Universiteit
Leiden
The Netherlands

Data Science & Artificial Intelligence

Evaluating the Impact of Pre-processing Choices on CNN-Based Dolphin Vocalization Classification in Passive Acoustic Monitoring

Dalia Kamalzadeh

1st Supervisor: Dr. Erwin M. Bakker
2nd Supervisor: Prof. Dr. Michael S.K. Lew

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

January 27, 2026

Abstract

Passive Acoustic Monitoring (PAM) enables non-invasive, long-term observation of marine mammals through underwater sound recordings. However, automated classification of dolphin species from PAM data remains challenging due to high acoustic variability, limited labeled datasets, and species-specific vocalization characteristics. Convolutional neural networks (CNNs), such as those used in the CetusID Framework proposed by Frainer et al. [16], have shown promise for this task. However, here the influence of signal-level and representation-level pre-processing choices remained underexposed.

In our work, we adapt the CetusID framework to investigate the impact of pre-processing choices on CNN-based (CNN2) dolphin species classification. Specifically, the effects of audio sampling rate, data augmentation, and signal-level band-pass pre-processing are evaluated. The analysis is restricted to three species (Delphinus delphis, Sousa plumbea, and Tursiops aduncus) and uses leave-one-recording-out (LORO) cross-validation on vocalizations from these species. Furthermore, the baseline CNN2 is compared with EfficientNetB0 using transfer-learning.

The results show that increasing the sampling rate yields a higher accuracy, with performance differences observed across species and recording conditions within the evaluated dataset. In this study, data augmentation did not improve accuracy. Signal-level band-pass pre-processing does not improve performance for Frainer et al's CNN2, but substantially increases the accuracy of the EfficientNetB0 model. The highest LORO validation performance is achieved by the EfficientNetB0 model combined with signal-level band-pass filtering and no data augmentation, reaching a mean validation accuracy of 0.849 ± 0.088 under LORO cross-validation.

Contents

Chapter 1

Introduction

Passive Acoustic Monitoring (PAM) has proven to be an effective technique for studying marine mammals [54]. To capture underwater soundscapes over extended periods, researchers use PAM to deploy hydrophones on fixed moorings or on the seafloor. As dolphins rely extensively on acoustic communication and echolocation, PAM provides a non-invasive and well-suited approach for monitoring their presence and activity [52, 44]. However, automated analysis of PAM data remains challenging [36, 49]. Marine soundscapes are acoustically complex due to variability introduced by both biotic factors, such as vocalizations from multiple species, and abiotic factors, including wind, rain, wave action, and ship noise [12]. Dolphin vocalizations are often short and included in highly variable background noise, complicating automated detection and classification [10].

Convolutional neural networks (CNNs) have demonstrated strong performance in bio-acoustic classification tasks [37, 26, 49] and are increasingly used for PAM-based dolphin monitoring. Despite this, CNN-based pipelines remain sensitive to the way acoustic signals are conditioned and represented prior to learning. In particular, pre-processing and representation choices, such as audio sampling rate, frequency filtering, and data augmentation, are often inherited from legacy speech processing conventions rather than systematically evaluated for bio-acoustic tasks. Research indicates that these "fixed" parameters can significantly impact classification accuracy, yet they are often used without optimizing performance across diverse species and acoustic environments [27, 49]. As a result, it remains unclear which pre-processing choices meaningfully improve classification performance, and which instead introduce unintended biases or instability under data-limited PAM conditions [32, 38].

Dolphins are highly vocal animals that produce a variety of different sounds, including whistles, echolocation clicks, and burst-pulsed calls [22]. Whistles and burst-pulsed calls are primarily associated with social communication, whereas echolocation clicks are broadband signals used for navigation and prey detection [22, 3, 24]. Although these vocalizations contain species-specific acoustic structure, their variability in duration, frequency range, and usage context complicates automated classification [2]. In natural soundscapes, dolphin calls frequently overlap with background noise and other biological signals [40, 12], motivating the use of time–frequency representations and learning-based methods that can capture complex spectro-temporal structure.

Prior to the use of deep learning approaches, automated dolphin vocalization analysis relied on rule-based detectors, handcrafted features, and shallow classifiers. While effective under controlled acoustic conditions, these approaches often degrade in noisy or variable environments [53]. This limitation has driven a transition toward deep learning models, particularly CNNs, which learn discriminative representations directly from spectrogram inputs. This raises the question of whether performance gains attributed to CNN architectures instead arise from unexamined pre-processing assumptions.

Frainer et al. [16] introduced the CetusID framework [15], a two-stage CNN-based system for automated analysis of dolphin vocalizations. In this framework, CNN1 performs binary detection of dolphin vocalizations, after which CNN2 assigns detected segments to predefined dolphin taxa. The publicly released implementation focuses on classifying three subspecies: Delphinus delphis, Sousa plumbea, Tursiops aduncus, and Orcinus orca. Frainer et al. demonstrated that representation-level parameters such as spectrogram resolution and window length significantly influence CNN performance. However, all recordings were standardized to a fixed sampling rate of 96 kHz, and no signal-level frequency filtering

was applied prior to spectrogram generation. While background-mixing augmentation was used during training, its isolated effect on classification accuracy and robustness was not systematically evaluated.

Frainer et al. used 723 minutes of annotated PAM recordings from South African coastal waters, with Orcinus orca representing approximately 1% of the dataset, and therefore reported both four-class and reduced three-class identification results. Their results show that CNN2 performance decreases when Orcinus orca is included, with the most stable and accurate models obtained in the two- and three-class configurations excluding this species. In this study, the dataset is more limited, as Frainer et al. did not release the full dataset publicly. Only a public subset is available for reproducibility, rather than the full 723-minute collection [15]. The focus of this thesis is on analyzing pre-processing effects rather than expanding species coverage; therefore, only the three-species configuration (D. delphis, S. plumbea, T. aduncus) was considered, and the fourth class was not evaluated.

Consequently, low-frequency regions of the spectrogram that were dominated by recording artifacts and background noise were retained as possible model inputs. Spectrogram inspection suggests that such non-vocal frequency regions can differ systematically across recordings and species [43]. When preserved, CNNs may exploit these regions as shortcut cues correlated with recording conditions rather than biologically meaningful vocalization structure [17, 49].

In this thesis, a signal-level band-pass pre-processing stage is introduced and evaluated prior to spectrogram generation to examine how removing background-dominated frequency regions affects CNN-based dolphin species classification under different model architectures. In particular, band-pass filtering is examined, which attenuates background-dominated components while limiting the signal to frequency ranges of dolphin vocalizations. Such filtering is commonly used in classical bio-acoustic analysis [4], but its effect on CNN-based species classification has not been evaluated in Fraier et al's paper [16].

The aim of this thesis is therefore to analyze the sensitivity of CNN-based dolphin species classification to the proposed signal-level pre-processing and representation-level design choices using the publicly available CetusID demo dataset, which contains annotated PAM recordings for three dolphin species (Delphinus delphis, Sousa plumbea, and Tursiops aduncus). Specifically, this work evaluates the effects of band-pass filtering, audio sampling rate, and data augmentation using experiments conducted on a reproduced CetusID demo baseline. The analysis focuses exclusively on the species classification network (CNN2); the detection network (CNN1) is validated only to confirm correct baseline behavior and is not further analyzed. Finally, a transfer learning approach using EfficientNetB0 has been taken and compared to CNN2. All experiments use leave-one-recording-out (LORO) cross-validation to account for recording-level variability and prevent segment-level data leakage.

Contributions

The main contributions of this thesis are as follows:

- Validation of the publicly available CetusID implementation, confirming correct baseline behavior of the CNN2 model using the default demo train–validation split, and establishing leave-one-recording-out cross-validation as the evaluation protocol for all subsequent experiments.
- A quantitative analysis of the effect of audio sampling rate on CNN2 classification performance, indicating small and species-dependent differences across sampling rates of 24, 48, and 96 kHz.
- An evaluation of data augmentation strategies, showing no performance improvement for CNN2 on the CetusID demo dataset and indicating class-dependent performance changes under an artificially noise-degraded dataset.
- An analysis of signal-level band-pass filtering prior to spectrogram generation, showing degraded performance for the baseline CNN2 model and improved performance, in this study, when combined with EfficientNetB0 using transfer-learning.

Research Questions

The main research question and subsequent sub-research questions of this study are formulated as follows:

MRQ.1: How do signal-level band-pass filtering, audio sampling rate, and data augmentation influence the performance of CNN-based dolphin species classification from PAM audio segments?

SRQ.1: How does applying band-pass filtering affect CNN2's classification performance and class-wise behavior under leave-one-recording-out cross-validation?

SRQ.2: How does changing the audio sampling rate impact CNN2's classification performance and stability across leave-one-recording-out cross-validation?

SRQ.3: How does data augmentation influence CNN2 classification performance under controlled signal degradation introduced by artificial noise?

The remainder of this thesis is structured as follows. Chapter 2 reviews related work on automated dolphin vocalization detection and classification, with emphasis on CNN-based approaches and the CetusID framework. Chapter 3 presents the theoretical and technical foundations relevant to this study, including cetacean acoustics, PAM, digital audio representations, spectrogram generation, data augmentation, and CNNs. Chapter 4 describes the datasets used in this thesis and clarifies the distinction between the full CetusID dataset and the reduced SA-Acoustics dataset used for experimentation. Chapter 5 introduces the baseline CetusID method and validates the publicly available demo implementation, which serves as a reference in this work. Chapter 6 details the experimental methodology, including recording-level cross-validation, audio segmentation, pre-processing pipelines, sampling-rate manipulation, data augmentation strategies, signal-level band-pass filtering, and model comparison. Chapter 7 presents the experimental results addressing the defined research questions. Chapter 8 discusses these results and their implications, including limitations and methodological considerations. Finally, Chapter 9 concludes the thesis with a summary of findings, contributions, and directions for future research. Additional analyses and supplementary results are provided in the appendices.

Chapter 2

Related Work

This chapter reviews prior work on automated detection and classification of dolphin vocalizations in PAM. The discussion progresses from traditional signal-processing approaches to contemporary CNN-based methods, with emphasis on how signal processing and representation choices are treated in existing literature. This review focuses on pre-processing, representation, and robustness factors relevant to CNN-based dolphin categorization, rather than being comprehensive. Particular attention is given to identifying gaps in prior work that motivate the experimental design of this thesis.

2.1 Traditional Signal Processing Approaches

Prior to the use of deep learning, automated cetacean vocalization analysis relied primarily on hand-crafted signal-processing techniques. Early studies focused on matched filtering, spectrogram correlation, and energy-based detectors to identify stereotyped calls within large PAM datasets [35, 48]. These methods exploited prior knowledge of call structure and were effective for species with highly regular vocalizations, such as blue and fin whales.

However, the frequency modulation, duration, and call structure of dolphin vocalizations vary significantly more, making template-based detection less successful [2]. To address this variability, image-processing techniques operating on spectrogram representations were introduced, including edge-detection methods designed to isolate tonal contours from background noise [18]. While these approaches improved detection of less stereotyped calls, they remained sensitive to noise, threshold selection, and recording conditions, limiting generalization across environments and species.

These limitations motivated a transition toward data-driven learning approaches. At the same time, they highlight the continued importance of pre-processing and signal conditioning, as both traditional and learning-based methods remain sensitive to how acoustic information is presented under noisy and variable PAM conditions.

2.2 CNN-based Methods for Dolphin Vocalization Classification

CNNs have become one of the dominant approaches for bio-acoustic classification due to their ability to learn hierarchical spectro-temporal representations directly from time–frequency inputs. Early work by Piczak [43] demonstrated that CNNs applied to spectrograms outperform traditional feature-based classifiers for environmental sound classification, particularly under variable noise conditions. Since then, this approach has matured into the dominant methodology in the field, as noted in recent comprehensive reviews of deep learning in bio-acoustics [49, 6]

Subsequent studies extended CNN-based approaches across a wide range of bio-acoustic domains, including marine mammals and birds. CNNs have been successfully applied to classify dolphin whistles and clicks [16, 10], as well as vocalizations of laying hens and cattle [25]. These cross-species applications demonstrate that CNNs provide a flexible end-to-end framework capable of learning complex acoustic structure across taxa.

Importantly, while CNN architectures are often reused across studies, signal processing and representation pipelines are frequently inherited from prior work and treated as fixed design choices. Parameters such as sampling rate, window length, spectrogram resolution, and data augmentation are rarely isolated and evaluated, despite their direct influence on the information presented to the network [13]. As a result, performance improvements reported in the literature are often difficult to separate from implicit pre-processing assumptions.

2.2.1 CetusID Framework by Frainer et al.

In 2023, Frainer et al. [16] introduced the CetusID framework, a two-stage CNN system designed for automated dolphin vocalization detection (CNN1) and species-level taxonomic identification (CNN2). The framework operates on spectrogram representations generated from short audio segments. CNN2 is used as the baseline; the pipeline and architecture of CNN2 are shown in Chapter 6.

In their study, Frainer et al. investigated several representation-level parameters, including spectrogram resolution and window length, and demonstrated that these choices influence classification performance. However, all recordings were standardized to a fixed sampling rate of 96 kHz, and the effects of alternative sampling rates were not examined. Likewise, while background-mixing augmentation was used to increase data diversity, its isolated impact on classification accuracy and controlled signal degradation was not systematically evaluated. Notably, no signal-level frequency filtering was applied prior to spectrogram generation, leaving low-frequency regions dominated by background noise and recording artefacts fully available to the CNN.

These omissions are significant because the sampling rate directly constrains the maximum representable frequency content, while data augmentation alters the effective training distribution and the model's robustness. Additionally, low-frequency spectrogram regions may further contain recording-specific artefacts rather than biologically meaningful vocalization structure. Without explicit analysis of these factors, it remains unclear whether reported performance gains generalize across recording conditions, species, or acquisition settings. This study attempts to address several aspects of the CetusID framework and to propose some improvements.

2.3 Datasets Used in Dolphin Research

The advancement of deep learning in marine bio-acoustics is constrained by the availability of labeled datasets. While PAM generates large volumes of raw audio, most recordings remain unlabeled or are stored within private institutional archives. For independent researchers, access to labeled dolphin vocalization data often requires institutional collaboration or closed access.

Many datasets cited in the cetacean literature are not publicly downloadable. The Sarasota Dolphin Whistle Database (SDWD), one of the most comprehensive collections of signature whistles with over 926 recording sessions of 293 individuals, is an institutional resource with restricted access [46]. In contrast, several smaller datasets are publicly available through open repositories. The SEANOE platform hosts freely downloadable dolphin vocalization datasets used in ecological and acoustic studies [47], while Figshare provides multiple open datasets, including bottlenose dolphin whistle recordings from the Adriatic Sea and Fremantle Inner Harbour [9, 33].

Even when data is accessible, the volume of labeled segments is often insufficient for training modern high-capacity architectures. For example, a dataset considered large in dolphin research may consist of fewer than 100 minutes of audio and several hundred annotated whistles [9]. While Frainer et al. released their code for reproducibility, only a small demo dataset was made publicly available, limiting independent experimentation. More recently, Lenhoff et al. [31] collected over 400 minutes of PAM recordings of Delphinus delphis, supplemented with detailed annotations, but this dataset has not yet been released publicly.

Overall, dolphin bio-acoustic research relies on a limited number of high-quality institutional datasets supplemented by small public repositories with heterogeneous annotation standards. These constraints hinder reproducibility, increase variance in experimental evaluation, and amplify sensitivity to pre-processing and data partitioning choices, motivating careful experimental design and conservative interpretation.

Frainer et al. [16] trained and evaluated CetusID on 723 minutes of annotated PAM recordings from South African coastal waters, of which only a limited subset is publicly available. This thesis uses

the publicly released CetusID demo dataset, which, despite its limited size, provides consistently annotated recordings, enabling a controlled comparison focused on isolating pre-processing effects rather than claiming population-level generalization.

2.4 The Importance of Pre-processing and Data Augmentation on Classification

CNN performance in bio-acoustics is influenced by signal conditioning and the diversity of training data. Salamon et al. [45] demonstrated that deep models trained on small labeled datasets are prone to overfitting in the absence of data augmentation. By applying label-preserving transformations such as pitch shifting, time stretching, and background noise mixing, models are encouraged to learn generalized spectro-temporal structure rather than memorizing dataset-specific noise patterns. These principles are directly applicable to dolphin vocalization classification, where labeled data is limited. In this thesis, data augmentation was restricted to background noise mixing, as this was the augmentation strategy used in the CetusID framework by Frainer et al. [16]. Additional techniques, such as pitch shifting or time stretching, were not applied in order to maintain methodological consistency with the baseline and to avoid introducing transformations that could alter biologically relevant temporal or spectral characteristics of dolphin vocalizations [24] in a data-limited setting.

