jz83

Abstract

An audio processing and classification pipeline is presented in this work. The main focus is on the classification of sounds in a marine acoustic environment, however, the presented approach can be applied to other audio data. Audio samples from heterogeneous sources automatically spliced, normalized and transformed into spectrogram based visual representation are tagged on the pipeline input. The said representation is then used to train a convolutional neural network that can identify the presented categories in future recordings.

Keywords:

audio processing, audio classification, convolutional neural network

1. Marine acoustic environment

The concept of a soundscape, initially proposed by Southworth [1] and subsequently popularized by Schafer [2], pertains to an auditory milieu or composite of sounds that emerge from an immersive environment. The notion of a soundscape encompasses not only the inherent acoustic characteristics of the natural surroundings, comprising animal vocalizations constituting an expression of a collective habitat, aptly referred to as *biophony*, but also includes the sounds emitted by weather phenomena and other natural elements, termed as *geophony*. Moreover, the environmental sounds generated by human activities find their place within the framework of *anthropophony*. While individuals may commonly associate the marine domain with serenity and silence, the veracity of this perception diverges significantly from the reality experienced by the numerous organisms inhabiting these aquatic realms. Human undertakings such as transportation, dredging, and drilling operations, as well as oceanographic investigations, invariably engender sounds and vibrations that impinge upon the behavioral patterns of marine life [3, 4], and in some instances, drive species to extinction. A conscientious effort within this research endeavor was directed towards amassing a highly versatile dataset, underpinning the aspiration for a classification model of optimal generality. Fig. 1 provides a diagram, presenting a comprehensive delineation of the diverse categories within the dataset. The primary dichotomy discerns between natural and anthropogenic sound sources, subsequently branching them into subdivisions that delineate between various natural sources, with a particular focus on the marine fauna. Further distinctions are made within the marine animal domain, encompassing pinnipeds, invertebrates, and cetaceans. However, it is crucial to acknowledge that the present work does not encompass the broad spectrum of piscine organisms, despite their capacity for generating an extensive repertoire of sounds.

1.1. Marine Mammals

Marine mammals encompass a diverse assemblage of mammals reliant upon the oceans or other marine ecosystems for survival. This group includes notable species such as seals, whales, manatees, sea otters, and polar bears. It is worth noting that marine mammals do not constitute a monophyletic group derived from a common ancestor. Rather, their shared characteristics result from a convergent evolution [5], as depicted in Fig. 2.

The extent of adaptation to the marine environment exhibits considerable variation across different species.

Cetaceans and sirenians are fully aquatic, whereas seals and sea lions are semi-aquatic creatures that spend a significant portion of their time in the water while still relying on the land for vital activities such as mating, breeding, and molting. Otters and polar bears, in contrast, exhibit lesser adaptations to an aquatic lifestyle. Despite their relatively lower numbers compared to their terrestrial counterparts, marine mammals hold a crucial ecological role in preserving marine ecosystems. However, it is disconcerting that 36% of marine mammal species face threats and are considered endangered [7]. Habitat degradation poses a significant risk to these creatures, hindering their ability to find sustenance. Furthermore, noise pollution detrimentally affects mammals reliant on echolocation, while the impacts of climate change pose significant challenges to Arctic ecosystems. More than one-third of marine mammal species face endangerment or extinction.

Baleen whales Mysticeti, colloquially known as baleen whales, comprise a parvorder within the infraorder Cetacea. They represent a widely distributed and diverse group of carnivorous marine mammals distinguished by the presence of baleen plates instead of teeth. The Mysticeti parvorder encompasses the families Balaenidae (right whales), Balaenopteridae (fin whales), Cetotheriidae (dwarf right whale), and Eschrichtiidae (gray whale). At present, fifteen species of mysticetes are recognized. Mysticetes emit a varied repertoire of vocalizations, often referred to as "songs", with humpback whales (Megaptera novaeangliae) particularly renowned for their melodic compositions.

Toothed Whales Odontoceti, commonly referred to as toothed whales, constitute a parvorder within the cetacean order, encompassing dolphins, porpoises, and other whales that possess teeth, including beaked and sperm whales. Toothed whales represent some of the most widely distributed mammals, with many species, particularly dolphins, exhibiting highly social behavior and forming pods comprising over a thousand individuals [8].

Pinnipeds Pinnipedia, commonly known as pinnipeds, form a diverse clade of marine mammals characterized by finned limbs and semi-aquatic lifestyle. The group comprises families of walruses, fur seals, and true seals. In addition to producing simple vocalizations like barking or walrus-like sounds, pinnipeds also have a repertoire of more complex vocalizations, including songs.

1.2. Other aquatic life

Apart from marine mammals, the world's oceans are home to numerous other organisms. Fish produce a wide range of underwater sounds for communication, mating, and territorial displays [9]. Some crocodile species, which

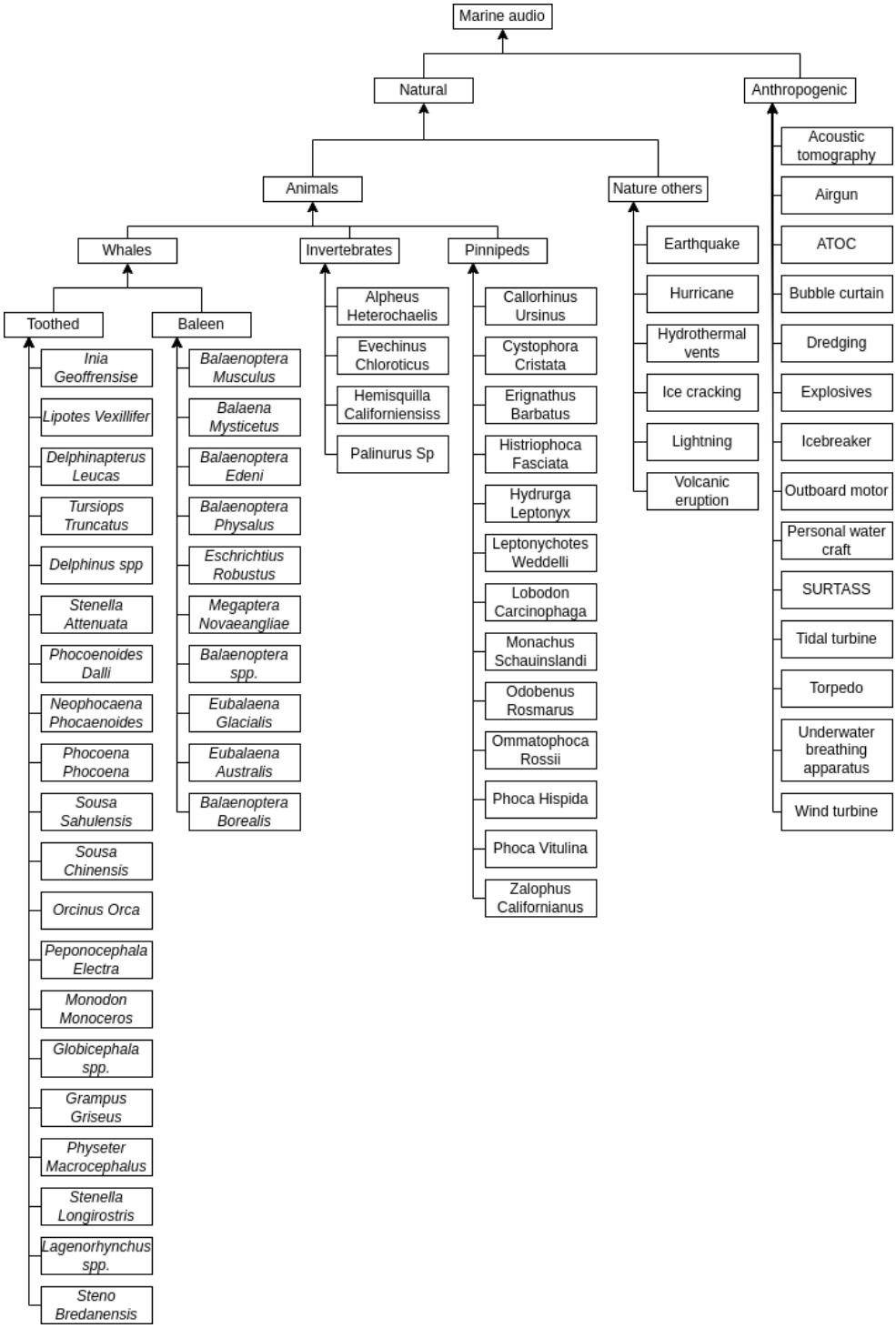


Figure 1: Overview of categories used in this project

can occasionally be found in coastal waters, emit low-frequency vocalizations, particularly during courtship or territorial displays [10]. While most bird species do not produce sounds underwater since their vocalizations are primarily adapted for the air, penguins are an exception. They communicate using a variety of vocalizations, including trumpeting, braying, and soft vocal calls, both on land and underwater [11]. Cephalopods like octopuses and squids can produce sounds through various mechanisms, including jet propulsion and muscle contractions. They use these sounds for communication, defense, and courtship. Some squid species produce high-pitched clicking sounds, while others emit low-frequency rumbling sounds [12]. Crustaceans generally do not produce vocalizations in the same way as other animals. However, some shrimp species can produce snapping sounds by rapidly closing their pincers. These sounds are typically used as defensive signals [13]. Jellyfish and cnidarians, such as corals, do not possess vocalization capabilities and therefore do not produce sounds in the traditional sense. However, some species of jellyfish can produce faint popping or crackling sounds due to the release of gas bubbles [14].

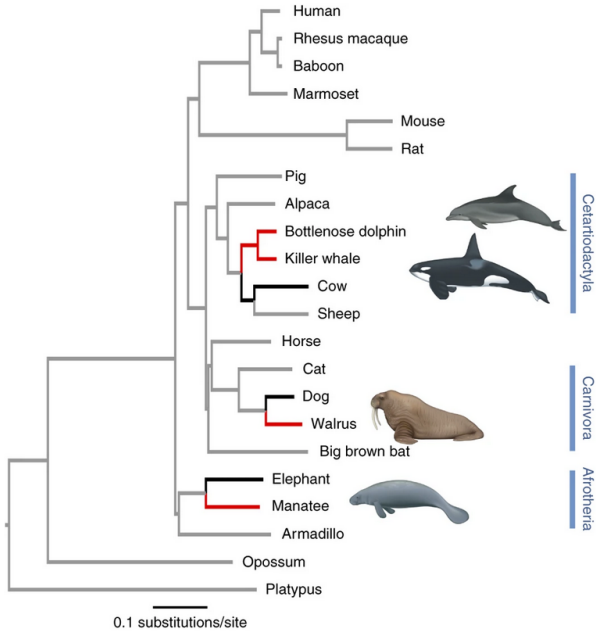


Figure 2: Evolutionary tree of marine mammals [6]

For the initial stages of the research, the fish category was dropped due to its extensive diversity. Birds and reptiles were also excluded due to limited diverse audio data. Cephalopods, crustaceans, cnidarians, echinoderms, and mollusks were grouped as invertebrates.

1.3. Natural Events

Underwater natural events produce a variety of sounds that contribute to the overall underwater acoustics. These events include:

Ice Ice-related sounds are common in regions with floating ice, such as the Arctic and Antarctic. Icebergs, which are large pieces of ice that have broken off from glaciers or ice shelves, create distinct sounds as they drift, collide, and grind against each other. These sounds can be loud and produce vibrations that register on seismometers as hydroacoustic Iceberg Harmonic Tremors (IHTs), also known as "iceberg songs." [15].

Volcanoes and thermal vents Underwater acoustics plays a crucial role in mapping, monitoring, and evaluating submarine volcanic eruptions. Acoustic data provides valuable information about the duration, frequency, intensity, and evolution of a volcanic activity over time. Gas-driven explosions associated with volcanic eruptions produce acoustic signals ranging from 1 to 80 Hz, peaking at approximately 30 Hz [16]. These signals exhibit a sudden onset, gradual rise in the amplitude over 30 seconds, and a subsequent sharp increase.

