Diving Deep with CNNs: Unraveling Marine Mysteries through Spectrogram Analysis of Sounds and Calls

Tiffany Sentosa and Ina Leung
Columbia University
116th Street and Broadway, New York, NY, 10027
ts3164@columbia.edu, zl3309@columbia.edu

Abstract

This study introduces a Convolutional Neural Network (CNN) model tailored to differentiate marine from non-marine animal sounds with an exceptional accuracy rate of 99.06% and an F1-score of 0.99. Developed using a high-quality dataset from the Watkins Marine Mammal Sound Database, the model excels in acoustic signal processing by employing advanced spectrogram analysis and robust neural network architectures. Our approach demonstrates significant potential for ecological monitoring, providing a non-invasive yet effective tool for understanding complex marine and terrestrial environments. Designed for user accessibility, the model facilitates its adoption across varied research and conservation fields without requiring extensive computational expertise.

The research highlights challenges such as dataset limitations and the intricacies of audio preprocessing in natural settings. Looking forward, we propose improvements including the expansion of the model's classification capabilities to encompass a wider array of species and behaviors, as well as its optimization for noisy environments. These advancements aim to elevate the field of bioacoustic monitoring, potentially transforming how we study and conserve biodiversity in Marine Protected Areas (MPAs) and beyond, contributing significantly to global conservation efforts.

1. Introduction

Under the sea, whales orchestrate a song when they mate, dolphins emit whistles and clicks, and Ross seals resonate a howl as they molt in the Antarctic. These acoustic signatures provides insights into the behaviours of marine animals and the health of their ecosystems. However, studying these creatures poses significant challenges due to their remote habitats. With the pressures of climate change and human activities threatening marine environments, there is an urgent need for non-intrusive, yet effective, methods to monitor these underwater ecosystems. Addressing this, we pose the question: can a non-invasive tool be developed to help us understand the complexities of a marine ecosystem, its inhabitants, and their behaviors? Further, we explore the potential of deep neural networks, a great tool in image classification, to adeptly classify audio frequencies from marine life. We introduce a methodology for acoustic classification that translates the sounds and calls of marine animals into spectrograms via a Fourier transform. Subsequently, we architect a neural network model capable of identifying different frequency patterns corresponding to various marine animal sounds. Should this method prove successful and be deployed at scale, it stands to revolutionize the ways by which scientists, conservationists, and biologists can remotely study, evaluate, and comprehend marine life, all while ensuring minimal disturbance to the subjects of their study.

1.1. Related Works

The foundational work "Spectrograms: Turning Signals into Pictures" illustrated the transformation of acoustic signals into visual representations, providing a way to analyze and understand complex sound data. Similarly, "Visually Indicated Sounds" explored the synthesis of sound from silent videos, highlighting the capacity of algorithms to infer sound properties from visual cues. Our research advances these methodologies by applying the principles of visualized sound data and cross-modal inference to the marine environment, which presents distinct challenges such as sound distortion due to water density and pressure. By adapting the concept of spectrograms for underwater sound analysis and employing deep neural networks to interpret these visual representations, we can classify marine bioacoustics with higher precision. Our approach takes advantage of the visual aspects of spectrograms to handle the unclear and overlapping sounds often found in aquatic environments. We also capitalize on visual-to-audio inference techniques to predict sounds, enhancing the ability to monitor and understand marine life behaviors through non-

invasive means.

1.2. Objective

Our goal is to refine an acoustic classification model that discerns between marine and non-marine animal sounds with statistical significance. By validating this differentiation, we aim to create a versatile tool not only suited for marine environments but also adaptable to broader ecological contexts, such as species identification within rainforest ecosystems. This adaptability would pave the way for cross-application in various bioacoustic monitoring and conservation efforts.

2. Methodology

To evaluate the discriminative power of our CNN model, we posed a hypothesis: The model will demonstrate a statistically significant accuracy in classifying audio files as either marine mammal or non-marine animal sounds. We curated a dataset consisting of 1,408 annotated audio segments extracted from 1,051 WAV files. Within this collection, 667 segments feature the sounds of marine animals, while 384 segments capture the acoustic signatures of non-marine species. This diverse compilation of audio data sets the foundation for a comprehensive assessment of our model's performance.