Similarly, Piczak [43] also explores the role of data augmentation in training high-capacity deep models on small bio-acoustic and environmental datasets. By artificially expanding the limited dataset, the model was forced to generalize beyond the specific noise profiles and temporal alignments of the original recordings, ultimately yielding performance gains comparable to human-level classification.

The influence of sampling frequency on the analysis of dolphin vocalizations has been explicitly investigated in bio-acoustic studies. In the paper by Papale et al. [41], they show that higher sampling frequencies preserve fine temporal and spectral characteristics of short-beaked common dolphin burst pulses, while downsampling progressively removes high-frequency components and alters measured acoustic features. Frainer et al. [16] recorded their PAM data at a sampling rate of 96 kHz, which is retained in the publicly available CetusID demo dataset [15]. In this thesis, the original 96 kHz recordings are downsampled to 48 kHz and 24 kHz, following a similar methodological approach, in order to explicitly evaluate how sampling-rate-induced information loss affects CNN-based dolphin species classification.

In contrast to sampling rate and data augmentation, signal-level frequency filtering is less commonly discussed in CNN-based classification studies of dolphin vocalizations. In PAM recordings, low-frequency spectrogram regions are often dominated by environmental noise sources such as wind, waves, and shipping activity rather than cetacean vocalizations [36, 23]. Consequently, traditional bio-acoustic analyses frequently apply band-pass filtering to suppress low-frequency background noise and improve the detectability of cetacean vocalizations in spectrogram-based analyses [34]. In contrast, these studies focus on signal detectability and manual or rule-based classification, whereas this thesis explicitly evaluates how such filtering affects learned feature representations and classification performance in CNN-based dolphin species identification. From a machine learning perspective, retaining background-dominated regions may introduce shortcut features correlated with recording conditions rather than vocalization characteristics, which deep classifiers are known to exploit [17]. Despite its widespread use in classical dolphin bio-acoustics, band-pass filtering is not applied or evaluated in existing CNN-based dolphin species classification studies reviewed in this thesis, including the CetusID framework proposed by Frainer et al. [16].

Given these recent advances, the combined influence of signal-level filtering, audio sampling rate, and data augmentation on CNN-based dolphin classification remains insufficiently documented, particularly under data-limited PAM conditions. This thesis addresses these gaps through controlled experiments conducted on a reproduced CetusID demo baseline, with explicit focus on pre-processing sensitivity and robustness.

Chapter 3

Fundamentals

This chapter provides conceptual background for the methods used in this thesis. It introduces the acoustic properties of cetacean vocalizations, the principles of PAM, digital audio representation, spectrogram generation, data augmentation, and CNNs [36, 54]. The purpose of this chapter is to establish the theoretical and technical foundation required to understand the experimental methodology described in Chapter 6. This chapter does not evaluate performance, justify experimental design choices, or interpret results, which are addressed in subsequent chapters.

3.1 Cetacean Acoustics

Cetaceans produce a wide range of acoustic signals that play essential roles in communication, navigation, and social interaction [22, 24]. Dolphins primarily emit whistles, echolocation clicks, and burst-pulsed sounds. Whistles are narrowband, frequency-modulated signals used mainly for social communication, coordination, and individual recognition. In several species, whistles function as individually distinctive signature calls [24, 46]. Reported whistle frequencies typically range between approximately 1 and 20 kHz, although substantial inter-species variability exists, with some species producing whistles extending beyond 40 kHz, as summarized in Table 3.1 [2].

Although some whistle energy can extend into the lower kilohertz range, the dominant frequency modulation patterns and harmonics used for species discrimination are typically concentrated above 2 kHz, while frequencies below this range are often dominated by low-frequency environmental noise [51]. Consequently, suppressing the 0–2 kHz band is expected to reduce background interference without removing the primary spectro-temporal cues used for classification [40].

Table 3.1: Reported whistle frequency ranges for dolphin species used in this thesis

Species	**Typical Range (kHz)**	**Upper Components (kHz)**	**References**
Delphinus delphis	5.0–20.0	48.0–50.0	[19, 42]
Sousa plumbea	2.3–16.0	33.0–35.0	[7, 11]
Tursiops aduncus	1.1–18.4	∼20.0	[50, 20]

Upper frequency components indicate reported maxima or harmonic content rather than dominant whistle energy, which remains concentrated within the typical frequency range. These reported frequency ranges motivate the use of 96 kHz recordings in the CetusID dataset [15] and provide a biological basis for evaluating the impact of downsampling to 48 kHz and 24 kHz in this thesis.

Echolocation clicks are short-duration, broadband signals used by odontocetes to locate and classify objects in their environment through echo analysis [3]. These clicks typically occupy higher-frequency bands, often between 40 and 130 kHz [3], and require high sampling rates to capture fine temporal detail. In addition to whistles and clicks, dolphins also produce burst-pulsed sounds that span a wide frequency range and serve communicative or behavioral functions [28].

Due to the diversity and overlap of dolphin vocalizations, automated classification is challenging. Whistles exhibit dynamic frequency modulation over time, while clicks are impulsive and broadband. In natural environments, these signals frequently overlap with background noise from other marine mammals, anthropogenic sources, and environmental processes. These characteristics motivate the use of time–frequency representations that capture both spectral structure and temporal evolution.

Taken together, the reported frequency characteristics of dolphin vocalizations indicate that a high sampling rate is required to preserve relevant whistle harmonics and broadband echolocation content. The use of 96 kHz recordings in the CetusID dataset [15] provides sufficient bandwidth for this purpose, while systematic downsampling to 48 kHz and 24 kHz. In addition, removal of the 0–2 kHz band is motivated by its dominance by environmental noise and its limited contribution to discriminative whistle structure.

3.2 Passive Acoustic Monitoring

PAM is a non-invasive technique for studying marine mammals by continuously recording underwater sound using hydrophones [36, 54]. As dolphins rely heavily on acoustic communication and echolocation, PAM enables long-term observation of their presence and activity without direct human interference.

PAM recordings typically consist of long-duration audio streams in which vocalizations are sparse and embedded within complex acoustic environments. Recordings may vary substantially across locations, sensors, and environmental conditions. As a result, PAM data are not independently and identically distributed at the segment level [49], which has implications for data partitioning and evaluation strategies in automated classification systems.

3.3 Digital Audio and Spectrograms

Underwater sound recorded by hydrophones is represented digitally as a discrete-time signal. This representation is obtained by sampling the continuous acoustic waveform at a fixed sampling rate, measured in hertz (Hz). The sampling rate determines the number of samples recorded per second and constrains the maximum frequency that can be represented digitally. According to the Nyquist theorem, a signal must be sampled at least twice its maximum frequency to avoid aliasing [39]:

$$f_{\max} = \frac{f_s}{2}. \tag{3.1}$$

Equation 3.1 represents the Nyquist theorem's formula, where f_s is the sampling rate and f_{max} is the highest frequency we can capture. Therefore, for example, a sampling rate of 48 kHz allows representation of frequencies up to 24 kHz. While a sampling rate of 48 kHz preserves the dominant fundamental contours of many dolphin whistles, higher-frequency whistle harmonics and broadband echolocation clicks cannot be fully represented at this rate [3, 2]. Consequently, higher sampling rates are commonly used in dolphin bio-acoustic research to preserve biologically relevant frequency content.

To analyze dolphin vocalizations, it is not enough to look only at the time-domain audio signal. The structure of whistles and clicks is best understood when viewed in terms of how their frequency content changes over time. This transformation is achieved using the STFT. The STFT divides the signal into short overlapping windows and computes a frequency spectrum for each window. The STFT is defined as:

$$STFT(t, f) = \sum_{n=-\infty}^{\infty} x[n]w[n-t]e^{-j2\pi f_n} \tag{3.2}$$

where $x[n]$ is the audio signal, $w[n-t]$ is the window function centered at time t, and f represents frequency. By performing this operation continuously with a sliding window over the signal, we obtain a two-dimensional time–frequency representation known as a spectrogram. The spectrogram displays time on the horizontal axis, frequency on the vertical axis, and amplitude as color intensity, allowing visualization of the curved shapes of whistles or the broadband bursts of clicks [34]. Using STFT spectrograms allows us to effectively "see" sounds. Dolphin whistles appear as smooth, curved lines,

while echolocation clicks appear as short vertical bursts of broadband energy. These patterns are complex to analyze directly from raw audio but become visually intuitive on spectrograms. Examples of this are shown in Section 3.5.

In this thesis, spectrogram representations serve as input to CNNs. STFT parameters and spectrogram resolution are inherited from the CetusID framework and are not optimized or modified unless explicitly stated in the experimental methodology.

3.4 Data Augmentation in Bio-acoustic Classification

Data augmentation refers to applying label-preserving transformations to existing training samples to increase dataset variability and mitigate overfitting. In bio-acoustic classification, augmentation is commonly used to address the limited availability of annotated data and to improve generalization to unseen recording conditions.

Augmentation can be applied at multiple stages of the audio processing pipeline. Audio augmentation modifies the raw time-domain signal, for example, by mixing background noise or applying temporal shifts. Spectrogram-level augmentation operates directly on time–frequency representations by altering visual characteristics such as time or frequency masking. By exposing models to controlled variability during training, augmentation encourages learning of robust spectro-temporal patterns rather than dataset-specific artefacts.

While augmentation is often used to reduce overfitting under limited training data, its effects are context-dependent and may affect performance differently under clean and degraded conditions. For this reason, augmentation strategies must be evaluated carefully within the intended application setting.

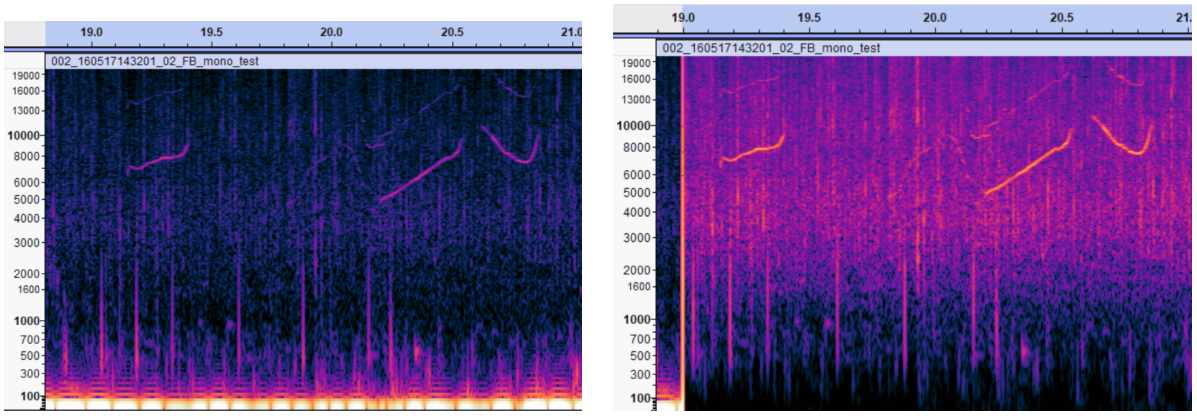
In this thesis, data augmentation is particularly relevant due to the limited size of the publicly available CetusID demo dataset [15] used for training and evaluation. This constraint motivates several methodological choices throughout the study, including a relatively simple CNN2 baseline, transfer learning with the EfficientNetB0 model, leave-one-recording-out cross-validation, and a conservative augmentation strategy. Consequently, all augmentation effects are interpreted in the context of data scarcity, and the scope of the findings is limited accordingly.

3.5 Spectrogram Inspection and Visualization Effects

This section presents a qualitative inspection of spectrograms to illustrate how dolphin vocalizations appear in time–frequency representations. The operations described here are used exclusively for visualization and human interpretation and do not form part of the pre-processing pipeline used for model training or evaluation.

3.5.1 Spectrogram Inspection and Call Visibility

In some recordings, dolphin vocalizations exhibit relatively low amplitude compared to broadband background noise, particularly in the low-frequency range below approximately 2–3 kHz, where environmental noise from wind, waves, and shipping is dominant [36]. As a result, tonal whistle structures may show limited visual contrast in default spectrogram displays, even though they are present in the signal.

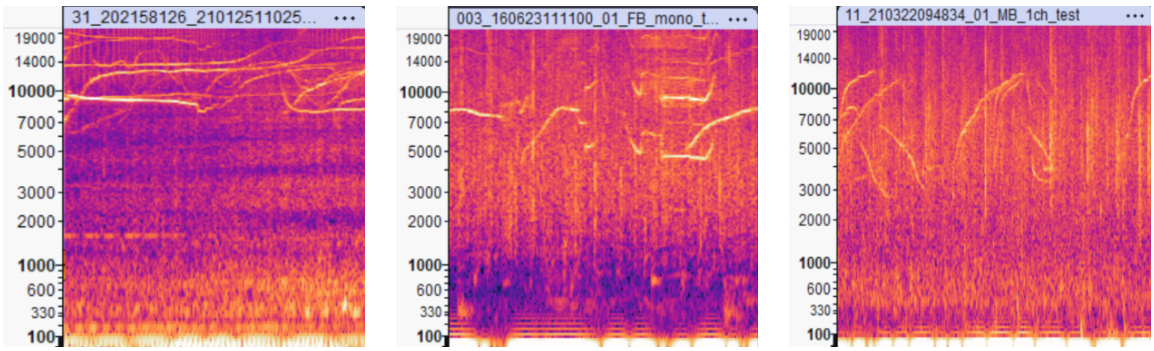


(a) Unprocessed spectrogram (b) Processed for visualization

Figure 3.1: Spectrogram segment (19–21 s) illustrating visualization-oriented signal processing. (a) Unprocessed spectrogram showing limited visual contrast between dolphin vocalizations and background noise. (b) Spectrogram after visualization-only processing, including normalization and amplitude scaling, which enhances the visibility of whistle contours. This processing is applied using Audacity [5], solely for interpretability and illustration, and is not used in the training or evaluation pipelines.

For illustrative purposes only, volume and compression methods were applied in Audacity [5] to improve visual interpretability. Specifically, peak normalization and mild amplification were used to increase overall contrast. After the filtering was applied, the vocalizations appeared more vibrant and easier to identify. The dolphin vocalizations in Figure 3.1 appear as light-colored streaks, mainly concentrated between 5 and 15kHz, and appear clearer after being pre-processed.

3.5.2 Species Variability in Spectrogram Patterns



(a) Delphinus Delphis (b) Sousa Plumbea (c) Tursiops Aduncus

Figure 3.2: Representative spectrograms illustrating inter-species variability in whistle structure for three dolphin species without band-pass filtering. All spectrograms are generated using the same visualization parameters as in section 3.5.3 in Audacity [5], except for band-pass filtering.

Spectrogram representations reveal qualitative differences in whistle structure across dolphin species. Figure 3.2 shows representative spectrogram segments for Delphinus delphis, Sousa plumbea, and Tursiops aduncus, generated using identical spectrogram parameters.

The spectrogram of Delphinus delphis is characterized by multiple overlapping whistle contours concentrated primarily between approximately 7 and 16 kHz. These contours appear as relatively thin, frequency-modulated curves with rapid changes in frequency over short time intervals, resulting in a dense time–frequency pattern.

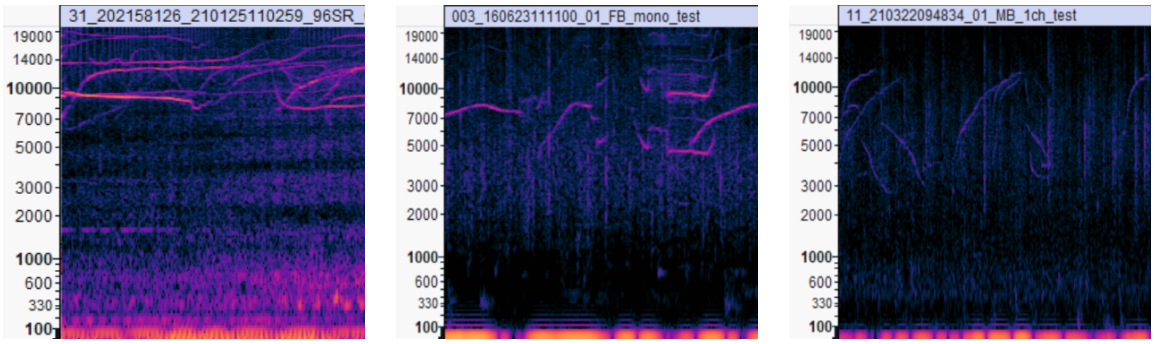
In contrast, Sousa plumbea exhibits fewer overlapping contours and a more confined frequency range. Whistles in this example appear smoother and more elongated in time, with slower frequency modulation and reduced contour density compared to Delphinus delphis.

