Preprints.org

Article

Not peer-reviewed version

Advancing Dolphin Acoustic Monitoring: A Comprehensive Whistle Classification Framework

Ming Xiang , Luobin Wang , Yankun Chen * , Kangrong Li * , Zhengqiao Zhao , Jie Chen

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1083.v1

Keywords: dolphin whistle signal; CNN; classification; marine biology; acoustic signal processing

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Advancing Dolphin Acoustic Monitoring: A Comprehensive Whistle Classification Framework

Ming Xiang [1,3], Luobin Wang [2,4] and Yankun Chen [1,3,*], Kangrong Li [1,3,*], Zhengqiao Zhao [2,4] and Jie Chen [2,4]

[1] South China Sea Marine Survey Center,Ministry of Natural Resources, Guangzhou 510300, China
[2] School of Marine Science and Technology,Northwestern Polytechnical University
[3] Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources, Guangzhou 510300, China
[4] Center of Intelligent Acoustics and Immersive Communications, Shaanxi Provincial Key Laboratory of Artificial Intelligence
* Correspondence: easiechen@qq.com, (Y.C.); 1556233203@qq.com (K.L.); Tel.: +0000-0001-5002-1510 (Y.C.); +0000-0001-5522-2733 (K.L.)

Abstract

Dolphins are widely recognized as intelligent marine mammals with sophisticated communication and echolocation. Accurately classifying their whistles is essential for understanding how they communicate and for tracking population size, structure, and distribution. Here, we assemble a large, high-quality dataset of dolphin whistle signals collected at the Chimelong Ocean Kingdom, including a whistle type not previously available to researchers. We then explore Convolutional Neural Networks (CNNs) for classifying whistles of the Indo-Pacific bottlenose dolphin (Tursiops aduncus), testing 5 CNN architectures to analyse the signals. Model performance is reported using mean Average Precision (mAP), showing that CNN approaches can reliably separate different whistle classes. To probe robustness, we also introduce noise at defined SNR levels to increase testing complexity and assess the stability of the classifier. We use Bellhop for channel simulation to construct the channel impulse response. The simulated data can be used as augmented data to add to the original data training set. The results did indicate that this can enhance the robustness of the classification model. This work provides valuable tools for marine biologists and researchers specialising in animal acoustics, enhancing the understanding of dolphin communication. It also contributes to the conservation and management efforts of dolphin populations, offering significant insights into their behaviour and ecological needs.

Keywords: dolphin whistle signal; CNN; classification; marine biology; acoustic signal processing

1. Introduction

Dolphins have an exceptional sonar system to adapt to their living environment. In their sonar system, there are three major acoustic signal types [1,2]. The first one is clicks or tickings, which are high-frequency, narrow pulse signals used for underwater navigation, localization, foraging, and obstacle avoidance. They have remarkable adaptive and anti-jamming abilities. The second type is whistles, which are a communication signal. The fundamental frequency of the communication signals (whistles) is mainly ranged from 500 Hz to 24 kHz, and the signal length is usually between 0.1 and 2 seconds; for those in the pregnancy period, the signal length might be longer than 3 seconds. They serve various functions, including population communication, emotional expression, long-distance communication, and individual identification [3]. The third type is burst pulses (distress signals), which usually show up when dolphins are tense, such as when they are angry, fearful, frustrated, or under acute stress. These sounds can come off like rough barks, drawn-out howls, or rapid trills. They feature shorter pulse intervals and lower intensity compared to echolocation clicks. These signals are usually challenging to characterize, which makes the research on burst

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

pulses difficult [4]. Hence, current studies on dolphin acoustic signals are mainly on echolocation and communication signals.

The Indo-Pacific bottlenose dolphin (*Tursiops aduncus*) is a medium-sized cetacean in the family Delphinidae. Compared to the common bottlenose dolphin, it is typically slimmer with more speckling on the belly. It usually inhabits warm coastal and shelf waters across the Indian Ocean and western Pacific [5]. It is brilliant and adaptable, it feeds on fish and cephalopods, and is known for cooperative hunting and frequent interactions with boats. In China, this species is identified as a nationally protected Class II animal. Extensive efforts have been devoted to studying their bioacoustic signals for detection and conservation.

Current public marine mammal acoustic datasets [6–8] are constrained by limited sample sizes, low signal-to-noise ratios, variable recording conditions, and ambiguous annotations. Consequently, models trained on these data often yield suboptimal performance in detection and recognition tasks. Furthermore, a systematic framework for training and evaluating different dolphin call detection models is lacking.

To address these issues, this study constructs a high-quality whistle signal dataset for the *Tursiops aduncus*. The acoustic signals are recorded in a quiet and controlled environment and subsequently annotated by human experts. We further develop a framework to comprehensively train and evaluate deep convolutional neural networks for robust whistle classification. Specifically, we explore the performance of various CNN architectures using different feature representations as input. To enhance model generalizability under complex, realistic conditions, we simulate underwater whistle sound propagation using Bellhop, a widely used underwater acoustic raytracing model and augment our dataset with *in situ* oceanic noise. The resulting synthetic signals are used for both training and evaluation, thereby improving the robustness of the classification models.

The remainder of the paper is organized as follows: Section II reviews research on marine-mammal sound recognition; Section III describes the collection and annotation of *Tursiops aduncus* signal records from Chimelong; Section IV reports experimental setup and classification results; and Section V concludes with a summary and directions for future work.

2. Literature Reviews

2.1. Dolphin Whistle Signal Types

The study of dolphin species classification and recognition algorithms typically begins with analyzing communication signal characteristics. Matthews et al. surveyed tonal calls across cetacean species, reporting summary statistics for acoustic parameters including start, end, minimum, maximum, and center frequencies, as well as call duration and the number of inflection points [9]. They also included information about the recordings, including location, encounter/group counts, and recording length. McCowan proposed a quantitative contour similarity based method to discriminate the whistle types [10]. Janick and Slater investigated the hypothesis that signature whistles serve to maintain group cohesion rather than being solely stress-induced [11]. The authors compared the whistle types of four captive bottlenose dolphins when they were together versus alone, observing whistle copying behavior in different experimental pools. They also studied individual differences among the dolphins based on this phenomenon. Beeman used a signal analysis system with fast Fourier transforms to analyze audio spectrograms and categorized communication signals into six types, which are ascending, fluctuating, sinusoidal, U-shaped, descending, and residual [12]. However, the production context of these signals and dolphin types remained unclear. Hawkins and Gartside conducted another study on classifying Indo-Pacific bottlenose dolphin whistles into five categories, including sinusoidal, ascending, descending, level, and concave [13]. They demonstrated that whistle types convey specific information to companions related to the caller's behavior or situation (behavioral context), establishing a connection between whistle types and behaviors. Azevedo et al. categorized the communication signals of Atlantic spotted dolphins into six types, namely, ascending, descending, ascending-descending, descending-ascending, smooth, and multiple [14]. They measured

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

nine acoustic parameters of each whistle's fundamental components (e.g., starting frequency, ending frequency, and frequency at 1/4 duration) and concluded that dolphins alter whistle structures based on behavioral states, confirming the communication function of whistles. Rui-chao et al. also defined dolphin whistles into six categories, which are constant, upsweep, downsweep, concave, convex, and sine [15]. Dolphin whistles vary with frequency over time, and individuals can use "signature whistles" to communicate their identity. Harley demonstrated that dolphins can generalize from trained signature whistles to previously unheard whistle instances produced by the same individuals, indicating that whistle classification is driven by contour-based representations rather than by specific acoustic parameters or voice cues [16].