Earthquakes Hydrophones, underwater microphones, are used to detect and measure submarine earthquakes. Seismic energy from these earthquakes is converted into acoustic energy at the seafloor-water boundary. The acoustic signals, known as Tertiary waves, typically range from 4 to 50 Hz. Hydrophones can detect earthquakes at significantly lower magnitudes than land-based seismometers, providing valuable information for earthquake monitoring and research [17].

1.4. Anthropogenic sounds

Anthropogenic, or human-generated noise in the marine environment is increasing at an alarming rate, posing a significant threat to marine ecosystems and the survival of marine organisms, including mammals, fish, and other ocean animals. Marine animals rely on sound for various essential activities such as navigation, finding food, locating mates, avoiding predators, and communication. The proliferation of underwater noise from human activities can have adverse effects on marine life, ranging from discomfort to injury and death. Additionally, noise can interact with other environmental threats, such as hiding acoustic cues used by animals to avoid ships or becoming entangled in fishing gear [18].

Commercial shipping Studies have shown that the emission of underwater noise from maritime transport, particularly commercial shipping, can have both short-term and long-term negative consequences for the marine fauna, especially marine mammals [19]. In 2014, the IMO approved guidelines to reduce underwater noise generated by merchant ships [20]. These guidelines primarily focus on the main sources of underwater noise, including propellers, hull shape, onboard machinery, and operational aspects. The guidelines provide recommendations for the ship design, construction, and maintenance practices, such as hull cleaning to mitigate underwater noise. The propeller cavitation, which generates noise across a broad frequency band and discrete peaks in harmonics, is a significant contributor to the underwater noise.

Dredging Dredging is an excavation activity that involves the removal of materials from the seafloor, lake bottoms, riverbeds, harbors, and other water bodies. During dredging operations, underwater noise is generated due to the movement and excavation of sediments. The noise can come from various sources, including the dredging equipment itself, such as suction pumps, cutterheads, and excavators, as well as the deposition of dredged material. The noise generated by dredging activities can have an impact on marine life, including fish, marine mammals, and invertebrates. The underwater noise generated by dredg-

ing can affect marine organisms in several ways. It can disrupt their normal behavior, including feeding, mating, and communication [21].

Wind turbines Underwater sound is generated during the construction, operation, and decommissioning of offshore wind turbines. Vibrations from the turbine's internal components, such as the generator and gearbox, are transmitted down the main shaft and into the foundation, propagating into the water column and seafloor [22]. This mechanical noise generated by offshore wind turbines is typically concentrated at low frequencies below 1 kHz, with a slight increase in the level as the wind speed increases [23]. Ongoing research is also focused on understanding the specific effects of the wind turbine noise on marine organisms and developing effective mitigation strategies [24].

Tidal turbines Tidal turbines are another form of the renewable energy technology that harnesses the kinetic energy from tidal currents to generate electricity. These turbines can be installed individually or as part of an array, and they consist of rotor blades that spin when exposed to tidal currents. Measurements of underwater sound produced by active tidal turbines have shown that the signals are tonal and low frequency, typically ranging from 50 to 8200 Hz [25]. Higher frequency signals, up to 20 kHz, have also been detected and linked to mechanical processes inside the turbine. Local marine animals may detect the underwater sound associated with operational tidal turbines. Studies have indicated that signals from tidal turbines can be detected by harbor seals, porpoises, grey seals, and bottlenose dolphins at varying distances from the turbines [26]. The impact of tidal turbine noise on marine life is still being studied, and research efforts are focused on understanding the potential effects on marine organisms and developing appropriate mitigation measures [25].

2. Data sources used in this project

2.1. Discovery of Sound in the Sea Website

Discovery of Sound in the Sea is a website dedicated to the popularization of marine acoustics, and it is available at www.dosits.org [27]. What is most important from the point of view of this paper is that DOSITS provides an audio gallery which aggregates recordings from a variety of sources [28].

The Discovery of Sound in the Sea website has been developed by the University of Rhode Island's Graduate School of Oceanography (GSO) in partnership with Inspire Environmental of Newport, RI. Funding was provided by the U.S. Office of Naval Research, the U.S. National Science Foundation, the U.S. National Oceanic and Atmospheric Administration, the E&P Sound & Marine Life Joint Industry Programme, and the International Association of Geophysical Contractors.

As DOSITS is an aggregation site, actual credit for recordings goes to other parties. Those will be referenced directly in the case of recordings described in more detail in this paper. However, due to its size, this paper will not provide a complete list of acknowledgments for recordings used in the model's training. It is to be assumed that all recordings present at DOSITS on 01.02.2023 and released under one of the variants of Creative Commons were used.

The complete list may be provided directly via email request from the corresponding author.

2.2. Watkins Marine Mammal Sound Database

One of the founders of marine mammal bioacoustics, William Watkins, carried out pioneering work with William Schevill at the Woods Hole Oceanographic Institution for more than four decades, laying the groundwork for our field today [29]. One of the lasting achievements of his career was the Watkins Marine Mammal Sound Database, a resource that contains approximately 2000 unique recordings of more than 60 species of marine mammals.

The archive[1] contains recordings that span seven decades, from the 1940s to the 2000s and includes the first recordings of 51 marine mammals [30]. This resource is entirely accessible online, as was Watkin's goal.

3. Audio data processing

Digital audio processing refers to the manipulation of digital audio signals using various algorithms and techniques. This processing can be done in real-time, as the audio is being captured or played back, or offline, as a post-processing step.

3.1. Downsampling

In audio processing, the sampling rate refers to the number of samples per second taken from an analog audio signal and converted into digital format. The sampling rate is typically measured in Hertz (Hz) and it determines the frequency range of the digital audio signal.

During the analog-to-digital conversion, the continuous analog waveform is sampled at regular intervals, and each sample is assigned a numerical value that represents the signal's amplitude at that moment in time. The sam-

[1] https://cis.whoi.edu/science/B/whalesounds/index.cfm

pling rate determines how often these samples are taken, and the higher the sampling rate, the more accurately the digital representation of the audio signal captures the original waveform.

The most common sampling rates used in audio processing are 44.1 kHz and 48 kHz, which are used for CD and DVD audio, respectively [31]. Other standard sampling rates include 96 and 192 kHz, used in high-resolution audio formats [31].

The choice of the sampling rate depends on the specific application and the desired level of audio fidelity.

Higher sampling rates produce a better sound quality but require more storage space and processing power. Additionally, some audio equipment may not support high sampling rates, so the hardware capabilities may limit the choice of the sampling rate.

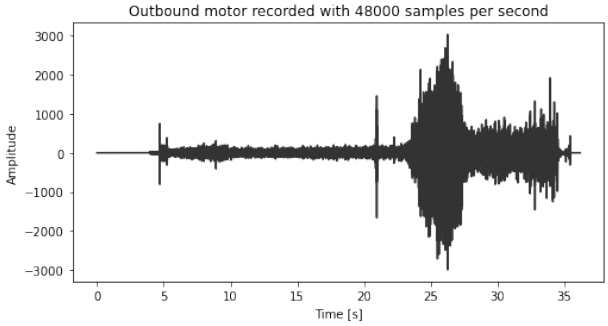


Figure 3: Recording of an outbound motor with original sampling frequency

The Nyquist theorem, also known as the Nyquist-Shannon sampling theorem, is a fundamental concept in the digital signal processing and communication theory. The theorem states that in order to reconstruct a continuous signal from its discrete samples accurately, the sampling rate must be at least twice the highest frequency component in the signal [32].

In other words, if a signal has the highest frequency component of f_{max}, then the sampling rate f_s must be greater than or equal to $2f_{max}$ in order to accurately reconstruct the original signal. This is because the discrete samples can only capture information about the signal up to a certain frequency, known as the Nyquist frequency f_n, which is half the sampling rate as described in equation 1.

$$f_n = \frac{f_s}{2} \tag{1}$$

If the sampling rate is too low, the Nyquist frequency will be lower than the highest frequency component of the signal, resulting in aliasing or distortion. Aliasing occurs when high-frequency components of the signal are folded back into the frequency range captured by the lower sampling rate, resulting in a distorted signal that cannot be accurately reconstructed.

As seen in Figs. 3 and 4 while downsampling causes

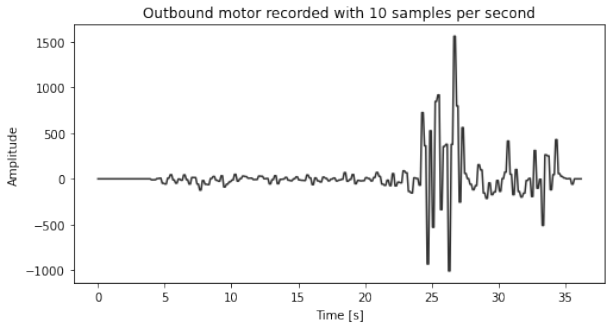


Figure 4: Recording of an outbound motor downsampled to 10 samples per second.

significant information loss, it can retain significant characteristics of a signal. This information loss, as well as the Nyquist theorem, may render downsampled recordings no longer fit for classification. However, requiring high sampling rates of recordings limits data sources to only professional institutions with expensive audio recording and processing equipment. Due to that, the choice of the target sampling frequency is a compromise between the quality and quantity of data.

One of the goals of this research is to prove that the audio CD quality sampling (44.1 kHz) is enough for classification purposes in the context of the marine audio.

3.2. Splicing and normalization

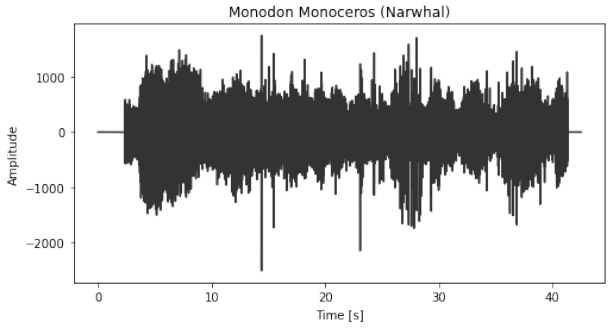


Figure 5: Source recording of Narwhal

Recordings will often contain a short silence before and after a meaningful signal. There can be many reasons, from human operators activating recording in anticipation of an event to automated sensors not recording the source of its initial trigger. Additionally, if the data is prepared by amateurs or small teams, they may often just splice the recording on specific timestamps and fail to trim silent periods.

This leading and trailing silence generally does not provide any significant information and can be removed without any loss.

As can be observed in Fig. 6, the difference between trimmed and source data from Fig. 5 is minimal. However,

it is also very significant as the leading and trailing silence can result in generation of splices that contain very little useful information.

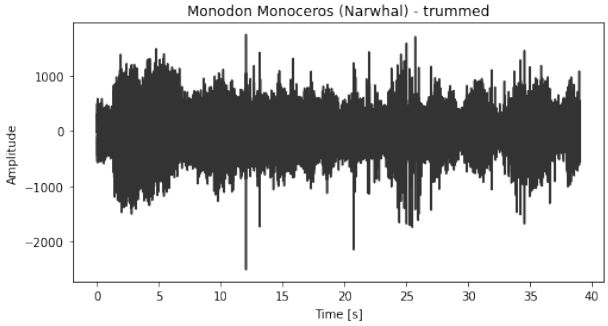


Figure 6: Source recording of Narwhal with leading and trailing silence trimmed

After downsampling, all inputs have the same number of samples per second. However, there is still a difference in the recording time. The other heterogeneity comes from the amplitude of signals. This difference can be significant for the training model, but it can also be a severe downside. For example, a normalized input works better in neural networks which are one of the more popular machine learning algorithms [33].

As visible in the plot in Fig. 5 the example recording of a Narwhal lasts for about 40 seconds and an amplitude with an absolute value over 2000 at peaks.