The spectrogram of Tursiops aduncus shows whistle contours spanning a broader frequency range and are more sparsely distributed in time. Individual calls appear more isolated, with clearer separation between successive vocalizations.

Across the three examples, differences are also visible in the extent of low-frequency, noise-dominated regions. These regions vary between recordings and are attributed to background noise and recording conditions rather than vocalization structure.

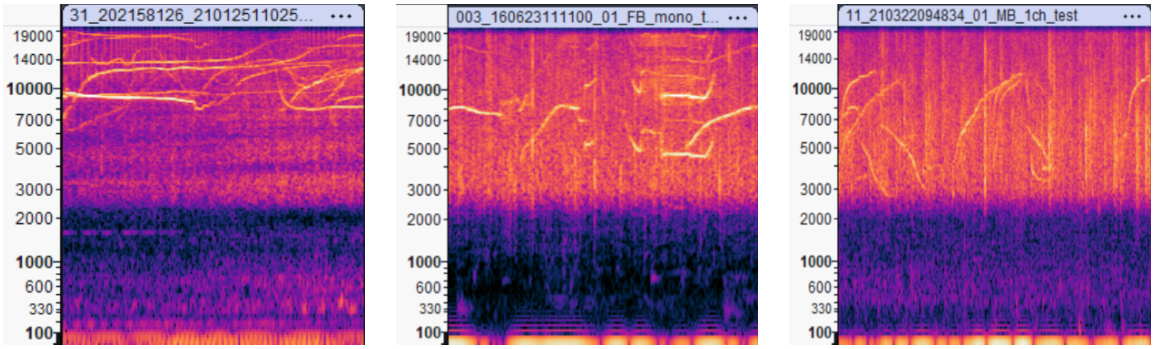
3.5.3 Band-pass Filtering

Band-pass filtering is a signal-processing technique that attenuates frequency components outside a predefined range while preserving frequencies of interest. In marine bio-acoustic analysis, band-pass filters are often used to suppress low-frequency environmental noise and sensor-related artefacts [34].



(a) *Delphinus delphis* (b) *Sousa plumbea* (c) *Tursiops aduncus*

Figure 3.3: Original recording spectrograms illustrating inter-species variability in whistle structure for three dolphin species prior to peak normalization, amplification, and band-pass filtering. All examples are generated in Audacity [5] and are shown for qualitative comparison with the filtered spectrograms in Figure 3.4.



(a) Delphinus Delphis (b) Sousa Plumbea (c) Tursiops Aduncus

Figure 3.4: Representative spectrograms illustrating inter-species variability in whistle structure for three dolphin species after band-pass filtering between 2 kHz and 19 kHz. All examples are generated using identical spectrogram parameters and band-pass filtering settings in Audacity [5] and are shown for qualitative inspection.

Comparing Figures 3.3 and 3.4 shows that band-pass filtering suppresses low-frequency background energy below 2 kHz, increasing the visual contrast of whistle contours across all three species. This effect is most pronounced in recordings where low-frequency noise dominates the unfiltered spectrogram, making tonal structures more clearly distinguishable after filtering.

To generate the filtered spectrograms shown in Figure 3.4, the following processing steps were applied sequentially in Audacity [5]:

- Removal of DC offset and peak normalization to −6 dB.
- Uniform amplification of the waveform by +6 dB to restore signal visibility after normalization.
- Application of a band-pass filter using the Filter Curve EQ tool, with attenuation below 2 kHz and above 19 kHz.
- Spectrogram visualization using Audacity's default spectrogram settings, applied consistently across all examples, including a Hann window with a window size of 2048 samples, a zero-padding factor of 2, Mel frequency scaling, and a fixed color scale (gain: 20 dB, range: 80 dB).

Although the applied band-pass filter attenuates frequencies below 2 kHz, residual energy remains visible in the 0–2 kHz region of the spectrograms. This is expected due to the finite roll-off of practical digital filters, which attenuate rather than perfectly eliminate frequency components, and to spectral leakage introduced by windowing during spectrogram computation. In addition, broadband noise components and visualization floor effects may remain visible at low amplitudes even after filtering, without contributing meaningful information to the retained signal content.

The spectrograms in this section were generated using Audacity [5], applying the same visualization settings as those used in Subsection 3.5.2, with the addition of a band-pass filter. After applying the band-pass filtering, the low-frequency, noise-dominated regions in Sousa plumbea and Tursiops aduncus become more similar in extent, resulting in a more uniform low-frequency appearance. In the context of CNN-based classification, filtering changes the input representation by removing selected frequency regions prior to spectrogram generation. Whether such filtering improves or degrades classification performance depends on how models utilize spectral information and is therefore evaluated empirically in later chapters.

3.6 Convolutional Neural Networks for Spectrogram Classification

CNNs are a class of deep learning models designed to learn hierarchical feature representations from structured input data [30]. Originally developed for image analysis [29], CNNs are well-suited to spectrogram-based audio classification because spectrograms can be interpreted as two-dimensional representations of local temporal and spectral patterns[43].

3.6.1 Convolutional Neural Network Architecture

Figure 3.5 illustrates the general structure of a CNN used for feature extraction and classification. The network consists of a sequence of convolutional, pooling, and fully connected (dense) layers. This figure is intended as a conceptual illustration of the CNN architecture rather than a layer-by-layer depiction of the exact model configuration.

The convolutional layers apply learnable filters that slide across the input, computing local dot products to detect spatially localized patterns such as edges, contours, or textures [29]. In the context of spectrograms, these filters learn to identify characteristic time–frequency structures such as harmonics, onsets, and frequency modulations [21].

Following convolution, max-pooling layers reduce the spatial resolution of feature maps by retaining the maximum activation within local neighborhoods. This operation introduces a degree of translational invariance and reduces computational complexity while preserving salient features [8]. After several convolutional and pooling stages, the resulting feature maps are flattened into a one-dimensional vector, which is then passed to one or more dense layers. These dense layers perform high-level reasoning and map the learned features to output class probabilities, typically using a softmax activation function for classification tasks [30].

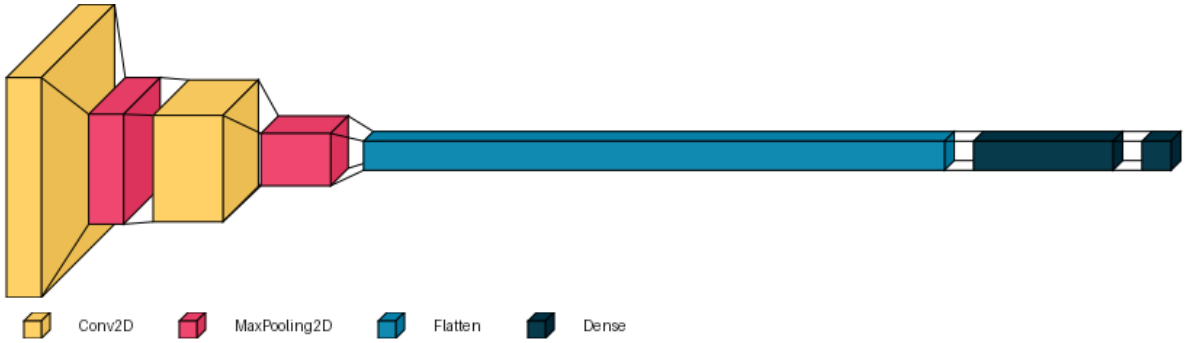


Figure 3.5: Illustration of a convolutional neural network architecture composed of convolutional layers, max-pooling layers, a flattening operation, and fully connected layers. The figure was generated using the VisualKeras library.

3.6.2 Spectrogram-Based Classification Pipeline

The complete audio classification pipeline used in this work is illustrated in Figure 3.6.

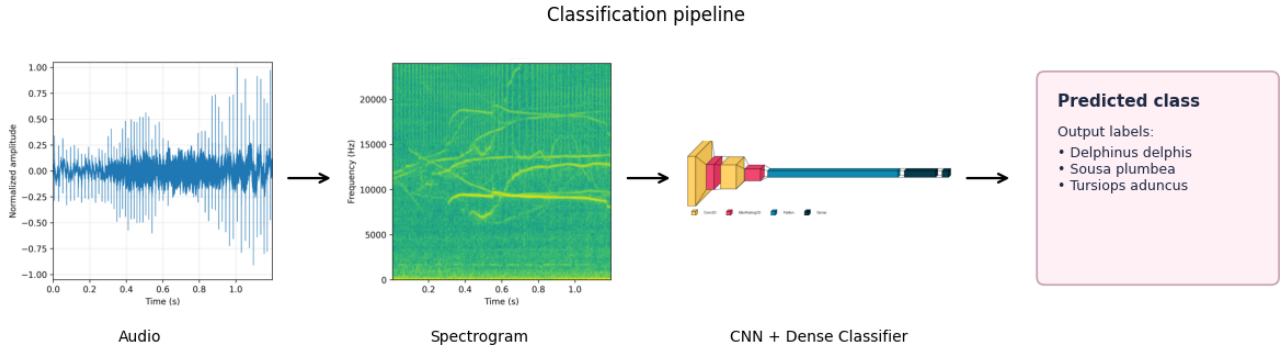


Figure 3.6: Overview of the spectrogram-based classification pipeline, showing the transformation from raw audio to spectrogram representation, followed by CNN-based classification.

The pipeline begins with a raw audio waveform, which is first normalized and segmented into fixed-duration samples. Each audio segment is then transformed into a spectrogram using a time–frequency analysis method such STFT. This representation captures how spectral energy evolves over time, making it well-suited for modeling non-stationary acoustic signals [39].

The resulting spectrogram is treated as a two-dimensional input image and fed into the CNN. Convolutional layers extract localized spectro-temporal patterns, while pooling layers progressively abstract these features into higher-level representations. The learned features are subsequently passed through dense layers to produce a final prediction over the target classes. In this case, the output layer assigns probabilities to multiple dolphin species, enabling automated classification of species from acoustic recordings.

This end-to-end pipeline leverages the representational power of CNNs to jointly learn discriminative features and perform classification, eliminating the need for handcrafted acoustic features and improving robustness across varying acoustic conditions [43, 21].

Chapter 4

Dataset Description

This chapter describes the datasets used in this study and clarifies the distinction between the dataset used in the original CetusID study by Frainer et al. [16] and the dataset used in this thesis. To avoid ambiguity, both datasets are explicitly named and referenced consistently throughout the remainder of this thesis.

4.1 Reference Dataset: CetusID-Full

The CetusID-Full dataset, established by Frainer et al. [16], served as the foundational acoustic library for developing an automated framework for dolphin detection and taxonomic identification in South African waters.

All acoustic recordings in the CetusID-Full dataset were standardized to a sampling frequency of 96 kHz. Recordings acquired at higher sampling rates were downsampled to this value to ensure consistency across sources.

Sample resolution was defined through the spectrogram transformation applied prior to CNN training. Spectrograms were computed using a Hann window with an FFT size of 1024 samples and a hop size of 128 samples, corresponding to a 75% overlap between successive frames. This configuration determines the effective time-frequency resolution presented to the CNN models and was kept constant across all experiments.

The dataset comprises 723 minutes of annotated dolphin acoustic recordings alongside 772 minutes of ambient soundscape data. Data were primarily sourced from boat-based focal follows, moored hydrophones (Mossel Bay and Fish Hoek), and free-drifting buoys.

The species distribution of the annotated dolphin vocalizations, including whistles, burst pulses, and echolocation clicks, is as follows:

- **Delphinus delphis:** 45.6%
- **Tursiops aduncus:** 39%
- **Sousa plumbea:** 14.4%
- **Orcinus orca:** 1%

To prepare the data for Convolutional Neural Network (CNN) training, the authors applied several specific processing steps:

- **Segmentation:** Audio was downsampled to 96 kHz and sliced using a sliding window approach with start times interspaced by one second.
- **Optimization:** Experiments with window sizes of 2, 3, 5, and 7 seconds determined that a 2-second window provided the best performance for both detection and identification.
- **Augmentation:** Dolphin vocalizations were randomly mixed with target soundscapes (90% dolphin, 10% background noise) to simulate real-world monitoring environments.

- **Spectrogram Conversion:** Segments were converted into 5×5 inch spectrogram images. The study found that 350×350 pixels was optimal for the detection model (CNN1), while 450×450 pixels was optimal for species identification (CNN2).

The dataset was utilized in a two-stage pipeline:

1. **Detection (CNN1):** A binary classifier to distinguish the presence of any dolphin vocalization from the background soundscape.
2. **Identification (CNN2):** A multi-class classifier applied only to the segments flagged by CNN1 to assign a specific taxonomic identity.

The framework was evaluated on a 24-hour "unseen" test set, achieving 84.4% detection accuracy and high species-specific identification rates.

To bridge the gap between raw audio and CNN input, CetusID-Full distinguishes between audio clips and spectrogram images. Total annotated data comprised 723 minutes of dolphin vocalizations. These were segmented using a sliding window, interspaced by one second, to sample vocalizations in varying contexts.

To ensure a balanced and robust training library, data augmentation was applied by mixing dolphin signals with local soundscapes. For the species with the highest representation, D. delphis, a total of 20,319 individual clips were generated, which were subsequently doubled through augmentation to produce 40,638 training spectrograms. This scaling approach allowed the total training library to reach up to 80,000 images depending on the model configuration (see Table 4.2).

4.2 Dataset Used in This Study: SA-Acoustics

In this study, a subset of the publicly released CetusID dataset was used. To clearly distinguish it from the full dataset employed by Frainer et al. [16], this dataset is hereafter referred to as the SA-Acoustics Training dataset.

The SA-Acoustics Training dataset was obtained from the CetusID demonstration repository on [14] and corresponds to the labeled training subset provided for demonstration and reproducibility. Unlike the full CetusID-Full dataset described by Frainer et al., this demo dataset is a highly reduced, curated subset intended primarily for method illustration rather than large-scale model training.

Only the training set was used in this work, as the accompanying test set is unlabeled and therefore unsuitable for supervised learning and quantitative evaluation.

The SA-Acoustics Training dataset contains labeled spectrogram segments from three dolphin species:

- **Delphinus delphis**
- **Sousa plumbea**
- **Tursiops aduncus**

For each dolphin species, exactly three labeled audio files are provided. The total duration of labeled recordings per species is summarized in Table 4.1. The recordings are not uniform in length, and the total amount of labeled data differs substantially between species.

Table 4.1: Composition of the SA-Acoustics Training dataset used in this study.

Species	**WAV files**	**Total duration (s)**	**Total duration (min)**
Delphinus delphis	3	92	1.53
Sousa plumbea	3	341	5.68
Tursiops aduncus	3	276	4.60
Total	**9**	**709**	**11.8**

In addition to the species-specific recordings, the dataset includes a single unlabeled background-soundscape recording of approximately 10 minutes. This soundscape contains environmental noise representative of

passive acoustic monitoring conditions, such as wind, wave noise, and other ambient sources. The soundscape recording is not associated with a specific dolphin species and is not used for supervised training or evaluation, but is used for augmentation and the dolphin vocalization detection (CNN1).

The SA-Acoustics Training dataset is extremely small, comprising only 9 labeled audio files totaling approximately 11.8 minutes of dolphin recordings across 3 species. As a result, the dataset does not support statistically robust performance estimation or strong generalization claims. Consequently, all results presented in this thesis should be interpreted as exploratory observations that illustrate the effects of pre-processing and model configuration choices within a constrained experimental setting.

This limited dataset size also motivated the exclusion of Orcinus orca from the analysis. In the CetusID-Full, this species represents only a very small fraction of the available recordings, resulting in extreme class imbalance. Given the already limited size of the demo training set, including this class would further exacerbate the imbalance and reduce the statistical reliability of model evaluation. For this reason, the analysis in this thesis focuses exclusively on species with sufficient representation in the available labeled data.

