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 enhancements 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 punctuate the depths with whistles and clicks, and Ross seals resonate a haunting howl as they molt in the Antarctic. The acoustic signatures of marine fauna are critical indicators of their behavior and the health of their ecosystems. Yet, the study of these enigmatic creatures is a formidable challenge due to their inaccessible habitats. As the specter of climate change approaches and human activities increasingly infringe upon marine environments, there is a need for techniques that offer non-intrusive yet effective monitoring of these underwater ecosystems. Addressing this, we pose the question: can a tool, non-invasive and proficient, 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 fast 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 modalities 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 means 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 attenuation and 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 not only utilizes the visual aspects of spectrograms to contend with the less distinct and more overlapping acoustic signals found in marine settings but also capitalizes on the visual-to-audio inference techniques to predict sounds from underwater imagery, 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 aquatic 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 posited 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 robust 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, Ross Seal, among other marine species. The bulk of the marine mammal sounds were contributed by the renowned Oceanographer Emeritus William A. Watkins. His recordings, along with contributions from other marine biologists, became the foundation of our marine audio repository. The selection focused on recordings with superior sound clarity to enable accurate analysis, primarily collected around the United States. Additionally, the non-marine mammal audio collection comprised ten distinct animal sounds from a global range, creating a diversified 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 exhaustive representation of the biosphere.

2.2. Preprocessing

For annotation, our process adhered to strict criteria to isolate the purest form of each animal's call, meticulously excluding any background or environmental noises to ensure signal clarity. Our dataset underwent a thorough annotation process using Raven Pro 1.6 software, meticulously cataloging each vocalization's key acoustic features. A detailed selection table was created, including the "Begin Time (s)", "End Time (s)", "Low frequency", "High frequency", and other pivotal acoustic characteristics for each call, as shown in Figure 4 in the Appendix. This granular detail laid the groundwork for accurate label attribution, distinguishing between "Marine animal" and "Non-marine animal" vocalizations, crucial for a robust framework for machine learning tasks.

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 balance data uniformity with fidelity. 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, a pragmatic balance between computational efficiency and retention of critical acoustic features. While this resolution optimizes input handling for our convolutional neural network, it may potentially smooth over subtler auditory nuances in favor of broader pattern recognition.

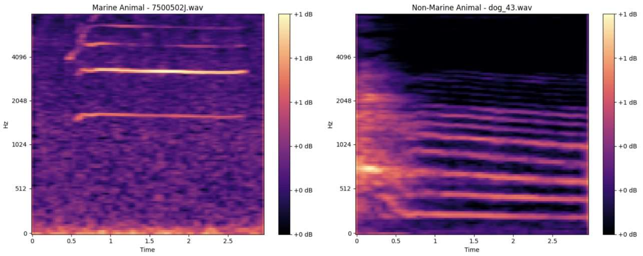

Figure 1. Spectrogram of Marine Animal (left) and Non-Marine Animal (right).

To ensure robust model evaluation, the dataset was divided into training, testing, and validation sets, with sizes of 985, 211, and 212 audio files, respectively. This split supports comprehensive learning and validation, allowing for a detailed assessment of the model's predictive accuracy and generalizability across unseen data.

2.3. CNN Model

Our CNN architecture is structured to extract distinct visual features from spectrograms. The network's initial segment comprises 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. Subsequent to 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 mitigate overfitting, a dropout layer with a rate of 0.5 precedes the final dense layer, which employs 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 with a mini-batch size of 32, 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 ascertain the efficacy of our CNN model, we utilized a comprehensive set of performance metrics, including test accuracy, F1-score for a balanced measure of precision and recall, and an analysis of mismatched samples to examine the model's discriminative capability. Further precision verification involved cross-referencing a subset of 120 predictions with their true labels. The results of these quantitative evaluations are complemented by visual tools—confusion matrices and Receiver Operating Characteristic (ROC) curves with corresponding Area Under the Curve (AUC) metrics—providing an intuitive and clear depiction of the model's performance, all of which are detailed in the results section.