In order to unify the data, source recordings are divided into splices. Each splice lasts for a predetermined amount of time. If the raw record time is not a multiple of the splice time, the trailing part of the record will be dropped. A raw record will be dropped altogether, if shorter than a single splice. Additionally, all values will be divided by the largest absolute value in raw recording resulting in signal values in a range between -1 and 1. It is important to remember that it is possible not to have a -1 or 1 value in a single splice, as normalization occurs for the scope of the whole raw recording. Normalization is an optional step, and it can be turned off if the original amplitude value is important in the model creation. The result of applying splicing and normalization to the recording from Fig. 5 can be seen in Fig. 7

Just like downsampling, this step results in data loss. The most egregious example of that is, of course, dropping the trailing parts of the recording. Another loss of data happens when removing information about absolute values of the amplitude, if normalization has been enabled. Finally, information about the relation between samples ending in separate splices is lost. This can be a significant issue, as splices are created in arbitrary places.

In this step, most data loss occurs and limits the proposed uses of the pipeline. For example, it cannot be used for speech recognition algorithms, as cutting up a sentence into small splices removes long-term relations between words necessary for understanding a sentence. It is even possible for a splicing mechanism to split the recording on a single word. It is, however, viable to use the presented pipeline for training algorithms that detects whether the recording contains speech or not, as most of the specific frequencies and patterns will remain even in spliced data.

3.3. Selection

After all categories have been filled with all possible samples from given recordings, a selection the step will be executed. The first part of the selection will be the removal of categories that have fewer samples than the minimum acceptable number which is given as a parameter to processing a script.

This is done because when the number of training samples is small, machine learning models can easily overfit the training data, which means that the model learns the noise in the data rather than in the underlying patterns. As a result, the model may perform well on the training data but poorly on new data, and this issue is called overfitting.

After removal of underrepresented categories, the smallest representation will be taken from the remaining pool. The resulting dataset sample distribution will look similar to the one in Fig. 8. Then, to avoid overrepresentation, a maximum number of samples per category will be calculated as the smallest representation multiplied by the maximum overrepresentation factor which is a value greater than the one given as an algorithm parameter.

This step is necessary when one category has significantly more samples than others, and it creates an imbalanced dataset. Imbalanced data can cause machine learning models to be biased toward the majority class and may result in poor performance in the minority class, as seen in the confusion matrix in Fig. 9.

After the complete selection step, a balanced dataset is generated, as shown in Fig. 10

3.4. Augmentation

Data augmentation can be used to increase the number of samples. It is not capable of creating new information. However, it can inject noise to make the machine learning model more robust [34]. One of the more common approaches to data augmentation is noise injection in which a random or pseudorandom noise is added to preexisting data [35]. The most common noise used for that purpose is white noise, a generic type of noise.

In this pipeline, the augmentation step is given a list of noise colors and factors. Colors of noise refer to different noise signals with specific spectral characteristics.

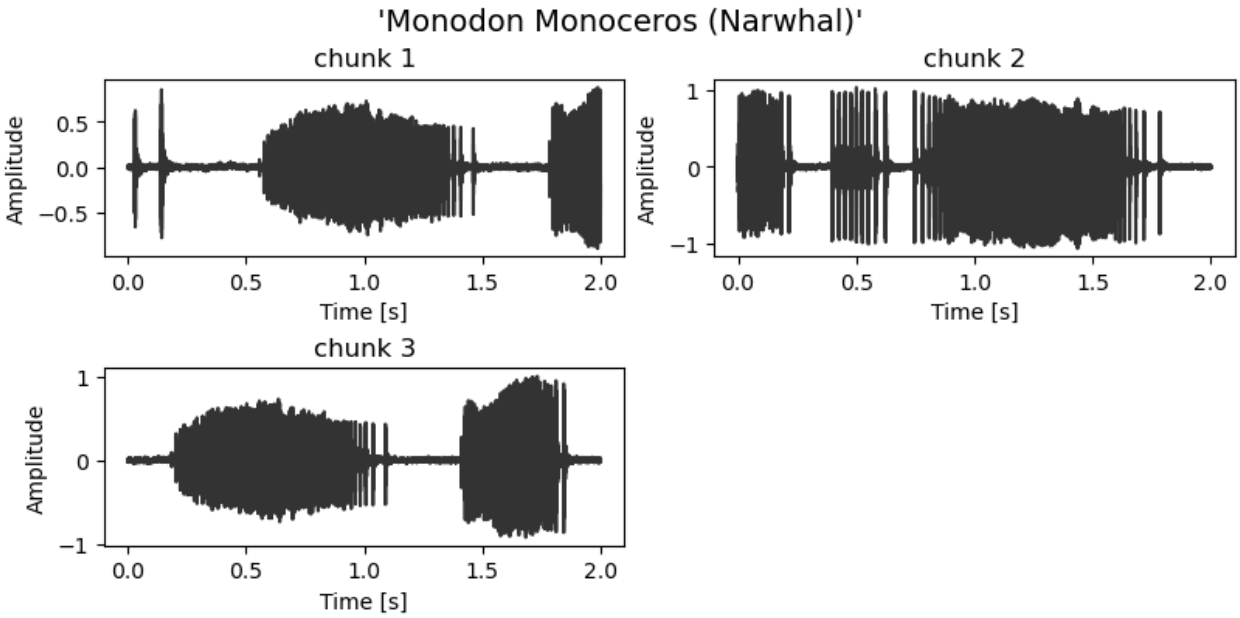


Figure 7: Recording of Narwhal split into splices

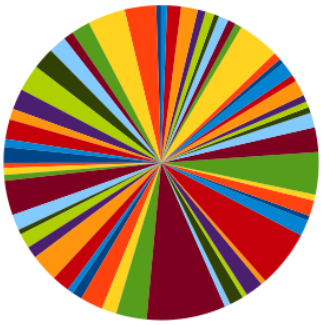


Figure 8: Category distribution in a dataset before selection, each sector represents a different category

1. White noise: White noise has equal energy across all frequencies in the audible spectrum. It is called white noise because it contains all the frequencies of the visible light spectrum, which combine to form white light. It has a flat power spectral density and sounds like a hissing sound.
2. Pink noise: Pink noise has a power spectral density that decreases by 3dB per octave as frequency increases. This means that higher frequencies have less energy than lower frequencies. Pink noise sounds like a waterfall and is often used in audio testing and calibration.
3. Brown noise: Brown noise has a power spectral density that decreases by 6dB per octave as frequency increases. This means that higher frequencies have significantly less energy than lower frequencies. Brown noise sounds like a deep rumble and is often used in audio testing and to mask unwanted sounds.
4. Blue noise: Blue noise has a power spectral density that increases by 3dB per octave as frequency increases. This means that higher frequencies have more energy than lower frequencies. Blue noise sounds like a hissing sound similar to white noise but with a higher frequency emphasis.
5. Violet noise: Violet noise has a power spectral density that increases by 6dB per octave as frequency increases. This means that higher frequencies have significantly more energy than lower frequencies. The violet noise has a sharp, hissing sound.
6. Grey noise: Grey noise is a type of noise with a power spectral density that is flat in the middle frequencies but has decreased energy towards the lower and higher frequencies. It is a more natural-sounding noise and is sometimes used in music production.

The noise factor refers to a relation between the highest noise amplitude and the highest amplitude of the the source signal. Noise with a factor of 1 will be at its most potent, as loud as the loudest sound in the source recording, while the noise of factor 0.5 will be half-powerful only.

Augmentation generates all possible combinations of source recordings, noise colors, and noise factors. Hence, for 200 unaugmented splices, three types of noises, and two factors, there will be a total of $200 \times 3 \times 2 = 1200$ augmented splices generated.

As shown in Figures 11 and 12, adding pink noise to a recording of a torpedo changes the exact values of a signal while retaining the patterns. It is important to remember that the noise cannot be stronger than the original signal as it will mask it. In other words, the original pattern will be lost. In this project, the values of noise to the source signal ratio were set in a range between 0.1 and 0.5.

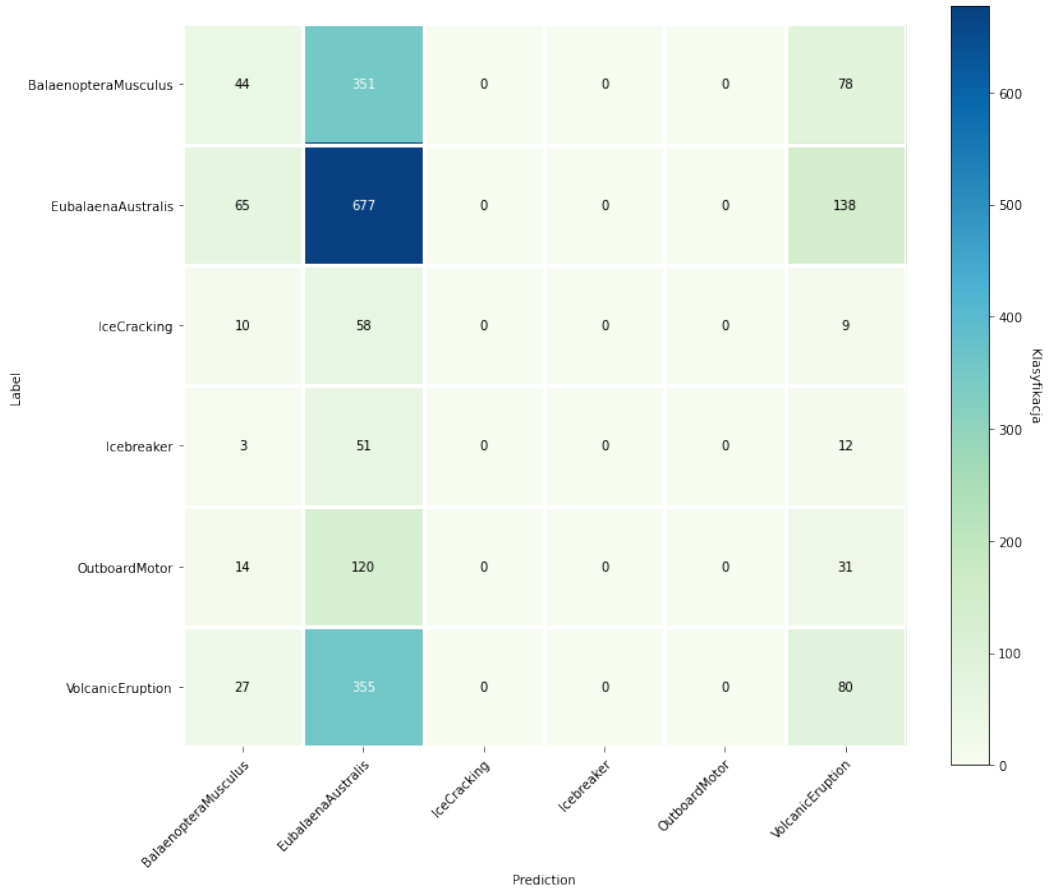


Figure 9: Confusion matrix for a model trained on data with overrepresentation

Figure 10: Category distribution in the dataset after selection, each sector represents a different category

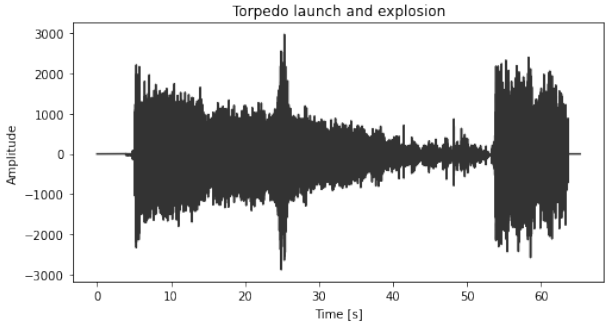


Figure 11: Recording of a torpedo launch and explosion

4. Visual data processing

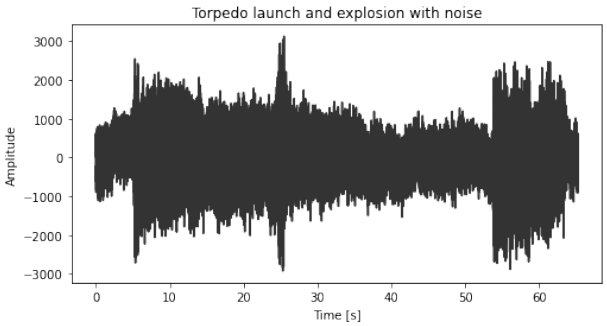
Sound can be represented visually through the use of spectrograms or waveforms. In this research, a spectrogram-based representation was used which is a visual representation of the frequency content of a sound over time. It is created by plotting the frequency spectrum of a sound signal as a function of time. The x-axis of a spectrogram represents time, while the y-axis represents frequency. The intensity of each point in the spectrogram represents the magnitude or power of the sound at that frequency and time. Spectrograms are commonly used to visualize and analyze animal vocalizations, musical signals, and speech.

