IST 691: Deep Learning in Practice

Final Project Report

Deep Learning for Acoustic Detection of North Atlantic Right Whale Calls: A Conservation and Maritime Safety Approach

Group 1:
Nandita Ghildyal
Mengqi Li

IST 691: Deep Learning in Practice

Final Project Report

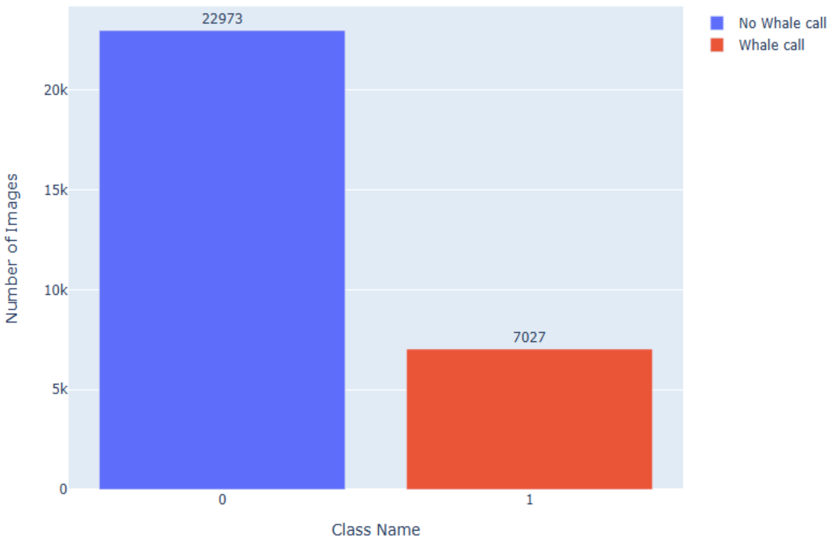
Introduction

The North Atlantic right whale, one of the most critically endangered marine species, faces significant threats from ship collisions and habitat degradation. This project aims to develop a deep learning-based algorithm to detect whale calls from audio, addressing challenges such as background noise, class imbalance, and complex vocalizations. Whale calls, including moans and clicks, provide crucial insights into behavior, migration, and habitats, enabling conservation efforts and real-time alerts to reduce ship collisions.

To tackle these challenges, the project employs advanced techniques like autoencoders, CNNs for spatial pattern recognition, and attention-based CNNs for any temporal tendencies. Raw audio waveform-based methods are also explored to improve detection performance. By automating whale call detection, this work contributes to marine conservation and supports global efforts to minimize the environmental impact of shipping while preserving endangered species.