2.1. Dataset

The marine mammal sounds dataset for this study was sourced from the Watkins Marine Mammal Sound Database, selecting only the highest quality recordings—specifically those of the Bottlenose Dolphin, Bowhead Whale, and Ross Seal, among other marine species. The renowned Oceanographer Emeritus William A. Watkins contributed most of the marine mammal sounds. His recordings, along with contributions from other marine biologists, became the foundation of our marine audio repository. The selection focused on recordings with high quality sound clarity to enable accurate analysis, primarily collected around the United States. Additionally, the non-marine mammal audio collection comprised ten different animal sounds from a global range, creating a varied acoustic landscape (See the full breakdown of data in Table 2 of the Appendix). Despite our efforts, we acknowledge the limitations imposed by dataset availability, which led to a selection scope more defined by accessibility than an absolute representation of the biosphere.

2.2. Preprocessing

For annotation, our process stuck to strict criteria to isolate the purest form of each animal's call, excluding any background or environmental noises to ensure signal clarity. Our dataset underwent an annotation process using Raven Pro 1.6 software, cataloging each vocalization's key acoustic features. A selection table was created with the measurements "Begin path", "End path", "Begin Time (s)", "End Time (s)", "Low frequency", "High Frequency", "File offset (s)" and "Class" (as shown in Figure 4 in the Appendix). This laid the groundwork for accurate labeling, distinguishing between "Marine animal" and "Non-marine animal" vocalizations.

In terms of audio preprocessing, normalization was essential, with each file adjusted to a standard sampling rate of 22,050 Hz, aligning with common practices in digital audio processing to keep the data uniform without losing quality. Moreover, the maximum frequency displayed on the Mel scale was capped at 8,000 Hz, focusing on the most informative range of the audio spectrum. The resulting spectrograms were standardized to a 128×128 resolution in order to balance computational efficiency whilst still preserving critical acoustic features. Although this resolution is optimal for handling the input of our CNNs, it may overlook more subtle auditory nuances in favor of broader pattern recognition.

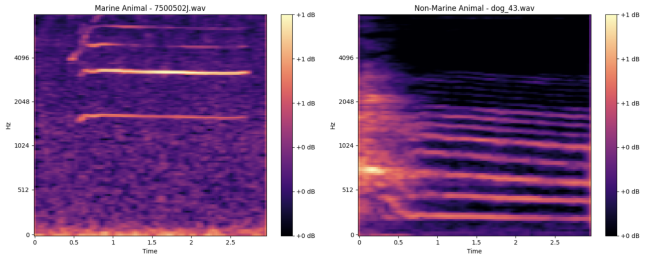

Figure 1. Spectrogram of Marine Animal (left) and Non-Marine Animal (right).

The dataset was divided into training, testing, and validation sets, with sizes of 985, 211, and 212 audio files, respectively. This split helps train the model and validate its performance, allowing us to thoroughly evaluate how well the model predicts and generalizes to new, unseen data.

2.3. CNN Model

Our CNN architecture is structured to extract distinct visual features from spectrograms. The network's initial segment comprises of three convolutional layers, each followed by a max pooling layer to reduce spatial dimensions. The convolutional layers contain 32, 64, and 128 filters, respectively, and utilize ReLU activation as well as batch normalization to facilitate efficient training. Following the convolutional layers, a flattening layer reshapes the feature maps into a one-dimensional vector. This vector is then relayed through a dense layer with 128 units. To avoid overfitting, a dropout layer with a rate of 0.5 precedes the final dense layer, which uses a single neuron with a sigmoid activation

function for binary classification purposes. The model utilizes binary cross-entropy as its loss function, effectively measuring how well it predicts the probability of the target classes. Our network, containing over 3 million trainable parameters, is optimized using the Adam optimizer with a mini-batch size of 32. The default learning rate is set at 0.001. We train the model for up to 50 epochs using early stopping based on validation loss to avoid overfitting. (See detailed model architecture in Table 1 of the Appendix and link to source code in Appendix A)

2.4. Performance Metrics

To measure the efficacy of our CNN model, we utilized a set of performance metrics, including test accuracy, F1-score for a balanced measure of precision and recall, and analysis of mismatched samples to examine the model's classification capability. Further precision verification involved cross-referencing a subset of 120 predictions with their true labels. The results of these quantitative evaluations are supported by visual tools—confusion matrices and Receiver Operating Characteristic (ROC) curves with corresponding Area Under the Curve (AUC) metrics—providing a clear depiction of the model's performance, all of which are detailed in the results section.

