



Article

# Detection of Typical Transient Signals in Water by XGBoost Classifier Based on Shape Statistical Features: Application to the Call of Southern Right Whale

Zemin Zhou <sup>1</sup>, Yanrui Qu <sup>2,\*</sup> , Boqing Zhu <sup>1</sup>  and Bingbing Zhang <sup>1</sup>

<sup>1</sup> College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China; zzm@nudt.edu.cn (Z.Z.); zhuboqing@outlook.com (B.Z.); zbbzb@nudt.edu.cn (B.Z.)

<sup>2</sup> School of Marine Science and Technology, Tianjin University, Tianjin 300072, China

\* Correspondence: 3021005147@tju.edu.cn

**Abstract:** Whale sound is a typical transient signal. The escalating demands of ecological research and marine conservation necessitate advanced technologies for the automatic detection and classification of underwater acoustic signals. Traditional energy detection methods, which focus primarily on amplitude, often perform poorly in the non-Gaussian noise conditions typical of oceanic environments. This study introduces a classified-before-detect approach that overcomes the limitations of amplitude-focused techniques. We also address the challenges posed by deep learning models, such as high data labeling costs and extensive computational requirements. By extracting shape statistical features from audio and using the XGBoost classifier, our method not only outperforms the traditional convolutional neural network (CNN) method in accuracy but also reduces the dependence on labeled data, thus improving the detection efficiency. The integration of these features significantly enhances model performance, promoting the broader application of marine acoustic remote sensing technologies. This research contributes to the advancement of marine bioacoustic monitoring, offering a reliable, rapid, and training-efficient method suitable for practical deployment.

**Keywords:** classified-before-detect; shape statistical features; southern right whale; XGBoost



**Citation:** Zhou, Z.; Qu, Y.; Zhu, B.; Zhang, B. Detection of Typical Transient Signals in Water by XGBoost Classifier Based on Shape Statistical Features: Application to the Call of Southern Right Whale. *J. Mar. Sci. Eng.* **2024**, *12*, 1596. <https://doi.org/10.3390/jmse12091596>

Academic Editor: Xinqiang Chen

Received: 10 June 2024

Revised: 6 September 2024

Accepted: 7 September 2024

Published: 9 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Underwater environmental monitoring, biological detection, and communication all depend heavily on transient signals in aquatic settings that come from both natural occurrences and human activity. Researchers working to comprehend and safeguard marine ecology depend on these signals, which include undersea explosions, hull collisions, fish vocalizations, and changes in ship velocity. They are essential to military and navigation systems in addition to ecological monitoring.

Among these, whale calls represent a key subset of marine transient signals. Cetaceans, as apex predators and vital ecological components, significantly contribute to biodiversity and the balance of marine ecosystems [1]. However, the escalation of marine activities, such as increased shipping and expansive marine construction, poses unprecedented threats to their natural habitats. These activities introduce substantial noise pollution and can alter whale behaviors, affecting crucial aspects of their life cycles, including breeding, migration, and foraging [2]. Consequently, the real-time monitoring and precise identification of these underwater acoustic signals, particularly cetacean vocalizations [3], are essential for evaluating and mitigating the impacts of human activities on these key marine species.

Passive acoustic monitoring (PAM) has become a cornerstone in marine biological surveys and environmental assessments [4,5]. This system allows researchers to remotely and continuously monitor the sounds of whales and other marine organisms without disturbing them. However, the complex nature of the marine environment and the presence of high background noise [6] pose significant challenges for traditional acoustic data

analysis methods for detecting specific transient signals, such as whale sounds. While the transient characteristics of these sounds—being short, non-periodic, and broadband—offer potential for signal classification, they also substantially increase the analytical complexity.

Given these difficulties, this paper presents a novel approach that uses shape statistical features to detect typical transitory signals in aquatic environments. These features, including standard deviation, take important information out of the unpredictability and distribution of sound waves, making it possible to identify individual passing occurrences in intricate ocean soundscapes—such as distinct whale calls—with accuracy. This work thoroughly analyzes the performance of these shape statistical features in different signal-to-noise ratio conditions with the goal of improving the automation and accuracy of sound signal processing models. The intention is to monitor maritime noises, including whale sounds, more efficiently.

The practical applications of our approach are numerous: It aids in marine biodiversity conservation by accurately identifying whale calls and other marine mammal vocalizations, which is essential for monitoring the health and behavior of these species. It also provides a reliable tool for assessing the impact of noise pollution from shipping and construction on marine life, enabling the development of strategies to mitigate these effects. In addition, accurate detection of underwater signals is crucial for naval operations and safe navigation, helping to avoid collisions and identify potential threats. Furthermore, our approach enhances the ability to monitor various underwater events, contributing to a better understanding of the marine environment and its changes over time.

The article is organized as follows: The relevant literature is compiled in the second section, while the theoretical approaches and statistical features are explained in the third section. The experimental data and the preparation techniques are presented in the fourth section. The experimental findings and analysis are presented in the fifth part. The study's primary conclusions are finally outlined in the sixth part, which also discusses the study's shortcomings and potential future research areas.

## 2. Related Work

### 2.1. Challenges of Traditional Methods

In traditional marine bioacoustic research, manual recognition and basic automatic detection techniques, such as energy detection and spectral feature matching, are effective for small-scale datasets and environments with low noise levels. However, as the volume of data expands and environmental complexity increases, these methods encounter significant challenges [7]. Particularly in scenarios requiring large-scale data processing and in complex acoustic environments, the limitations of these traditional approaches become increasingly apparent, often leading to low processing efficiency and high error rates.

The energy detection method relies on sound intensity and is relatively effective in environments where background noise is stable and predictable. However, in a dynamic marine setting, factors such as ship activity and wave impact introduce unstable noise, significantly increasing the false alarm rate of energy detection methods. Moreover, due to this method's sensitivity to environmental noise, it struggles to accurately identify target sounds in conditions of high sound variability.

The spectral feature matching method relies on a predefined sound feature library, which often becomes a bottleneck when analyzing the diverse sounds of marine life. This method struggles to maintain high accuracy for species with constantly changing sound characteristics or those that have not been extensively studied. The inability of traditional spectral feature matching methods to accommodate sound diversity renders them inefficient in practical applications.

### 2.2. Innovations and Challenges in Deep Learning Models

The detection and categorization of marine biological sounds has been greatly improved in recent years by the use of deep learning technologies in acoustic signal processing [8], especially convolutional neural networks (CNN) and recurrent neural networks

(RNN). Based on signal properties, the “classification-before-detection” approach has become more and more popular. With their precise identification and detection of marine organism noises, these technologies prove to be excellent in the real marine environment. High degrees of automation and reliable performance are displayed by them, especially when handling big information.