4.1. Fourier transform

Fourier analysis [36] is a mathematical technique used to decompose a complex waveform into its component frequencies. In the context of audio signals, Fourier analysis is used to analyze the spectral content of an audio signal, which can provide helpful information about the signal's timbre, pitch, and harmonic structure.

The Fourier transform works by expressing a signal as a sum of sine and cosine waves of different frequencies, each with its amplitude and phase. It can be calculated by solving the equation 2.

$$F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-i\tau\nu t} dt \tag{2}$$



g

Figure 12: Recording of a torpedo launch and explosion with added pink noise

The Fourier transform algorithm calculates the amplitudes and phases of the Fourier series coefficients for a given signal, which allows the signal to be expressed in the frequency domain.

The Discrete Fourier Transform (DFT) is a version of the Fourier transform that is used for analyzing digital signals. Unlike the continuous Fourier transform, which is defined over an infinite time domain, the DFT is defined over a finite time domain, which makes it more suitable for processing digital signals that are stored as discrete samples.

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{i\tau}{N}kn} \tag{3}$$

The DFT works by taking a finite sequence of N samples and calculating the Fourier transform of that sequence. The result is a sequence of N complex numbers, which represent the frequency content of the input signal at discrete frequency intervals.

The DFT can be computed using a matrix multiplication operation, but this approach can be computationally expensive. The Fast Fourier Transform (FFT) [37] is a more efficient algorithm for computing the DFT and is commonly used in practice.

The FFT works by exploiting the symmetry properties of the DFT to reduce the number of computations required. Specifically, the FFT is based on the divide-and-conquer approach, which divides the input sequence into smaller sub-sequences and recursively computes their DFTs.

The most commonly used implementation of the FFT is the Cooley-Tukey algorithm [38], which recursively splits the input sequence into two sub-sequences of length $N/2$, computes their DFTs using the FFT algorithm, and then combines the results to compute the final DFT of the original sequence.

The Short-time Fourier transform (STFT) [37] is a signal processing technique that analyzes a signal's frequency content over time. It is a modification of the Fourier transform that allows us to examine the frequency content of a signal as it changes over time.

The STFT works by dividing a more extended signal into shorter segments, known as windows. Each window is multiplied by a window function, such as a Hamming or Hanning window, to reduce the spectral leakage that can occur with the Fourier transform. The Fourier transform is applied to each windowed segment to obtain its frequency spectrum. Using overlapping windows, we can obtain a time-frequency representation of the signal that shows how the frequency content changes over time.

The STFT is commonly used in audio signal processing, speech processing, and other areas of signal analysis where the frequency content of a signal changes over time. It is an essential tool for understanding the time-frequency characteristics of a signal and is widely used in many applications, including audio compression, noise reduction, and audio synthesis.

4.2. Psychoacoustics and signal scaling

Psychoacoustics is a subdiscipline of psychophysics that studies the relationship between sound stimuli and the auditory sensations resulting from these stimuli. It can be divided into external psychoacoustics and internal psychoacoustics.

The prediction of the influence of these cognitive processes is a matter of the discipline of subjective acoustics. Psychoacoustics uses psychological and physical measurement methods to measure auditory sensations. The methods of psychological measurement quantify the auditory sensations from two stories or the psychological reactions of the listener. Classical methods are the Békesy tracking, Magnitude Estimation, and the Limits Method [39].

One of the premises of psychoacoustics is that a slight variation in the magnitude of the stimulus does not necessarily lead to a variation in the magnitude of the sensation. This occurs because it is necessary for there to be a minimum variation in the stimulus so that there is a variation in the sensation that can be differentiated. The difference in the magnitude of the stimulus that causes a variation that can be justly perceived is known as the Just Noticeable Difference, the Difference Threshold, or the Nonlimiting Difference [39]. In this way, psychoacoustics provides essential information for product engineering or for noise control because of, in both cases, the alteration of a physical characteristic of the signal. The mel scale is the pitch scale measured by the method of psychological acoustics, which determines the subjective perception of the sound level by the human ear concerning the objective scale of measuring the sound frequency in hertz. The scale was defined in 1937 by Stevens, Volkman, and Newman [40]. The name comes from the first three letters of the English word melody. In 1946, Stevens published a work entitled On the Theory of Scales of Measurement,

which initiated the development of psychophysical measurements using scaling methods. The unit of frequency in this scale is mel. The relationship between the mel and Hz scales is non-linear and is defined by the formula:

$$m = 2595 \log_{10}(1 + \frac{v}{700}) \tag{4}$$

Based on the measurements, it was assumed that a tone with a frequency of 1000 Hz at a sound pressure level of 40 dB above the threshold of hearing had 1000 mels. The number of mels is proportional to the pitch of a given sound at a given frequency and loudness.

4.3. Visual representation of audio data

STFT converts a sequence of signal segments into a sequence of harmonics, where: f_k is the value of a signal sample, k is the number of this sample, N is the number of samples, n is the number of the harmonic component i $n = 0, \ldots, N-1$, i is the imaginary unit.

One of the problems that can be encountered when dealing with visual data is the memory required for storing initial images as well as responses of convolution layers. The memory requirement for storing a single image is described in equation 5.

$$M = W \times H \times C \times \frac{R}{8} \tag{5}$$

where M is the memory usage in bytes, W is the width and H the height of image in pixels, C is the number of channels, and R is the bit resolution of a single channel. Hence, for example, a simple amateur digital photo that will be printed on a 4×6 inch card with 1200 PPI (Pixels Per Inch) resolution will have the following values.

$$M = 4 \times 1200 \times 6 \times 1200 \times 3 \times \frac{24}{8} \approx 311MB \tag{6}$$

This imposes high limitations on the architecture of convolutional networks as a layer that processes such an image will output a result where the number of channels is equivalent to the number of kernels. This means that a single response will take approximately 1 GB of memory for nine kernel layers. In the case of spectrograms, each image parameter is directly tied to information about the signal stored in it. The number of channels reflects how many different spectrograms are stored in a single image. The scope of this work will always be 1. However, it is possible to store multiple spectrograms, for example, pure STFT and MEL in a single image. The image height is dependent on the minimum and maximum frequency registered as well as on the frequency resolution. This relation is described by the equation.

$$H = \frac{v_{max} - v_{min}}{\Delta v}. \tag{7}$$

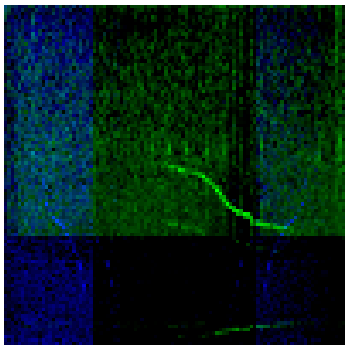


Figure 13: A 3-second splice visualized with linear and MEL scales combined

5. Machine learning

The problem with the classification of data by experts is its cost. Not only do those experts require a significant amount of time to process all the gathered data, but they also have to be paid, which in case of such a specific field of experience usually bumps up the project costs significantly. The second issue is data storage or transfer. If experts need to analyze spectrograms all data recorded by the robot must be stored on a local device or transferred to a remote location.

Due to that, an alternative data-based model generated with a machine learning algorithm will be implemented. As it can be run on robot hardware, only significant data must be stored for revision, and noise can be ignored.

6. Artificial neural networks

Artificial Neural Networks (ANNs) are a type of a machine learning model that is inspired by the structure and function of biological neurons in the human brain. ANNs are composed of interconnected nodes or "neurons", organized into layers which process input data and produce output predictions.

In ANNs, the input data is processed through a series of layers, where each layer performs a mathematical operation on the input and passes the result to the next layer. The first layer of the network is called the input layer, and the last layer is called the output layer. The intermediate layers are known as hidden layers because their computations are not directly visible in the output.

Each neuron in the network receives input from neurons in the previous layer and computes a the weighted sum of the inputs, passing through an activation function

to produce an output. The weights and biases of the neurons are learned during training using optimization algorithms such as the gradient descent to minimize the error between the predicted and actual output.

6.1. Dense layer

This is the most basic variant of the layer found in neural networks present from Rosenblatt [41] perceptron models. Such a layer consists of individual neurons, each of which is connected to all of the previous layer outputs. This means that the number of tunable parameters in a given layer is described according to the formula

$$\phi_l = n_l * n_{l-1} + n_l \tag{8}$$

where n_l is the number of neurons in a given layer and n_{l-1} is the number of neurons in the preceding layer [42]. Since this layer directly uses neurons based on the McCulloch-Pitts model [43], the response of each neuron is a linear function of the previous layer's response that has been further subjected to the activation function.

The dropout technique in neural networks relies on the random shutdown of specific neurons during the learning runs. This technique was initially proposed by Hinton in 2012 [44] then further developed by Srivastava [45]. They proved that the use of dropout could improve the accuracy of a neural network by about two percentage points. This is especially important in the case of networks capable of achieving high precision because improving the score from 96% to 98% is more important than from 60% to 62%.

The use of dropout prevents neurons in the model from relying on neighbors in the same layer. Furthermore, relying on a small fraction of the previous layer's outputs.

The parameter that defines the dropout is the probability of dropout, which tells what the probability is that every single neuron will be skipped in a learning cycle. Most often, this parameter is set as 0.5. The retention coefficient described by the formula shown in equation 9 also results from the dropout probability.

$$k = 1 - P_d, \tag{9}$$

The TensorFlow [46] library uses a particular layer called the drop layer. Adding it behind the regular neural layer gives the application of abandonment to that layer. After training, these layers are removed from the network. Within the framework of the described project, all densely connected layers appear as a folding pair from the more dense layer and the drop layer.

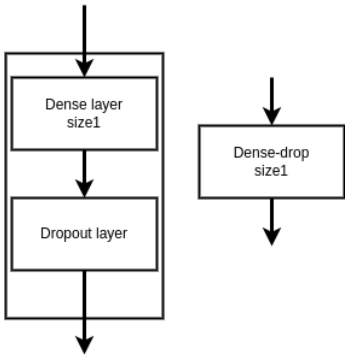


Figure 14: A pair of dense and dropout layers

6.2. Convolutional layer

Much of the growth of deep learning is primarily due to advances made in computer vision. One of the most used algorithms in this field is convolution, from which the convolutional neural network (CNN) is derived. In addition to being a system inspired by the primary visual cortex, the network can decipher or learn the patterns more complexly existing in a set of images, and it does it utilizing said the convolution. Fundamentally, the convolution comprises an operator with two functions; the image and the filter or kernel. The function takes a part of the image and highlights patterns by multiplying each point of the image fragment with the filter elements. The result is weighted into a sum, and the generated values are located at the position corresponding to the position of the image fragment. The process is repeated, moving the filter across the entire image, creating an image with highlighted features that depend on the filter structure. In the case of CNN, the convolution is performed similarly. However, the images generated by the convolution are known as feature maps.

In this context, the parameters of training the network are the weights associated with all the filters, i.e., the network learns the optimal filters to highlight the high-level features that converge to the desired task. The process is repeated layer after layer creating more features which are more abstract each time. In addition, the CNN architecture can also be implemented with other types of networks, such as fully connected models. Fig. 15 shows a graphical description of the convolutional process where the input image generates one or more maps of characteristics that depend on the number of filters for that layer. In the same way, each layer can have the filter size and the desired number of strides. The stride is the jump in which the filter moves along the images.