2.2. Acoustic Signal Classification

Early signal classification methods relied heavily on traditional statistical models to extract whistle segments, but a key limitation of these statistical models was the incomplete detection of whistle signals, resulting in compromised classification results. Douglas and Gillespie addressed this by aggregating statistics over many fragmented whistle signals, which yielded a classification accuracy of over 94% [17]. Researchers combined cepstral coefficients and the Gaussian mixture model to detect and recognize different bioacoustic signals, namely, background noise, communication signals, burst pulse signals, and combined signals, and identify species [18,19]. Yang Wuyi et al. proposed a method for classifying broad-snouted dolphin communication signals using syntactic patterns [20]. This method involves extracting the trajectory curves of the fundamental frequency of dolphin communication signals over time, followed by identifying the primitive sequences of fundamental frequency changes. Based on the categorization criteria for these signals, the grammar that generates the primitive sequences for each category is then summarized. With these established syntactic patterns, they can extract the primitive sequence features and classify them to achieve automatic classification of dolphin communication signals, advancing marine mammal acoustic research.

In recent years, Convolutional Neural Networks (CNNs), renowned for their success in computer vision, have also proven highly effective for audio processing tasks like audio tagging and sound event detection, demonstrating versatile applicability [21]. Gao et al. utilized deep neural networks to classify and identify echolocation signals and pulse noise from three typical marine mammal species [22]. Their results showed that the fully connected network surpassed the spectral energy algorithm, achieving a 30% higher accuracy. Zhang Yu et al. applied CNN for the classification of multi-species echolocation click signals [23]. To predict species labels, they applied majority voting and Maximum A Posteriori (MAP) methods across m consecutive clicks, finding that classification accuracy improved as m increased. While deep learning methods excel at human speech recognition when trained on large datasets, their application to dolphin vocalizations remains challenging due to the limited size of available samples.

2.3. Underwater Acoustic Propagation Modeling

To evaluate and enhance model robustness in complex marine environments, it is essential to understand underwater signal propagation and simulate the acoustic communication channel. Research on underwater acoustic propagation modeling theory began in the 1960s [24], initially focusing on ray theory and normal mode theory for horizontally invariant environments. To handle complex, horizontally varying acoustic propagation, parabolic equation (PE) theory and coupled normal mode theory emerged in the 1970s. Propagation models include ray models, normal mode models, parabolic equation models, multipath expansion models, fast field models, and hybrid algorithms [25]. Notably, BELLHOP [26], a widely-used ray model particularly suitable for shallow water and high-frequency scenarios, employs Gaussian beam tracing for computing ray paths and acoustic fields in horizontally non-uniform environments. RAY [27], another prominent ray model, evaluates the impact of seabed parameters such as compressional and shear wave speeds, attenuation, and density on broadband signal propagation, making it highly valuable in detailed acoustic environmental studies. KRAKEN [28], a frequently used normal mode model, efficiently computes underwater acoustic fields

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

by solving eigenvalue problems under horizontally invariant or mildly varying conditions, ideal for low-frequency propagation modeling. PE methods effectively address horizontally varying scenarios, but their computational load increases significantly at higher frequencies or with complex seabed interactions, limiting their application. Consequently, underwater acoustic propagation models must be carefully selected based on specific scenarios, such as signal frequency, environmental complexity, and computational efficiency, to ensure accurate representation of underwater acoustic channels.

3. Methodology

In this section, we describe the data collection and annotation procedures, the deep learning classification methods, and the data simulation process.

3.1. Data Collection and Annotation

The raw data were collected at the dolphin aquarium of Chimelong Ocean Kingdom in Zhuhai, China. Two self-contained hydrophones (SoundTrap 300 HF) were used, each enclosed in a custom acrylic housing to prevent dolphins from accessing the devices. As shown in Figure 1, the cylindrical housing has two sections: the upper section holds adjustable suction-cup mounts for the hydrophone and an underwater camera, and the lower section contains lead weights to ensure stability underwater. The hydrophones were deployed at two locations in the pool, as shown in Figure 2: one at the center to capture signals from all directions, and another in a corner, positioned safely out of the dolphins' reach. The hydrophones record signals at a sampling frequency of 144kHz from five *Tursiops aduncus*—two adults and three infants—in a single large pool.

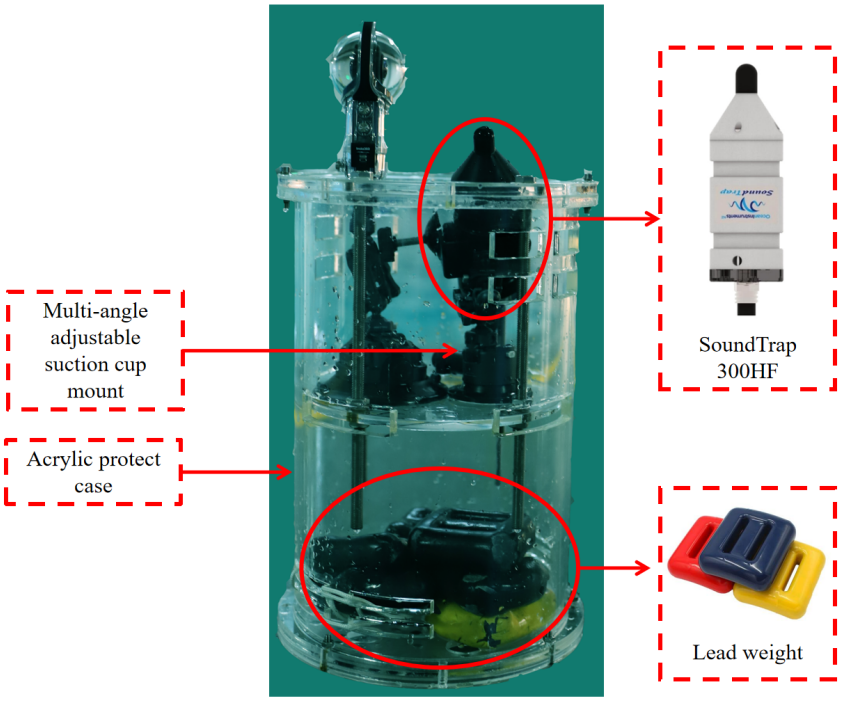


Figure 1. Hydrophone setup.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

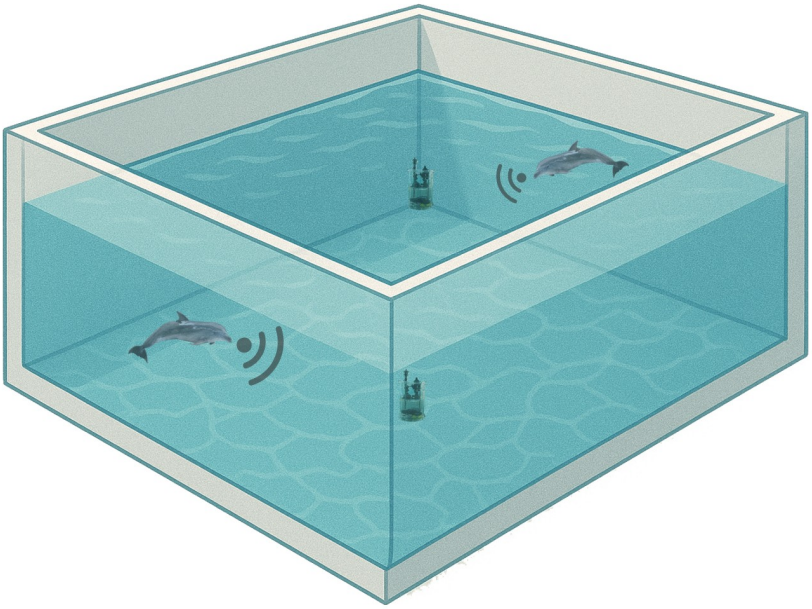


Figure 2. Underwater hydrophone deployment.