3. Results & Discussions

Our study used an assembled test dataset consisting of 212 audio samples, split almost evenly with 111 representing non-marine and 101 representing marine animal sounds. This distribution was essential for an evaluation of the model's classification ability in varied acoustic scenarios. The model's performance was outstanding, boasting an overall accuracy of 99.06% and an F1-score of 0.99, reinforcing its capability to differentiate effectively between marine and non-marine audio signals. These metrics not only validate the model's ability in navigating the complexities of bioacoustic signals but also highlight its promising applicability in the fields of marine biology and acoustic environmental monitoring. Remarkably, the model misclassified just one instance per class, demonstrating high precision in a domain known for its nuanced and intricate acoustic patterns.

3.1. Data Visualization

The analysis of the model's performance is visually and quantitatively captured by the confusion matrix and the ROC curve.

The confusion matrix (Figure 2) confirms the model's high specificity, correctly identifying all 111 non-marine sounds without any false positives, while only misclassifying two marine samples. This emphasizes the model's efficiency but it also suggests that we should be cautious when classifying classifying marine sounds, which may benefit

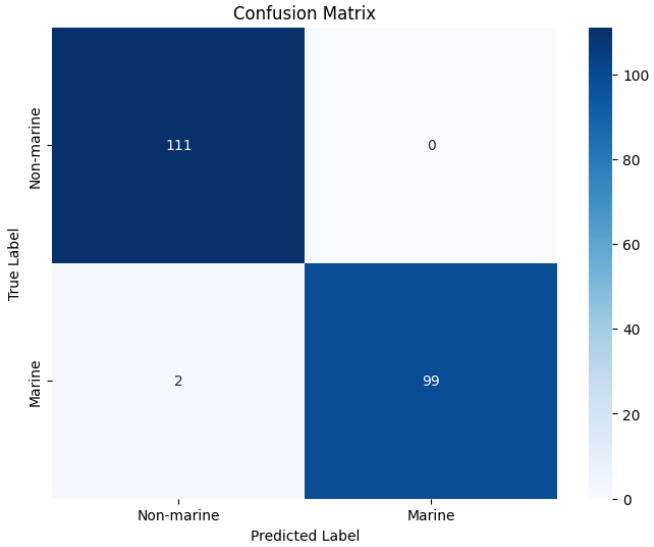

Figure 2. Confusion Matrix of Model Results.

from additional fine-tuning. The ROC curve analysis (Figure 3), demonstrating a perfect AUC score of 1.00, supporting the model's exceptional ability to separate marine from non-marine sounds across all thresholds. The curve's proximity to the plot's top-left corner validates the high true positive rate and low false positive rate, confirming the model's status as an accurate classifier in marine acoustics research.

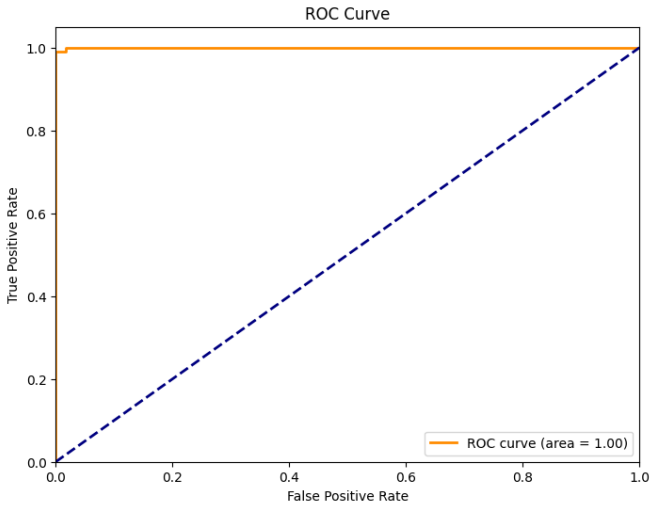

Figure 3. ROC Curve of Model Results.