The deep learning model can extract multi-dimensional features from audio signals [9,10], enhancing the accuracy and robustness of sound recognition through techniques such as data augmentation and transfer learning. For instance, the two-stream convolutional network model (TSCA) integrates signal processing in both time and frequency domains [11], effectively reduces background noise through the use of an attention mechanism, and significantly speeds up the detection and recognition of marine biological sounds. Additionally, in [12], the ORCA-SPOT toolkit demonstrated remarkable generalization capabilities and effectiveness in processing large-scale datasets, particularly in detecting killer whale sounds.

On the other hand, the classification model with a multi-channel parallel structure [13] achieves highly accurate sound classification by integrating audio features from various channels and employing a trainable, fully connected layer to fuse these features. This model consistently demonstrates high accuracy across multiple experiments, achieving an average accuracy of 95.21% with a standard deviation of 0.65%. These results underscore the advancement and effectiveness of deep learning technology in acoustic signal processing.

Furthermore, using convolutional neural networks to automatically recognize humpback whale songs in large passive acoustic datasets [14], deep learning’s tremendous potential for spectrogram categorization and audio event identification is revealed. The use of an active learning technique in this study allows the model to more correctly correct the coverage error of the initial batch labels, resulting in improved performance through iterative improvement of the model configuration.

Despite this progress, traditional methods for detecting marine biological sounds, such as pattern recognition techniques based on specific sound features, still require a large amount of manually labeled data and lack flexibility when dealing with unknown or mixed signals. Commonly employed sound recognition technologies, including sound spectrum analysis, energy detectors, and monitoring of specific frequency bands, often perform poorly with real marine environmental data.

Although deep learning models theoretically offer superior performance [15], they encounter several challenges in practical applications. These include the high costs associated with data labeling, demands for processing large-scale data, and maintaining stability in extreme noise conditions. Additionally, these models typically require substantial computing resources, which can limit their deployment in real-time monitoring and scenarios where computing resources are constrained.

### *2.3. New Methodology for Enhanced Detection*

To address these problems, this study presents a novel methodology that combines traditional acoustic feature analysis with modern machine learning techniques to improve the accuracy and efficiency of marine biological sound detection. Our method focuses on efficiently using basic shape statistical aspects to promote adaptability to complex marine sound environments while keeping the model simple. This work elucidates the significance of each feature to marine biological sound detection by testing the model’s performance under various signal-to-noise ratio settings. It also provides novel design insights for future acoustic monitoring technology.

The objectives of this paper are as follows: Firstly, the audio is preprocessed for noise reduction, and basic sound features are extracted. These features are then used to detect Southern right whale (SRW) calls. The XGBoost classifier model is employed to determine the weight of each feature in the automatic detection process, and simulated audio is utilized to assess the performance of each feature under different signal-to-noise ratios. We will thoroughly detail the model design, data processing methods, and experimental

results, and discuss the challenges and prospects of this approach. Additionally, we will explore the potential applications of the proposed methods in marine biological protection and biodiversity monitoring. Figure 1 illustrates the workflow for whale sound detection and classification using XGBoost (Version: 2.0.3).

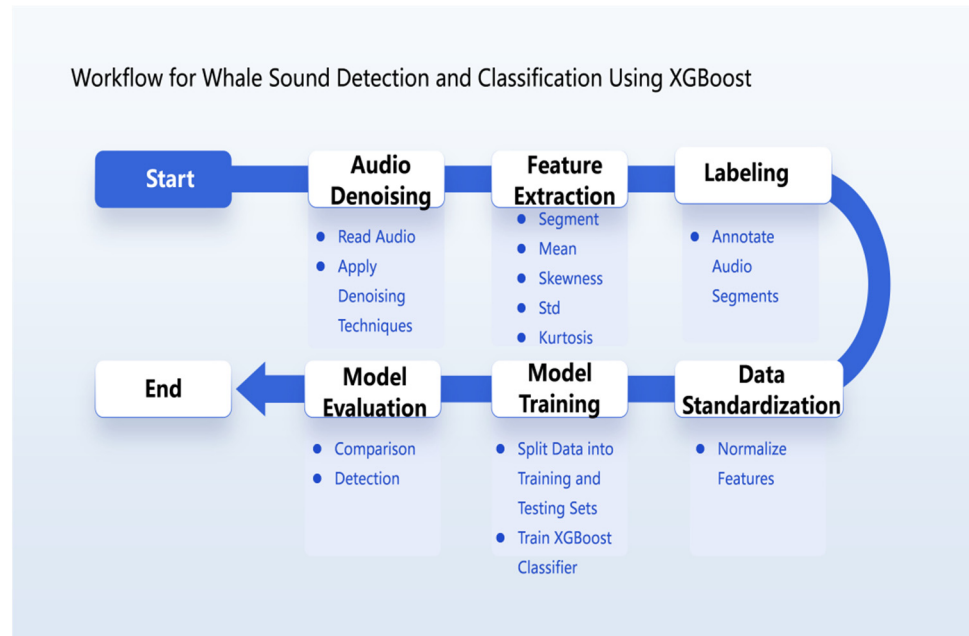


Figure 1. The workflow for whale sound detection and classification using XGBoost.

### 3. Methodology

#### 3.1. Pretreatment

Five popular denoising methods are used to preprocess the audio, which are spectral subtraction, Wiener filtering, EMD, EEMD, and VMD. The brief introduction is as follows:

##### 3.1.1. Spectral Subtraction with Maximum Noise Residue Smoothing

Spectral subtraction is a widely used technique for noise reduction in speech signals, functioning by [16–18] manipulating the signal’s spectrum. This method fundamentally estimates noise within silent segments, typically calculated by averaging multiple silent segments. Enhanced by maximum noise residue smoothing, this approach further improves the denoising effect.

The mathematical expressions for noise estimation (Equation (1)) and spectral subtraction (Equation (2)) are as follows:

$$N_{est}(f, t) = \alpha \cdot \text{median}(X(f, t)), \tag{1}$$

$$Y(f, t) = \max(X(f, t) - N_{est}(f, t), \epsilon). \tag{2}$$

Here,  $(X(f, t))$  represents the signal spectrum at frequency  $f$  and time  $t$ ,  $N_{est}(f, t)$  is the estimated noise spectrum, and  $\alpha$  and  $\epsilon$  are adjustment parameters and a small positive number to ensure operational stability. This method effectively reduces residual noise and enhances the clarity of the speech signal.