Data

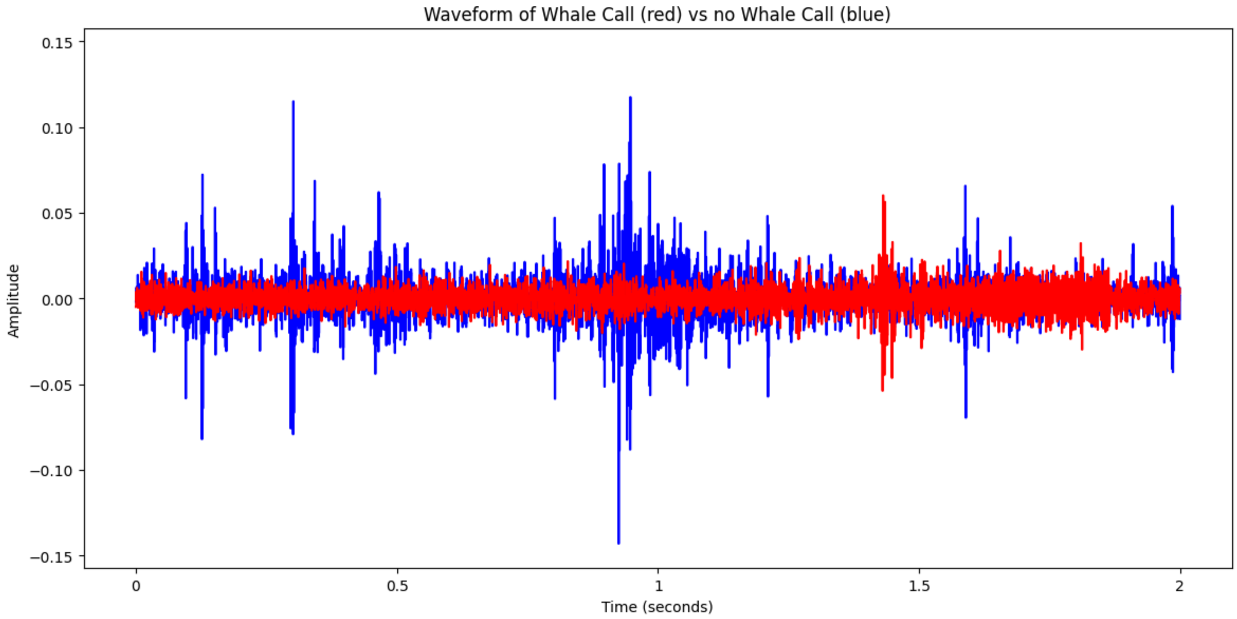
The dataset comprises approximately 30,000 training samples and 54,503 testing samples, each as a 2-second .aiff sound clip sampled at 2 kHz. Each clip is labeled as 1 (right whale call) or 0 (absence of a call) and may include a mix of right whale calls, non-biological noise, or other whale sounds. Right whale signals typically range between 100 and 500 Hz, with energy concentrated around 200 Hz, and often feature harmonics and overtones extending up to 500 Hz.



2.1 Class Imbalance in the Dataset

To address the class imbalance, where the 'No whale call' class dominates, we implemented two techniques. For clustering tasks, we used SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class, creating a balanced dataset and improving pattern recognition. For the CNN model, we applied class weighting, adjusting the loss function to give greater importance to underrepresented classes. This approach ensured the model prioritized learning features of the minority class, enhancing its performance on imbalanced data and improving generalization.

We visualized audio patterns to understand the differences between whale calls and background noise. The waveform comparison shows whale calls (red) as consistent with lower amplitude variation, reflecting their structured and periodic nature. In contrast, background noise (blue) exhibits frequent, pronounced spikes, indicative of random interference. Both signals share a similar baseline noise level, likely due to shared environmental conditions or recording artifacts. This analysis underscores the need for robust methods to effectively distinguish whale calls from noise.