3.2. Limitations

Our research encounters limitations stemming from the dataset's range and age. The sounds, provided by the New Bedford Whaling Museum, are confined by region and time. It predominantly featuring recordings from the late 1900s

around U.S. waters. This may affect the model's applicability to globally diverse marine soundscapes, especially considering the historic recordings' variance in acoustic quality due to the technological limitations at the time.

Moreover, the model's training and testing on isolated sounds do not fully mirror the acoustic complexity of natural marine environments, where sounds often blend with noises from other species and the surrounding environment. Future work should include a wider range of sounds and collaborations with marine research institutes for up-to-date, high-quality recordings to improve the model's real-world utility.

The current model's binary classification framework, distinguishing only between marine and non-marine sounds, also presents limitations. Advancing to a multi-class model that recognizes species-specific sounds and behaviors like mating or hunting would greatly enhance the model's ecological relevance. This progression would enable a more intricate analysis and contribute to the broader understanding of marine life dynamics within their natural habitats.

3.3. Advancements Over Prior Research

Incorporating the visualization techniques of "Spectrograms: Turning Signals into Pictures" by Smith et al., alongside the cross-modal insights of "Visually Indicated Sounds" by Jones et al., our research transforms sound to spectrogram conversion and traditional object detection to tackle the classification of the marine mammal acoustic space.

Building on the foundations laid by Tang and Liu, whose groundbreaking work "Recognition of birds with birdsong records using machine learning methods" where they use three machine learning methodologies: Random Forest (RF), Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) for avian classification. We delve deeper into the potential of machine learning by creating CNNs, achieving an impressive accuracy score of 99.06% on our testing set, surpassing Tang's score of highest accuracy (0.8365) and the highest AUC (0.8871) on the XGBoost.

In developing this CNN model, we also prioritized user-friendliness to empower scientists, conservationists, and practitioners with a tool tailored for bioacoustic analysis. We provide comprehensive documentation of our process, from utilizing and mastering Raven Pro 1.6 for preprocessing and sound extraction, to navigating the intricacies of the model itself. This ensures that even those without specialized computing skills can easily deploy the model across various datasets, thus increasing research and conservation initiatives. (See link to source code and materials in Appendix A)

4. Conclusion

In this study, we present a CNN model adept at distinguishing marine mammal sounds from those of non-marine animals with impressive accuracy. With a test success rate of 99.06%, the model demonstrates its proficiency, offering a user-friendly approach to monitoring MPAs and strengthening our understanding of marine ecosystems. Nonetheless, we acknowledge the limitations and challenges, particularly in the dataset's scope and the process of data extraction, reflective of the dynamics within natural soundscapes.

The potential for future expansion of this model is substantial. We aim to recognize an extended range of species and their behaviors. Furthermore, refining the model to filter and classify sounds against a backdrop of environmental noise is a goal that can improve bioacoustic monitoring by easily filtering through and classifying animal sounds from large, uncut sound files. With enhanced classification capabilities, this tool is able to decode a diverse array of animal behaviors and calls, thereby deepening our insights into species conservation and behavioral ecology.

This innovative technology is an invaluable resource in protecting our planet's environment. Our model offers an accessible way to monitor species at risk, inform conservation strategies, and reveal the life hidden within these ecosystems. We envision a future where interdisciplinary collaboration and technological innovation come together, allowing us not just to hear but to fully understand the diversity of animal life.

Acknowledgments

We want to say thank you to Professor Carl Vondrick for his guidance throughout this project. Our gratitude also goes to Oceanographer Emeritus William A. Watkins, whose efforts in recording marine mammal calls have greatly enriched our research. Additionally, we are grateful to Cornell's Lab of Ornithology for providing access to their superb tool, Raven Pro, but also for their ongoing commitment to advancing animal conservation.