Our approach consists of performing an STFT with a window length of 1024 and a skip length of 512 on the noisy audio to extract frequency domain information. The noise spectrum was estimated using the first 5 min of audio segments in which no whale calls occurred, which are assumed to contain predominantly background noise, and are then used to calculate the average power spectrum for noise estimation. The main parameters include alpha ( $\alpha$ ), set to 4 for aggressive noise reduction; beta ( $\beta$ ), set to 0.02 to ensure

signal distortion is minimized; and smoothing coefficient ( $k$ ), set to 2 for averaging the power spectrum and reducing artefacts. The residual threshold is 0.1 to keep the processed sound natural. The enhanced amplitude spectrum is recombined with the original phase, and after inverse STFT for time domain reconstruction, the output is saved.

### 3.1.2. Wiener Filtering

Wiener filtering, a classical signal denoising method based on statistical properties, designs filters by minimizing the expected mean square error. For theoretical foundations and practical applications of Wiener filtering, readers may consult references [19,20]. The frequency domain expression (Equation (3)) of the Wiener filter is:

$$H(f, t) = \frac{S_x(f, t)}{S_x(f, t) + S_n(f, t)}. \quad (3)$$

Here,  $S_x(f, t)$  and  $S_n(f, t)$  represent the power spectral densities of the signal and noise at frequency  $f$  and time  $t$ , respectively.

For the short-time Fourier transform (STFT), we chose a window size of 1024 and a hop length of 512, which are critical for balancing the frequency resolution and temporal resolution of our signal analysis. The noise spectrum was estimated using the first 5 min of audio segments, in which no whale calls occurred, which are assumed to contain predominantly background noise. This estimation helps to set a baseline for the noise level of the entire audio file.

### 3.1.3. Empirical Mode Decomposition

Compared with spectral subtraction and Wiener filtering, the modal decomposition denoising method has the advantages of stronger adaptability, a wide application range, and retaining signal characteristics, so it may be a better choice in some cases. The modal decomposition method can decompose the signal adaptively according to the characteristics of the signal and can selectively remove some modes to achieve denoising. This adaptability can better adapt to different types and complexity of signals. The modal decomposition method is not limited by the stability of the signal and can deal with non-stationary signals, such as speech signals, music signals, etc. In contrast, spectral subtraction and Wiener filtering have higher requirements for signal stability. The modal decomposition method can usually retain the important features and structural information of the signal while removing the noise and avoiding the signal distortion problem that may be introduced by the spectral subtraction and Wiener filtering methods. It is usually not necessary to accurately estimate the statistical characteristics of the signal and noise, so it is less dependent on prior knowledge and easier to apply in practical scenarios. In addition, the parameters of the modal decomposition method are usually lower and easier to adjust. In contrast, the spectral subtraction and Wiener filtering methods may involve more parameter selection and tuning, and more experience and experiments are needed to adjust the parameters.

Empirical mode decomposition (EMD) is a method for analyzing nonlinear and non-stationary time-series data [21,22]. It reveals the essential characteristics and trends in the data by decomposing the data into a series of intrinsic mode functions (IMFs) and a residual term. EMD denoising is the process of using this decomposition technique to reduce the noise of the signal. The original signal is decomposed into multiple IMFs and a residual. Each IMF should satisfy two conditions: in the entire dataset, the number of extreme points and the number of zero-crossing points must be equal or, at most, one; at any point, the mean value of the upper envelope defined by the local maximum and the lower envelope defined by the local minimum is zero. The IMFs that are noise components are then identified. In general, high-frequency IMFs (usually the first few IMFs) are considered to contain more noise. Threshold processing is performed on IMFs identified as noise to reduce noise. This can be achieved by the soft threshold or hard threshold method. The



processed IMFs and the unprocessed residuals are used to reconstruct the signal to obtain the denoised signal. The definition of IMF (Equation (4)) is:

$$\text{mean}(\text{envelope}_{\max}(x(t)) + \text{envelope}_{\min}(x(t))) = 0, \tag{4}$$

where  $\text{envelope}_{\max}(x(t))$  and  $\text{envelope}_{\min}(x(t))$  are the upper and lower envelopes composed of local maximum and local minimum points, respectively. The signal reconstruction formula (Equation (5)) is:

$$x_{\text{reconstructed}}(t) = \sum_{i=1}^N \text{IMF}_i(t) + r(t), \tag{5}$$

$N$  is the number of IMFs, and  $r(t)$  is the final residual.

In our study, EMD adaptively decomposes the raw data into 10 different IMFs. This decomposition is achieved by iteratively separating the intrinsic characteristics of the oscillatory modes directly from the data by EMD without the need for predefined basis functions. By analyzing the frequency content and energy distribution of each IMF, we identify three specific IMFs that are essential for capturing the fundamental characteristics of the original signal while minimizing noise. We then reconstruct the signal using these selected IMFs, utilizing their combined information content to maintain the integrity and fidelity of the original audio while effectively reducing background noise.

### 3.1.4. Ensemble Empirical Mode Decomposition

Ensemble empirical mode decomposition (EEMD) is an improved version of EMD. EEMD adopts the idea of set average. By adding random noise to the original signal and performing EMD decomposition multiple times, the random error is finally reduced by averaging the results of each decomposition. Refs. [23–25] improve the stability and reliability of decomposition. Given a signal with noise, the signal is decomposed by EEMD many times, and different random noise is added to each decomposition. The average IMF is obtained by averaging the IMF obtained by each decomposition. The average IMF is combined into a denoised signal. The formula of EEMD is similar to that of EMD, and the specific decomposition process (Equation (6)) can be expressed by the following basic formulas:

$$c^{(k)}(t) = x(t) + \varepsilon^{(k)}(t) - \sum_{i=1}^n h_i^{(k)}(t). \tag{6}$$

Here,  $c^{(k)}(t)$  is the residual signal obtained by the  $k$ th decomposition,  $x(t)$  is the original signal,  $\varepsilon^{(k)}(t)$  is the random noise added by the  $k$ th decomposition,  $\sum_{i=1}^n h_i^{(k)}(t)$  is the  $i$ th IMF obtained by the  $k$ th decomposition, and  $n$  is the number of IMFs. This process is repeated many times. The average IMF is obtained by averaging the IMF obtained by each decomposition. Then, the average IMF is added to obtain the denoised signal (Equation (7)) as:

$$x_{\text{denoised}}(t) = \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^m h_i^{(k)}(t), \tag{7}$$

where  $K$  is the number of decompositions and  $m$  is the number of IMFs. The selection of the appropriate IMF and the number of decompositions are usually determined based on the energy distribution of the signal and noise in the IMF.