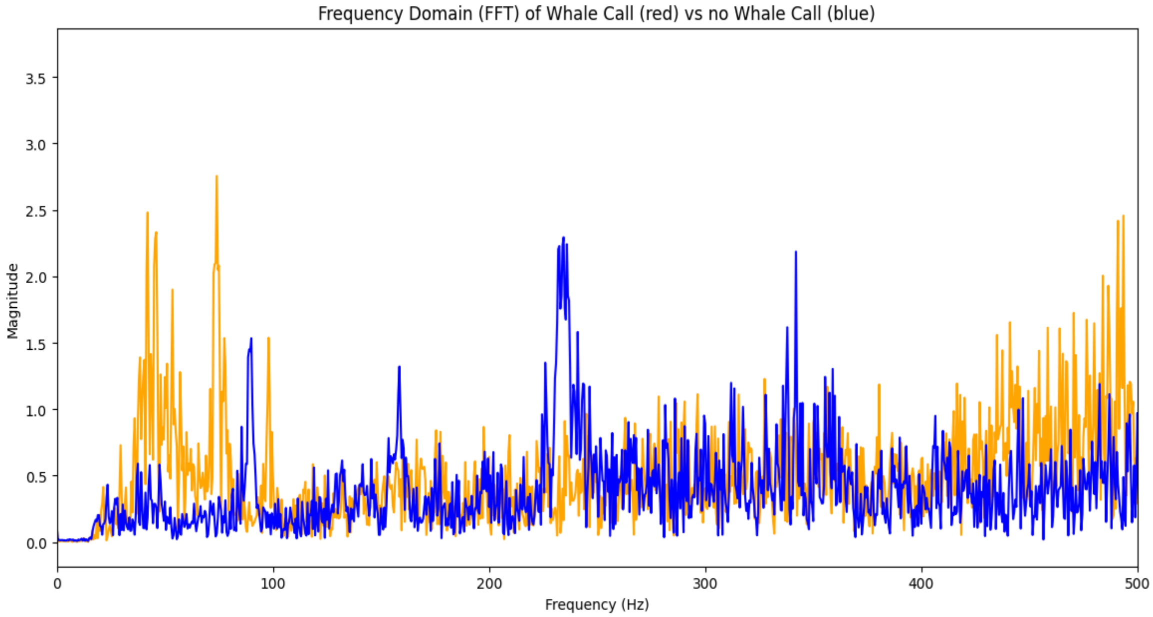


2.2 Waveform of Whale Call (red) vs no Whale Call (blue)

Notably, the regions where the red waveform diverges from the blue corresponds to distinctive features of the whale call, such as tonal patterns or harmonics, which are absent in the noise-only signal. Identifying subtle differences in signal behavior is critical for developing filters or algorithms to effectively isolate whale calls in noisy environments.

Next, we performed a frequency domain analysis of whale calls (orange) and background noise (blue) using a Fast Fourier Transform (FFT). The x-axis represents frequency components in Hertz (Hz), while the y-axis shows their respective magnitudes.

The orange curve, corresponding to whale calls, exhibits distinct peaks at certain frequencies, reflecting dominant tones or harmonics characteristic of whale vocalizations. In contrast, the blue curve, representing noise, has a flatter profile with lower magnitudes across the spectrum, indicating the absence of prominent tonal features.

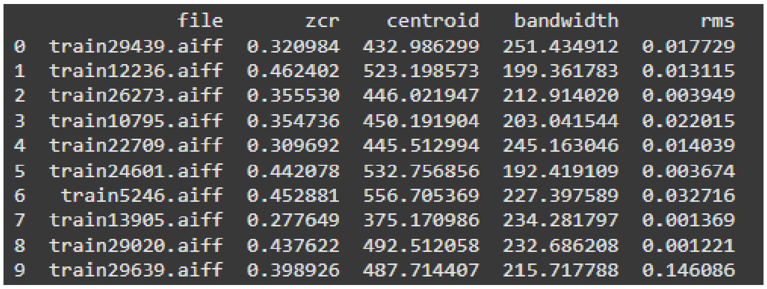


2.3 FFT of Whale Call (red) vs no Whale Call (blue)

Methods

To build our model, we extracted key acoustic features from the audio data, including Zero Crossing Rate (ZCR), Spectral Centroid, Spectral Bandwidth, and Root Mean Square (RMS). ZCR measured the signal's zero crossings, Spectral Centroid represented the weighted average of the frequency spectrum, Spectral Bandwidth captured the spread around the centroid, and RMS quantified the average squared amplitudes.