The raw dataset contains approximately 72 hours of recordings. Since the signals were recorded in a large pool with low ambient noise, the whistle signals have relatively high signal-to-noise ratios. In this study, we manually annotated 19.5 hours of recordings following the category definitions of Xue Rui-chao et al. [15], with the addition of a new class termed double concave. The seven whistle types are defined as:

- Constant: nearly flat contour with frequency variation under 1 kHz across the time span of the signal.
- Upsweep: the fundamental frequency increases over time.
- Downsweep: the fundamental frequency decreases over time.
- Concave: the fundamental frequency first decreases, then increases over time.
- Convex: the fundamental frequency first increases, then decreases over time.
- Sine: sinusoidal-like contour.
- Double concave: two consecutive concave contours concatenated together.

All these seven whistle samples are visualized in Figure 3.

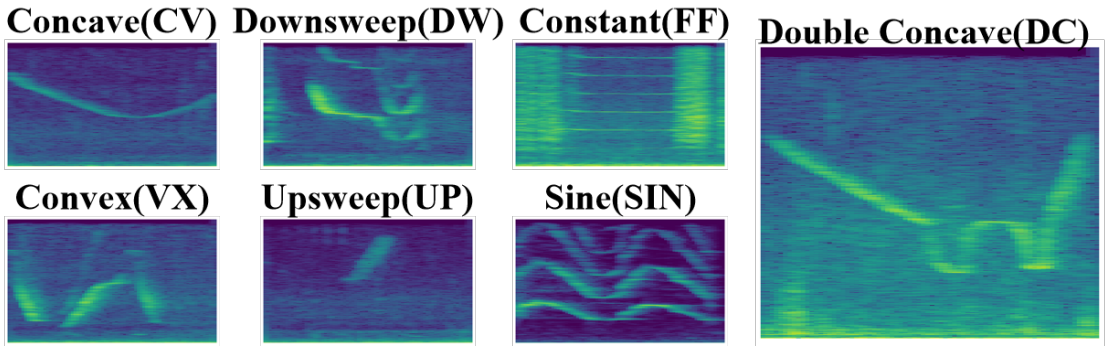


Figure 3. Seven typical *Tursiops aduncus* whistle signals.

The annotated dataset includes the start and end times of each whistle in the raw recordings, along with its call type. In total, 3913 samples were labeled, and the dataset statistics are summarized in Table 1. The classes are imbalanced—for example, the convex type has 912 samples, whereas the constant type has only 256. Overall, the whistles span a frequency range of approximately 3–40 kHz.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Table 1. Annotated Dataset Details.

ID	Label	Mean Duration	Variance	Count
CV	Concave	0.5103	0.0710	586
DC	Double Concave	0.7600	0.0069	462
DW	Downsweep	0.3430	0.0078	549
FF	Constant	0.3690	0.0291	256
SIN	Sine	0.8787	0.1064	278
UP	Upsweep	0.1909	0.0055	870
VX	Convex	0.3603	0.0205	912
Total Whistle Count				3913

3.2. Deep Learning Classification Methods

For whistle classification, we employed five widely used CNN network architectures [21], including MobileNet, Xception, ResNet, ResNeXt and SE-ResNeXt. The descriptions of the implemented models are detailed as follows.

- MobileNet [29]: a lightweight CNN that uses depthwise separable convolutions to reduce computation complexity while maintaining strong performance, which is suitable for mobile applications.
- Xception [30]: the "Extreme Inception" model that decouples spatial and channel-wise convolutions for more efficient feature extraction.
- ResNet (Residual Network) [31]: introduces residual connections (shortcuts) to enable effective training of very deep networks.
- ResNeXt [32]: extends ResNet with grouped convolutions and a cardinality parameter, capturing a broader range of feature interactions.
- SE-ResNeXt [33]: combines ResNeXt with Squeeze-and-Excitation blocks to model channel dependencies and enhance feature recalibration.

In the computer vision field, input data typically consists of three 2-dimensional channels (red, green, and blue). In contrast, our mono-channel acoustic recordings are 1-dimensional time-series signals, making it difficult to apply standard vision-based CNN architectures directly. To ensure compatibility with these CNN models, we explore two types of feature representations for the whistle signals, namely, the spectrogram-based representations and the waveform-based representations.

For the spectrogram-based representations, the audio signal is first transformed into either a 2D log-scaled Mel spectrogram (Log-Mel) or a 2D Mel-frequency cepstral coefficient (MFCC) sequence. Both features are widely used in speech and sound recognition because they reflect human auditory perception. MFCCs characterize the short-term spectral envelope of the signal, while Log-Mel spectrograms provide a detailed time–frequency representation. To match the 3-channel input format of common CNNs, we then compute the first- and second-order delta coefficients for each feature map. These deltas capture temporal changes in the features, allowing the 3-channel representation to encode both spectral structure and its evolution over time.

For the 1D waveform-based input, the CNN architectures are adapted to accommodate raw audio directly. Specifically, a multi-scale stacking module is introduced to learn a 2D representation for the downstream 2D-convolutional layers, as shown in Table 2. Unlike the Log-Mel or MFCC-based method, this module requires no explicit frequency-domain transformations. Instead, the convolutional blocks implicitly learn the spectral features, providing an end-to-end solution for the classification task.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Table 2. Network Architecture for Waveform Processing.

waveform (bs, 1, $sr \times duration$)		
Conv1d out channels=32, kernel size=11, stride=1, padding=5 (bs, 32, $sr \times duration$)	Conv1d out channels=32, kernel size=51, stride=5, padding=25 (bs, 32, $sr \times duration/5$)	Conv1d out channels=32, kernel size=101, stride=10, padding=50 (bs, 32, $sr \times duration/15$)
BatchNorm1d	BatchNorm1d	BatchNorm1d
ReLU		
Conv1d out channels=32, kernel size=3, stride=1, padding=1	Conv1d out channels=32, kernel size=3, stride=1, padding=1	Conv1d out channels=32, kernel size=3, stride=1, padding=1
BatchNorm1d	BatchNorm1d	BatchNorm1d
ReLU		
MaxPool1d kernel size=150, stride=150	MaxPool1d kernel size=30, stride=30	MaxPool1d kernel size=10, stride=10
unsqueeze (bs,1, 32, $sr \times duration/150$) cat (bs,1, 96, $sr \times duration/150$)		
Conv2d kernel size=(7,7), stride=(2, 2), padding=(3, 3), bias=False (bs, 64, 48, $sr \times duration/150$)		

We evaluate the models using mean Average Precision (mAP), a standard metric in information retrieval and object detection. It measures the average precision of a model across multiple classes. For each class, a precision–recall curve is computed, where precision (P) is defined as true positive (TP) divided by the sum of TP and false positive (FP) and recall (R) is the ratio of true positive detections to the total number of ground truth positives. The average precision (AP) is calculated as the area under the interpolated precision–recall curve, and mAP is obtained by averaging the AP values across all classes or queries.