In our study, the EEMD was employed to adaptively decompose the original data into seven distinct IMFs. We utilized the EEMD algorithm with specific parameters, setting the number of ensemble members to 100 and the ensemble noise level at 0.1 of the data standard deviation. These parameters were carefully chosen to enhance the robustness of the decomposition while ensuring that each IMF reliably represented inherent signal characteristics without being overly influenced by the added noise. By analyzing the frequency content and energy distribution of each IMF, we identified three specific IMFs that were critical for signal reconstruction. These selected IMFs contained the essential

characteristics of the original signal, effectively capturing its core dynamics while minimizing noise components. This selective reconstruction approach allows us to preserve the integrity and fidelity of the signal, highlighting the efficacy of the EEMD in handling complex signal environments.

### 3.1.5. Variational Mode Decomposition

Variational mode decomposition (VMD) solves the process of signal decomposition by optimizing the problem. Refs. [26,27] decomposed the signal into multiple modes, with each mode representing a frequency band of the signal. The bandwidth of each mode is determined by constraints or parameters in the optimization process. Given a signal with noise, the appropriate parameters are selected, including the number of modes and constraints. By solving the optimization problem of VMD, the signal is decomposed into multiple modes. According to the energy distribution of the signal and noise, the appropriate mode is selected for reconstruction to obtain the denoised signal. The optimization problem of VMD (Equation (8)) can be expressed in the following basic forms:

$$\min_{\{u_j, \omega_j\}_{j=1}^K} \sum_{j=1}^K |u_j|_2^2 + \lambda |\nabla T u_j - \omega_j|_2^2, \tag{8}$$

where  $u_j$  is the  $j$ th mode,  $\omega_j$  is the frequency parameter,  $K$  is the number of modes,  $\lambda$  is the equilibrium parameter, and  $\nabla T$  is the gradient of operator  $T$ . When selecting the appropriate mode for reconstruction, the denoised signal (Equation (9)) can be obtained by adding the selected modes, that is:

$$x_{\text{denoised}}(t) = \sum_{i=1}^m u_i(t), \tag{9}$$

where  $m$  is the number of selected modes. The selection of which modes is usually based on the energy distribution of the signal and noise in the mode.

In our study, VMD was configured to decompose the original data into 5 IMFs. These intrinsic mode functions were chosen to balance complexity and computational efficiency, and the decomposition process allowed for a maximum of 300 iterations to ensure complete convergence and stability of the results. Through our analysis, we selected two IMFs for reconstructing the signal.

In this study, we selected variational mode decomposition (VMD) for the processing of our datasets. VMD offers superior control over the frequency range of each mode by adjusting the bandwidth parameters, thus providing better frequency resolution. In contrast, empirical mode decomposition (EMD) and ensemble EMD (EEMD) have weaker control over frequency resolution and may experience modal aliasing. VMD exhibits strong robustness against signal changes and noise, enabling more stable decomposition of nonlinear and non-stationary signals. Conversely, EMD and EEMD may result in more unstable outcomes when dealing with noisy and nonlinear signals. The bandwidth parameters in VMD allow for tailored signal decomposition to fit specific application scenarios and requirements, ensuring more satisfactory results. Unlike methods that rely on accurate statistical signal characteristics estimation, such as Wiener filtering and spectral subtraction, VMD simplifies the process without requiring detailed tuning of filter parameters, making it more user-friendly in practical applications. Figure 2 shows the signal-to-noise ratio (SNR) results after applying five different noise reduction methods to the processing. These data come from a one-hour-long live audio selected from dataset 1, which recorded a whale call and its background noise in a natural environment. Each part of the box plot represents different statistical characteristics of the data distribution. The signal-to-noise ratio (SNR) is calculated by first identifying the audio segments that contain the signal and distinguishing these from the background noise. Then, the average power of both the signal and the noise segments is calculated. Finally, the SNR is determined by

comparing these power levels to assess how much clearer the signal is in relation to the noise. The line in the middle of the box indicates the median, while the bottom and top of the box denote the lower and upper quartiles, respectively. The height of the box represents the middle 50% range of the data (interquartile range, IQR). Extensions beyond the box indicate maximum and minimum values, usually capped at 1.5 times the IQR, with outliers marked separately on the plot. "SSS" in the figure refers to smoothed spectral subtraction.

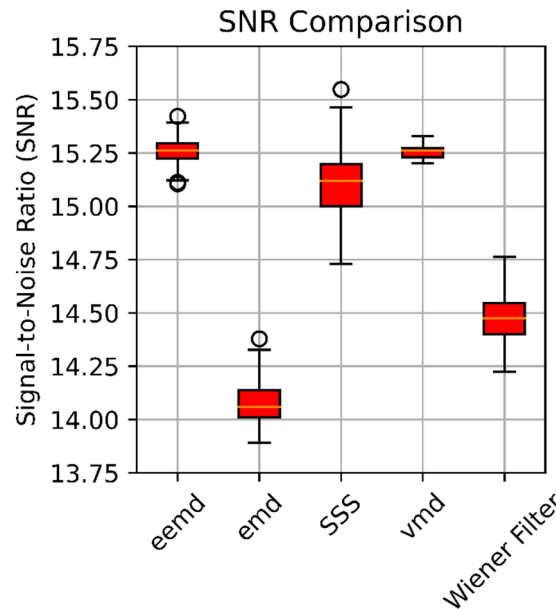


Figure 2. SNR comparison among different noise reduction methods.

### 3.2. Shape Statistical Features

These features provide the statistical properties of the spectrum, including the mean, standard deviation (Std), skewness, and kurtosis. It focuses on extracting the statistical features of the original audio directly [28]. The statistical [29] calculation of the shape includes:

In audio analysis, the mean (Equation (10)) indicates the segment’s average amplitude level, which reflects the overall intensity of the audio signal, whereas the standard deviation (Equation (11)) measures the dispersion of amplitude values around the mean, with a higher standard deviation suggesting more amplitude variance. Skewness (Equation (12)) is a statistical measure of the asymmetry of a probability distribution, with positive skewness indicating higher amplitude values and negative skewness indicating lower amplitude values in the segment. Kurtosis (Equation (13)) describes the sharpness of the amplitude distribution; high kurtosis indicates more extreme amplitude values, whereas low kurtosis indicates a flatter distribution. These metrics work together to help assess and understand the features of audio data.