Table 4.2: Comparison of Dataset Characteristics: CetusID-Full vs. SA-Acoustics

Feature	**CetusID-Full (Frainer et al.)**	**SA-Acoustics (This Study)**
Total annotated audio duration	723 min (dolphin vocalizations) + 772 min (background noise)	11.8 min (dolphin recordings) + 10 min soundscape
Number of species	4 (*Delphinus delphis, Tursiops aduncus, Sousa plumbea, Orcinus orca*)	3 (*Delphinus delphis, Tursiops aduncus, Sousa plumbea*)
Number of encounters	43 boat-based acoustic encounters	Not defined (standalone demo recordings)
Species distribution	*D. delphis* (45.6%), *T. aduncus* (39.0%), *S. plumbea* (14.4%), *O. orca* (1.0%)	Equal number of recordings per species (3 WAV files each), but unequal total duration
Number of labeled WAV files	Hundreds of long-duration recordings	9 labeled WAV files (3 per species)
Training units used	Up to ∼80,000 spectrogram images	700 spectrogram images derived from 9 WAV files
Intended purpose	Large-scale training and benchmark evaluation	Demonstration dataset for method illustration and exploratory analysis

Table 4.2 shows the substantial scale difference between the CetusID-Full dataset used by Frainer et al. and the SA-Acoustics dataset employed in this study. CetusID-Full comprises hundreds of long-duration recordings collected across 43 annotated acoustic encounters, enabling large-scale training and the generation of up to approximately 80,000 spectrogram images.

In contrast, SA-Acoustics consists of only nine labeled WAV files totaling 11.8 minutes of dolphin recordings across three species, supplemented by a single 10-minute unlabeled soundscape. While the number of recordings per species is balanced, the total recording duration is not, resulting in unequal amounts of training data after spectrogram segmentation.

Due to these limitations, the approximately 700 spectrogram images derived from SA-Acoustics support only exploratory analysis. Consequently, results presented in this study focus on comparative effects of pre-processing and model configuration rather than generalizable classification performance.

Chapter 5

Baseline Method

In this section, the implementation of the baseline method is introduced, and its results are presented using the CetusID framework [15] proposed by Frainer et al. [16]. Baseline results are reported on two datasets: the full-scale CetusID-Full dataset used in the original study, and the reduced SA-Acoustics dataset used in this thesis. Unlike Frainer et al., who report encounter-level performance after post-processing and majority voting, all results in this thesis are reported at the segment level unless explicitly stated otherwise.

The complete processing pipeline is illustrated across Figures 5.1 and 5.2, covering data preparation, augmentation, and two-stage CNN inference.

Raw hydrophone recordings are first resampled to a fixed sampling rate of 96 kHz and segmented using a sliding window of 2–7 s with a step size of 1 s. Each segment is converted into a time-frequency representation using an STFT with a Hann window, FFT length of 1024, hop size of 128 samples, and 75% overlap. The resulting spectrograms are formatted as images with a fixed physical size and a resolution of 200×200 to 500×500 pixels. No signal-level pre-processing, such as band-pass or high-pass filtering, is applied in the baseline configuration.

CNN1 is trained as a binary classifier to detect the presence or absence of dolphin vocalizations in each spectrogram segment. The network consists of three convolutional layers with 32 filters of size 4×4 and ReLU activations, followed by max pooling, dropout, and a fully connected softmax layer producing dolphin/no-dolphin predictions. Segments classified as non-dolphin are discarded.

Segments classified as dolphin by CNN1 are passed to CNN2, which performs multi-class species classification. CNN2 uses the same core convolutional architecture as CNN1 but outputs species labels via a softmax layer with C classes. CNN1 and CNN2 are trained independently, following the original CetusID design.

During training only, data augmentation was applied in all baseline experiments by mixing dolphin vocalization segments with background soundscape recordings at a 90% dolphin signal to 10% background noise ratio, following the procedure described by Frainer et al. [16]. As illustrated in Figure 5.1(b), augmentation was combined with dataset balancing at both the species and acoustic encounter levels. Augmentation was not applied during validation or inference. Specifically, each annotated dolphin audio segment was augmented by randomly selecting a background-soundscape segment of equal duration and linearly combining the two. Background segments were temporally aligned to match the dolphin segment length prior to mixing.

As shown in Figure 5.2(b), CNN2 outputs segment-level species probabilities that are subsequently aggregated using post-processing steps. Final species identification followed the same post-processing strategy as the baseline method, including confidence thresholding of segment-level predictions, grouping detections into acoustic encounters based on temporal proximity, and assigning a single species label to each encounter via majority voting across segments.

The baseline results were obtained using the publicly available CetusID implementation released by Frainer et al. [16], without modifying any part of the code, including the data augmentation and post-processing components.

In CetusID, CNN1 denotes the binary dolphin detection network, whereas CNN2 denotes the multi-class species identification network [16], both trained independently using the same customised CNN architecture (Figure 5.2). Both networks use the customised CNN architecture reported by Frainer et al., consisting of three Conv2D layers (32 filters, 4×4 kernels, ReLU), each followed by dropout (0.4) and max-pooling (4×4), and a fully connected layer with 64 ReLU units and dropout (0.4), followed by a softmax output layer (2 outputs for CNN1; C outputs for CNN2) [16]. Training was performed for 50 epochs using the Adam optimiser (learning rate 10^{-3}), batch size 32, and categorical cross-entropy loss, as specified in the original implementation [16]. Frainer et al. trained CNN2 for 50 epochs in the original study. However, the publicly released CetusID demo implementation defaults to 20 training epochs, which is used consistently in all experiments reported in this thesis.

It is important to note that the reproduced baseline results reported in this chapter are obtained using the SA-Acoustics dataset, whereas the reference results from Frainer et al. [16] are based on the substantially larger CetusID-Full dataset. As a result, absolute performance values are not directly comparable, and differences should be interpreted in the context of the dataset scale and the evaluation protocol.

Baseline split validation was performed using the default CetusID demo evaluation protocol (random 80/20 train–validation split) to confirm correct reproduction of the reference implementation. This split was used exclusively for baseline verification and is not intended to provide a robust estimate of generalization performance. All subsequent analyses and comparisons in this thesis employ LORO cross-validation to enforce recording-level separation and prevent segment-level data leakage.

Tables 5.1 and 5.2 summarize the baseline split validation results for CNN1 and CNN2, respectively. CNN1 achieves perfect accuracy and precision on the internal validation set. This result should be interpreted with caution, as it reflects the small size and controlled nature of the demo dataset rather than generalizable detection performance. CNN2 shows consistently high performance across species, with lower accuracy in Delphinus delphis than in Sousa plumbea and Tursiops aduncus, mirroring the relative performance trends reported by Frainer et al.

Table 5.1: Baseline split validation performance of the dolphin detection network (CNN1). Frainer et al. [16] report results on their test set; the reproduced result is the internal validation performance obtained on the CetusID demo dataset using the same CNN1 configuration (96 kHz, window size 2 s, Pixel = 350 × 350.

Model	**Accuracy (%)**	**Precision (%)**
Frainer et al. Custom CNN1	84.4	87.6
Validated Demo Custom CNN1	100	100

Table 5.2: Baseline split validation performance of the species identification network (CNN2). Frainer et al.'s [16] results are reported on their test set; reproduced results are obtained using the unmodified CetusID implementation on the demo dataset. The configuration used has a sampling rate of 96 kHz, a window = 2 s, and a pixel resolution of 450 × 450. Per-class accuracy is computed at the acoustic-encounter level.

Model	**Species**	**Per-class accuracy (%)**
Frainer et al. CNN2	*Sousa plumbea*	96.9
Frainer et al. CNN2	*Tursiops aduncus*	100.0
Frainer et al. CNN2	*Delphinus delphis*	78.0
Frainer et al. CNN2	**Average**	**91.6**
Own implementation using [16] CNN2	*Sousa plumbea*	86.0
Own implementation using [16] CNN2	*Tursiops aduncus*	100.0
Own implementation using [16] CNN2	*Delphinus delphis*	67.0
Own implementation using [16] CNN2	**Average**	**84.3**

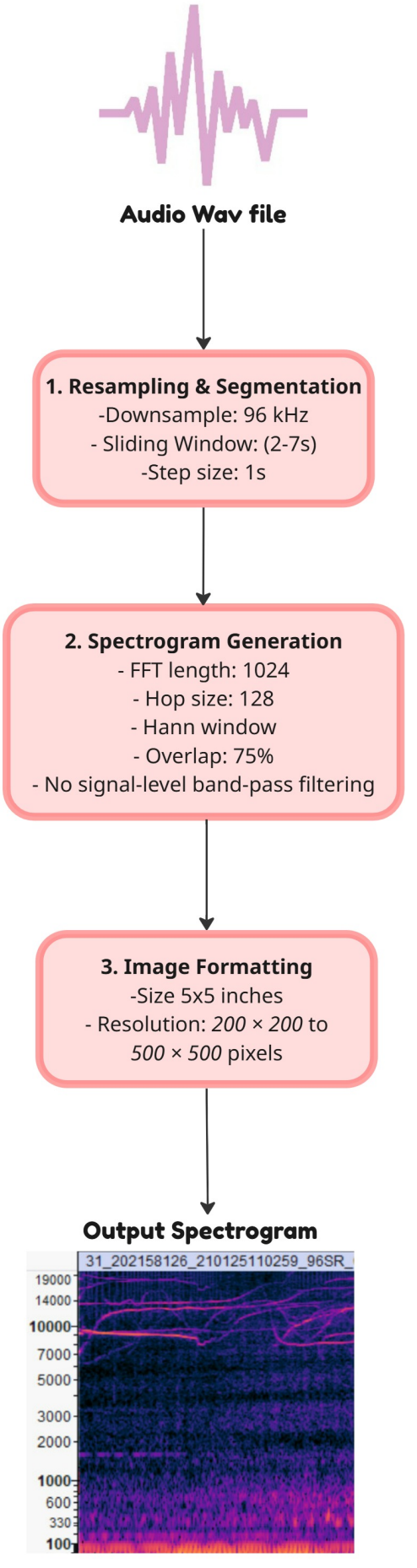
The absolute performance values differ from those reported in the original study, which is expected given the substantially reduced dataset size, the use of a demo-level training set, and differences in the evaluation protocol. Importantly, the reproduced baseline exhibits qualitatively similar behavior to the reference implementation and therefore provides a valid and stable reference configuration. This baseline enables subsequent experiments to isolate the relative effects of sampling rate variation, data augmentation, and signal-level pre-processing without confounding architectural or implementation differences.

For the baseline experiments, model performance was evaluated using a standard random 80/20 train–validation split, consistent with the evaluation procedure reported by Frainer et al. [16]. This strategy was retained

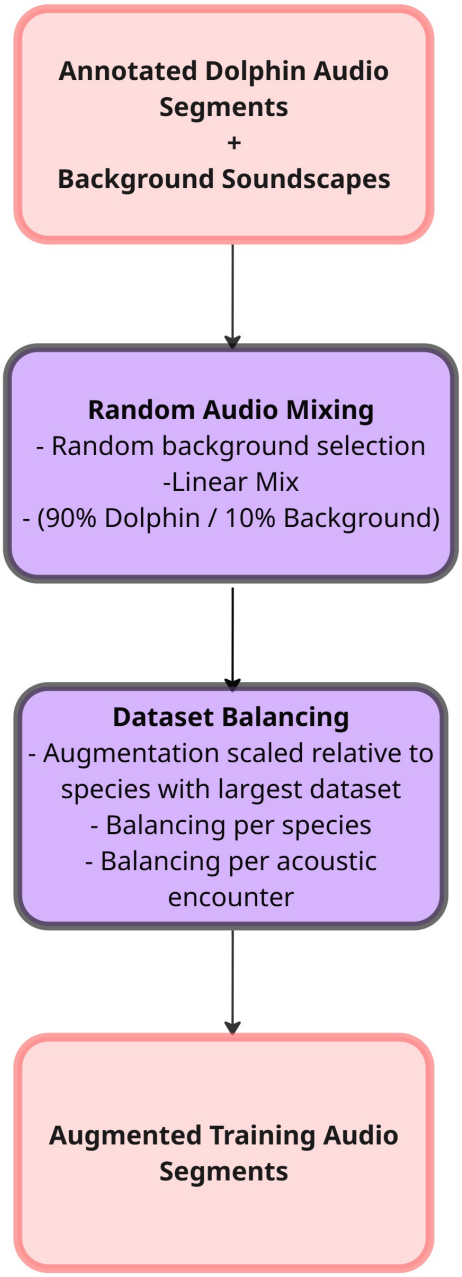
to ensure a faithful reproduction of the baseline method and comparability with the originally reported results.

In contrast, the proposed methodology employs a LORO validation strategy to assess generalization across recording sessions. The use of LORO is therefore treated as a methodological extension and is described in detail in Chapter 6.

To avoid ambiguity, results obtained with the default 80/20 random split in the reproduced CetusID demo are referred to as baseline split validation accuracy, whereas all subsequent results in this thesis are obtained using LORO validation accuracy, which enforces recording-level generalization.

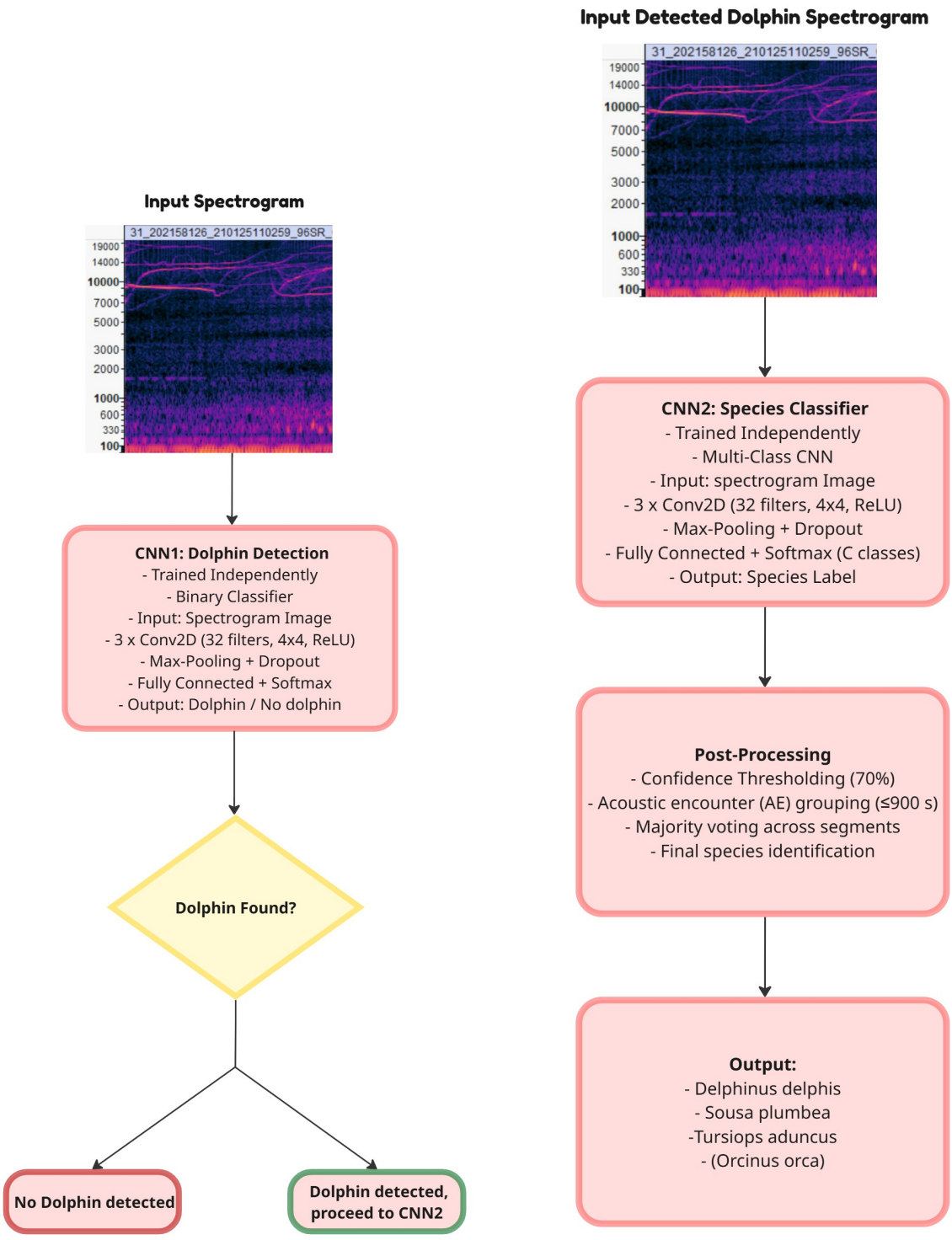


(a) Spectrogram Generation.