3.3. Simulated Marine Whistle Signals

To further evaluate and enhance the robustness of the whistle classification models, we simulate whistle propagation through a marine acoustic channel and add *in-situ* ambient noise at controlled SNR levels for both training and testing.

Specifically, we use BELLHOP to compute the impulse response of the marine acoustic channel between the source and receiver. The simulated signals are then generated by convolving the whistle signal with this channel response:

$$r(t) = s(t) * h(t) \tag{1}$$

where *r(t)* is the received signal from the receiver in simulation, *s(t)* is the source signal, and *h(t)* is channel impulse response. To simplify the simulation model, it is assumed that the channel is linear and time invariant during the vocalization. It is worth noting that the underwater acoustic channel is complex, with multiple propagation paths generating many arrival pulses, most of which have very small amplitudes and minimal impact. For simplicity, pulses below 1% of the maximum amplitude are discarded. Additionally, because the impulse response is very weak, it is scaled by 100 to mimic hydrophone gain and ensure a signal intensity comparable to the original. Cross-correlation analysis is then used to determine the offset introduced by convolution, ensuring that the simulated whistle signal is centered within the segment.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Ambient noise is assumed to be additive and is directly added to the signal:

$$y(t) = x(t) + n(t) \tag{2}$$

where $y(t)$ denotes the resultant noisy signal, $x(t)$ denotes the clean whistle signal, and $n(t)$ is a noise signal randomly sampled from the ocean noise recordings in the SHIPSEAR dataset [34].

4. Experiment and Results

In this section, we will explain the experiment setup and then present the results.

4.1. Input Data Pre-Processing and Experiment Set Up

The whistle signals are cropped into 0.75-second segments. For signals longer than 0.75 seconds, the central 0.75-second segment of the original recording is used. For shorter signals, the segment is extended by including additional portions from the original recording at both ends to reach the target length. As shown in Figure 4, this corresponds to roughly the 85th percentile, ensuring that most whistle contours are complete.

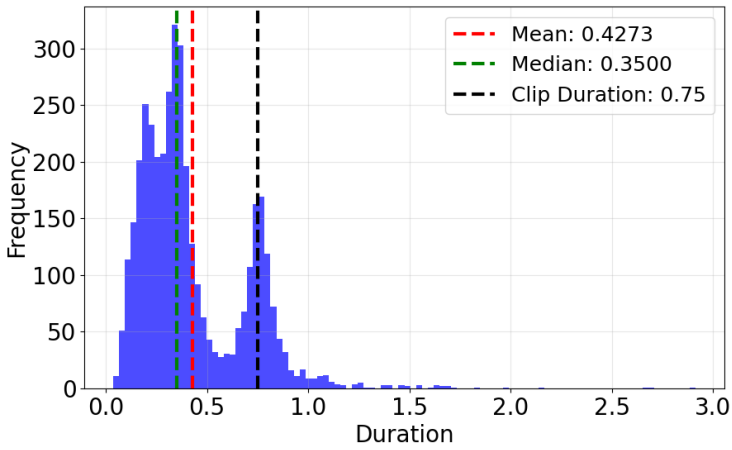
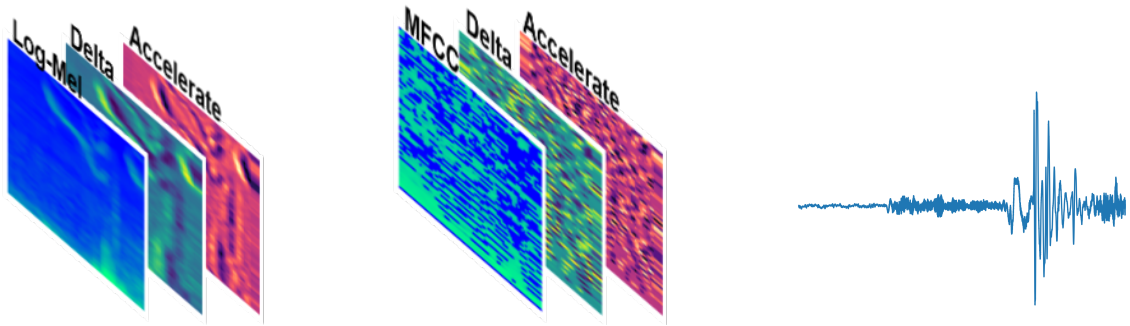


Figure 4. The distribution of whistle signal lengths.

While the raw audio is recorded at 144 kHz, the fundamental frequencies of dolphin whistles lie primarily below 20 kHz. We therefore downsample the data to 44.1 kHz using the Kaiser algorithm. The downsampled signals are then transformed into Log-Mel and MFCC spectrograms, and their first-order (delta) and second-order (accelerate) coefficients are added as additional channels. This results in three input representations: Log-Mel + delta + accelerate, MFCC + delta + accelerate, and raw waveform data, as illustrated in Figure 5.



(a) Log-Mel (b) MFCC (c) Wave

Figure 5. Input data formats.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

We use 64 Mel filter banks for the Log-Mel and MFCC transformation. The frame length is 80 ms, and the frameshift is 10 ms. In such a case, the shape of the MFCC/Log-mel would be 64×76. Additionally, we use a window size of 9 to calculate the delta and acceleration coefficients. This value is often used as a reasonable compromise between capturing sufficient temporal context and avoiding over-smoothing. It has been found to work well empirically in many audio processing tasks.

To evaluate model performance, we perform 5-fold cross-validation. The dataset is divided into five stratified folds. In each of five iterations, four folds are used for training, while the remaining fold is split in half for validation and testing, resulting in an 8:1:1 ratio for training, validation, and test sets. The CNN models described in the Methodology section are trained from scratch for 50 epochs using cross-entropy loss and the stochastic gradient descent (SGD) optimizer. The model achieving the highest validation accuracy is retained for subsequent testing.

In the BELLHOP simulation, standard environmental settings are used, including seabed topography, acoustic parameters, and the sound velocity profile. A source and receiver location are selected to compute the acoustic channel impulse response. Both the source and receiver are positioned 100 meters underwater, consistent with typical dolphin activity and hydrophone deployment. Figure 6 illustrates the locations of the source and receiver, along with the surrounding underwater terrain.

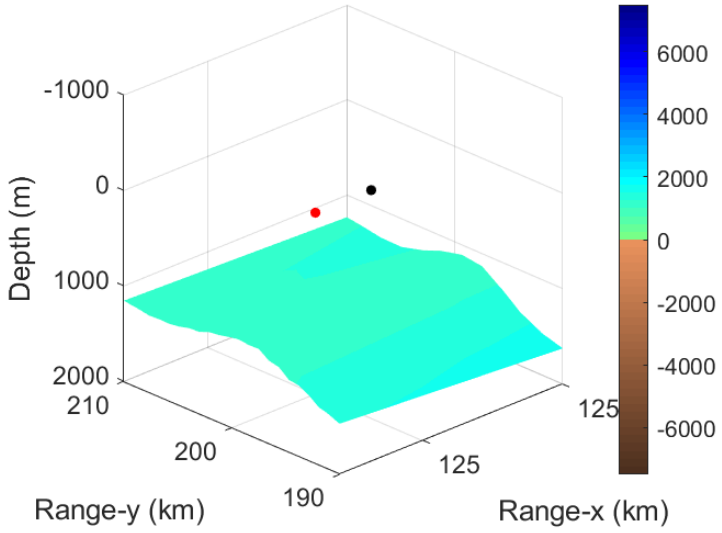


Figure 6. Positions of the source (red) at (125 km, 200 km, 100m) and receiver (black) at (125.2 km, 200 km, 100 m).