6.3. Pooling layer

In Convolutional Neural Networks, a pooling layer is a component that plays a role in reducing the spatial

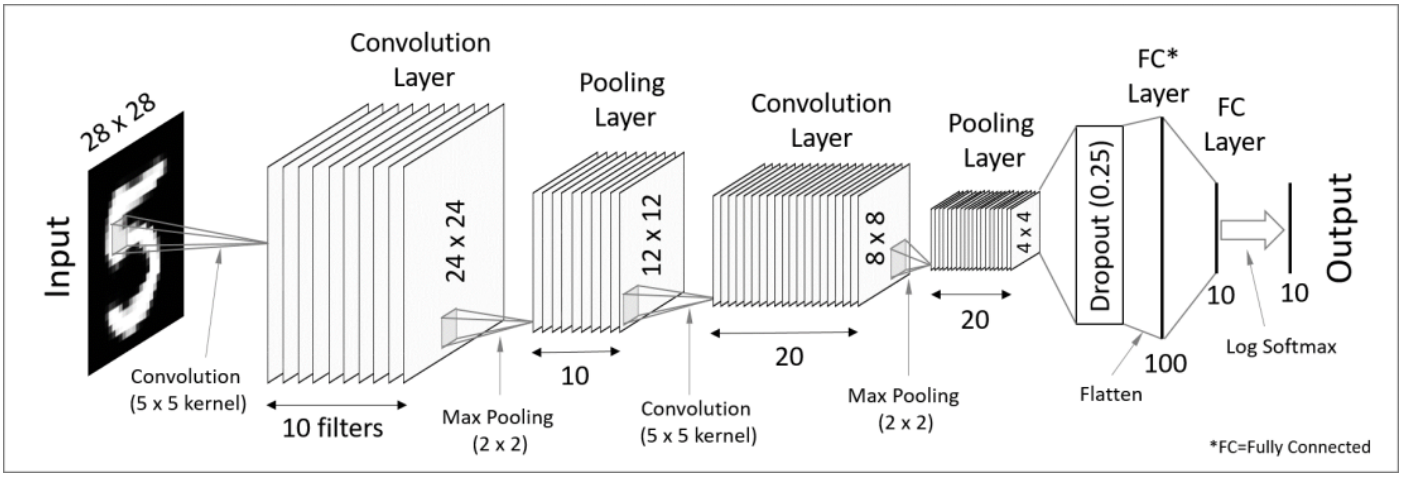


Figure 15: Convolutional neural network (source: wordpress.com)

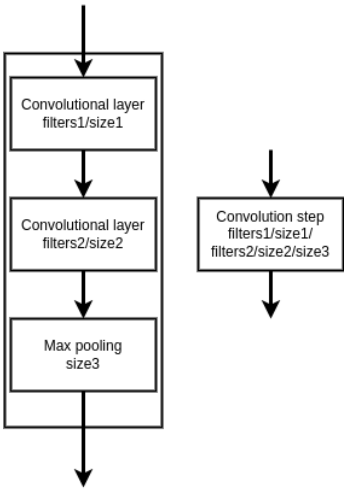


Figure 16: A double convolution layer

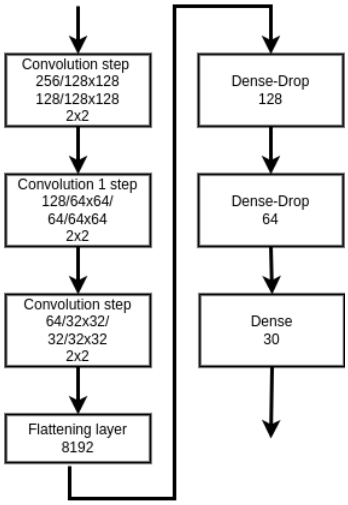


Figure 17: An overview of the architecture of network used

dimensions of the input volume, which, in turn, helps in reducing the number of parameters and controlling overfitting. The pooling layer works by downsampling the feature maps created by the previous convolutional layers. Max pooling is the most widely used form of pooling in CNNs, it takes a small rectangular region (usually 2x2 or 3x3) and slides it over the feature map. For each region, it selects the maximum value. This maximum value becomes the new value for that region in the output.

7. Model used in this work

The proposed network architecture was designed for spectrogram images with dimensions 128x128, hence, the number of convolutional neurons in each layer. The use of this structure allowed us to obtain 1,633,278 training parameters. The algorithm Adam (*Adaptive Moment Estimation*) was selected as an optimizer responsible for minimizing the loss function by modifying the weights. It is a combination of two other optimizers: momentum optimization and the RMSProp (*Root Mean Square Propagation*) algorithm. The former changes the weights based on the exponential distribution of the mean of the earlier gradients. The latter only considers the gradient from the most recent steps by decomposing exponentially on the mean of the squares of the gradients. Cross-entropy which examines the correspondence between the prediction of probabilities and the target classes was used as a function of losses in the model. A pooling layer was used to reduce the computational load. This helps to reduce the resolution of the image by sub-sampling it. It is crucial to use this layer to define the size of the connecting kernel, which is the window from which the combined value will be calculated. Two of the available methods of computing this value are finding the maximum value in the window - Max Pooling and computing the average value - Average Pooling [47]. The first method was selected, as it was shown to produce slightly better results. Fig. 17 illustrates a complete architecture of the created neural network.

8. Generated datasets

These categories were provided by both DOSITS as well as the Watkins Database and covered both toothed and baleen whales. There are 20 categories, and most of them have approximately 70 recordings with a 3 second sample time. A larger number of recordings would be possible for most of them, however, the overrepresentation factor was set at 2.0, which means no more than 70, since the lowest representation was at 35 splices. A detailed list of categories with the number of splices per category is shown in Table 2.

The second dataset created used all the gathered recordings as the input. However, due to the limit of no less than five unaugmented splices, a total of 26 categories were dropped. (Needs rewriting)

9. Experiments

A series of experiments were run to verify the correctness of the proposed approach.

9.1. Proof of concept

The first experiment was dedicated to verifying if the use of visual classifiers for marine audio identification was a correct approach. In this initial stage of the project, a dataset was prepared manually.

This implementation was able to work only with smaller datasets as it stored whole data in memory and, as such, was not scalable. The dataset created covered only marine mammals provided by the Watkins database. The collection of sound signals belonging to 30 classes corresponding to animal species was built based on the Watkins Marine Mammal Sound Database.

Evaluating the test data and making predictions using a trained neural network further confirmed the greater effectiveness of spectrograms in which the frequency was transformed into the Mel scale. 128 samples were classified correctly and 32 incorrectly for the test set of 160 sound samples. The accuracy of the classification of samples for this set was 80.0%. The prediction made for the entire data set, i.e., 1594 sound samples, was accurate at the level of 94.8%, giving correct results for 1511 signals and incorrect for 83. The confusion matrix was used to assess the accuracy of the classification of the neural network model. It is represented by matrices with dimensions KxK, where K is the number of the available class labels, where the rows correspond to the actual sample classes and the columns to the predicted classes. Fig. 28 illustrates the error matrix for the classification of Mel spectrograms from the test set. Deep neural networks are

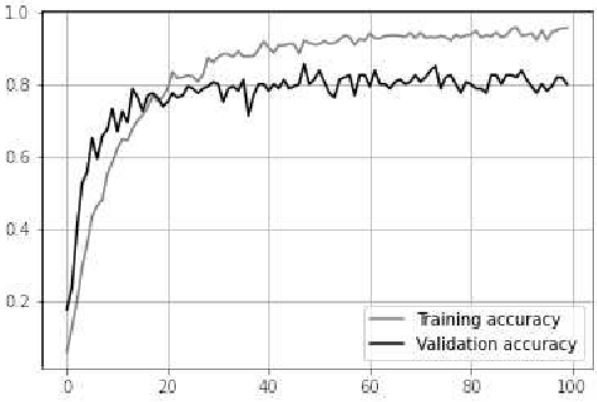


Figure 18: Neural network training accuracy function for test and validation sets in the MEL trained model

capable of creating exact matches based on up to a million parameters which can sometimes result in overfitting the model. The neural network is excessively attached to the training set and does not give effective classification results on the test set. Various regularization methods are used to prevent such phenomena. The most common of these is *Dropout*, which involves the omission of specific neurons during subsequent learning runs. Using this method on the neural network model improved its classification accuracy on the Mel test set from 75.0% to 80.0%. Fig. 18 illustrates how close the validation corresponds to the training checks.

Recording of the audio signals emitted by marine animals plays a crucial role in analyzing their behavior as well as the state of a specific area in the ocean. Automatic classification of these signals opens up new opportunities for institutions specializing in oceanography. A network with an accuracy of 80% was created.

9.2. Verification of new implementation

The MNIST (Modified National Institute of Standards and Technology) database [48] is an extensive database of handwritten digits that is commonly used for training various image processing systems. The database is also widely used for training and testing in machine learning. It was created by "re-mixing" samples from the original datasets of the NIST. The creators felt that since the dataset of the NIST was taken from American Census Bureau employees, while the testing dataset was taken from American high school students, it was not well-suited for machine learning experiments. Furthermore, the black and white images from the NIST were normalized to fit into a 28x28 pixel bounding box and anti-aliased, which introduced grayscale levels.

The MNIST database contains 60,000 training images and 10,000 testing images.

A specific version of the dataset is provided on Kag-

Figure 19: Example of data from MNIST dataset

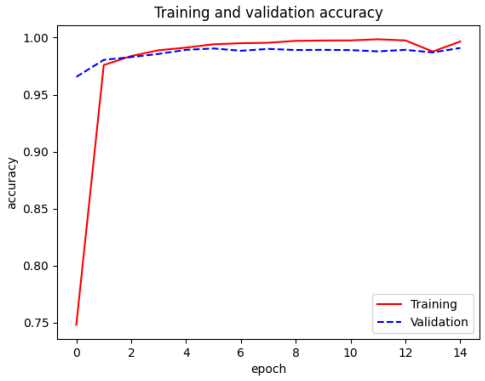


Figure 20: Accuracy for MNIST dataset

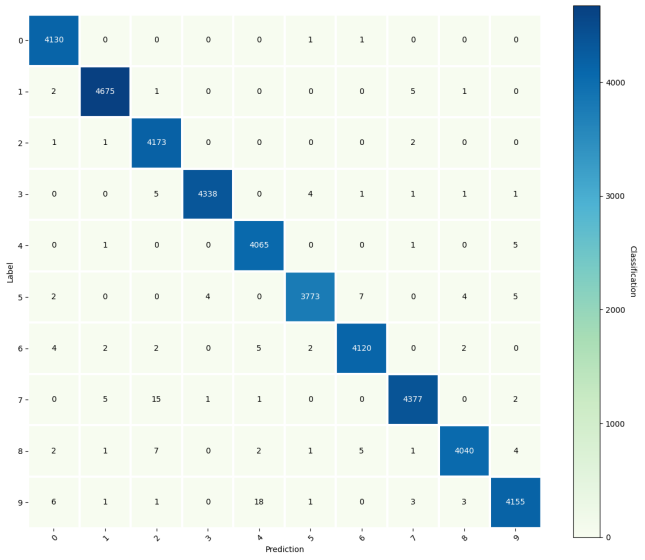


Figure 21: Confusion matrix for MNIST dataset

gle by Colianni at https://www.kaggle.com/datasets/scolianni/mnistasjpg. This version is similar in structure to the output of the visual dataset generator.

As can be told by examining the learning history from Fig. 20 and the confusion matrix resulting from Fig. 21 classifier, the algorithm works correctly.

With that part of the pipeline having been verified, it was possible to start experiments that were supposed to verify, if the dataset generation worked appropriately.