The calculation formula of these features is:

$$\text{Mean} : \mu = \frac{1}{N} \sum_{i=1}^N x_i, \tag{10}$$

$$\text{Std} : \sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2}, \tag{11}$$

$$\text{Skewness} = \frac{1}{N} \sum_{i=1}^N \left( \frac{x_i - \mu}{\sigma} \right)^3, \tag{12}$$



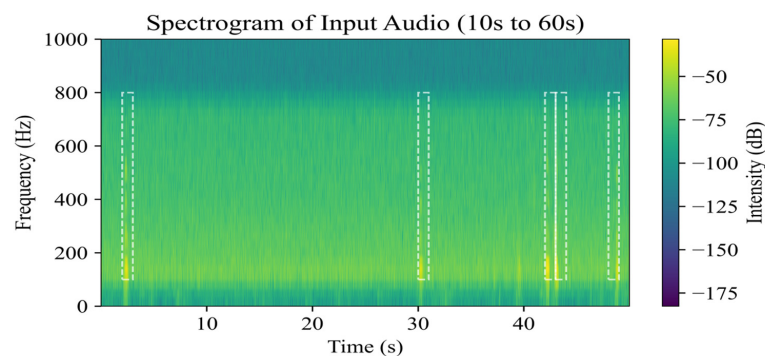
$$\text{Kurtosis} = \frac{1}{N} \sum_{i=1}^N \left( \frac{x_i - \mu}{\sigma} \right)^4 - 3. \tag{13}$$

Here,  $\mu$  denotes the mean, which is the average value of the dataset.  $N$  represents the total number of observations or data points in the dataset.  $x_i$  stands for the individual data points in the dataset.  $\sigma$  symbolizes the standard deviation, a measure of the dispersion or spread of the data points around the mean.

When we merely want to know the overall statistical properties of the audio or reduce the processing flow, extracting the statistical features of the original audio is a more direct and easy way. This method is especially useful for situations that do not require extensive time-frequency information and only need to grasp the overall features of the audio.

By directly extracting the shape statistical features of the original audio, we can quickly obtain the general information of the entire audio segment, such as the mean. These features can help us understand the overall distribution of the audio, the shape characteristics of the waveform, etc., which is of great significance for quickly analyzing the characteristics of the audio and comparing the differences between different audio [30].

First, the audio data are normalized so that the maximum absolute value is 1. For each audio segment, the following features are calculated: (mean) (Std) (skewness) (kurtosis). Then, the values of each feature at each time frame are considered separately. If the value exceeds a certain threshold, the signal frame is classified as the target sound. Figure 3 is an example of the detection results of dataset 1. The white box indicates that the time frame containing the call of the whale has been identified.



**Figure 3.** Detection results of dataset 1.

### 3.3. XGBoost Model

Extreme gradient boosting (XGBoost) is an optimized distributed gradient boosting library that implements gradient boosting algorithms [31,32]. It is a tree-based model that iteratively trains multiple decision tree models and combines them to improve prediction performance. XGBoost uses the gradient boosting algorithm as the basis to continuously improve the performance of the previous model by training a new model in each iteration. In each round, the new model will try to fit the residuals of the previous model. The basic learner in XGBoost is the decision tree. Typically, classification and regression trees (CARTs) are used as the base decision tree, but other types of tree models are also supported. Each decision tree is gradually constructed by continuously splitting nodes to minimize the loss function. A regularization term is introduced to control the complexity of the model and prevent overfitting. This includes pruning the leaf nodes of the tree and limiting the node splitting to avoid overly complex models. XGBoost can provide the importance score of each feature in the model. These scores [33] can help understand how the model predicts and help with feature selection and model interpretation. For the classification problem, XGBoost outputs the probability of the category.

XGBoost is an independent library that implements the gradient boosting algorithm. It optimizes calculation speed and memory usage and provides multiple language interfaces, including Python. In this study, four statistical features were extracted from audio data and

used as feature vectors for input to the XGBoost model. After each feature is calculated, they are aggregated into a feature vector and added to the feature list, which is then converted into a vector with the label instrument so that it can be directly used for training and testing of machine learning models. In order to improve the training efficiency and performance of the model, the feature vectors are standardized. Next, an instance of the XGBoost classifier is created. During the training process, logloss is used as an indicator to measure the difference between the predicted probability and the actual label. In the experiment, the value of logloss is 0.071, which is a relatively low value, indicating that on average, the model provides a high probability for correct labels. During the training phase, the model is trained using both features and corresponding labels. Once training concludes, the test set is processed, yielding probabilities that represent the likelihood of the positive category—in this case, whale calls. These probabilities serve as the basis for evaluating the model’s performance and for comparative analysis. Additionally, by examining feature importance, we gain insights into the relative contributions of individual features to the model’s predictions. Figure 4 shows the flow chart describing the XGBoost optimization process [34]. The direction of the flowchart is from top to bottom. Each step is represented by a box with corresponding labels describing the content of the step. Arrows between nodes indicate the order of execution. We extracted four statistical features from the audio data—mean, standard deviation, skewness, and kurtosis—and normalized them to improve model performance. First, we defined the objective function, including the mean square error and the regularization term of the model complexity. Then, the model parameters were initialized and iterative training starts. In each iteration, the gradient of the objective function was calculated and the model parameters were updated using gradient descent. We also applied regularization techniques to control the complexity of the model and prevent overfitting. By monitoring the convergence criterion, when the convergence condition was satisfied, the iteration stopped and the final model was output.

Here is an explanation of the symbols used in the formulas within the flowchart:

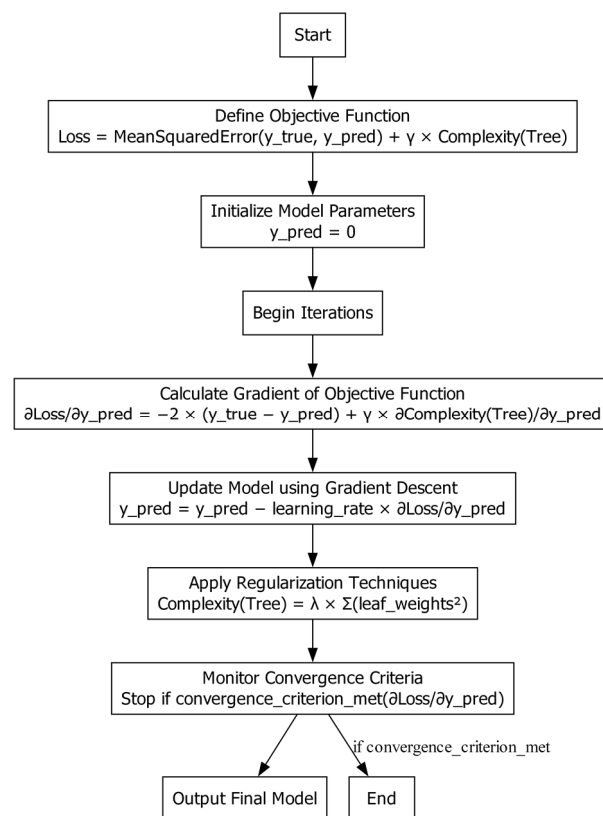


Figure 4. XGBoost optimization process.