Figure 7 shows spectrograms of an example original signal (left), its corresponding simulated signal (middle), and the difference between them (right).

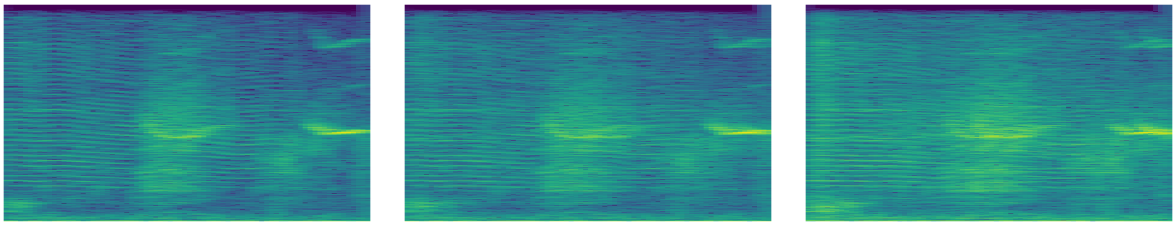


Figure 7. Spectrogram of the original slice (left), simulated slice (middle) and the difference value between them (right).

The models are trained on three types of datasets: (1) original training data (org), (2) simulated training data (sim), and (3) a combination of original and simulated data (all). They are evaluated on corresponding test sets: (1) original (org), (2) simulated (sim), and (3) combined (all), with varying SNR levels to assess model generalizability.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

4.2. Results

Our analysis begins with an evaluation of all input feature representations across all models using the clean datasets. The results are summarized in Figure **??** and Table 3. All three input representations, namely, Log-Mel (logmel), MFCC (mfcc), and the raw waveforms (wave), yield strong performance, with every model achieving an mAP above 0.80. Although the waveform representation performs slightly below the MFCC and Log-Mel inputs across most architectures, the gap is modest. For example, the largest mAP difference between waveform and Log-Mel inputs is approximately 0.04 on the Xception model. Several factors may account for the performance advantage of spectrogram-based features. First, converting the waveform to a time-frequency representation can reduce the burden on the network to learn low-level spectral decomposition. Second, MFCC and Log-Mel features emphasize perceptually salient components, which may improve the training efficiency under limited data. In contrast, waveform-based models must learn these transformations implicitly, which often requires larger datasets or deeper architectures. In general, MFCC and Log-Mel inputs perform similarly. The highest mAP is achieved by the Xception architecture with the Log-Mel input, while the MFCC input lags behind by only 0.01, indicating that both representations are highly effective for whistle classification.

In Table 3, we examine the class level AP values. For each model, Log-Mel and MFCC demonstrate generally comparable performance across all categories. Interestingly, within the ResNet architecture, MFCC lags behind Log-Mel by roughly 0.05 across the CV, SIN, and UP categories. And the waveform input in general doesn't perform well on CV and SIN classes. This degradation is likely influenced in part by the imbalanced class distribution.

Table 3. Class-level model performance on various input features.

Model	Input	mAP	CV	DC	DW	FF	SIN	UP	VX
MobileNet	logmel	0.8584	0.6348	0.9651	0.9531	0.8483	0.8059	0.8345	0.9672
	mfcc	0.8572	0.6529	0.9809	0.9600	0.8386	0.7816	0.8234	0.9626
	wave	0.8319	0.5290	0.9491	0.9510	0.8785	0.7554	0.8066	0.9539
Xception	logmel	**0.9231**	**0.7643**	**0.9896**	**0.9866**	**0.9373**	**0.9105**	**0.8924**	**0.9812**
	mfcc	**0.9130**	**0.7345**	**0.9901**	**0.9901**	**0.9382**	0.8751	**0.8834**	**0.9797**
	wave	0.8842	0.6709	0.9662	0.9598	0.9171	0.8322	0.8748	0.9684
ResNet	logmel	0.8341	0.5819	0.9237	0.9470	0.7960	0.7998	0.8303	0.9598
	mfcc	0.8086	0.5368	0.9359	0.9235	0.7902	0.7474	0.7696	0.9568
	wave	0.8073	0.5071	0.9246	0.9364	0.8233	0.7066	0.8136	0.9398
ResNeXt	logmel	0.8468	0.5994	0.9367	0.9616	0.8249	0.8010	0.8374	0.9663
	mfcc	0.8282	0.5608	0.9365	0.9576	0.7520	0.8287	0.8051	0.9570
	wave	0.8412	0.5545	0.9719	0.9595	0.8570	0.7561	0.8330	0.9563
SE-ResNeXt	logmel	0.8790	0.6605	0.9752	0.9711	0.8689	0.8469	0.8600	0.9705
	mfcc	0.8925	0.6954	0.9834	0.9833	0.8828	**0.8950**	0.8385	0.9687
	wave	0.8445	0.5758	0.9591	0.9410	0.8893	0.7762	0.8249	0.9453

We further examine the confusion matrices of the three input feature representations on the first fold of the best-performing Xception model. A large number of CV samples are misclassified as UP, leading to a reduced average precision for the CV class. This confusion is likely driven by the short duration of certain CV signals, causing models to interpret their rapid, transient frequency changes as resembling UP patterns. In contrast, the DW class—despite its limited sample size—shows more distinctive characteristics, as reflected in its consistently high AP scores. This observation indirectly supports the validity of the newly defined DW category, which is clearly separable from the six previously established classes.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

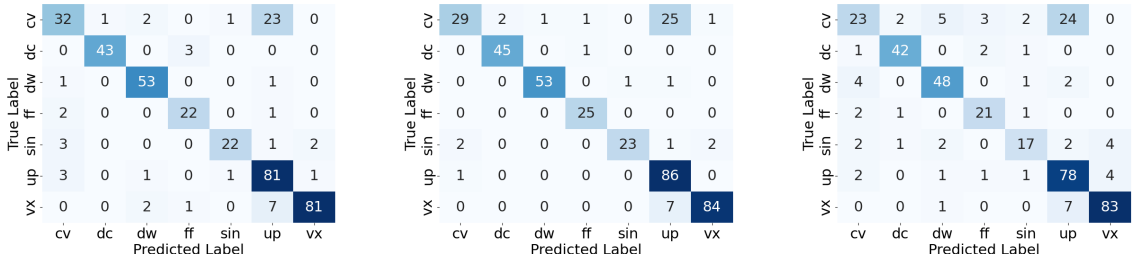


Figure 8. Confusion matrices of the Xception model on first-fold data using three input representations: Log-Mel (left), MFCC (middle), and Wave (right).

We further investigate fine-tuning ImageNet-pretrained CNN models [21] for the dolphin whistle classification task. The training dataset, methodology, model architectures, and hyperparameters are kept identical to those used for models trained from scratch. The resulting performance is summarized in Table 4 and illustrated in Figure 9.