Table 1: Coarse dataset

Category	Samples
Anthropogenic	167
Natural	115
Whales	228
Pinnipeds	223

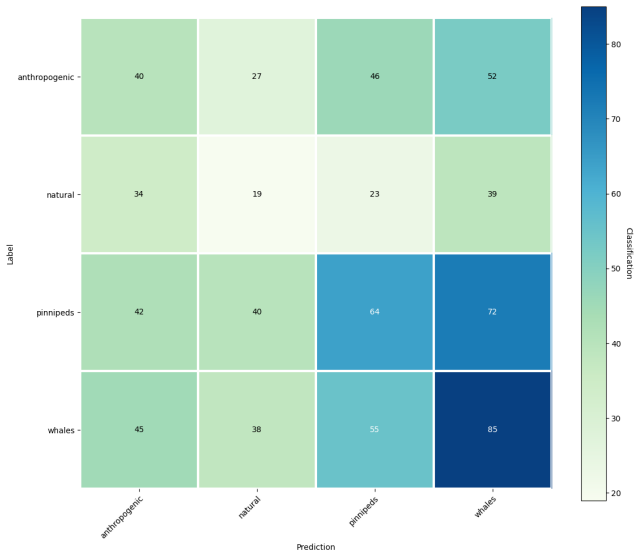


Figure 22: Confusion matrix for small dataset with overrepresentation

9.3. Coarse classification

A coarse dataset was created for the first test of the whole data processing pipeline. Unlike the target classification, it was only the most general category that was associated with the recordings as described in Table 1.

After the initial run, a problem with overrepresentation appeared. There was a strong preference of the predictor model towards whales and pinniped, as seen in the confusion matrix in Fig. 22.

This causes a much worse result in a validation set than in the prediction set. In other words an overfitting of the model. This can be seen in the plot in Fig. 23

9.4. Whales only, long splice

Following the most coarse classification, the next attempt was focused on reproducing the initial experiment with the whale classification. However, this time, the result of a prepared pipeline will be used instead of a hand generated dataset. All the categories in the dataset as well as the number of samples per category are listed in Table 2.

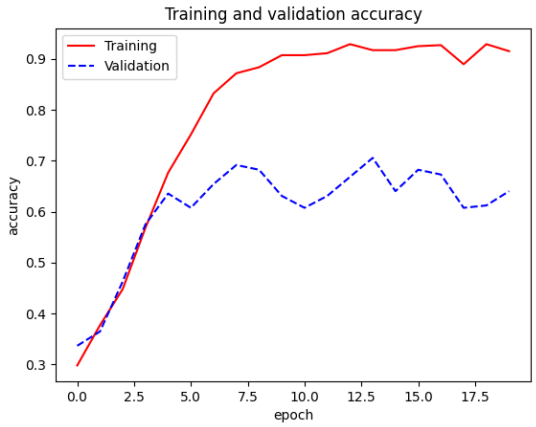


Figure 23: Accuracy for small dataset with overrepresentation

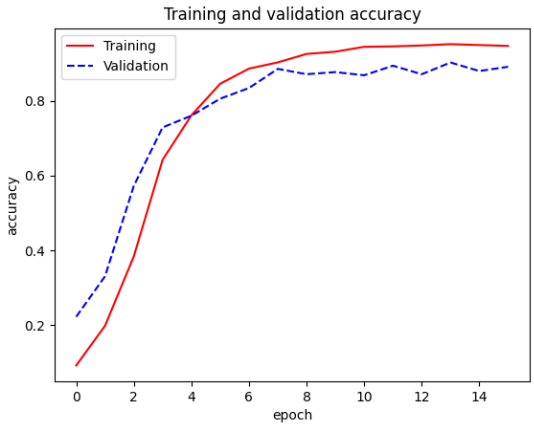


Figure 24: Accuracy for whales only dataset with 3-second splice

Table 2: Whales only dataset

Category	Samples
EubalaenaAustralis	70
DelphinusSpp	63
PhyseterMacrocephalus	70
MegapteraNovaeangliae	70
EubalaenaGlacialis	70
NeophocaenaPhocaenoides	63
BalaenopteraSpp	70
BalaenaMysticetus	70
BerardiusBairdii	56
BalaenopteraMusculus	70
OrcinusOrca	70
BalaenopteraEdeni	49
MonodonMonoceros	70
StenellaLongirostris	35
DelphinapterusLeucas	42
SousaChinensis	35
EschrichtiusRobustus	70
PhocoenoidesDalli	42
StenoBredanensis	70
BalaenopteraBorealis	35

9.5. Complete set, long splice

The next experiment was focused on generation as well as classification of a complete set created from all the gathered recordings. Nevertheless, it is important to remember that due to the presence of the selection step, it is possible that not all of the categories present in the source recordings will get through to the final dataset. All of the categories present in the dataset as well as the number of samples per category are listed in Table 3. 26 categories were dropped due to an insufficient, smaller than five, number of unagumented splices. While this longer, five second, time of splice results in the noticeable number of the categories dropped, also allows for a better classification. The resulting accuracy is shown by the plot in Fig .26alongside the corresponding confusion matrix in Fig. **??**.

As this classification still produces results of acceptable quality, further experiments with more limited splice time can be run.

9.6. Complete set, short splice

Finally, a dataset with a 1-second splice was generated. Such a short splice results in 48 thousand measurements per splice, which is a relatively small value. However, even short recordings can generate an acceptable amount of splice, resulting in the most diverse dataset as seen in Table 4.

The network used in this research can still categorize those short recordings. As shown in Fig. 27, the accuracy reaches over 90% for the training data. However, unlike in the previous examples, the ability of the network to generalize drops significantly which results in a drop in the validation data accuracy down to a range of 60%. This is not only a much lower validation accuracy than in the previous steps, but also a much more significant difference between the train and validation accuracies. By shortening the observation window, the model loses some crucial contextual information that was present in the longer window. While it still recognizes the patterns in the training data, due to shorter-term dependencies, it struggles to accurately predict the future values in the validation data that require longer-term dependencies. Consequently, the accuracy of the model decreases on the validation data but remains high on the training data since it has learned to capture the short-term patterns effectively. Due to that, experiments stopped at a one-second splice with a confusion matrix for the whole dataset, training, and validation, visualized in Fig. 27.

10. Conclusion

Selected approaches for dataset generation were proven to be successful. A CNN based classifier achieved a classification with an accuracy of approximately 90% on the

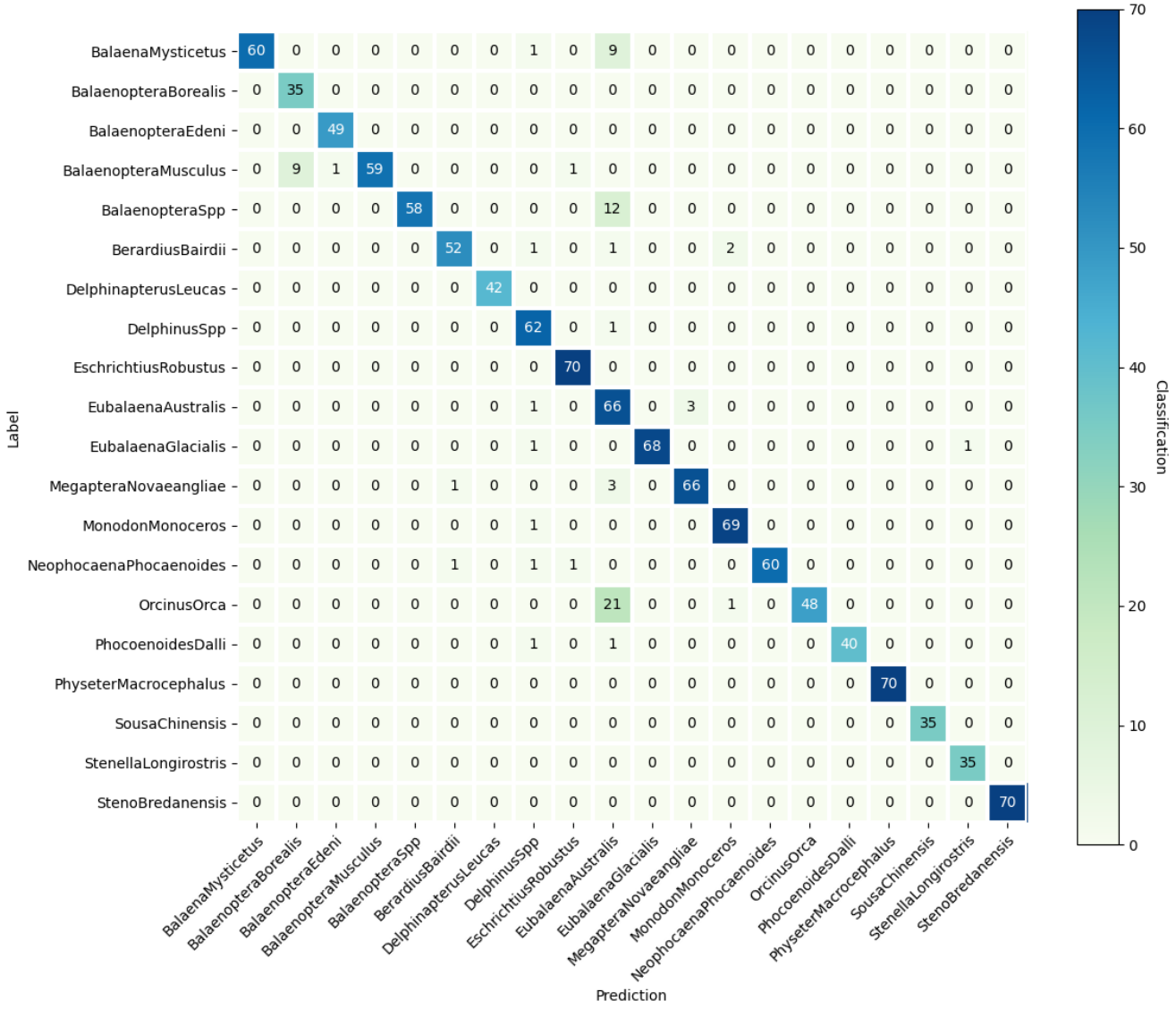


Figure 25: Confusion matrix for whales only dataset with 3-second splice

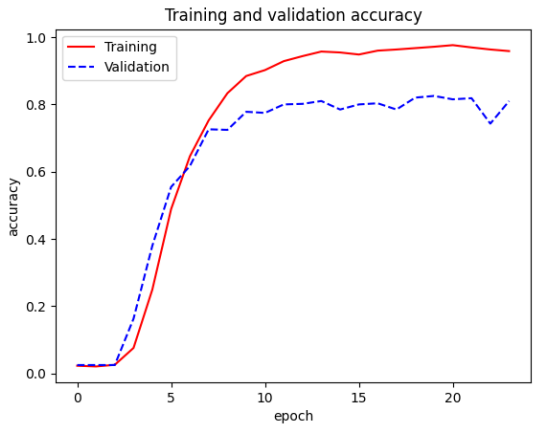
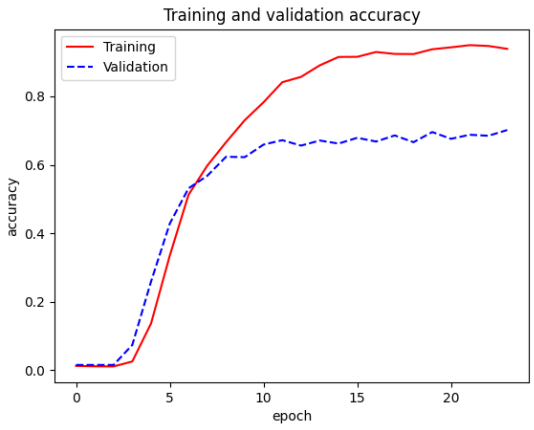


Figure 26: Accuracy for dataset with 3-second splice



a

Figure 27: Accuracy for dataset with 1-second splice

training set for every tested variant. In most cases, the classification on the test set had an accuracy above 80%, which dropped to 60% in the worst case.