**Loss:** Represents the objective function, which combines the mean squared error of the predictions and a regularization term to control model complexity.

$y_{true}$ : The true label or actual value in the dataset.

$y_{pred}$ : The predicted value generated by the model.

$\gamma$ : A regularization parameter that controls the weight of the complexity term in the objective function.

**Complexity (Tree):** Represents the complexity of the decision tree model, which includes terms to prevent overfitting.

$\frac{\partial \text{Loss}}{\partial y_{pred}}$ : The gradient of the loss function with respect to the predicted values, indicating the direction and rate of change needed to minimize the loss.

**learning\_rate:** A parameter that controls the step size during the gradient descent update process.

$\lambda$ : A regularization parameter used to control the impact of the regularization term on the model complexity.

$\sum(\text{leaf\_weights}^2)$ : The sum of the squared weights of the leaf nodes in the decision tree, used to measure the complexity of the tree and prevent overfitting.

In our study, we chose XGBoost instead of a deep neural network. This superiority can be attributed to several factors. Firstly, XGBoost is inherently robust to uninformative features due to its built-in feature selection mechanism, which iteratively adds the most informative features and ignores the irrelevant ones. In contrast, deep neural networks are highly sensitive to all input features, including uninformative ones, which can lead to overfitting and reduced generalization performance. Secondly, XGBoost demonstrates greater robustness to low signal-to-noise ratio (SNR) data. The boosting process in XGBoost effectively minimizes the impact of noise by combining multiple weak classifiers to form a strong classifier, whereas deep neural networks tend to overfit noise in the data, despite the use of regularization and data augmentation techniques. Lastly, deep neural networks often produce overly smooth solutions due to their reliance on continuous activation functions, making them less capable of capturing complex, non-linear patterns inherent in shape statistical features. XGBoost, on the other hand, constructs more flexible decision boundaries through its ensemble of decision trees, enabling it to better fit complex data distributions without over-smoothing. These advantages make XGBoost a more suitable choice for modeling with shape statistical features in our experiments.

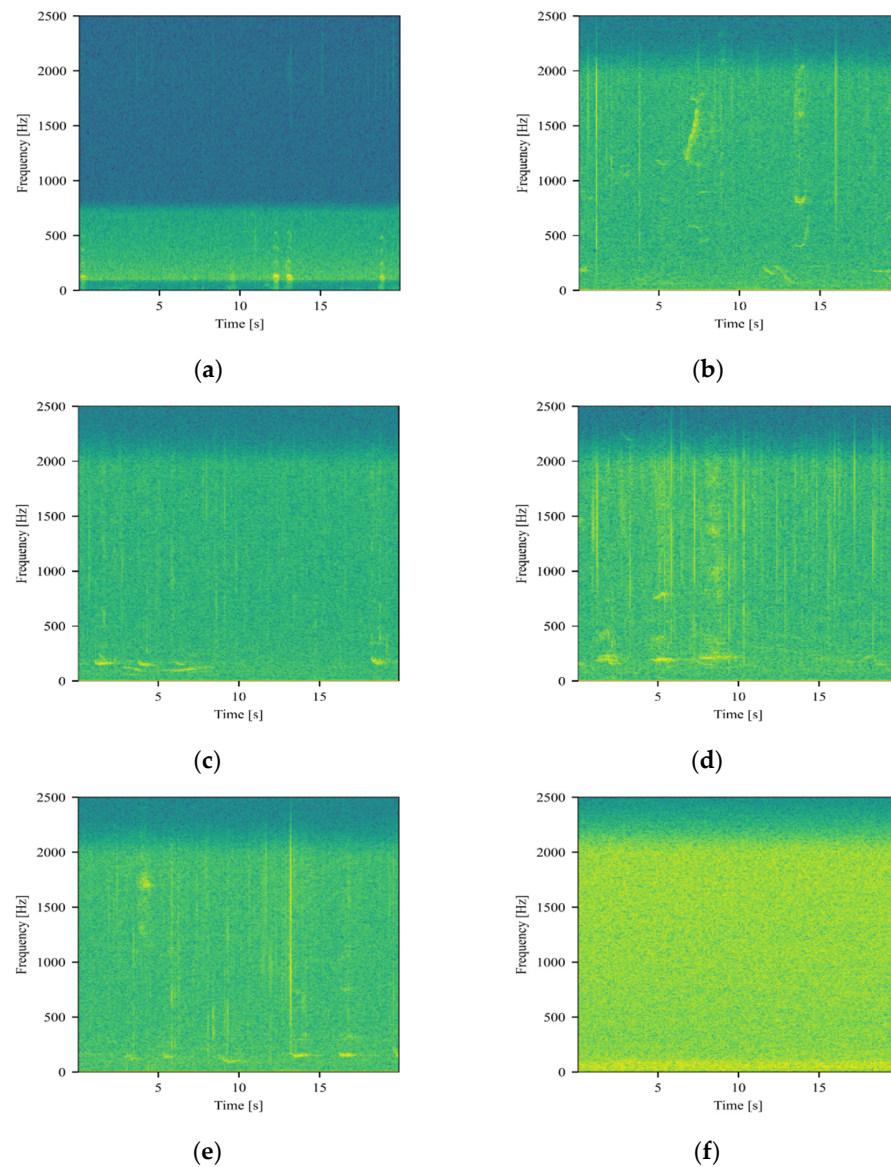
## 4. Experiment

### 4.1. Data

The dataset in this paper consists of two parts. Both contain the sound of whales. Dataset 1 is audio data captured by a hydrophone off the coast of Falls Bay, South Africa. Dataset 2 is simulated audio that uses whale calls to embed background noise in the marine environment. The following describes them in detail.