3. Results & Discussions

Our study engaged with a carefully assembled test dataset, encompassing 212 audio samples split evenly, with 111 representing non-marine and 101 marine animal sounds. This equitable distribution was essential for an extensive evaluation of the model's classification prowess 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 skill in navigating the intricacies 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, showcasing a remarkable precision within a domain replete with nuanced and complex acoustic patterns.

3.1. Data Visualization

The fine-grained analysis of the model's performance is visually and quantitatively captured by the confusion matrix and the Receiver Operating Characteristic (ROC) curve.

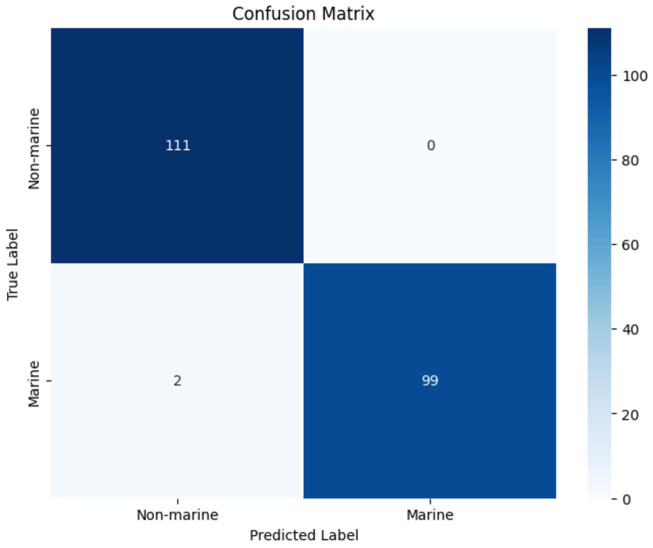

Figure 2. Confusion Matrix of Model Results.

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 precise distinction underscores the model's efficiency, although it also indicates a cautious approach to classifying marine sounds that may benefit from further calibration. The ROC curve analysis (Figure 3), demonstrating a perfect AUC score of 1.00, supports 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, cementing the

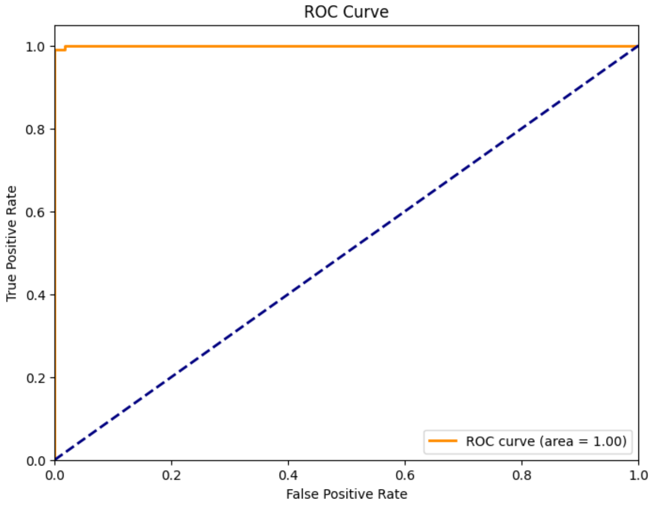

Figure 3. ROC Curve of Model Results.

model's status as an ideal classifier in marine acoustics research.

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 regionally and temporally confined, predominantly featuring recordings from the late 1900s around U.S. waters. This specificity may affect the model's applicability to contemporary, globally diverse marine soundscapes, especially considering the historic recordings' potential variance in acoustic quality due to the technological limitations of the time.

Additionally, the model's training and testing on isolated sounds do not fully mirror the acoustic complexity of natural marine environments, where sounds often intermingle with other species and ambient noises. Future work should include a richer array of sounds and collaborations with marine research institutes for up-to-date, high-quality recordings to improve the model's real-world utility and robustness.

The current model's binary classification framework, distinguishing only between marine and non-marine sounds, also presents limitations. Advancing to a multi-class, behavior-centric 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 acoustic habitats.