Table 3: Categories

Category	Samples	Category	Samples
EubalaenaAustralis	52	OutboardMotor	52
DelphinusSpp	52	PhyseterMacrocephalus	52
IceCracking	35	PhocaVitulina	52
HydrothermalVents	52	MegapteraNovaeangliae	52
OmmatophocaRossii	52	CallorhinusUrsinus	52
Earthquake	52	EubalaenaGlacialis	52
NeophocaenaPhocaenoides	52	BalaenopteraSpp	52
BalaenaMysticetus	52	BerardiusBairdii	52
BalaenopteraMusculus	52	LeptonychotesWeddelli	52
OrcinusOrca	52	OdobenusRosmarus	52
Torpedo	52	CystophoraCristata	52
HydrurgaLeptonyx	52	BalaenopteraEdeni	49
MonodonMonoceros	52	MonachusSchauinslandi	52
WindTurbine	52	PersonalWaterCraft	42
StenellaLongirostris	35	HemisquillaCaliforniensiss	52
Hurricane	52	Dredging	52
DelphinapterusLeucas	42	SousaChinensis	35
EschrichtiusRobustus	52	VolcanicEruption	52
PhocoenoidesDalli	42	BubbleCurtain	52
ZalophusCalifornianus	35	TidalTurbine	52
StenoBredanensis	52	BalaenopteraBorealis	35

Table 4: Classes

Category	Samples	Category	Samples
EubalaenaAustralis	52	OutboardMotor	52
DelphinusSpp	52	PhyseterMacrocephalus	52
Explosive	49	IceCracking	52
PhocaVitulina	52	HydrothermalVents	52
MegapteraNovaeangliae	52	OmmatophocaRossii	52
GlobicephalaSpp	52	CallorhinusUrsinus	52
Earthquake	52	StenellaAttenuata	52
EubalaenaGlacialis	52	NeophocaenaPhocaenoides	52
BalaenopteraSpp	52	Icebreaker	52
BalaenaMysticetus	52	BerardiusBairdii	52
BalaenopteraMusculus	52	EvechinusChloroticus	52
LeptonychotesWeddelli	52	OrcinusOrca	52
OdobenusRosmarus	52	LobodonCarcinophaga	52
UnderwaterBreathingApparatus	52	Torpedo	52
CystophoraCristata	52	HydrurgaLeptonyx	52
BalaenopteraEdeni	52	MonodonMonoceros	52
LipotesVexillifer	52	MonachusSchauinslandi	52
WindTurbine	52	ATOC	52
BalaenopteraPhysalus	52	PeponocephalaElectra	52
SURTASS	52	PersonalWaterCraft	52
LagenorhynchusSpp	35	StenellaLongirostris	52
HemisquillaCaliforniensiss	52	Hurricane	52
Dredging	52	PhocoenaPhocoena	52
PhocaHispida	52	Airgun	52
GrampusGriseus	52	HistriophocaFasciata	52
DelphinapterusLeucas	52	SousaChinensis	52
EschrichtiusRobustus	52	AcousticTomography	52
VolcanicEruption	52	Lightning	52
TursiopsTruncatus	52	SousaSahulensis	52
PhocoenoidesDalli	52	BubbleCurtain	52
ErignathusBarbatus	52	PalinurusSp	35
ZalophusCalifornianus	52	TidalTurbine	52
IniaGeoffrensis	52	StenoBredanensis	52
AlpheusHeterochaelis	52	BalaenopteraBorealis	52

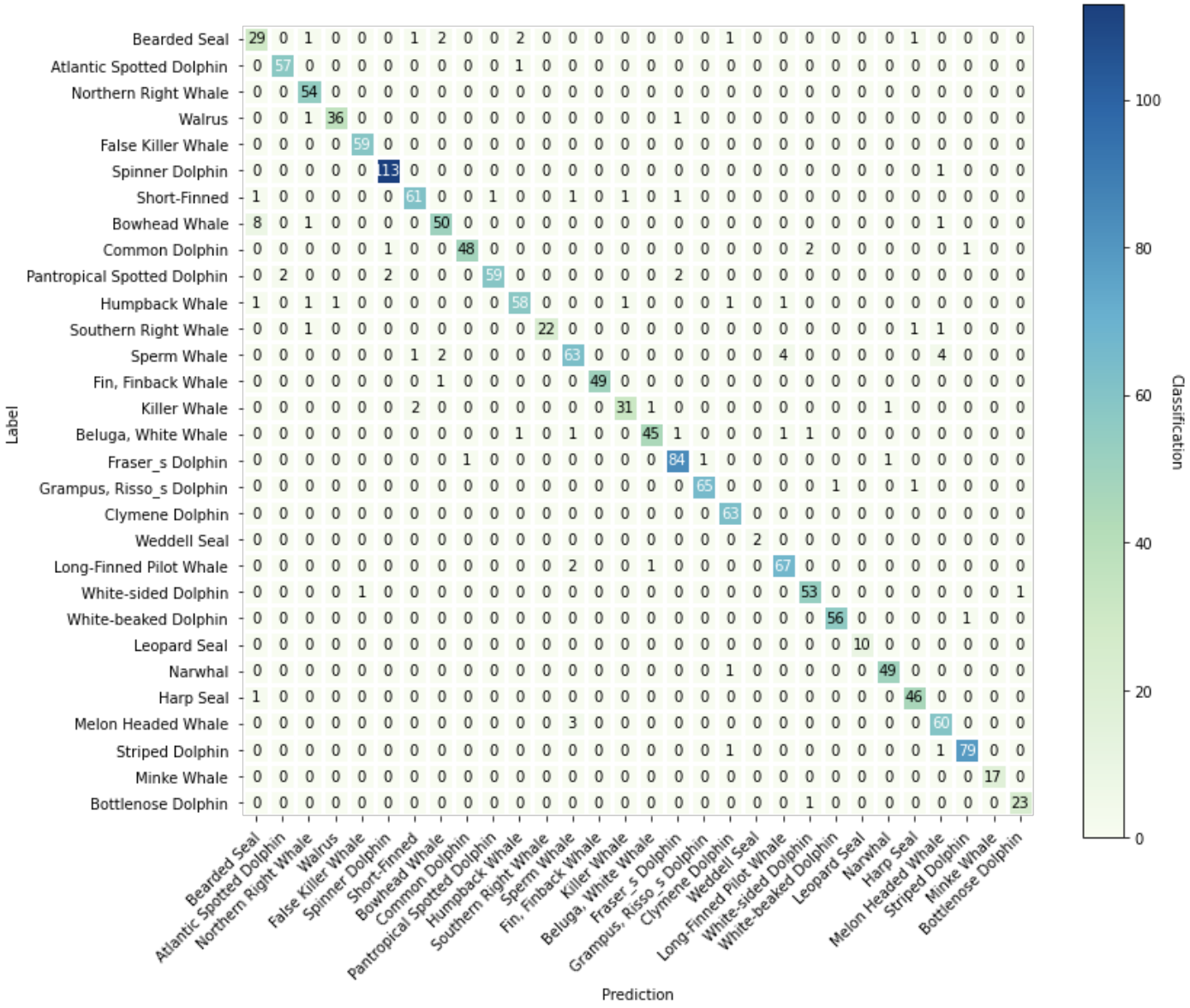


Figure 28: Confusion matrix with predictions for the whole data set

11. Contributions

This study presents several notable contributions, delineated as follows:

1. The first contribution establishes empirical evidence to support the notion that the downsampling of audio recordings to 44.1 kHz, a standard representative of the Audio CD quality, does not render the recordings unsuitable for classification purposes. It has been verified that such downsampling does not undermine the viability of classification endeavors.
2. The second contribution demonstrates the efficacy of splicing and normalization techniques employed on the recordings, thereby yielding uniformly segmented data that remains amenable to classification across the spectrum of categories explored within the purview of this investigation. These techniques have proven to sustain the integrity of the data despite the processing applied.
3. A pivotal contribution materialized in the form of a meticulously devised software pipeline, meticulously crafted to facilitate the generation of the aforementioned datasets. This pipeline has been made publicly accessible, engendering an invaluable resource for fellow researchers in the field, augmenting the collective progress and fostering further exploration.
4. Furthermore, an innovative model founded on convolutional neural networks has been engendered to classify marine audio. Rigorous scrutiny and validation have ensued, affirming the model's commendable performance and boasting an accuracy surpassing 90%

References

[1] M. Southworth, "The sonic environment of cities," *Environment and Behavior*, vol. 1, 1969.

[2] R. M. Schafer, *The soundscape : our sonic environment and the tuning of the world*. Destiny Books, 1994.

[3] E. D. Franco, P. Pierson, L. D. Iorio, A. Calò, J. M. Cottalorda, B. Derijard, A. D. Franco, A. Galvé, M. Guibbolini, J. Lebrun, F. Micheli, F. Priouzeau, C. R. de Faverney, F. Rossi, C. Sabourault, G. Spennato, P. Verrando, and P. Guidetti, "Effects of marine noise pollution on mediterranean fishes and invertebrates: A review," *Marine Pollution Bulletin*, vol. 159, 2020.

[4] E. Spanier and D. Zviely, "Key environmental impacts along the mediterranean coast of israel in the last 100 years," 2023.

[5] A. D. Foote, Y. Liu, G. W. Thomas, T. Vinař, J. Alföldi, J. Deng, S. Dugan, C. E. V. Elk, M. E. Hunter, V. Joshi, Z. Khan, C. Kovar, S. L. Lee, K. Lindblad-Toh, A. Mancia, R. Nielsen, X. Qin, J. Qu, B. J. Raney, N. Vijay, J. B. Wolf, M. W. Hahn, D. M. Muzny, K. C. Worley, M. T. P. Gilbert, and R. A. Gibbs, "Convergent evolution of the genomes of marine mammals," *Nature Genetics*, vol. 47, 2015.

[6] A. Foote, Y. Liu, G. Thomas, T. Vinař, J. Alföldi, J. Deng, S. Dugan, C. Van elk, M. Hunter, V. Joshi, Z. Khan, C. Kovar, S. Lee, K. Lindblad-Toh, A. Mancia, R. Nielsen, X. Qin, J. Qu, B. Raney, and R. Gibbs, "Convergent evolution of the genomes of marine mammals," *Nature Genetics*, 01 2015.

[7] P. J. Clapham, S. B. Young, and R. L. Brownell, "Baleen whales: Conservation issues and the status of the most endangered populations," *Mammal Review*, vol. 29, 1999.

[8] S. Hooker, *Toothed Whales, Overview*, pp. 1173–1179. 12 2009.

[9] M. C. Amorim, R. Vasconcelos, and P. Fonseca, *Fish Sounds and Mate Choice*, vol. 4, pp. 1–33. 03 2015.

[10] I. Charrier, L. Jeantet, L. Maucourt, S. Régis, N. Lecerf, A. Benhalilou, and C. Damien, "First evidence of underwater vocalisations in green sea turtles chelonia mydas," *Endangered Species Research*, vol. 48, pp. 31–41, 05 2022.

[11] A. T. G. D. P. P. Thiebault A, Charrier I, "First evidence of underwater vocalisations in hunting penguins," *PeerJ*, 2019.

[12] Y. ZHANG, F. SHI, J. SONG, X. ZHANG, and S. YU, "Hearing characteristics of cephalopods: Modeling and environmental impact study," *Integrative Zoology*, vol. 10, no. 1, pp. 141–151, 2015.

[13] M. W. Johnson, F. A. Everest, and R. W. Young, "The role of snapping shrimp (crangon and synalpheus) in the production of underwater noise in the sea," *The Biological Bulletin*, vol. 93, no. 2, pp. 122–138, 1947.

[14] M. J. Vermeij, K. L. Marhaver, C. M. Huijbers, I. Nagelkerken, and S. D. Simpson, "Coral larvae move toward reef sounds," *PloS one*, vol. 5, no. 5, p. e10660, 2010.

[15] D. R. MacAyeal, E. A. Okal, R. C. Aster, and J. Bassis, "Seismic and hydroacoustic tremor generated by colliding icebergs," *Journal of Geophysical Research: Earth Surface*, vol. 113, no. F3, 2008.