Dataset 1 consists of two audio segments, one lasting 45 min and the other 60 min. These segments contain approximately 300 Southern right whale (SRW) vocalizations. Southern right whales can emit a variety of types of sounds, including but not limited to low-frequency calls; "up-calls", a rising-tone sound often used in social and foraging settings; pulse sound; and continuous sound. The SRW's call was captured by a drifting buoy 9 m below the sea surface at 96 kHz. The drifting buoy was placed for two to four hours in the summer and then recovered. The Southern right whale's data are downsampled to 8000 Hz, and most calls show good signal-to-noise ratios (between 15 and 20 dB). The duration of the call is between 250 and 700 milliseconds. It also includes recording the time when the manually marked whale call appears in the audio. The first 45 min segment is used for training the model, and the subsequent 60 min segment is used for testing. Both segments include manually marked timestamps indicating the presence of whale calls, providing ground truth for training and evaluation.

Dataset 2 consists of hydrophone recordings sourced from eight different NOAA NCEI stations, capturing diverse marine environmental background noises and vocalizations from various marine mammals. The above stations are part of a series of passive acoustic data monitoring stations deployed by the National Oceanic and Atmospheric Administration's (NOAA) Pacific Environment Laboratory (PMEL). These stations cover the data collection period from 2014 to 2021 and are located along the coast of North America, in the Caribbean Sea, and in the North Pacific Ocean. From these recordings, we specifically selected 50 segments containing whale calls with high signal-to-noise ratios and 150 segments without whale calls, each 5 s in duration. Figure 5 contains example calls from the datasets.



**Figure 5.** Spectrogram images: (a) spectrogram image of dataset 1; (b) spectrogram image of dataset 2 (deployment name: NRS\_01\_2020-2022); (c) spectrogram image of dataset 2 (deployment name: NRS\_01\_2014-2015); (d) spectrogram image of dataset 2 (deployment name: NRS\_08\_2016-2018); (e) spectrogram image of dataset 2 (deployment name: NRS\_04\_2015-2016); (f) spectrogram image of dataset 2 (deployment name: NRS\_03\_2017-2019).

#### 4.2. Detection Methods

This section describes the bioacoustic recording and its application in detection experiments. The performance of the shape statistical detection method is compared with the

other three popular detection methods to evaluate their efficiency in identifying bioacoustic signals from stable and broadband noise.

Convolutional neural network (CNN) can automatically extract important features from audio data, which is very suitable for processing data with highly spatio-temporal structural features, such as audio and images. It is suitable for complex audio classification tasks that require automatic feature extraction. The input to our CNN comprises spectrograms derived from audio recordings. Our CNN architecture includes three convolutional layers that help in emphasizing different aspects of the input features, each followed by max-pooling layers to reduce spatial dimensions while preserving essential features. The network continues with two fully connected layers and incorporates dropout regularization to mitigate overfitting. The final layer uses a sigmoid activation function to classify the presence of bioacoustic events. The model is trained using a dataset of labeled audio segments, ensuring broad representation under various conditions. We utilized the Adam optimizer and a learning rate determined by validation set performance, conducting training over 10 epochs with batch sizes of 32. Similar architectures have been applied to tasks like environmental sound classification [35,36] and audio event detection.

Mel frequency cepstral coefficients (MFCCs) are a feature commonly used in speech and audio processing. They can capture the main energy distribution of audio and are especially suitable for tasks such as speech recognition and music classification. MFCCs can effectively characterize the characteristics of audio. MFCCs are used as the audio feature and are combined with logistic regression model for classification. Logistic regression is a simple and effective classification algorithm. For this study, we computed the first 13 MFCCs from each audio frame, using these coefficients as features for logistic regression—a straightforward yet powerful classification method. The logistic model was trained on a robust dataset of annotated audio recordings, applying feature scaling and L2 regularization to optimize performance. This combination has been extensively used in speech and audio processing [37,38], proving effective in tasks like speech recognition and music classification.

The band-limited energy detection (BLED) obtains the input record and calculates the energy within a specific frequency bandwidth, and then compares the energy with the detection threshold to determine whether it contains the signal of interest. It successfully detects potential vocalizations quickly in the presence of weak background noise, and is often used as a detection benchmark.

For the shape statistical features, the XGBoost classifier is used to combine the four features for model training. In order to optimize the performance of the model, the parameters are tuned by GridSearchCV. Use the best model trained before to predict it.

For each method, a time frame of 1 s is used, with a 50% overlap. The detection results of the four methods are compared in the case of a training set of 30 min and 45 min.

#### 4.3. Evaluation

In this study, we utilized a refined evaluation system to quantify the detection performance of whale calls. We calculated true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs), based on the comparison between the automatically detected events and manually labeled time periods. This comparison enabled us to assess whether each time frame was correctly identified as containing whale calls.

The metrics used for evaluating the model's performance—precision (Equation (14)), recall (Equation (15)), and the F1 score (Equation (16))—are commonly employed in machine learning to measure the accuracy of classification models. The formulas used are:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

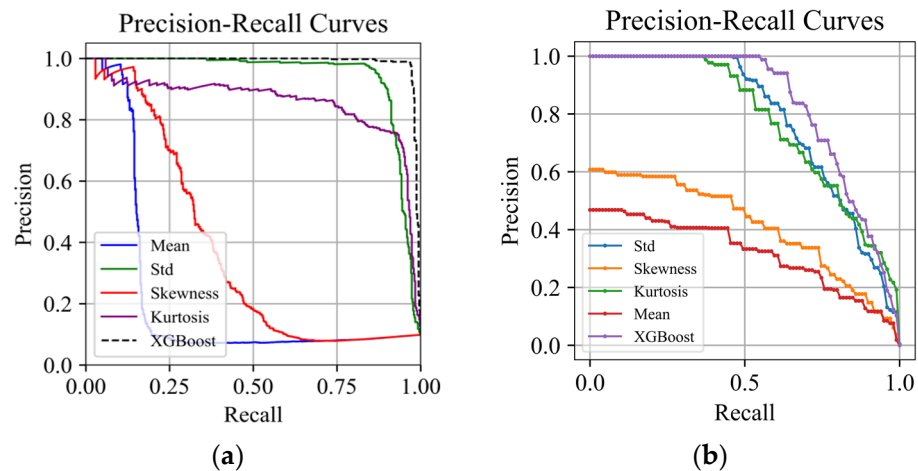


$$F1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

## 5. Results

### 5.1. Detection Results of Two Datasets

In Figure 6, the curves of the XGBoost model on dataset 1 consistently maintains a high level of precision, indicating that the XGBoost model has good stability and generalization ability on dataset 1. In comparison to dataset 1, the curves of the XGBoost model on dataset 2 still maintain a high level of precision when the recall rate is low, but as the recall rate increases, the precision decreases significantly.