3.3. Advancements Over Prior Research

Incorporating the groundbreaking 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 transcends sound to spectrogram conversion and traditional object detection to tackle the classification of the marine mammal acoustic sphere.

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 mobilizing Convolutional Neural Networks (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 versatile 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 customize and deploy the model across various datasets, thus enhancing research and conservation initiatives. (See link to source code and materials in Appendix A)

4. Conclusion

In this study, we present a Convolutional Neural Network (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 Marine Protected Areas and enhancing our understanding of marine ecosystems. Nonetheless, we acknowledge ongoing challenges, particularly in the dataset's scope and the intricate process of data extraction, reflective of the complex dynamics within natural soundscapes.

The potential for future expansion of this model is substantial. We aim to recognize an extended array of species and their behaviors, from the tumult of monkey confrontations to the intimate rituals of whale mating. Furthermore, refining the model to filter and classify sounds against a backdrop of environmental noise is an aspirational 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 poised to decode a diverse array of animal behaviors and calls, thereby deepening our insights into species conservation and behavioral ecology.

This innovative technology is poised to become an invaluable resource in safeguarding our planet's most treasured environments. From the verdant expanses of the Amazon rainforest to the icy seascapes of the Ross Sea Region Marine Protected Area, our model offers an accessible means to monitor at-risk species, inform conservation strategies, and reveal the rich tapestry of life hidden within these ecosystems. We envision a future where interdisciplinary collaboration and technological innovation converge, allowing us to not only hear but fully comprehend the vast symphony of animal life in all its richness and diversity. — This revised conclusion is designed to be more fluid and cohesive, ensuring a compelling summary of your work and a vision for its future application.

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

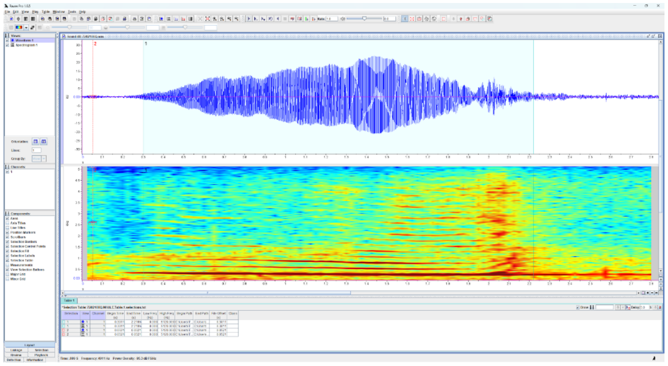

Figure 4. Whale annotations selection on Raven Pro.

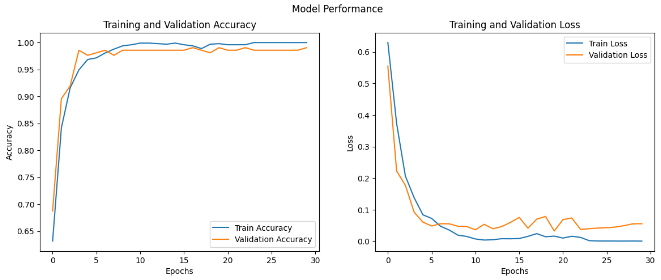

Figure 5. Model Learning Curve.

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

Table 1. Model Architecture
Total params: 3,304,193 (12.60 MB)
Trainable params: 3,304,193 (12.60 MB)
Non-trainable params: 0 (0.00 B)

A. Supplementary Materials

The source code for the models and additional materials in this paper is available at the following Google Drive: `https://drive.google.com/drive/folders/1NSe36FOMtt4MaHTv53XvGa8Ya4ZrwFuJ?usp=drive_link`

B. Contributions

Task	**Contributor(s)**
Dataset Collection	Tiffany Sentosa, Ina Leung
Dataset Annotations	Tiffany Sentosa, Ina Leung
Code Base	Ina Leung, Tiffany Sentosa
Documentation	Tiffany Sentosa, Ina Leung
Writing	Tiffany Sentosa and Ina Leung

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, please note this does not mean number of selected sounds