(b) Data Augmentation.

Figure 5.1: Data preparation pipeline used to construct the training dataset. (a) Spectrogram generation from raw audio segments. (b) Data augmentation is applied at the audio level during training only.



(a) CNN1: dolphin detection.

(b) CNN2: species classification.

Figure 5.2: Two-stage CNN inference pipeline for dolphin detection and species identification. (a) CNN1 performs dolphin presence detection on input spectrograms. (b) CNN2 performs multi-class species classification on detected dolphin segments.

Chapter 6

Methodology

This chapter describes the experimental methodology used to investigate the sensitivity of CNN-based dolphin species classification to signal-level pre-processing and representation choices. All experiments are conducted using a reproduced CetusID demo baseline validated in Chapter 5. In this thesis, the term reproduced CetusID demo baseline refers to the execution of the publicly released CetusID demonstration code provided by Frainer et al. [16], without re-implementing or modifying the original model architecture, training procedure, or data processing pipeline. Reproduction therefore consists of running the authors' provided scripts on the publicly available demo dataset to verify baseline behavior prior to extending the methodology.

Within each experimental block, the model architecture and training protocol are fixed, while representation and pre-processing factors are varied. In addition to experiments conducted with the baseline CetusID CNN (CNN2), the EfficientNetB0-based transfer-learning model is evaluated in the final experimental block to assess the interaction between signal-level pre-processing and model capacity.

Note that in the remainder, SA-Acoustics recordings refer to the original dataset as described in Chapter 4, while noise-degraded refers exclusively to recordings corrupted by additive Gaussian noise ($\sigma = 0.05$). These terms are used consistently and do not imply real-world noise conditions. Moreover, "signal-level pre-processing" refers exclusively to operations applied in the time domain prior to spectrogram computation (e.g., filtering). Normalization and amplitude-to-decibel conversion are considered representation-level processing.

6.1 Methodological Overview

An experiment in this thesis is defined as a controlled modification of a single pre-processing or representation parameter while holding all other components of the pipeline constant. The methodology consists of three experimental blocks: sampling-rate experiments (addressing SRQ.2), data augmentation and controlled signal degradation experiments (addressing SRQ.3), and final pre-processing sensitivity experiments combining band-pass filtering with model capacity comparison (addressing SRQ.1). All experiments use LORO cross-validation.

6.1.1 Experimental overview

The preliminary experiments are designed to characterize the behavior of the reproduced CetusID baseline and to motivate the configuration choices adopted in the final experiments. These experiments are structured into three components:

Preliminary Experiments

(i) **Dataset and cross-validation setup** The data and evaluation protocol used in all preliminary experiments are defined as follows:

- The SA-Acoustics dataset is used, consisting of three labeled recordings per dolphin species and one background soundscape recording.

- LORO cross-validation is applied with $k = 3$, where one complete recording per species is held out for validation in each fold.
- This recording-level partitioning prevents segment-level data leakage and enforces generalization across recording sessions.

(ii) **Preliminary experimental interventions** Using the fixed baseline processing pipeline, two exploratory studies are conducted:

- *Sampling-rate study:* CNN2 is trained and evaluated at sampling rates of 24, 48, and 96 kHz without data augmentation.
- *Augmentation and controlled signal degradation:* Performance is evaluated using SA-Acoustics and SA-Acoustics (noise-degraded, $\sigma = 0.05$), with training-time background-mixing augmentation turned ON or OFF.

In all cases, augmentation is applied to training data only, while validation data remains unaugmented.

(iii) **Fixed processing and evaluation setup** All remaining components of the pipeline are kept constant across preliminary experiments:

- Audio segmentation using 2-second sliding windows with a 1-second overlap.
- Spectrogram generation using baseline CetusID parameters, including STFT and amplitude-to-decibel conversion.
- CNN2 architecture trained from scratch for 20 epochs using the Adam optimizer (lr = 0.001), batch size 32, and dropout rate 0.4.
- Evaluation using mean accuracy across folds, row-normalized confusion matrices, recall, and F1-score.

Final Experiments

The final experiments constitute the main contribution of this thesis and extend the baseline pipeline to evaluate the interaction between signal-level pre-processing, data augmentation, and model capacity. The experimental design is structured as follows:

(i) **Fixed data and pre-processing baseline**

- All experiments use the same recordings as in the preliminary stage and are evaluated using LORO cross-validation ($k = 3$).
- Raw audio is fixed at the highest available sampling rate (96 kHz) to preserve maximal frequency information.
- Validation data always consists of the SA-Acoustics recordings.

(ii) **Signal-level and training-time interventions** Two experimental factors are varied in a controlled ON/OFF manner:

- *Signal-level band-pass filtering:* A proposed band-pass filter targeting dolphin vocalization frequencies is applied to both training and validation audio when enabled.
- *Training-time data augmentation:* Background-mixing augmentation is either applied or omitted during training; Validation data is never augmented.

(iii) **Model comparison and evaluation**

- Two model architectures are evaluated under identical data splits and training protocols:
  - the baseline CNN2 architecture from CetusID;
  - an EfficientNetB0-based transfer-learning classifier.
- All conditions are evaluated using identical metrics and reporting procedures.
- The full-factorial design results in eight experimental conditions, as summarized in Table 6.1.

Table 6.1: Overview of the final experimental conditions evaluated in this thesis. All experiments use SA-Acoustics LORO validation data, while training augmentation, signal-level band-pass filtering, and model architecture are varied in a full-factorial design.

Dataset	Augmentation (train)	Band-pass filter	Model
SA-Acoustics	OFF	OFF	Frainer et al.'s CNN2
SA-Acoustics	OFF	OFF	EfficientNet
SA-Acoustics	ON	OFF	Frainer et al.'s CNN2
SA-Acoustics	ON	OFF	EfficientNet
SA-Acoustics	OFF	ON	Frainer et al.'s CNN2
SA-Acoustics	OFF	ON	EfficientNet
SA-Acoustics	ON	ON	Frainer et al.'s CNN2
SA-Acoustics	ON	ON	EfficientNet

6.1.2 Data Description

All experiments in this thesis are conducted using the *SA-Acoustics training dataset*, which is described in detail in Chapter 4. This dataset is a curated subset of the publicly released CetusID demo data and is used consistently across all experimental stages.

In summary, the SA-Acoustics dataset contains annotated passive acoustic monitoring recordings from three dolphin species: Delphinus delphis, Sousa plumbea, and Tursiops aduncus. For each species, three labeled WAV files containing dolphin vocalizations are provided, for a total of nine. In addition, a single long-duration background-soundscape recording without dolphin vocalizations is included.

Each dolphin recording consists of continuous underwater audio captured by a stationary hydrophone and is accompanied by a CSV annotation file that specifies the start and end times of labeled vocalizations and background segments. These annotations are used to associate extracted audio segments and spectrograms with species labels during training and evaluation.

The dolphin vocalization recordings are used exclusively for species classification. The background-soundscape recording is used solely for data augmentation and is never included in validation data. Due to the limited number of recordings per species, all experiments addressing SRQ.1, SRQ.2, and SRQ.3 employ LORO cross-validation to prevent segment-level data leakage between training and validation sets.

6.1.3 Recording-level Data Partitioning

LORO cross-validation is used to ensure independence between training and validation data. Since three recordings are available per species, the number of folds is $k = 3$. In each fold, one complete recording per species is reserved for validation, while the remaining recordings are used for training.

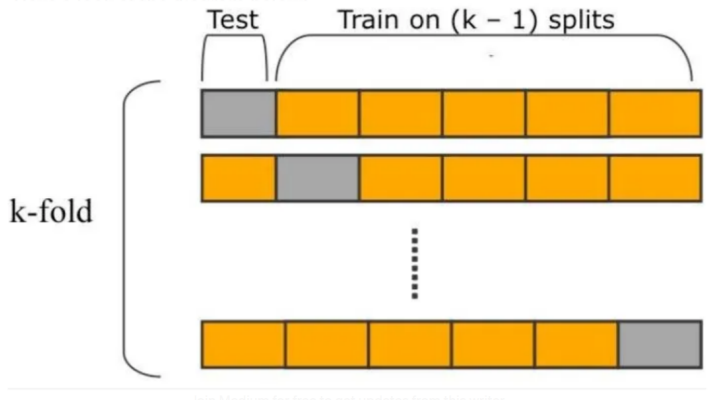


Figure 6.1: Illustration of k-fold cross-validation at the recording level. In LORO cross-validation, each fold holds out an entire recording for validation (grey), while the remaining recordings are used for training (orange). This strategy ensures independence between training and validation data by preventing segment-level overlap across folds. Adapted from [1].

As illustrated in Figure 6.1, one entire recording is held out as the validation set while all remaining recordings are used for training, and this process is repeated across folds. By enforcing separation at the recording level, LORO prevents spectrogram segments derived from the same recording from appearing in both training and validation sets, thereby avoiding artificial performance inflation caused by shared acoustic context. In each fold, one recording per species is used for validation, while the remaining two recordings per species are used for training, resulting in three folds in total.

Since only a single background soundscape recording was available, it was divided into three non-overlapping temporal segments of equal duration. These segments were treated as independent background samples and consistently assigned across folds for data augmentation.

6.1.4 Audio segmentation

All recordings are segmented into fixed-length audio windows of 2 seconds using a sliding-window strategy with a 1 second hop size, resulting in overlapping segments. Segmentation is applied uniformly across entire recordings and is not restricted to annotated vocalization intervals.

CSV annotation files are used to associate recordings with species labels and to provide reference vocalization timing information; however, these annotations are not used to define or align the segmentation windows. Instead, fixed-length segments are extracted across the full recording duration.

Each segment inherits the species label of its source recording and is treated as an independent sample during training and evaluation. Consequently, some segments may not contain dolphin calls of the assigned species and may consist solely of background or ambient acoustic content.

6.1.5 Spectrogram generation and data augmentation

All audio segments are converted into time-frequency representations prior to CNN processing. Two spectrogram-generation pipelines are evaluated that differ only in whether data augmentation is applied during training. All other processing parameters are kept identical across pipelines.

In both configurations, spectrograms are generated using the default CetusID settings, with a fixed segment duration of 2 seconds and a spectrogram resolution of 450×450 pixels. Identical STFT parameters and amplitude-to-decibel scaling are applied in all experiments.

In the non-augmented pipeline, waveform segments are first serialized into audio.pkl files and subsequently converted into static spectrogram images using the STFT. The resulting spectrograms are stored in img.pkl files and used directly as CNN inputs during training and evaluation.

In the augmented pipeline, background-mixing augmentation is applied during training by combining dolphin vocalization segments with temporally shifted background soundscape segments using a fixed mixing ratio, following the procedure described by Frainer et al. (see Chapter 5). In this case, spectrograms are generated dynamically during training from the augmented waveforms. Validation data are never augmented and are generated once using the non-augmented pipeline.

When background-mixing augmentation is enabled, the effective training set size increases to 37–41 samples per class per fold, depending on the held-out recording, compared to the non-augmented configuration.

6.1.6 Sampling Rate Experiments

To evaluate the effect of audio sampling rate (SRQ.2), experiments were conducted at 24, 48, and 96 kHz. Lower sampling rates were obtained by downsampling the original recordings, thereby discarding high-frequency information. For each sampling rate, the full pre-processing and training pipeline was executed independently using identical fold assignments, architectures, and hyperparameters.

CNN2 was trained from scratch for each sampling rate and fold. Data augmentation was disabled in these experiments to isolate the effect of sampling rate on classification performance and stability.

6.1.7 Augmentation and Controlled Noise Sensitivity Experiments

To evaluate model sensitivity to controlled signal degradation, an additional dataset was generated by injecting zero-mean Gaussian noise with a standard deviation of $\sigma = 0.05$ into all audio recordings prior

to segmentation. Gaussian noise was selected as a controlled and distribution-agnostic perturbation that enables repeatable sensitivity testing without modeling specific underwater noise sources. This noise injection is used solely as a controlled degradation mechanism to probe how classification performance changes under artificial signal perturbations, and is not intended as a formal robustness evaluation or as a label-preserving augmentation applied only during training.

Experiments were conducted across four configurations: SA-Acoustics without augmentation, SA-Acoustics with augmentation, SA-Acoustics (noise-degraded) without augmentation, and SA-Acoustics (noise-degraded) with augmentation. Performance differences between SA-Acoustics and SA-Acoustics (noise-degraded) validation sets were used to characterize performance sensitivity to artificial noise, while the effect of data augmentation was assessed by comparing changes in classification accuracy across these controlled degradation conditions.

Because the injected noise is synthetic and limited in scope, these experiments are interpreted as a sensitivity analysis under controlled signal degradation, rather than as evidence of robustness to real-world underwater noise conditions.

6.1.8 Final Experiments: Band-pass pre-processing and Model Comparison

The final experiments constitute the main contribution of this thesis and address SRQ.1. Based on the preliminary experiments, all final evaluations are conducted at a fixed sampling rate of 96 kHz. A signal-level band-pass filter is introduced as a pre-processing step and evaluated with and without filtering. When enabled, the filter is applied identically to both training and validation audio prior to spectrogram generation.

Before generating spectrograms, only the dolphin species audio recordings were filtered to retain the frequency range relevant to dolphin sounds. This was done using a Butterworth band-pass filter, which removes low-frequency and high-frequency components while preserving the structure of the remaining signal. Frequencies between 2 kHz and 19 kHz were retained, as this range contains the most informative acoustic features for discriminating among dolphin species in the dataset. The same filtering was applied to all training and validation recordings. The soundscapes were not filtered because there was nothing relevant to dolphins within the recordings.

Although echolocation clicks occupy higher frequency bands, the publicly released SA-Acoustics demo dataset is dominated by whistle-based segments. As a result, CNN2 performance in this study is primarily driven by whistle structure rather than broadband click energy, making the selected band-pass range appropriate for the evaluated experiments.

Training is performed with and without augmentation after spectrogram generation, while the validation data remains unaugmented. Two model architectures are evaluated under identical pre-processing and data splits: the baseline CNN2 architecture from CetusID, trained from scratch, and the EfficientNetB0 transfer-learning classifier. This allows the sensitivity of both a lightweight CNN and a higher-capacity pretrained model to pre-processing choices to be assessed, as higher-capacity pretrained models may exploit filtered representations differently than shallow CNNs.

6.1.9 Model Training and Evaluation

CNN2 was trained from scratch for each experimental condition and cross-validation fold for 20 epochs using the Adam optimizer (learning rate = 0.001), categorical cross-entropy loss, a batch size of 32, and a dropout rate of 0.4. A fixed random seed (42) was used to ensure reproducibility.

Although Frainer et al. trained for 50 epochs, the publicly released CetusID implementation defaults to 20 epochs. Training curves indicated stable convergence within this limit, and extending training was neither necessary nor appropriate given the limited dataset size.

Performance was evaluated using LORO validation accuracy, reported as mean $\pm$ standard deviation across folds. Additional analyses using row-normalized confusion matrices, class-wise recall, and F1-scores were conducted to assess classification stability and inter-class behavior.

Chapter 7

Experimental Results

This chapter reports the experimental results obtained for the three sub-research questions defined in Chapter 1, focusing specifically on CNN2. Results are presented using LORO validation accuracy, confusion matrices, and class-wise recall, reported as mean ± standard deviation over three recording-level cross-validation folds. The analysis in this chapter is descriptive and comparative; causal interpretation and broader implications are discussed in Chapter 8. For completeness, class-wise F1-scores are reported in Appendix A.3.

To improve traceability, the chapter is organized by sub-research question: Section 7.0.1 addresses SRQ.2 (sampling rate), Section 7.0.2 addresses SRQ.3 (augmentation and controlled signal degradation), and Section 7.0.3 addresses SRQ.1 (band-pass pre-processing and model comparison).

7.0.1 Effect of Sampling Rate on Classification Performance

This subsection evaluates CNN2 classification performance using LORO validation accuracy under LORO cross-validation. As described in Chapter 6, data augmentation is disabled in these experiments to isolate sampling-rate effects. Performance is assessed using aggregate validation accuracy, confusion matrices, and class-wise recall to capture both overall trends and species-specific behavior across folds.