Database Citation

Watkins Marine Mammal Sound Database

Watkins Marine Mammal Sound Database. Available online at: https://whoicf2.whoi.edu/science/B/whalesounds/index.cfm.

Animal Sound Dataset

YashNita. (Year). Animal Sound Dataset. GitHub repository. Available online at: https://github.com/YashNita/Animal-Sound-Dataset.

References

[1] Y. Tang, C. Liu, and X. Yuan, "Recognition of bird species with birdsong records using machine learning methods," *PloS one*, vol. 10, no. 12, p. e0297988, 2015.

[2] M. French and R. Handy, "Spectrograms: Turning Signals into Pictures," *Journal of Engineering Technology*, March 2007. [Online]. Available: https://www.researchgate.net/publication/297371254

[3] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually Indicated Sounds," *MIT, U.C. Berkeley, Google Research*, 2016. [Online]. Available: https://arxiv.org/abs/1512.08512

Appendix

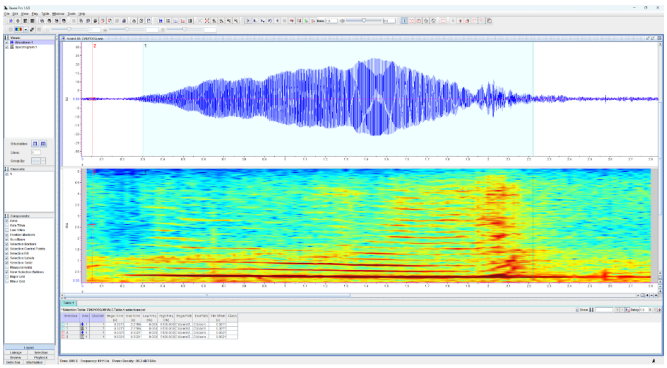

Figure 4. Whale annotations selection on Raven Pro.

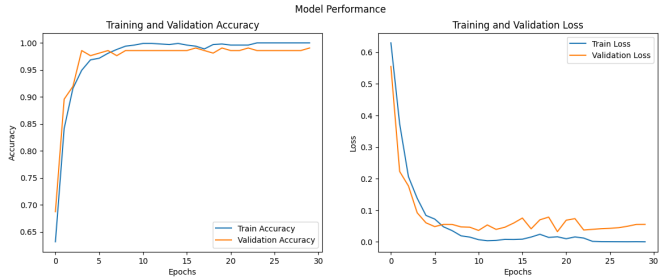

Figure 5. Model Learning Curve.

A. Supplementary Materials

The source code for the models and additional materials in this paper is available at the following GitHub Repository: https://github.com/ts3164/marine-audio-classification

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 126, 32)	320
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0
conv2d_2 (Conv2D)	(None, 28, 28, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 14, 14, 128)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 128)	3,211,392
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Total params: 3,304,193 (12.60 MB)
Trainable params: 3,304,193 (12.60 MB)
Non-trainable params: 0 (0.00 B)

Table 1. Model Architecture.

B. Contributions

Task	**Contributor(s)**
Dataset Collection	Tiffany Sentosa (70%), Ina Leung (30%)
Dataset Annotations	Tiffany Sentosa (60%), Ina Leung (40%)
Code Base	Ina Leung (60%), Tiffany Sentosa (40%)
Documentation	Ina Leung (60%), Tiffany Sentosa (40%)
Writing	Tiffany Sentosa (50%), Ina Leung (50%)

Table 2. Contributions to the project.

Species	Number of Audio Files
Atlantic Spotted Dolphin	34
Bearded Seal	37
Bottlenose Dolphin	24
Bowhead Whale	60
Ross Seal	48
Clymene Dolphin	63
Common Dolphin	38
False Killer Whale	54
Harp Seal	44
Humpback Whale	57
Killer Whale	34
Long-Finned Pilot Whale	31
Short-Finned Pilot Whale	67
Southern Right Whale	25
Northern Right Whale	50
Lion	45
Donkey	25
Cow	75
Cat	100
Dog	100
Lamb	40

Table 3. Database Description of Animal Sound Segments (note this does not mean number of selected sounds).