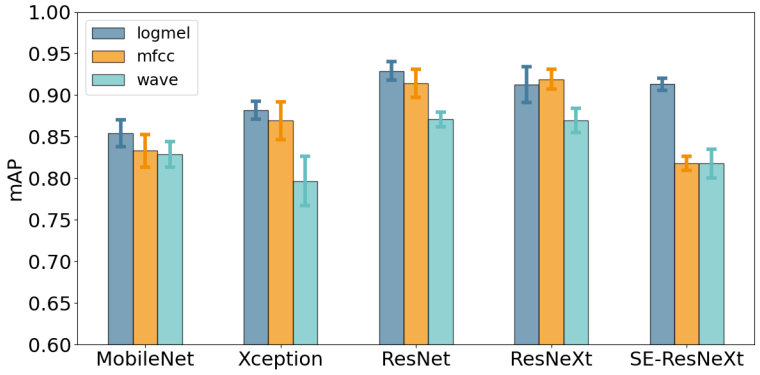


Figure 9. Comparison of fine-tuned pretrained model performance on various input features.

Table 4. Class level performance of fine-tuned pretrained models across input features.

Model	Input	mAP	CV	DC	DW	FF	SIN	UP	VX
MobileNet	logmel	0.8541	0.5812	0.9545	0.9563	0.8793	0.8351	0.8171	0.9555
	mfcc	0.8331	0.5371	0.9571	0.9588	0.8185	0.8217	0.7904	0.9481
	wave	0.8290	0.5040	0.9234	0.9256	0.8824	0.7948	0.8257	0.9471
Xception	logmel	0.8820	0.6667	0.9738	0.9659	0.8853	0.8769	0.8528	0.9530
	mfcc	0.8692	0.6218	0.9769	0.9664	0.9024	0.8569	0.7992	0.9608
	wave	0.7968	0.4716	0.8944	0.9329	0.8016	0.7563	0.7791	0.9417
ResNet	logmel	**0.9291**	**0.7582**	0.9855	0.9888	0.9398	**0.9436**	**0.9058**	**0.9822**
	mfcc	0.9142	0.7365	0.9829	**0.9907**	0.9392	0.8840	**0.8883**	0.9779
	wave	0.8709	0.6389	0.9558	0.9504	0.8616	0.8500	0.8658	0.9739
ResNeXt	logmel	0.9123	0.7342	0.9821	**0.9894**	0.9291	0.9099	0.8617	**0.9800**
	mfcc	**0.9191**	**0.7614**	**0.9870**	0.9879	**0.9477**	0.8962	0.8789	0.9748
	wave	0.8697	0.6102	0.9580	0.9511	0.8552	0.8658	0.8748	0.9729
SE-ResNeXt	logmel	0.9131	0.7166	**0.9860**	0.9734	**0.9431**	**0.9223**	0.8774	0.9731
	mfcc	0.8180	0.4824	0.9511	0.9329	0.8096	0.7952	0.7990	0.9558
	wave	0.8177	0.5190	0.9149	0.9096	0.8337	0.7780	0.8166	0.9524

As shown in Figure 9, the ResNet-family models benefit from pretraining and fine-tuning, whereas models with fewer trainable parameters, such as MobileNet and Xception, show no improvement. This advantage is likely due to the residual connections and larger model capacity in ResNet architectures, which facilitate more effective parameter updates and adaptation to new domains. It is also worth noting that the fine-tuned SE-ResNeXt performs poorly with MFCC inputs. This may be because

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

MFCC processing generates a compact set of largely uncorrelated coefficients, whereas the pre-trained SE modules are designed to capture channel correlations important for visual recognition. Since these learned channel relationships are not present in MFCC features, fine-tuning on them results in degraded performance. Conversely, starting from scratch to capture these MFCC features might yield better results than the original spectral features. Log-Mel features more closely resemble image-like representations, allowing the fine-tuned model to better leverage prior knowledge and achieve superior results. In contrast, models fine-tuned on waveform inputs show smaller performance gains, as waveform data fundamentally differ from the two spectral representations.

We then assess model robustness on simulated datasets. The mAP results for Log-Mel, MFCC, and waveform inputs are presented in Tables 5, 6, and 7, respectively. Models trained on the original data with all three input types show only a slight decrease in mAP (0.01–0.02) when evaluated on simulated data, indicating that they can still accurately classify the simulated signals. These results also suggest that, despite visual similarity on spectrograms, the simulated data introduce slightly more confusion in classification than the original recordings. When tested on signals with different SNR levels, models perform comparably to the clean dataset at high (40 dB, 30 dB) and moderate (20 dB) signal-to-noise ratio (SNR) levels. At lower SNR (10 dB), performance declines moderately but remains acceptable. When the SNR drops to 0 dB, i.e., noise intensity matches the signal, model performance decreases sharply, as the noise overwhelms the signal. Overall, all five models maintain adequate classification accuracy at SNRs of 10 dB or higher, suggesting potential applicability to real-world data. For lower-quality signals, further processing, such as signal enhancement, would be needed, though this is beyond the scope of the current study.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Table 5. Model robustness on Log-Mel input.

Model	Data		SNR					
	Test on	Train on	Pure	40	30	20	10	0
MobileNet	org	org	**0.8584**	**0.8567**	**0.8571**	0.8341	0.7916	**0.6091**
		sim	0.8351	0.8341	0.8297	0.8164	0.7497	0.5519
		all	0.8527	0.8527	0.8506	**0.8397**	**0.7962**	0.6052
	sim	org	0.8450	0.8450	0.8440	0.8262	0.7767	0.5615
		sim	0.8446	0.8442	0.8411	0.8279	0.7494	0.5377
		all	**0.8532**	**0.8538**	**0.8519**	**0.8366**	**0.7887**	**0.5798**
Xception	org	org	**0.9231**	**0.9226**	**0.9202**	**0.9109**	**0.8649**	**0.6987**
		sim	0.9055	0.9059	0.9048	0.8921	0.8368	0.6727
		all	0.9124	0.9127	0.9103	0.9010	0.8450	0.6763
	sim	org	**0.9148**	**0.9152**	**0.9139**	**0.8989**	**0.8394**	**0.6579**
		sim	0.9091	0.9090	0.9074	0.8936	0.8272	0.6530
		all	0.9063	0.9063	0.9042	0.8965	0.8296	0.6448
ResNet	org	org	0.8341	0.8341	0.8287	0.8148	0.7263	0.5082
		sim	0.8075	0.8078	0.8031	0.7841	0.6987	0.4925
		all	**0.8416**	**0.8407**	**0.8404**	**0.8255**	**0.7627**	**0.5427**
	sim	org	0.8217	0.8214	0.8169	0.7944	0.6996	0.4949
		sim	0.8158	0.8156	0.8112	0.7936	0.7057	0.4886
		all	**0.8444**	**0.8441**	**0.8425**	**0.8249**	**0.7514**	**0.5322**
ResNeXt	org	org	0.8468	0.8463	0.8463	0.8369	0.7810	**0.5579**
		sim	0.8378	0.8374	0.8333	0.8272	0.7549	0.5186
		all	**0.8554**	**0.8538**	**0.8521**	**0.8429**	**0.7812**	0.5364
	sim	org	0.8292	0.8301	0.8254	0.8213	0.7495	**0.5389**
		sim	0.8473	0.8480	0.8467	0.8286	**0.7652**	0.5155
		all	**0.8532**	**0.8539**	**0.8521**	**0.8395**	0.7649	0.5304
SE-ResNeXt	org	org	0.8790	0.8785	0.8777	0.8628	0.7856	0.5532
		sim	0.8596	0.8592	0.8605	0.8434	0.7498	0.5930
		all	**0.8862**	**0.8864**	**0.8863**	**0.8696**	**0.7967**	**0.5955**
	sim	org	0.8693	0.8693	0.8637	0.8474	0.7687	0.5271
		sim	0.8721	0.8715	0.8717	0.8500	0.7513	**0.5880**
		all	**0.8865**	**0.8873**	**0.8877**	**0.8676**	**0.7841**	0.5745

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Table 6. Model robustness on MFCC input.