Table 7.1 summarizes the validation accuracy of CNN2 for the three evaluated sampling rates over three folds.

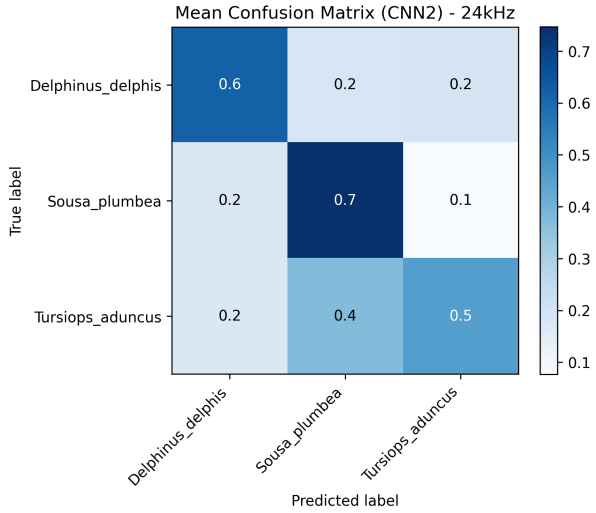
Table 7.1: CNN2 LORO validation accuracy for different audio sampling rates.

Sampling Rate	Fold Accuracies	Validation Accuracy
24 kHz	[0.800, 0.571, 0.455]	0.609 ± 0.143
48 kHz	[0.767, 0.667, 0.424]	0.619 ± 0.144
96 kHz	[0.867, 0.571, 0.545]	**0.661 ± 0.146**

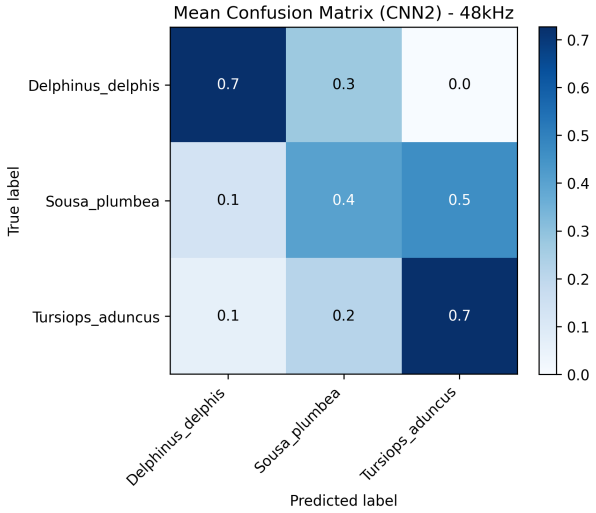
The mean LORO validation accuracy increases slightly with the sampling rate, reaching a maximum at 96 kHz. However, standard deviations remain high across all sampling rates, indicating substantial fold-to-fold variability. Specifically, the performance depends on which recordings are held out in the LORO evaluation.

Confusion Matrix Analysis

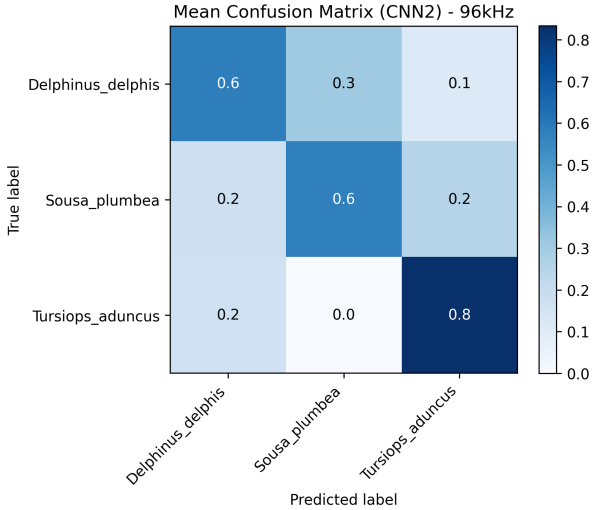
For each sampling rate, confusion matrices were computed per fold and aggregated by element-wise mean across folds; the resulting mean matrix was then row-normalized. Figure 7.1 shows the resulting row-normalized mean confusion matrices. Numeric mean confusion matrices are provided in Appendix A.1.



(a) 24 kHz



(b) 48 kHz



(c) 96 kHz

Figure 7.1: Row-normalized mean confusion matrices for CNN2 at sampling rates of 24 kHz, 48 kHz, and 96 kHz. Each matrix represents the element-wise mean across three folds, with rows corresponding to true labels and columns to predicted labels. Color intensity indicates the proportion of predictions per true class.

At 24 kHz (Figure 7.1a), CNN2 achieves its highest recall for Sousa plumbea at 0.747, while Tursiops aduncus shows substantial confusion with Sousa plumbea. Delphinus delphis shows moderate recall (0.620) and is misclassified at similar rates into the other two classes, indicating limited separability at this sampling rate.

At 48 kHz (Figure 7.1b), recall improves for Delphinus delphis and Tursiops aduncus, both exceeding 0.72. In contrast, Sousa plumbea exhibits increased confusion, particularly with Tursiops aduncus, suggesting that intermediate sampling rates do not uniformly improve class separability.

At 96 kHz (Figure 7.1c), CNN2 achieves its strongest performance for Tursiops aduncus, with a recall of 0.8 and no confusion with Sousa plumbea. Sousa plumbea remains partially confusable with Tursiops aduncus, while Delphinus delphis continues to show moderate recall and confusion with Sousa plumbea. Overall, increased sampling rate improves discrimination for specific species pairs but does not eliminate inter-species confusion.

Class-wise Recall Analysis

Table 7.2 reports class-wise recall across sampling rates, which is the mean ± standard deviation over three folds.

Table 7.2: Class-wise recall of CNN2 across sampling rates

Class	24 kHz	48 kHz	96 kHz
Delphinus delphis	0.620 ± 0.147	**0.727 ± 0.473**	0.580 ± 0.356
Sousa plumbea	**0.747 ± 0.274**	0.403 ± 0.464	0.573 ± 0.498
Tursiops aduncus	0.457 ± 0.412	**0.727 ± 0.473**	**0.833 ± 0.181**

Sampling-rate effects are species-dependent. Tursiops aduncus exhibits higher recall at 96 kHz compared to 24 kHz, whereas Delphinus delphis exhibits its highest mean recall at 48 kHz. Sousa plumbea shows its highest mean recall at 24 kHz. Standard deviations are large for multiple classes and sampling rates, indicating strong fold-to-fold variability in recall.

As 96 kHz yields the highest mean validation accuracy among the tested sampling rates (Table 7.1), the subsequent experiments evaluate controlled signal degradation and augmentation effects under fixed pre-processing settings, and the final experimental block uses a fixed raw audio sampling rate at 96 kHz as described in Chapter 6.

7.0.2 Effect of Data Augmentation Under Controlled Signal Degradation

This subsection addresses SRQ.3 by evaluating the effect of background-mixing augmentation (Chapter 6) in SA-Acoustics and SA-Acoustics (noise-degraded) datasets. Performance is summarized using training accuracy, validation accuracy, confusion matrices, and class-wise recall.

Table 7.3 summarizes training and validation accuracy across experimental conditions.

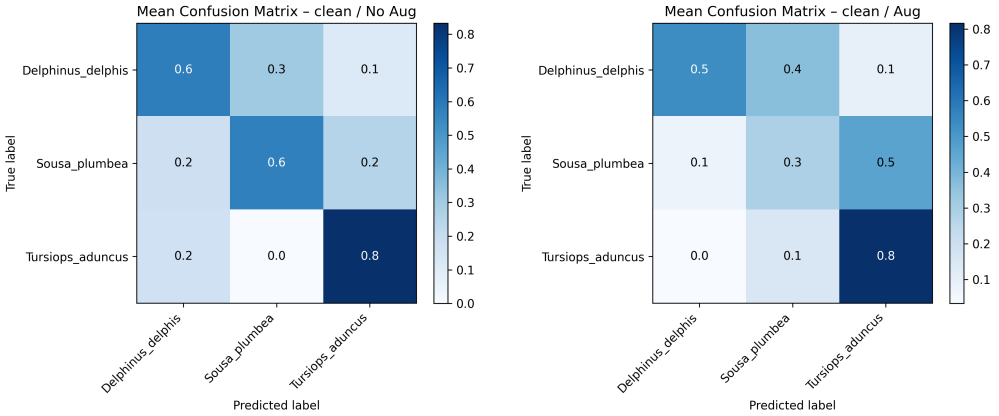
Table 7.3: CNN2 performance in the SA-Acoustics and SA-Acoustics (noise-degraded) datasets with and without data augmentation

Dataset	Setting	Train Acc	Val Acc
SA-Acoustics	No Augmentation	0.735 ± 0.118	**0.661 ± 0.178**
SA-Acoustics	Augmented	0.769 ± 0.150	0.603 ± 0.172
SA-Acoustics (noise-degraded)	No Augmentation	0.388 ± 0.051	0.429 ± 0.005
SA-Acoustics (noise-degraded)	Augmented	0.348 ± 0.097	0.483 ± 0.131

In the SA-Acoustics dataset, validation accuracy is lower with augmentation than without augmentation (0.603 vs. 0.661). In the SA-Acoustics (noise-degraded) dataset, validation accuracy is higher with augmentation than without augmentation (0.483 vs. 0.429). Notably, in the SA-Acoustics (noise-degraded) dataset, mean training accuracy is lower than mean validation accuracy in both settings. On the other hand, the SA-Acoustics dataset shows higher training than validation accuracy.

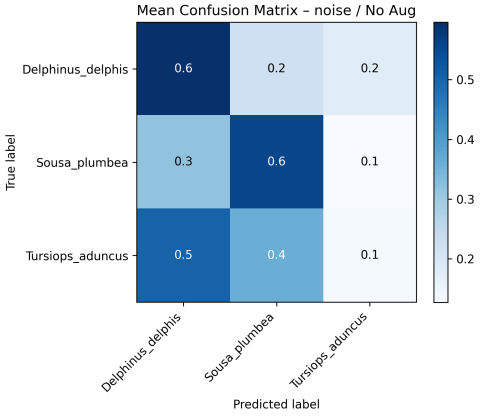
Confusion Matrix Analysis

Figure 7.2 presents row-normalized mean confusion matrices for SA-Acoustics and SA-Acoustics (noise-degraded) datasets, with and without augmentation, computed by first averaging confusion matrices across three folds and then row-normalizing. Numeric mean confusion matrices are provided in Appendix A.2.

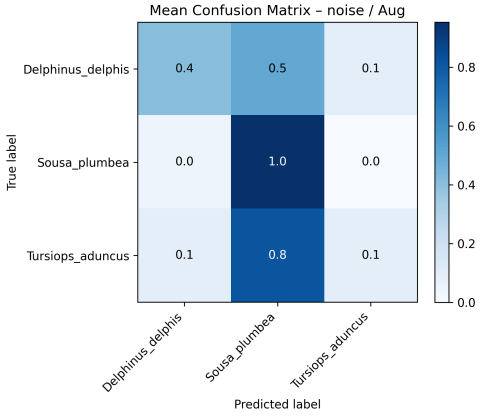


(a) SA-Acoustics No Aug

(b) SA-Acoustics Aug



(c) SA-Acoustics (noise-degraded) No Aug



(d) SA-Acoustics (noise-degraded) Aug

Figure 7.2: Average row-normalized confusion matrices of CNN2 under SA-Acoustics and SA-Acoustics (noise-degraded) datasets, with and without data augmentation. Each matrix represents the mean across all cross-validation folds; diagonal entries indicate per-class recall.

Under SA-Acoustics, non-augmented conditions (Figure 7.2a), the confusion matrix shows the highest diagonal values for Tursiops aduncus. Under SA-Acoustics augmented conditions (Figure 7.2b), confusion increases between Delphinus delphis and Sousa plumbea relative to the non-augmented condition.

In the SA-Acoustics (noise-degraded) dataset, non-augmented conditions (Figure 7.2c), diagonal values decrease, and misclassifications increase across all classes. In the SA-Acoustics (noise-degraded) augmented dataset, the diagonal entry for Sousa plumbea increases substantially (1.0), while diagonal entries for Delphinus delphis and Tursiops aduncus decrease (0.4 and 0.1, respectively), indicating a strong shift in per-class recall patterns under this configuration.

Class-wise Recall Analysis

Tables 7.4 and 7.5 report class-wise recall across SA-Acoustics and SA-Acoustics (noise-degraded) conditions.

Table 7.4: Class-wise recall of CNN2 under SA-Acoustics conditions, with and without data augmentation

Class	No Aug	Aug
Delphinus delphis	0.580 ± 0.356	0.537 ± 0.474
Sousa plumbea	0.573 ± 0.498	0.290 ± 0.387
Tursiops aduncus	**0.833 ± 0.181**	**0.817 ± 0.236**

Table 7.5: Class-wise recall of CNN2 under SA-Acoustics (noise-degraded) conditions, with and without data augmentation

Class	No Aug	Aug
Delphinus delphis	**0.597 ± 0.365**	0.408 ± 0.432
Sousa plumbea	0.557 ± 0.389	**0.953 ± 0.081**
Tursiops aduncus	0.133 ± 0.231	0.091 ± 0.157

Class-wise recall results show that augmentation effects are condition- and class-dependent. In the SA-Acoustics data, the recall for Tursiops aduncus is very high, ≈0.8, for both augmented and non-augmented data. Both Tursiops aduncus and Delphinus delphis show little difference when augmentation is applied to the data. However, Sousa plumbea shows a significant drop in recall from 0.573 to 0.290 when augmentation is applied.

On the other hand, in the SA-Acoustics (noise-degraded) dataset, recall for Sousa plumbea increases from 0.557 to 0.953 with augmentation, while recall for Tursiops aduncus remains low in both SA-Acoustics (noise-degraded) dataset configurations (0.133 without augmentation and 0.091 with augmentation). Standard deviations remain large for multiple classes, indicating substantial fold-to-fold variability.

7.0.3 Final evaluation: band-pass pre-processing and model comparison

This subsection addresses SRQ.1 and reports the main experimental results evaluating the combined effect of signal-level band-pass pre-processing and model architecture. The baseline CNN2 classifier is compared with the EfficientNetB0 transfer-learning model under no-band-pass-filtered and band-pass-filtered conditions, with and without augmentation. All conditions are evaluated using the same LORO folds ($k = 3$), and validation data remain unaugmented in all configurations (Chapter 6). Table 7.6 summarizes mean training and validation accuracy over three folds.

Table 7.6: Classification performance of the custom CNN2 from Frainer et al. [16] and EfficientNetB0 model over three recording-level cross-validation folds.

Model / Pre-processing	Train accuracy	Validation accuracy
CNN2 from Frainer et al.		
No-Band-pass (No Augmentation)	0.735 ± 0.118	$\mathbf{0.661 \pm 0.178}$
No-Band-pass + Augmentation	0.769 ± 0.150	0.603 ± 0.172
Band-pass (No Augmentation)	0.566 ± 0.090	0.572 ± 0.027
Band-pass + Augmentation	0.516 ± 0.169	0.463 ± 0.101
EfficientNetB0		
No-Band-pass + Augmentation	0.901 ± 0.171	0.778 ± 0.192
No-Band-pass (No Augmentation)	0.970 ± 0.053	0.700 ± 0.058
Band-pass (No Augmentation)	0.976 ± 0.014	$\mathbf{0.849 \pm 0.088}$
Band-pass + Augmentation	0.942 ± 0.026	0.836 ± 0.020

Table 7.7: Comparison of validation accuracies obtained under different models and evaluation protocols. The baseline CNN2 result is reported from a single random 80/20 validation split used solely for reproduction verification. On the other hand, the EfficientNetB0 result is obtained under LORO cross-validation. The values are shown to provide scale context only and are not directly comparable.

Source	Configuration	Validation Accuracy
Table 5.2	Own implementation using [16] CNN2 (Average)	0.843
Table 7.6	EfficientNet, band-pass (No Augmentation)	0.849

Table 7.7 places the baseline CNN2 validation accuracy reported in Table 5.2 alongside the best-performing EfficientNetB0 configuration from Table 7.6 to provide scale context. Both experiments used the SA-Acoustics dataset. The baseline value of approximately 0.84 is of similar magnitude to the LORO validation accuracy of 0.849 ± 0.088 observed in the final experiments.