Correlation analysis and ANOVA were then performed, leading to the removal of Spectral Bandwidth and RMS as they contributed minimally to distinguishing whale calls from background noise. RMS, which measures loudness, was discarded because the audio recordings were captured at controlled decibel levels, resulting in minimal loudness variation across samples. Similarly, Spectral Bandwidth, which reflects the range of frequencies in a signal, was not a strong differentiator due to the relatively narrow frequency range of whale calls compared to background noise. With these features removed, we focused on the more informative ZCR and Spectral Centroid for subsequent clustering and classification tasks, as they provided clearer distinctions between whale calls and noise.



```
              file       zcr    centroid   bandwidth       rms
0  train29439.aiff  0.320984  432.986299  251.434912  0.017729
1  train12236.aiff  0.462402  523.198573  199.361783  0.013115
2  train26273.aiff  0.355530  446.021947  212.914020  0.003949
3  train10795.aiff  0.354736  450.191904  203.041544  0.022015
4  train22709.aiff  0.309692  445.512994  245.163046  0.014039
5  train24601.aiff  0.442078  532.756856  192.419109  0.003674
6   train5246.aiff  0.452881  556.705369  227.397589  0.032716
7  train13905.aiff  0.277649  375.170986  234.281797  0.001369
8  train29020.aiff  0.437622  492.512058  232.686208  0.001221
9  train29639.aiff  0.398926  487.714407  215.717788  0.146086
```

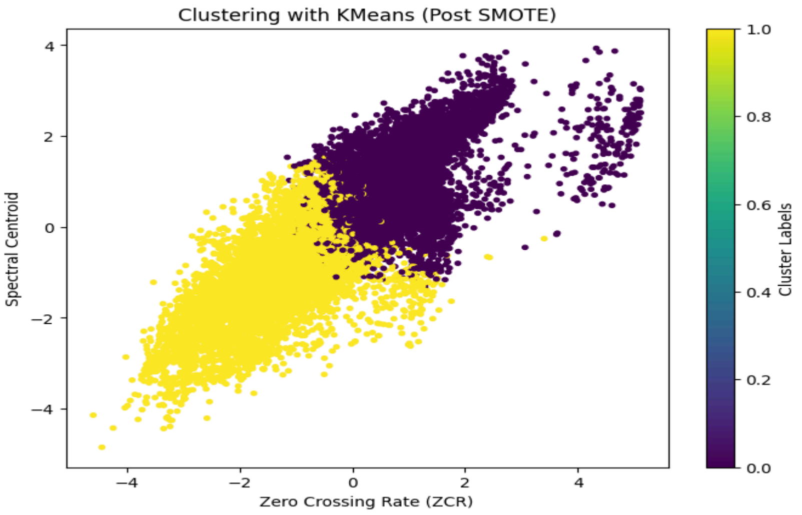
3.1 Numerical Features extracted from audio files

We then employed these relevant features for clustering and classification tasks. For dimensionality reduction and pattern learning, we utilized Convolutional Autoencoders (CAEs). We then experimented with CNNs to fine-tune the model for optimal classification by incorporating acoustic features learned before.

Clustering

After performing correlation analysis and ANOVA to remove irrelevant features, we used the remaining relevant features for clustering and classification. The clustering was performed using KMeans, with the data divided into two main clusters. The scatter plot below visualizes the clustering results after applying Synthetic Minority Over-sampling Technique (SMOTE), which helped balance the data by generating synthetic samples.

The plot below shows two distinct clusters, represented by yellow and purple points, with some overlap in the middle. This overlap suggests that the ZCR and Spectral Centroid alone are not sufficient to fully separate whale calls from non-whale sounds, especially when considering more complex environmental noises.



3.2 Numerical Features clustered for Classification

The clustering analysis using Zero Crossing Rate (ZCR) and Spectral Centroid revealed distinct differences between whale calls and non-whale sounds. Non-whale sounds showed higher ZCR values and broader bandwidths, reflecting noisier, more complex signals. However, the clustering performance highlighted limitations. Class 0 (non-whale calls) achieved better metrics, with a precision of 0.76, recall of 0.44, and an F1 score of 0.56, while Class 1 (whale sounds) had lower precision (0.23), recall (0.56), and an F1 score of 0.33. The overall clustering accuracy was 46%, indicating that these features alone were insufficient for robust classification, particularly for the minority class.

While ZCR and Spectral Centroid provide some separation between classes, they do not fully capture the dataset's complexity. These features will be integrated into a Convolutional Neural Network (CNN) alongside more discriminative features like MFCCs to improve classification accuracy and better differentiate whale calls from environmental noise.