Model	Data		SNR					
	Test on	Train on	Pure	40	30	20	10	0
MobileNet	org	org	0.8572	0.8560	0.8564	0.8312	0.7571	**0.5550**
		sim	0.6775	0.6762	0.6675	0.6266	0.5084	0.3511
		all	**0.8685**	**0.8688**	**0.8691**	**0.8485**	**0.7853**	0.5469
	sim	org	0.8489	0.8484	0.8496	0.8187	0.7179	0.4989
		sim	0.6852	0.6836	0.6748	0.6287	0.5168	0.3441
		all	**0.8629**	**0.8625**	**0.8590**	**0.8366**	**0.7656**	**0.5142**
Xception	org	org	0.9130	0.9136	0.9135	0.9003	0.8463	0.6774
		sim	0.8997	0.8992	0.8971	0.8799	0.8121	0.6520
		all	**0.9290**	**0.9292**	**0.9273**	**0.9073**	**0.8590**	**0.7059**
	sim	org	0.9043	0.9040	0.9013	0.8825	0.8188	0.6262
		sim	0.9106	0.9101	0.9049	0.8859	0.8107	0.6256
		all	**0.9261**	**0.9265**	**0.9232**	**0.9005**	**0.8376**	**0.6702**
ResNet	org	org	0.8086	0.8083	0.8022	0.7709	0.6603	0.4151
		sim	0.7704	0.7699	0.7700	0.7533	0.6511	0.4011
		all	**0.8415**	**0.8406**	**0.8357**	**0.8294**	**0.7266**	**0.4885**
	sim	org	0.7835	0.7846	0.7803	0.7476	0.6338	0.4037
		sim	0.7842	0.7832	0.7825	0.7651	0.6537	0.3975
		all	**0.8418**	**0.8422**	**0.8379**	**0.8290**	**0.7201**	**0.4843**
ResNeXt	org	org	0.8282	0.8277	0.8284	0.8105	0.7276	0.5182
		sim	0.7842	0.7829	0.7801	0.7585	0.6634	0.4478
		all	**0.8470**	**0.8470**	**0.8472**	**0.8299**	**0.7419**	**0.5198**
	sim	org	0.8155	0.8151	0.8160	0.7930	0.7025	0.4737
		sim	0.8046	0.8040	0.8022	0.7764	0.6756	0.4501
		all	**0.8328**	**0.8333**	**0.8349**	**0.8177**	**0.7247**	**0.5015**
SE-ResNeXt	org	org	0.8925	0.8926	0.8919	0.8727	0.8013	0.5903
		sim	0.8694	0.8707	0.8680	0.8584	0.7818	0.5858
		all	**0.9031**	**0.9030**	**0.9021**	**0.8908**	**0.8180**	**0.6100**
	sim	org	0.8822	0.8823	0.8768	0.8530	0.7670	0.5355
		sim	0.8829	0.8827	0.8824	0.8694	0.7873	0.5749
		all	**0.9006**	**0.9003**	**0.8958**	**0.8823**	**0.8009**	**0.5813**

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

Table 7. Model robustness on waveform input.

Model	Data		SNR					
	Test on	Train on	Pure	40	30	20	10	0
MobileNet	org	org	0.8319	0.8325	0.8312	0.8236	0.7878	0.5671
		sim	0.7752	0.7751	0.7737	0.7688	0.7293	0.4885
		all	**0.8531**	**0.8530**	**0.8530**	**0.8497**	**0.8072**	**0.5968**
	sim	org	0.7927	0.7928	0.7941	0.7896	0.7435	0.5272
		sim	0.7804	0.7799	0.7790	0.7719	0.7196	0.4777
		all	**0.8473**	**0.8475**	**0.8484**	**0.8407**	**0.7936**	**0.5689**
Xception	org	org	0.8842	0.8837	0.8855	0.8794	0.8428	0.6524
		sim	0.8514	0.8518	0.8512	0.8503	0.8152	0.6325
		all	**0.8949**	**0.8945**	**0.8946**	**0.8897**	**0.8532**	**0.6877**
	sim	org	0.8639	0.8638	0.8643	0.8553	0.8118	0.6085
		sim	0.8620	0.8627	0.8616	0.8591	0.8102	0.6156
		all	**0.8898**	**0.8895**	**0.8901**	**0.8856**	**0.8377**	**0.6519**
ResNet	org	org	0.8073	0.8074	0.8078	0.7992	0.7555	0.5308
		sim	0.7479	0.7485	0.7490	0.7537	0.7096	0.4904
		all	**0.8610**	**0.8611**	**0.8605**	**0.8593**	**0.8198**	**0.6349**
	sim	org	0.7834	0.7835	0.7838	0.7762	0.7171	0.4918
		sim	0.7539	0.7540	0.7572	0.7573	0.6999	0.4799
		all	**0.8598**	**0.8600**	**0.8607**	**0.8590**	**0.8111**	**0.6193**
ResNeXt	org	org	0.8412	0.8407	0.8426	0.8387	0.8083	0.6170
		sim	0.8189	0.8190	0.8188	0.8176	0.7810	0.5695
		all	**0.8678**	**0.8679**	**0.8683**	**0.8647**	**0.8354**	**0.6616**
	sim	org	0.8232	0.8235	0.8226	0.8172	0.7748	0.5544
		sim	0.8276	0.8275	0.8260	0.8246	0.7794	0.5563
		all	**0.8619**	**0.8621**	**0.8611**	**0.8594**	**0.8266**	**0.6428**
SE-ResNeXt	org	org	0.8445	0.8446	0.8443	0.8361	0.7856	0.5685
		sim	0.7909	0.7910	0.7896	0.7839	0.7413	0.5734
		all	**0.8702**	**0.8704**	**0.8693**	**0.8621**	**0.8272**	**0.6070**
	sim	org	0.8183	0.8177	0.8177	0.8063	0.7457	0.5179
		sim	0.8076	0.8075	0.8050	0.7998	0.7474	0.5537
		all	**0.8688**	**0.8688**	**0.8684**	**0.8589**	**0.8069**	**0.5787**

Finally, we explore the potential of using simulated data for data augmentation. Models are trained using both the original training data and the corresponding simulated signals, and tested under the same conditions as described above. The results, shown in Tables 5–7, indicate that incorporating simulated data generally improves mAP scores. The augmented training also enhances model robustness to noise, particularly at 0 dB SNR. Overall, these experiments demonstrate that the proposed data simulation process effectively improves model generalizability.