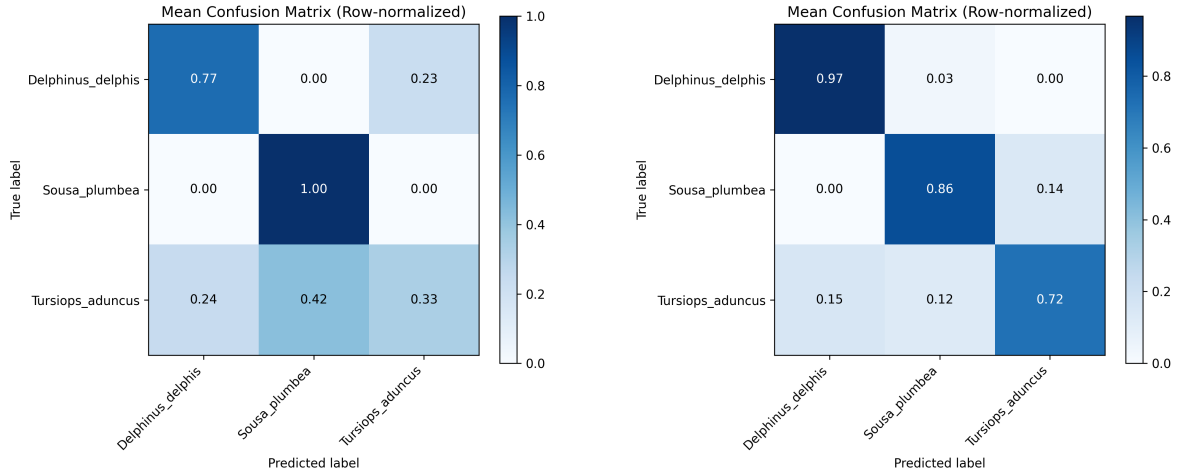
These values are not directly comparable. The baseline CNN2 result is obtained using a single random 80/20 train–validation split during baseline reproduction, whereas the EfficientNetB0 result is evaluated using LORO cross-validation with $k = 3$, which enforces recording-level generalization.

Across all experiments, three consistent tendencies emerge. First, data augmentation has a limited and highly conditional effect. Augmentation improves validation performance only for the EfficientNetB0 model under no-band-pass conditions. In all other configurations, augmentation either has no benefit or degrades validation performance, particularly for the baseline CNN2. This indicates that augmentation interacts with both model capacity and input representation rather than providing a uniform robustness gain.

Second, signal-level band-pass pre-processing consistently benefits EfficientNetB0 while consistently degrading CNN2 performance. For EfficientNetB0, band-pass filtering yields higher validation accuracy in both augmented and non-augmented settings and leads to the best-performing configuration observed in this study. In contrast, CNN2 shows reduced training and validation accuracy under band-pass filtering across all configurations, indicating that the baseline architecture fails to effectively exploit the filtered representation.

Third, EfficientNetB0 outperforms CNN2 across all evaluated configurations. This performance gap persists regardless of sampling rate, augmentation, or band-pass pre-processing, suggesting that model capacity and pretrained feature representations play a dominant role under the evaluated data constraints.

Moreover, to analyze class-level behavior for the best-performing model under the two relevant conditions, Figure 7.3a and Figure 7.3b show the row-normalized mean confusion matrices for EfficientNet under no-band-pass and band-pass filtered conditions without augmentation.



(a) No-Band-pass conditions without augmentation

(b) Band-pass pre-processing without augmentation

Figure 7.3: Row-normalized mean confusion matrices for EfficientNet with no augmentation. Figure (a) represents the result without pre-processing, and Figure (b) with band-pass pre-processing.

Under no-band-pass conditions (Figure 7.3a), EfficientNet already exhibits strong performance for Delphinus delphis and Sousa plumbea, while Tursiops aduncus remains partially confused with Sousa plumbea. After applying band-pass pre-processing (Figure 7.3b), predictions become more concentrated along the diagonal for all three species. In particular, misclassifications between Tursiops aduncus and Sousa plumbea are substantially reduced, indicating improved separability after suppressing background-dominated frequency components.

Table 7.8 reports class-wise recall for EfficientNet under these two conditions, derived from the row-normalized confusion matrices.

Table 7.8: Class-wise recall (mean ± standard deviation) for EfficientNet under no-band-pass and band-pass filtered conditions without augmentation.

Species	**No-Band-pass (No Aug)**	**Band-pass (No Aug)**
Delphinus delphis	0.767 ± 0.404	$\mathbf{0.967 \pm 0.058}$
Sousa plumbea	$\mathbf{1.000 \pm 0.000}$	0.857 ± 0.248
Tursiops aduncus	0.333 ± 0.577	$\mathbf{0.724 \pm 0.394}$

Band-pass pre-processing leads to substantial recall improvements for Delphinus delphis and Tursiops aduncus, while recall for Sousa plumbea decreases slightly but remains high. Overall, these results indicate that signal-level band-pass filtering enhances class separability when combined with a transfer-learning model, indicating that these signals are not effectively exploited by the baseline CNN2. Full class-wise recall results for all models and pre-processing configurations are provided in Appendix A.11.

Chapter 8

Discussion

This chapter interprets the experimental results presented in Chapter 7 in relation to the sub-research questions defined in Chapter 1. Given the limited size of the SA-Acoustics dataset and the resulting variability observed across recording-level cross-validation folds, the discussion is intentionally conservative. The focus is on observations consistently supported by the reported results, while avoiding generalization beyond the evaluated experimental setting.

Overall, the results highlight three findings. First, changes in the audio sampling rate produce only small differences in mean validation accuracy compared with the observed fold-to-fold variability. Second, the effect of background-mixing augmentation depends strongly on the acoustic condition and redistributes classification errors across classes rather than uniformly improving performance. Finally, signal-level band-pass pre-processing interacts strongly with model architecture, degrading performance for the baseline CNN2 while substantially improving performance for the EfficientNetB0 transfer-learning model.

8.1 Effect of Sampling Rate on Species Classification

The sampling-rate experiments show that increasing the audio sampling rate from 24 kHz to 96 kHz results in only modest changes in mean LORO validation accuracy (Table 7.1). These differences are consistently smaller than the fold-to-fold variability observed across recording-level splits, indicating that performance is dominated by recording-specific effects rather than by sampling rate alone in this dataset.

Importantly, all experiments use fixed-size spectrogram images as input to the CNN. As a result, changes in sampling rate do not alter the dimensionality of the input representation but only affect the frequency content encoded within it. Within this constrained setting, no sampling rate yields consistently superior performance across folds.

Confusion-matrix analysis (Figure 7.1 and Appendix A.1) shows that misclassification patterns vary across sampling rates; however, these variations are accompanied by large standard deviations and are not stable across folds. Consequently, apparent class-wise differences in recall across sampling rates should be interpreted with caution and are likely driven by recording-level variability rather than systematic effects of sampling rate.

These observations are consistent with prior bio-acoustic findings showing that higher sampling rates preserve fine temporal and spectral structure without necessarily yielding improved downstream classification performance. Papale et al. [41], for example, report that higher sampling rates retain high-frequency components of dolphin burst pulses, while downsampling alters measured acoustic features. In the present experiments, however, preservation of higher-frequency detail does not consistently translate into improved classification performance, which aligns with the strong recording-level variability and data limitations observed in this study.

Overall, the results indicate that, within the SA-Acoustics dataset and under fixed spectrogram representations, sampling rate is not a dominant factor governing CNN2 classification performance. These findings should be interpreted as dataset-specific and exploratory rather than as evidence for an optimal sampling rate for dolphin species classification.

8.2 Impact of Data Augmentation Under Controlled Signal Degradation

The augmentation experiments show that background-mixing augmentation does not have a uniform effect on CNN2 performance and instead interacts with the acoustic condition of the data. In the SA-Acoustics dataset, augmentation increases training accuracy while reducing validation accuracy relative to the non-augmented configuration (Table 7.3). This suggests a mismatch between the augmented training distribution and the unaugmented validation data used in this study.

In contrast, in the SA-Acoustics (noise-degraded) dataset, augmentation increases mean validation accuracy compared to the non-augmented condition. This indicates that augmentation partially mitigates the specific synthetic degradation introduced by additive Gaussian noise. However, this improvement should not be interpreted as general robustness, as the applied noise does not reflect the complexity of real underwater acoustic environments.

Class-wise analyses indicate that background-mixing augmentation redistributes classification errors rather than uniformly improving performance. In the SA-Acoustics (noise-degraded) dataset, augmentation leads to a strong shift in class-wise recall, with one class achieving very high recall while recall for the remaining classes decreases substantially (Table 7.5). This pattern reflects a concentration of predictions toward a dominant class in some folds, as also visible in the corresponding confusion matrices, and indicates a failure mode in which the classifier collapses toward a single prediction under controlled signal degradation.

Taken together, these results indicate that background-mixing augmentation does not provide consistent performance gains in this setting. Instead, its effect depends on the acoustic condition and can introduce class-conditional trade-offs. Given the limited dataset and the use of a single background-soundscape recording, these findings should be interpreted as exploratory sensitivity analyses rather than as general conclusions about augmentation effectiveness in PAM.

8.3 Effect of Band-Pass Pre-processing and Model Capacity

The final experiments demonstrate a strong interaction between signal-level band-pass pre-processing and model architecture. For the baseline CNN2, applying band-pass filtering consistently reduces both training and validation accuracy across all evaluated configurations (Table 7.6). The reduction in training accuracy indicates that CNN2 struggles to fit the filtered representation under the constraints of the available data.

In contrast, the EfficientNetB0 transfer-learning model benefits substantially from band-pass pre-processing. Validation accuracy increases under band-pass filtering in both augmented and non-augmented settings, with the highest performance observed for the band-pass, no-augmentation configuration. Confusion-matrix analysis shows improved diagonal concentration and reduced inter-class confusion under this condition.

Class-wise recall results indicate that band-pass filtering improves recall for multiple classes when combined with EfficientNetB0, although some class-to-class trade-offs remain (Table 7.8). Given the small number of recordings per species and the observed variability across folds, these improvements should be interpreted as indicative trends rather than as reliable performance estimates.

A plausible interpretation of these results is that the baseline CNN2 relies partly on background-dominated spectral regions when performing species classification. In the unfiltered setting, low-frequency regions that contain little dolphin vocalization but reflect recording-specific characteristics remain present in the input. When band-pass filtering removes these regions, CNN2 performance degrades consistently, suggesting that the model does not effectively exploit vocalization-dominant frequency content alone.

In contrast, EfficientNetB0 benefits from band-pass pre-processing, indicating that the higher-capacity pretrained model is better able to extract species-relevant information from vocalization structure while being less dependent on background cues. This interpretation is consistent with the observed trends: band-pass filtering is always beneficial for EfficientNetB0, never beneficial for CNN2, and EfficientNetB0 consistently outperforms CNN2 across all configurations.

Overall, these findings suggest that signal-level pre-processing cannot be evaluated independently of model capacity. In this dataset, the baseline CNN2 does not effectively exploit the filtered representation, whereas a higher-capacity pretrained model can benefit from the removal of background-dominated frequency regions. This interaction underscores the importance of jointly considering pre-processing choices and model architecture in PAM-based classification pipelines.

8.4 Limitations

Several limitations constrain the interpretation and generalizability of these results.

Limited data and unstable estimates. The dataset contains only three annotated recordings per species, requiring LORO cross-validation with $k = 3$. While this prevents segment-level leakage, it results in high variance and unstable class-wise estimates. Reported mean differences should therefore be interpreted as indicative trends rather than reliable population-level estimates.

No independent labeled test set. All evaluations are restricted to cross-validation splits drawn from the same small recording pool. Generalization to unseen recording sites, sensors, or environmental conditions cannot be directly assessed.

Demo framework constraints. Model architectures, training hyperparameters, and core processing steps were inherited from the publicly released CetusID demo implementation. The full datasets and training configurations used in the original study are not publicly available, limiting architectural exploration and systematic hyperparameter optimization.

Segment-level evaluation. Performance is evaluated at the segment level using overlapping 2 s windows. This differs from encounter-level aggregation in the original CetusID study and may emphasize different error patterns than those encountered in deployment scenarios.

Noise and augmentation representativeness. Controlled signal degradation is evaluated using additive Gaussian noise and augmentation based on a single background-soundscape recording. These conditions provide a controlled stress test but do not capture the diversity of real underwater acoustic environments.

Robustness is evaluated only with respect to additive Gaussian noise and background-mixing augmentation derived from a single soundscape recording. These conditions do not reflect the spectral, temporal, or non-stationary characteristics of real underwater noise

Single-seed training. All experiments use a single fixed random seed. While this ensures reproducibility, it does not quantify variability due to stochastic optimization.

Downsampling confound. Lower sampling rates are obtained through downsampling, which removes high-frequency content and may alter spectral statistics. Observed sampling-rate effects therefore reflect reduced-bandwidth representations within this pipeline rather than optimal acquisition settings.

Chapter 9

Conclusion

This thesis examined how pre-processing and representation choices affect the performance and behavior of CNNs for dolphin species classification in PAM within the scope of the evaluated dataset and experimental framework. Using the publicly available CetusID framework, controlled experiments were conducted to evaluate the effects of audio sampling rate, data augmentation, and signal-level band-pass pre-processing on species classification. All experiments followed a controlled design in which pre-processing and representation choices were varied while model training and evaluation protocols were kept fixed. Results were evaluated using LORO cross-validation on a limited dataset comprising three dolphin species.

9.1 Answer to the Research Question

The main research question addressed in this thesis was:

How do signal-level band-pass filtering, audio sampling rate, and data augmentation influence the performance of CNN-based dolphin species classification from PAM audio segments?

The results indicate that, in this study, pre-processing choices do not provide uniform performance improvements and instead exhibit context-dependent effects. Changes in audio sampling rate lead to only modest differences in validation accuracy, with variations that are small relative to the observed fold-to-fold variability. Class-wise analyses show that sampling-rate effects differ across species, but no sampling rate consistently outperforms others across all folds and classes.

Data augmentation exhibits strongly condition-dependent behavior. In the SA-Acoustics dataset, augmentation reduces validation performance, whereas in the SA-Acoustics (noise-degraded) dataset, it improves validation accuracy. However, this improvement is class dependent and primarily reflects a redistribution of classification errors rather than a uniform robustness gain.

Signal-level band-pass pre-processing does not improve performance for the baseline CNN2 architecture under the evaluated conditions and is associated with reduced training and validation accuracy. In contrast, when combined with a higher-capacity transfer-learning model, band-pass pre-processing yields a substantial increase in validation accuracy and reduces inter-class confusion. This indicates that the effectiveness of signal-level pre-processing depends strongly on model capacity and the ability to exploit vocalization-dominant spectral information.

Overall, the findings suggest that pre-processing choices in PAM-based dolphin classification should be evaluated jointly with model architecture and data constraints, rather than assumed to be universally beneficial.

9.2 Contributions

This thesis makes the following contributions:

- A reproducible validation of the CetusID framework under recording-level cross-validation.
- A controlled analysis of audio sampling rate effects, indicating species-dependent performance behavior.
- An evaluation of data augmentation for the SA-Acoustics and SA-Acoustics (noise-degraded) datasets, pointing to class-conditional trade-offs.
- A investigation of signal-level band-pass pre-processing, showing that when combined with an EfficientNetB0 transfer-learning model, it achieves the highest validation accuracy (0.849 ± 0.088) under LORO cross-validation in this study.
- A preliminary class-wise performance analysis of aggregate accuracy in PAM classification.

9.3 Future Work

Future work should evaluate these findings on larger and more diverse datasets to reduce variance and assess generalization across recording sites and environmental conditions. Encounter-level evaluation and aggregation should be explored to better reflect deployment scenarios. In addition, pre-processing strategies should be developed in a species-aware manner, and robustness testing should be extended to more realistic underwater noise sources. Finally, multi-seed training and broader architectural comparisons would improve statistical reliability.

Bibliography

[1] 100 Days of ML Code. Day 50 of 100daysofml: Cross-validation. https://medium.com/100daysofmlcode/day-50-of-100daysofml-b1b14949dfe7, 2018. Accessed: 2026-01-09.