Convolutional Autoencoder

We implemented a Convolutional Autoencoder (CAE) to extract high-level, lower-dimensional features from spectrogram data. Preprocessing involved converting spectrograms to grayscale, resizing to 50 × 101 pixels, and normalizing pixel values to [0, 1], with the dataset split into 80% training and 20% validation subsets. The CAE architecture utilized convolutional and max-pooling layers in the encoder to extract hierarchical features, and upsampling layers in the decoder to reconstruct spectrograms while preserving essential data integrity. Trained over 40 epochs with MSE loss and the Adam optimizer, the CAE effectively compressed spectrograms into compact latent representations.

These 3D features were flattened using Global Average Pooling and further reduced via PCA for improved interpretability, enabling effective clustering and feature separability. Reconstruction quality and latent space visualization validated the model’s performance, demonstrating its ability to retain critical information and facilitate seamless integration into downstream classification tasks.

Convolutional Neural Network

Building on insights from clustering and autoencoder analysis, we developed a Convolutional Neural Network (CNN) to classify whale calls. Features such as Zero Crossing Rate (ZCR), Spectral Centroid, and Mel-Frequency Cepstral Coefficients (MFCCs) were utilized to enhance performance and distinguish whale calls from environmental noise. The CNN architecture included two convolutional layers with 32 and 64 filters, followed by max-pooling layers to reduce dimensionality.

A dense layer with 128 neurons and a dropout rate of 0.5 was added to prevent overfitting, with a sigmoid activation function for binary classification. This model achieved a classification accuracy of 75% and a recall of 0.56, but its limitations in handling complex patterns in noisy data highlighted the need for a more advanced design.

The Enhanced CNN improved on this by introducing batch normalization for stable training and deeper convolutional blocks with progressively increasing filters (32 to 256) and larger kernels to capture broader spatial patterns. Image features were combined with numerical data representations for a comprehensive approach. Weighted loss functions addressed class imbalance by prioritizing the minority class, while dropout and batch normalization improved regularization. This architecture achieved a validation accuracy of 91.73% and an F1 score of 0.84, demonstrating a strong balance between precision (0.75) and recall (0.79).

While the Attention-Based CNN prioritized key regions in spectrograms, it struggled with the short 2-second clips, failing to capture the long-term dependencies of whale calls. Similarly, LSTM models were unsuitable due to the lack of significant sequential information. In contrast, the Enhanced CNN emerged as the best-performing model, effectively leveraging diverse features to generalize across noisy and imbalanced datasets, making it the optimal choice for separating whale calls from environmental noise.

Finally, we experimented with SincNet, a deep neural network designed for audio signal processing, to evaluate its potential in identifying whale calls. Unlike traditional models relying on predefined features like MFCCs or Mel spectrograms, SincNet employs learnable Sinc function-based filters to adaptively capture frequency responses during training. This flexibility makes it effective for speech recognition tasks.

However, applying SincNet to whale call data posed challenges, as whale calls occupy low-frequency ranges (10 Hz to 10 kHz) distinct from the human speech range (300 Hz to 8 kHz) for which SincNet is optimized. The model struggled to learn filters that captured the subtle, low-frequency characteristics of whale calls effectively.