5. Conclusions and Future Work

This work presents a novel *Tursiops aduncus* whistles dataset, consisting of 3913 manually annotated signals across seven whistle types. Using this dataset, we develop a comprehensive framework to train and evaluate various CNN architectures and input feature representations. The experiments demonstrate that both the training-from-scratch model and the fine-tuned ImageNet-pretrained models achieve strong performance, with mAP consistently above 0.8. Xception performs best from scratch using Log-Mel and MFCC features, while pretrained ResNet-family models deliver comparable or slightly better accuracy. To assess the robustness of the models in a complex environment, we introduce simulated data generated with the Bellhop acoustic channel and added real marine noise from the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

SHIPSEAR dataset. The models maintain reliable performance on simulated data and remain effective under moderate noise. Training with simulated data further improves accuracy and noise robustness, demonstrating its value as a practical augmentation strategy for whistle signal classification. Future work will focus on efficient dolphin signal detection methods and lightweight classification models suitable for deployment on resource-limited devices.

Acknowledgments: We would like to express our sincere gratitude to Chimelong Ocean Kingdom for their generous support in collecting dolphin acoustic signals. Their assistance has been invaluable to this study. We also thank Jingwen Pang and Kefei Zhu for their efforts in the data annotation process.

References

1. Au, W. Echolocation signals of wild dolphins. *Acoustical Physics* **2004**, *50*, 454–462.
2. Au, W.W. *The sonar of dolphins*; Springer Science & Business Media, 1993.
3. Janik, V.M.; Sayigh, L.S. Communication in bottlenose dolphins: 50 years of signature whistle research. *Journal of Comparative Physiology A* **2013**, *199*, 479–489.
4. Luís, A.R.; Couchinho, M.N.; Dos Santos, M.E. A quantitative analysis of pulsed signals emitted by wild bottlenose dolphins. *PLOS one* **2016**, *11*, e0157781.
5. Haughey, R.; Hunt, T.N.; Hanf, D.; Passadore, C.; Baring, R.; Parra, G.J. Distribution and habitat preferences of Indo-Pacific bottlenose dolphins (Tursiops aduncus) inhabiting coastal waters with mixed levels of protection. *Frontiers in Marine Science* **2021**, *8*, 617518.
6. Sayigh, L.S.; Janik, V.M.; Jensen, F.H.; Scott, M.D.; Tyack, P.L.; Wells, R.S. The Sarasota Dolphin Whistle Database: A unique long-term resource for understanding dolphin communication. *Frontiers in Marine Science* **2022**, *9*, 923046.
7. Di Nardo, F.; De Marco, R.; Lucchetti, A.; Scaradozzi, D. A WAV file dataset of bottlenose dolphin whistles, clicks, and pulse sounds during trawling interactions. *Scientific Data* **2023**, *10*, 650.
8. Wall, C.C.; Haver, S.M.; Hatch, L.T.; Miksis-Olds, J.; Bochenek, R.; Dziak, R.P.; Gedamke, J. The next wave of passive acoustic data management: How centralized access can enhance science. *Frontiers in Marine Science* **2021**, *8*, 703682.
9. Matthews, J.; Rendell, L.E.; Gordon, J.C.D.; Macdonald, D. A review of frequency and time parameters of cetacean tonal calls. *Bioacoustics* **1999**, *10*, 47–71.
10. McCowan, B. A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (Delphinidae, Tursiops truncatus). *Ethology* **1995**, *100*, 177–193.
11. Janik, V.M.; Slater, P.J. Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls. *Animal behaviour* **1998**, *56*, 829–838.
12. Beeman, K. SIGNAL Sound Analysis System. *Belmont, MA* **1996**.
13. Hawkins, E.R.; Gartside, D.F. Whistle emissions of Indo-Pacific bottlenose dolphins (Tursiops aduncus) differ with group composition and surface behaviors. *The Journal of the Acoustical Society of America* **2010**, *127*, 2652–2663.
14. Azevedo, A.F.; Flach, L.; Bisi, T.L.; Andrade, L.G.; Dorneles, P.R.; Lailson-Brito, J. Whistles emitted by Atlantic spotted dolphins (Stenella frontalis) in southeastern Brazil. *The Journal of the Acoustical Society of America* **2010**, *127*, 2646–2651.
15. Rui-chao, X.; Fu-qiang, N.; Yan-ming, Y.; Yue-kun, H.; Wei, L. Study on automatic extraction of bottlenose dolphin whistles from the background of ocean noise. *2nd International Conference on Information, Communication and Engineering* **2019**.
16. Harley, H.E. Whistle discrimination and categorization by the Atlantic bottlenose dolphin (Tursiops truncatus): A review of the signature whistle framework and a perceptual test. *Behavioural processes* **2008**, *77*, 243–268.
17. Gillespie, D.; Caillat, M.; Gordon, J.; White, P. Automatic detection and classification of odontocete whistles. *The Journal of the Acoustical Society of America* **2013**, *134*, 2427–37. https://doi.org/10.1121/1.4816555.
18. Lopez-Otero, P.; Docio-Fernandez, L.; Cardenal-L pez, A. Using Discrete Wavelet Transform to Model Whistle Contours for Dolphin Species Classification. *Proceedings* **2018**, *2*. https://doi.org/10.3390/proceedings2181183.
19. Roch, M.; Soldevilla, M.; Burtenshaw, J.; Henderson, E.; Hildebrand, J. Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California. *The Journal of the Acoustical Society of America* **2007**, *121*, 1737–48. https://doi.org/10.1121/1.2400663.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 April 2026 doi:10.20944/preprints202604.1083.v1

20. YANG, W.; SUN, X.; ZHANG, Y.; WEI, C.; YANG, Y.; NIU, F. An automatic classification method for whistles of bottlenose dolphin (Tursiops truncates). *ACTA ACUSTICA* **2016**, *41*, 181–188. https://doi.org/10.15949/j.cnki.0371-0025.2016.02.005.
21. Xu, K.; Zhu, B.; Kong, Q.; Mi, H.; Ding, B.; Wang, D.; Wang, H. General audio tagging with ensembling convolutional neural networks and statistical features. *The Journal of the Acoustical Society of America* **2019**, *145*, EL521–EL527. https://doi.org/10.1121/1.5111059.
22. GAO, D.; GAO, D.; LI, X. Deep learning-based recognition of click signals of typical marine mammals. *Journal of Shaanxi Normal University,Natural Science Edition* **2019**, *47*, 37–437.
23. Yang, W.; Luo, W.; Zhang, Y. Classification of odontocete echolocation clicks using convolutional neural network. *The Journal of the Acoustical Society of America* **2020**, *147*, 49–55. https://doi.org/10.1121/10.0000514.
24. Etter, P.C. *Underwater acoustic modeling and simulation*; CRC press, 2018.
25. Wang, L.; Heaney, K.; Pangerc, T.; Theobald, P.; Robinson, S.P.; Ainslie, M. Review of underwater acoustic propagation models. **2014**.
26. Porter, M.B.; Bucker, H.P. Gaussian beam tracing for computing ocean acoustic fields. *The Journal of the Acoustical Society of America* **1987**, *82*, 1349–1359.
27. Bowlin, J.B.; Spiesberger, J.L.; Duda, T.F.; Freitag, L.F. Ocean acoustical ray-tracing software RAY. Technical report, 1992.
28. Porter, M.B. The KRAKEN normal mode program. *Unknown* **1992**.
29. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* **2017**.
30. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
32. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
34. Santos-Domínguez, L.; Vázquez, M.; Llorca, J.M.; Carballo, A. ShipsEar: An underwater vessel noise database. *Applied Acoustics* **2016**, *114*, 155–163. Accessed: 2025-06-08, https://doi.org/10.1016/j.apacoust.2016.06.008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.