[16] R. P. Dziak, E. T. Baker, A. M. Shaw, D. R. Bohnenstiehl, W. W. Chadwick, J. H. Haxel, H. Matsumoto, and S. L. Walker, "Flux measurements of explosive degassing using a yearlong hydroacoustic record at an erupting submarine volcano," *Geochemistry, Geophysics, Geosystems*, vol. 13, 2012.

[17] C. E. Nishimura, *Monitoring whales and earthquakes by using SOSUS*. Naval Research Laboratory Washington, DC, USA, 1994.

[18] J. A. Hildebrand, "Impacts of anthropogenic sound," *Marine mammal research: conservation beyond crisis*, pp. 101–124, 2005.

[19] W. M. Zimmer, *Passive acoustic monitoring of cetaceans*. 2011.

[20] I. M. Organization, *Guidelines for the reduction of underwater noise from commercial shipping to address adverse impacts on marine life*, 2014.

[21] K. J. Reine, D. Clarke, and C. Dickerson, "Characterization of underwater sounds produced by hydraulic and mechanical dredging operations," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3280–3294, 2014.

[22] J. C. Wilson, M. Elliott, N. D. Cutts, L. Mander, V. Mendão, R. Perez-Dominguez, and A. Phelps, "Coastal and offshore wind energy generation: Is it environmentally benign?," *Energies*, vol. 3, 2010.

[23] J. Tougaard, P. T. Madsen, and M. Wahlberg, "Underwater noise from construction and operation of offshore wind farms," *Bioacoustics*, vol. 17, 2008.

[24] P. T. Madsen, M. Wahlberg, J. Tougaard, K. Lucke, and P. Tyack, "Wind turbine underwater noise and marine mammals: implications of current knowledge and data needs," *Marine ecology progress series*, vol. 309, pp. 279–295, 2006.

[25] M. K. Pine, P. Schmitt, R. M. Culloch, L. Lieber, and L. T. Kregting, "Providing ecological context to anthropogenic subsea noise: Assessing listening space reductions of marine mammals from tidal energy devices," *Renewable and Sustainable Energy Reviews*, vol. 103, 2019.

[26] L. Palmer, D. Gillespie, J. D. MacAulay, C. E. Sparling, D. J. Russell, and G. D. Hastie, "Harbour porpoise (phocoena phocoena) presence is reduced during tidal turbine operation," *Aquatic Conservation: Marine and Freshwater Ecosystems*, vol. 31, 2021.

[27] G. Scowcroft, C. Knowlton, H. Morin, and D. Mclntire, "Discovery of sound in the sea," *University of Rhode Island*, 2012.

[28] K. J. Vigness-Raposa, G. Scowcroft, J. H. Miller, D. R. Ketten, and A. N. Popper, "Discovery of sound in the sea: Resources for educators, students, the public, and policymakers," 2016.

[29] J. Lynch, "The voices of marine mammals—william e. schevill and william a. watkins: Pioneers in bioacoustics," *The Journal of the Acoustical Society of America*, vol. 148, 2020.

[30] L. Sayigh, M. A. Daher, J. Allen, H. Gordon, K. Joyce, C. Stuhlmann, and P. Tyack, "The watkins marine mammal sound database: An online, freely accessible resource," 2017.

[31] D. Lavry, "Sampling theory for digital audio," *Lavry Engineering, Inc. Available online: http://www. lavryengineering. com/documents/Sampling_Theory. pdf (checked 24.5. 2010)*, 2004.

[32] E. Por, M. van Kooten, and V. Sarkovic, "Nyquist–shannon sampling theorem," *Leiden University*, vol. 1, no. 1, 2019.

[33] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Transactions on Nuclear Science*, vol. 44, 1997.

[34] "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, 2019.

[35] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, 2020.

[36] E. M. Stein and R. Shakarchi, *Fourier analysis: an introduction*, vol. 1. Princeton University Press, 2011.

[37] H. Medwin and C. S. Clay, *Fundamentals of Acoustical Oceanography*. Academic Press, San Diego, 1998.

[38] R. Mersereau and T. Speake, "A unified treatment of cooley-tukey algorithms for the evaluation of the multidimensional dft," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 5, pp. 1011–1018, 1981.

[39] W. Ellermeier, M. Eigenstetter, and K. Zimmer, "Psychoacoustic correlates of individual noise sensitivity," *The journal of the acoustical society of America*, vol. 109, no. 4, pp. 1464–1473, 2001.

[40] K. Rangra and M. Kapoor, "Exploring the mel scale features using supervised learning classifiers for emotion classification," *In-*

ternational Journal of Applied Pattern Recognition, vol. 6, no. 3, pp. 232–253, 2021.

[41] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, 1958.

[42] S. Osowski, *Sieci Neuronowe.* Oficyna Wydawnicza Politechniki Warszawskiej, 1994.

[43] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 1943.

[44] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 7 2012.

[45] "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[46] *Tensorflow API Documentation for Python. [Online].* `https://www.tensorflow.org/api_docs/python/tf`.

[47] A. Géron, *Hands-On Machine Learing With Scikit-Learn & Tensor Flow.* 2017.

[48] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[49] J. H. Steele, S. A. Thorpe, and K. K. Turekian, *Encyclopedia of ocean sciences.* Elsevier ScienceDirect, 2008.

[50] R. Tadeusiewicz, *Problemy Biocybernetyki.* 1994.

[51] F. e. a. Pedregosa, "Scikit learn: Machine learning in Python," vol. 12, pp. 2825–2830, 2011.

[52] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," 2017.

[53] D. Gillespie, D. Mellinger, J. Gordon, D. Mclaren, P. Redmond, R. McHugh, P. Trinder, X.-Y. Deng, and A. Thode, "Pamguard: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans.," *The Journal of the Acoustical Society of America*, vol. 125, p. 2547, 05 2009.

[54] T. P. Zieliński, *Cyfrowe przetwarzanie sygnałów - od teorii do zastosowań.* 2 ed., 2005.

[55] H. A. Ghani, M. R. A. Malek, M. F. K. Azmi, M. J. Muril, and A. Azizan, "A review on sparse fast fourier transform applications in image processing," 2020.

[56] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, 1937.

[57] J. A. Hildebrand, "Anthropogenic and natural sources of ambient noise in the ocean," *Marine Ecology Progress Series*, vol. 395, 2009.

[58] I. Herrera, M. Carrillo, M. C. de Esteban, and R. Haroun, "Distribution of cetaceans in the canary islands (northeast atlantic ocean): Implications for the natura 2000 network and future conservation measures," *Frontiers in Marine Science*, vol. 8, 2021.

[59] E. C. M. Parsons, "The negative impacts of whale-watching," *Journal of Marine Biology*, vol. 2012, 2012.

[60] S. Baulch and C. Perry, "Evaluating the impacts of marine debris on cetaceans," *Marine Pollution Bulletin*, vol. 80, 2014.

[61] A. Maglio, C. Soares, M. Bouzidi, Y. zabel, F. Souami, and G. Pavan, "Mapping shipping noise in the pelagos sanctuary (french part) through acoustic modelling to assess potential impacts on marine mammals.," *Sci. Rep. Port-Cros National Park*, pp. 167–185.

[62] M. Randone, M. Bocci, C. Castellani, C. Laurent, and C. Piante, "Safeguarding marine protected areas in the growing mediterranean blue economy—recommendations for the maritime transport sector," *International Journal of Design and Nature and Ecodynamics*, vol. Volume 14 (2019), Issue 4, no. 9/2020, p. 10.

[63] W. M. X. Zimmer, *Passive Acoustic Monitoring of Cetaceans.* Cambridge University Press, 2021.

[64] I. Herrera, M. Carrillo, and R. Haroun, "Conservación de cetáceos y planificación del espacio marino en las islas canarias," *Okeanos.*

[65] V. Sessions and M. Valtorta, "The effects of data quality on machine learning algorithms," 2006.

[66] P. Gnyś, "Mereogeometry based approach for behavioral robotics," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 70–80, 2017.

[67] P. Gnyś, "Mereogeometry based approach for behavioral robotics," pp. 70–80, 2017.

[68] G. P and P. P, "Application of long short term memory neural networks for gps satellite clock bias prediction," *TASK Quarterly*, vol. 25, pp. 381–395, 2021.

[69] J. Cabrera-Gámez, A. C. Domínguez-Brito, F. Santana-Jorge, D. Gamo, D. Jiménez, A. Guerra, and J. J. Castro, "Acoustic detection of tagged angelsharks from an autonomous sailboat," vol. 1092 AISC, 2020.

[70] W. W. Chadwick, R. W. Embley, E. T. Baker, J. A. Resing, J. E. Lupton, K. V. Cashman, R. P. Dziak, V. Tunnicliffe, D. A. Butterfield, and Y. Tamura, "Spotlight 10: Northwest rota-1 seamount," *Oceanography*, vol. 23, 2010.

[71] W. W. Chadwick, K. V. Cashman, R. W. Embley, H. Matsumoto, R. P. Dziak, C. E. de Ronde, T. K. Lau, N. D. Deardorff, and S. G. Merle, "Direct video and hydrophone observations of submarine explosive eruptions at nw rota-1 volcano, mariana arc," *Journal of Geophysical Research: Solid Earth*, vol. 113, 2008.

[72] W. Europe, "Offshore statistics in europe."

[73] D. Risch, N. van Geel, D. Gillespie, and B. Wilson, "Characterisation of underwater operational sound of a tidal stream turbine," *The Journal of the Acoustical Society of America*, vol. 147, 2020.

[74] M. J. Kennish, *Practical handbook of marine science.* 2019.

[75] P. Clapham, "Why do baleen whales migrate?," *Marine Mammal Science*, vol. 17, pp. 432–436, 04 2001.

[76] H. Medwin, C. S. Clay, and S. M. Flatte, "Fundamentals of acoustical oceanography," *Physics Today*, vol. 52, 1999.

[77] H. Urban, *Handbook of Underwater Acoustic Engineering.* STN-Atlas-Elektronik GmbH, 2002.

[78] J. D. Macaulay and D. Gillespie, "Pamguard: Open-source detection, classification, and localization software," *The Journal of the Acoustical Society of America*, vol. 151, 2022.

[79] T. Webber, D. Gillespie, T. Lewis, J. Gordon, T. Ruchirabha, and K. F. Thompson, "Streamlining analysis methods for large acoustic surveys using automatic detectors with operator validation," *Methods in Ecology and Evolution*, vol. 13, 2022.

[80] B. N. Korkmaz, R. Diamant, G. Danino, and A. Testolin, "Automated detection of dolphin whistles with convolutional networks and transfer learning," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[81] Microsoft, "Multimedia programming interface and data specifications 1.0," *International Business*, 1991.

[82] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models.* 2007.

[83] K. S. Norris, R. M. Goodman, B. Villa-Ramirez, and L. Hobbs, "Behavior of california gray whale, eschrichtius robustus, in southern baja california, mexico," *Fish. Bull*, vol. 75, no. 1, pp. 159–72, 1977.

[84] C. J. Deutsch, J. P. Reid, R. K. Bonde, D. E. Easton, H. I. Kochman, and T. J. O'Shea, "Seasonal movements, migratory behavior, and site fidelity of west indian manatees along the atlantic coast of the united states," *Wildlife monographs*, pp. 1–77, 2003.

[85] *Sub-Committee on Ship Design and Construction (SDC 8), 17-21 January 2022*, 2022.

[86] E. Chou, B. L. Southall, M. Robards, and H. C. Rosenbaum, "International policy, recommendations, actions and mitigation efforts of anthropogenic underwater noise," *Ocean & Coastal Management*, vol. 202, p. 105427, 2021.

[87] J. Gordon, D. Thompson, D. Gillespie, M. Lonergan, S. Calderan, B. Jaffey, and V. Todd, "Assessment of the potential for acoustic deterrents to mitigate the impact on marine mammals of underwater noise arising from the construction of offshore windfarms," *Cowrie Ltd, July*, 2007.