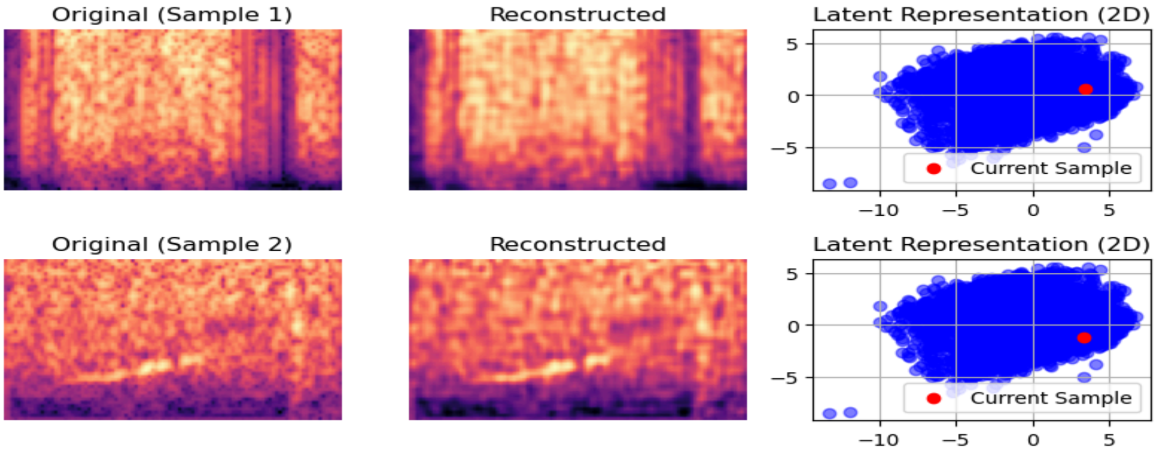
Additionally, the complex, long-duration acoustic patterns of whale calls, which differ from the predictable nature of human speech, further hindered SincNet's performance. While its innovative design makes it a robust tool for speech recognition, it proved less effective for marine mammal sound detection. The specialized demands of whale call identification highlight the need for models tailored to marine acoustics. Ultimately, the Enhanced CNN, with its ability to integrate diverse features and generalize across noisy datasets, emerged as the most effective model for this study.

Results

Convolutional autoencoder (CAE)

The performance of the Convolutional Autoencoder (CAE) was evaluated through spectrogram reconstruction and latent space visualization, demonstrating its effectiveness in compressing and reconstructing data while preserving critical information. The reconstructed spectrograms closely resembled the original inputs, with key patterns and features retained, although some fine-grained details were lost during compression.

Latent space representations, visualized using PCA, revealed compact and distinct clustering, indicating the CAE's ability to encode spectrograms into meaningful lower-dimensional features. These results highlight the CAE's capability to generate high-quality representations suitable for downstream tasks like classification and clustering.



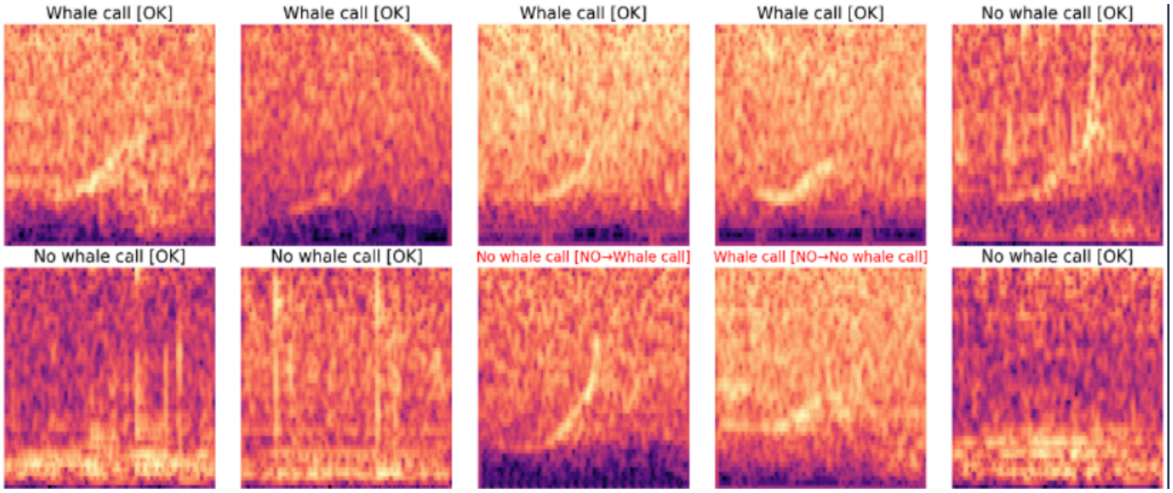
4.1 Results from the Autoencoder Reconstruction