[2] Thiago OS Amorim, Franciele R de Castro, Giovanne A Ferreira, Fernanda M Neri, Bruna R Duque, João P Mura, and Artur Andriolo. Acoustic identification and classification of four dolphin species in the brazilian marine area affected by the largest tailings dam failure disaster. *The Journal of the Acoustical Society of America*, 152(6):3204–3215, 2022.

[3] Whitlow WL Au. *The sonar of dolphins*. Springer Science & Business Media, 1993.

[4] Whitlow WL Au and Mardi C Hastings. *Principles of marine bioacoustics*, volume 510. Springer, 2008.

[5] Audacity Team. Audacity®: Free audio editor and recorder, 2025. URL https://www.audacityteam.org. Accessed: January 2026.

[6] Michael J Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A Roch, Sharon Gannot, and Charles-Alban Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628, 2019.

[7] Isha Bopardikar, Dipani Sutaria, Mihir Sule, Ketki Jog, Vardhan Patankar, and Holger Klinck. Description and classification of indian ocean humpback dolphin (sousa plumbea) whistles recorded off the sindhudurg coast of maharashtra, india. *Marine Mammal Science*, 34(3):755–776, 2018.

[8] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[9] Rocco De Marco, Francesco Di Nardo, Alessandro Lucchetti, Massimo Virgili, Daniel Li Veli, Andrea Petetta, Laura Screpanti, and David Scaradozzi. Bottlenose dolphin's (tursiops truncatus, montagu 1821) acoustic emissions recorded during interaction with bottom trawl nets in the central-northern adriatic sea. https://figshare.com/collections/Bottlenose_dolphin_s_Tursiops_truncatus_Montagu_1821_acoustic_emissions_recorded_during_interaction_with_bottom_trawl_nets_in_the_central-northern_Adriatic_Sea/6313308, 2023. Version 2, accessed: 2025-12-27.

[10] Francesco Di Nardo, Rocco De Marco, Daniel Li Veli, Laura Screpanti, Benedetta Castagna, Alessandro Lucchetti, and David Scaradozzi. Multiclass cnn approach for automatic classification of dolphin vocalizations. *Sensors*, 25(8):2499, 2025.

[11] S Dines, R Probert, A Gullan, S Elwen, G Frainer, and T Gridley. Case study: Evidence of long-term stability in a stereotyped whistle in a single free-ranging humpback dolphin (sousa plumbea) found in sympatry (tursiops aduncus). *JASA Express Letters*, 4(12), 2024.

[12] Carlos M Duarte, Lucille Chapuis, Shaun P Collin, Daniel P Costa, Reny P Devassy, Victor M Eguiluz, Christine Erbe, Timothy AC Gordon, Benjamin S Halpern, Harry R Harding, et al. The soundscape of the anthropocene ocean. *Science*, 371(6529):eaba4658, 2021.

[13] Andres Ferraro, Dmitry Bogdanov, Xavier Serra Jay, Ho Jeon, and Jason Yoon. How low can you go? reducing frequency and time resolution in current cnn architectures for music auto-tagging. In *2020 28th European signal processing conference (EUSIPCO)*, pages 131–135. IEEE, 2021.

[14] G. Frainer et al. Cetusid: Demo dataset for automatic dolphin identification, 2023. URL `https://zenodo.org/records/8074949`.

[15] Guilherme Frainer. Cetusid: Deep learning framework for dolphin vocalization identification. `https://github.com/Gui-Frainer/CetusID`, 2023. Accessed: 20 December 2025.

[16] Guilherme Frainer, Emmanuel Dufourq, Jack Fearey, Sasha Dines, Rachel Probert, Simon Elwen, and Tess Gridley. Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring. *Ecological Informatics*, 78:102291, 2023.

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[18] Douglas Gillespie. Detection and classification of right whale calls using an'edge'detector operating on a smoothed spectrogram. *Canadian Acoustics*, 32(2):39–47, 2004.

[19] Emily T Griffiths. Whistle repertoire analysis of the short-beaked common dolphin, delphinus delphis, from the celtic deep and the eastern tropical pacific ocean. *Bangor University*, 2009.

[20] Elizabeth R Hawkins. Geographic variations in the whistles of bottlenose dolphins (tursiops aduncus) along the east and west coasts of australia. *The Journal of the Acoustical Society of America*, 128 (2):924–935, 2010.

[21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[22] Denise L Herzing. Clicks, whistles and pulses: Passive and active signal use in dolphin communication. *Acta Astronautica*, 105(2):534–537, 2014.

[23] John A Hildebrand. Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, 395:5–20, 2009.

[24] Vincent M Janik and Laela S Sayigh. Communication in bottlenose dolphins: 50 years of signature whistle research. *Journal of Comparative Physiology A*, 199(6):479–489, 2013.

[25] Dae-Hyun Jung, Na Yeon Kim, Sang Ho Moon, Hyoung Seok Kim, Taek Sung Lee, Jung-Seok Yang, Ju Young Lee, Xiongzhe Han, and Soo Hyun Park. Classification of vocalization recordings of laying hens and cattle using convolutional neural network models. *Journal of Biosystems Engineering*, 46 (3):217–224, 2021.

[26] Ivan Kiskin, Davide Zilli, Yunpeng Li, Marianne Sinka, Kathy Willis, and Stephen Roberts. Bioacoustic detection with wavelet-conditioned convolutional neural networks. *Neural Computing and Applications*, 32(4):915–927, 2020.

[27] Elly C Knight, Sergio Poo Hernandez, Erin M Bayne, Vadim Bulitko, and Benjamin V Tucker. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 29(3):337–355, 2020.

[28] Marc O Lammers, Whitlow WL Au, and Denise L Herzing. The broadband social acoustic signaling behavior of spinner and spotted dolphins. *The Journal of the Acoustical Society of America*, 114 (3):1629–1639, 2003.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[31] Loïc Lehnhoff, Hervé Glotin, Yves Le Gall, Eric Menut, Hélène Peltier, Jérôme Spitz, Olivier Van Canneyt, and Bastien Mérigot. High resolution acoustic recordings of wild free-ranging short-beaked common dolphins for etho-acoustical and repertoire studies. *Earth System Science Data Discussions*, 2025:1–22, 2025.

[32] Jennifer MacIsaac, Stuart Newson, Adham Ashton-Butt, Huma Pearce, and Ben Milner. Improving acoustic species identification using data augmentation within a deep learning framework. *Ecological Informatics*, 83:102851, 2024.

[33] Sarah A. Marley, Christine Erbe, and Chandra P. Salgado Kent. Underwater recordings of the whistles of bottlenose dolphins in fremantle inner harbour, western australia. *Figshare Dataset*, 2017. doi: 10.6084/m9.figshare.5011637. Accessed: 2025-12-27.

[34] David K Mellinger and Christopher W Clark. Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, 107(6):3518–3529, 2000.

[35] David K. Mellinger, Kathleen M. Stafford, Sue E. Moore, Robert P. Dziak, and Haru Matsumoto. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography*, 20(4):36–45, 2007. doi: 10.5670/oceanog.2007.03. URL `https://doi.org/10.5670/oceanog.2007.03`.

[36] David K Mellinger, Kathleen M Stafford, Sue E Moore, Robert P Dziak, and Haru Matsumoto. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography*, 20(4):36–45, 2007.

[37] Fernando Merchan, Ariel Guerra, Héctor Poveda, Héctor M Guzmán, and Javier E Sanchez-Galan. Bioacoustic classification of antillean manatee vocalization spectrograms using deep convolutional neural networks. *Applied Sciences*, 10(9):3286, 2020.

[38] Marius Miron, David Robinson, Milad Alizadeh, Ellen Gilsenan-McMahon, Gagan Narula, Emmanuel Chemla, Maddie Cusimano, Felix Effenberger, Masato Hagiwara, Benjamin Hoffman, et al. What matters for bioacoustic encoding. *arXiv preprint arXiv:2508.11845*, 2025.

[39] Alan V Oppenheim and George C Verghese. *Signals, systems & inference*. Pearson London, UK:, 2017.

[40] Elena Papale, Marco Gamba, Monica Perez-Gil, Vidal Martel Martin, and Cristina Giacoma. Dolphins adjust species-specific frequency parameters to compensate for increasing background noise. *PLoS One*, 10(4):e0121711, 2015.

[41] Elena Papale, Giuseppe Alonge, Francesco Caruso, Rosario Grammauta, Salvatore Mazzola, Barbara Mussi, Daniela S Pace, and Giuseppa Buscaino. The higher, the closer, the better? influence of sampling frequency and distance on the acoustic properties of short-beaked common dolphins burst pulses in the mediterranean sea. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31: 51–60, 2021.

[42] Vincenzo Petrella, Emmanuelle Martinez, Michael G Anderson, and Karen A Stockin. Whistle characteristics of common dolphins (delphinus sp.) in the hauraki gulf, new zealand. *Marine Mammal Science*, 28(3):479–496, 2012.

[43] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015.

[44] Nicholas Rasmussen, Rodrigue Rizk, Omera Matoo, and KC Santosh. Deepwhalenet: Climate change-aware fft-based deep neural network for passive acoustic monitoring. *International Journal of Pattern Recognition and Artificial Intelligence*, 38(14):2459014, 2024.

[45] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283, 2017.

[46] Laela S Sayigh, Vincent M Janik, Frants H Jensen, Michael D Scott, Peter L Tyack, and Randall S Wells. The sarasota dolphin whistle database: A unique long-term resource for understanding dolphin communication. *Frontiers in Marine Science*, 9:923046, 2022.

[47] SEANOE Dataset. Bottlenose dolphin vocalizations in controlled environments. `https://www.seanoe.org/data/00979/109081/`, 2025. Accessed: 2025-12-27.

[48] Kathleen Mary Stafford. Characterization of blue whale calls from the northeast pacific and development of a matched filter to locate blue whales on the us navy sosus (sound surveillance system) arrays. 1995.

[49] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10: e13152, 2022.

[50] Rhianne Ward, Iain Parnum, Christine Erbe, and Chandra Salgado-Kent. Whistle characteristics of indo-pacific bottlenose dolphins (tursiops aduncus) in the fremantle inner harbour, western australia. *Acoustics Australia*, 44(1):159–169, 2016.

[51] Gordon M Wenz. Acoustic ambient noise in the ocean: spectra and sources. *The journal of the acoustical society of America*, 34(12):1936–1956, 1962.

[52] Ellen L White, Paul R White, Jonathan M Bull, Denise Risch, Susanna Quer, and Suzanne Beck. Evaluating the performance of automated detection systems for long-term monitoring of delphinids in diverse marine soundscapes. *PLoS One*, 20(6):e0323768, 2025.

[53] Xiang Zhou, Ru Wu, Wen Chen, Meiling Dai, Peibin Zhu, and Xiaomei Xu. Thresholding dolphin whistles based on signal correlation and impulsive noise features under stationary wavelet transform. *Journal of Marine Science and Engineering*, 13(2):312, 2025.

[54] Walter MX Zimmer. *Passive acoustic monitoring of cetaceans*. Cambridge University Press, 2011.

Appendix A

Additional Results

A.1 Numeric Mean Confusion Matrices for Sampling-Rate Experiments

Tables A.1–A.3 report the numeric mean confusion matrices for the sampling-rate experiments. Each entry corresponds to the element-wise mean across three cross-validation folds. Rows represent true labels and columns represent predicted labels.

Table A.1: Numeric mean confusion matrix for CNN2 at 24 kHz (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.620**	0.183	0.197
Sousa plumbea	0.177	**0.747**	0.080
Tursiops aduncus	0.173	0.370	**0.457**

Table A.2: Numeric mean confusion matrix for CNN2 at 48 kHz (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.727**	0.273	0.000
Sousa plumbea	0.133	**0.403**	0.463
Tursiops aduncus	0.060	0.213	**0.727**

Table A.3: Numeric mean confusion matrix for CNN2 at 96 kHz (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.580**	0.307	0.113
Sousa plumbea	0.177	**0.573**	0.250
Tursiops aduncus	0.167	0.000	**0.833**

A.2 Numeric Mean Confusion Matrices for Augmentation and Noise Experiments

Tables A.4–A.7 report the numeric mean confusion matrices for the augmentation and controlled signal degradation experiments. Each entry corresponds to the element-wise mean across three cross-validation folds. Rows represent true labels and columns represent predicted labels.

Table A.4: Numeric mean confusion matrix for CNN2 with SA-Acoustics dataset and no augmentation (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.580**	0.307	0.113
Sousa plumbea	0.177	**0.573**	0.250
Tursiops aduncus	0.167	0.000	**0.833**

Table A.5: Numeric mean confusion matrix for CNN2 with SA-Acoustics dataset and augmentation (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.537**	**0.367**	0.100
Sousa plumbea	0.080	0.290	**0.463**
Tursiops aduncus	0.030	0.150	0.817

Table A.6: Numeric mean confusion matrix for CNN2 with the SA-Acoustics (noise-degraded) dataset and no augmentation (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.597**	0.223	**0.180**
Sousa plumbea	0.317	**0.557**	0.127
Tursiops aduncus	0.503	0.363	0.133

Table A.7: Numeric mean confusion matrix for CNN2 with the SA-Acoustics (noise-degraded) dataset and augmentation (mean over three folds).

True / Predicted	Delphinus	Sousa	Tursiops
Delphinus delphis	**0.407**	0.500	0.090
Sousa plumbea	0.047	**0.953**	0.000
Tursiops aduncus	**0.090**	0.817	**0.090**

A.3 Class-wise F1-Scores

A.3.1 F1-Scores Across Sampling Rates

Table A.8: Class-wise F1-score of CNN2 across sampling rates (mean ± standard deviation over three folds).

Class	24 kHz	48 kHz	96 kHz
Delphinus delphis	0.625 ± 0.074	**0.699 ± 0.386**	0.573 ± 0.296
Sousa plumbea	0.444 ± 0.396	0.607 ± 0.298	**0.761 ± 0.130**
Tursiops aduncus	**0.667 ± 0.244**	0.334 ± 0.292	0.507 ± 0.461

A.3.2 F1-Scores for the SA-Acoustics and SA-Acoustics (noise-degraded) Datasets

Table A.9: Class-wise F1-score of CNN2 in the **SA-Acoustics dataset** with and without data augmentation (mean ± standard deviation over three folds).

Class	SA-Acoustics (No Aug)	SA-Acoustics (Aug)
Delphinus delphis	0.573 ± 0.296	0.556 ± 0.486
Sousa plumbea	0.507 ± 0.461	0.236 ± 0.230
Tursiops aduncus	**0.761 ± 0.130**	**0.684 ± 0.059**

Table A.10: Class-wise F1-score of CNN2 in the **SA-Acoustics (noise-degraded) dataset** with and without data augmentation (mean ± standard deviation over three folds).

Class	SA-Acoustics (noise-degraded, No Aug)	SA-Acoustics (noise-degraded, Aug)
Delphinus delphis	0.476 ± 0.092	0.435 ± 0.430
Sousa plumbea	**0.497 ± 0.087**	**0.595 ± 0.108**
Tursiops aduncus	0.133 ± 0.231	0.118 ± 0.204

A.4 Recall Table for Main Experiment

Table A.11: Per-class recall (mean ± standard deviation) over three recording-level cross-validation folds, derived from the confusion matrices.

Model / Pre-processing	***Delphinus delphis***	***Sousa plumbea***	***Tursiops aduncus***
CNN2 from Frainer et al.			
No-Band-pass (No Augmentation)	$\mathbf{0.580 \pm 0.356}$	0.573 ± 0.498	$\mathbf{0.833 \pm 0.181}$
No-Band-pass + Augmentation	0.537 ± 0.474	0.290 ± 0.387	0.817 ± 0.236
Band-pass (No Augmentation)	0.390 ± 0.361	$\mathbf{0.639 \pm 0.140}$	0.688 ± 0.290
Band-pass + Augmentation	0.100 ± 0.173	0.606 ± 0.533	0.684 ± 0.471
EfficientNet			
No-Band-pass (No Augmentation)	0.767 ± 0.404	$\mathbf{1.000 \pm 0.000}$	0.333 ± 0.577
No-Band-pass + Augmentation	$\mathbf{1.000 \pm 0.000}$	0.333 ± 0.577	$\mathbf{1.000 \pm 0.000}$
Band-pass (No Augmentation)	0.967 ± 0.058	0.857 ± 0.248	0.724 ± 0.394
Band-pass + Augmentation	0.903 ± 0.100	0.852 ± 0.150	0.753 ± 0.230