Enhanced CNN

The best model, the Enhanced CNN, achieved significant improvements with a validation accuracy of 91.73% by the fifth epoch and an F1 score of 0.84, indicating a strong balance between precision and recall. With a precision of 0.75, the model correctly identified 75% of positive predictions, and with a recall of 0.79, it detected 79% of actual positive cases. These results underscore the model's ability to handle noisy, imbalanced datasets while leveraging diverse acoustic features for superior performance compared to the baseline.

Key factors contributing to this performance include batch normalization, which stabilized and accelerated training, and a deeper architecture with larger filters in later convolutional blocks to capture complex spatial patterns.

Despite its robustness, misclassifications occurred due to overlapping features and conflicting weights in one case, and potential manual labeling errors or insufficient data in another. These insights highlight opportunities to refine the model further to address feature overlap and data inconsistencies effectively.



4.2 Results from the Enhanced CNN Classification

Addressing Challenges in Whale Call Classification

1. **Class Imbalance**

   Class imbalance negatively impacted recall and F1 scores for whale calls, the minority class. To address this, SMOTE was applied to generate synthetic samples, balancing the dataset. Weighted loss functions in the CNN assigned higher penalties to misclassified whale calls, improving recall while maintaining precision.

2. **Noise and Feature Overlap**

   Noise and overlapping features, particularly in ZCR and Spectral Centroid, led to misclassification between whale calls and background noise due to shared characteristics in the dataset.

3. **Limitations Due to Decibel Levels and SincNet Performance**

   SincNet, optimized for human speech (300 Hz to 8 kHz), struggled to detect the low-frequency components of whale calls (10 Hz to 10 kHz). Additionally, low-decibel whale calls and the complexity of underwater noise hindered feature extraction, limiting SincNet's effectiveness.

Conclusion

In this study, we explored various approaches to improve whale call detection amidst environmental noise. Using resources from Babel Lab, which provided 5-hour whale call recordings with labeled call stamps, we faced computational limitations that restricted analysis to 2-3 recordings. Despite this, the recordings and accompanying environmental data (e.g., temperature and salinity) offered valuable insights for future enhancements. These data also revealed distinct sound types, suggesting potential for clustering different call structures in future studies.

Initial models like SincNet and Attention-Based CNN faced challenges, including feature overlap, frequency mismatches, and the complexity of underwater acoustics. SincNet struggled with the low-frequency nature of whale calls, while the Attention-Based CNN was limited by short audio lengths and low-decibel data.

The Enhanced CNN emerged as the most promising model, achieving a validation accuracy of 91.73% and an F1 score of 0.84. It effectively handled noisy, imbalanced data by leveraging both traditional acoustic features and advanced spectrogram-based analysis. Batch normalization and larger filters enabled the model to capture complex patterns, even in challenging underwater environments.

Future work could incorporate multimodal approaches, integrating environmental data like temperature and salinity, and explore LSTM and attention-based mechanisms to better understand sequential and low-frequency whale call patterns. These advancements offer promising directions for further improving whale call detection.

Citations

1.“The Marinexplore and Cornell University Whale Detection Challenge.” Kaggle,

https://www.kaggle.com/c/whale-detection-challenge/data

2.“AI decodes the calls of the wild.” Nature,

https://www.nature.com/immersive/d41586-024-04050-5/index.html. 10 Dec. 2024

3.“SMOTE for Imbalanced Classification with Python.” Machine Learning Mastery,

https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

4.“Convolutional Autoencoders.” Digital Ocean,

https://www.digitalocean.com/community/tutorials/convolutional-autoencoder

5.“MFCC’s Made Easy.” Medium,

https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040. 15 Jun. 2019