**Figure 6.** Precision–recall curves: (a) four features and the XGBoost model of dataset 1; (b) XGBoost model of dataset 2.

The curves of the mean and skewness models for both datasets reveal poorer precision, particularly in the region with higher recall. The Std model’s curves are flatter, indicating that its accuracy is generally constant over varying recall rates; on dataset 2, the curves of the Std, skewness, and kurtosis models show a similar performance trend, indicating relatively low accuracy. The steeper slopes of these models on dataset 2 compared to dataset 1 suggest that these models had higher difficulty when dealing with the dataset 2 with various maritime ambient sounds.

### 5.2. Test Results of Four Methods

The detection results of the data with a training set of 45 min are shown in Figure 7. In our study, we systematically divided the audio data from dataset 1, which consists of two audio recordings, into training and testing sets to ensure robust model evaluation. Specifically, the first audio segment was used for training. The entire second audio segment, which lasts for one hour, was then utilized as the testing dataset.

It can be seen from the figure that in the case of fewer data in the training set, the detection results using the four shape features still obtain F1 scores of more than 90%. The performance of the deep learning convolutional network and shape statistical features is better, while the performance of band-limited energy detection is slightly worse.



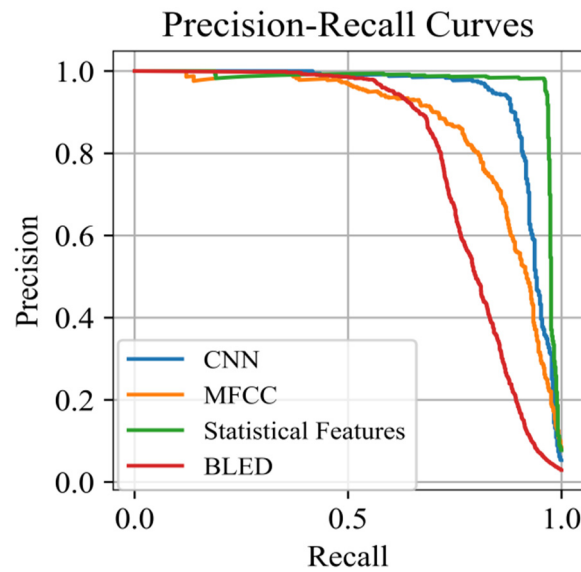


Figure 7. Precision–recall curves: the training data: 45 min.

5.3. Detection Results of Two Features

The five noise reduction methods were applied to dataset 1, and then the Std and kurtosis features were used. The detection results of dataset 1 are shown in Figure 8. In the figure, “SSS” refers to smoothed spectral subtraction. In our study, we analyzed the effectiveness of individual statistical features, specifically standard deviation, in distinguishing bioacoustic signals from background noise. For each time frame, we calculated the standard deviation, using these values as the decision-making basis by setting various thresholds. Decisions were labeled as positive if the standard deviation of a segment surpassed a predetermined threshold, aligning these against ground truth annotations to calculate precision and recall for various thresholds.

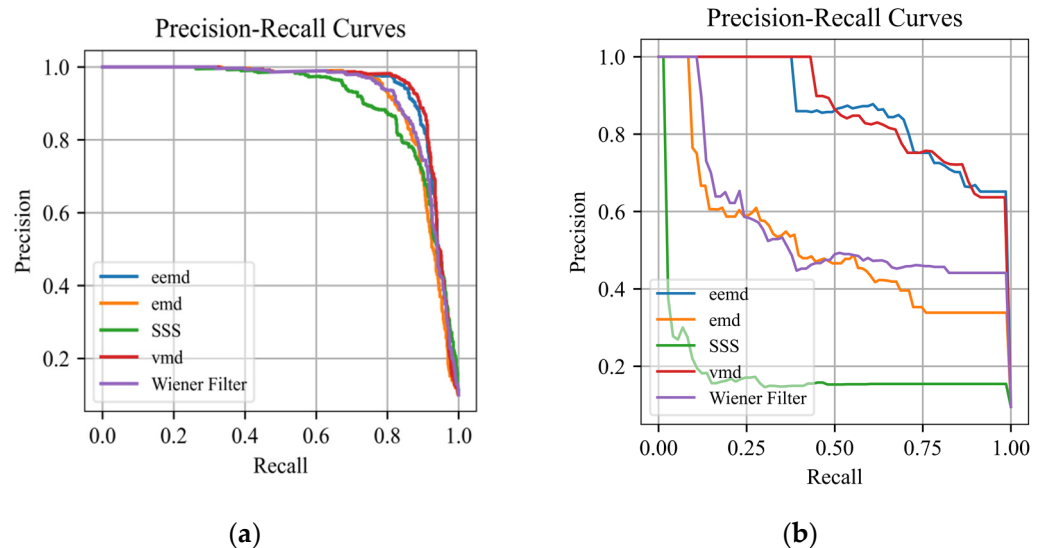
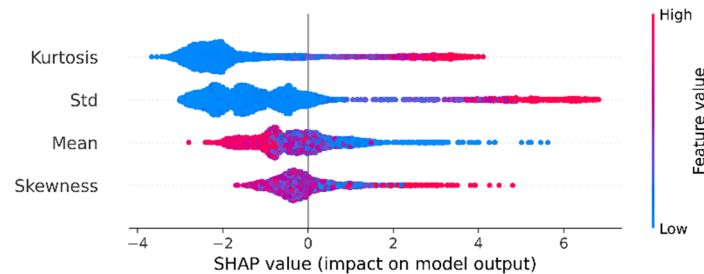


Figure 8. Precision–recall curves: (a) using Std; (b) using kurtosis.

5.4. Feature Importance

Figure 9 is a Shapley additive explanations (SHAP) summary diagram, which provides a visual representation of the influence of each feature on the model output, and can be used to analyze the contribution of features to model prediction. Each point represents the SHAP value of a data point, indicating the contribution of the feature to the model

prediction. The size and sign (positive and negative) of the SHAP value tell us how the feature affects the model prediction. The positive value indicates that the feature promotes the model prediction to the positive class on this data point, and the negative value is the opposite. The color represents the size of the eigenvalues, from low (blue) to high (pink). This helps to explain why some eigenvalues drive the predicted value to shift in a specific direction.



**Figure 9.** SHAP summary diagram.

This SHAP value distribution graph provides a clear visualization of how each feature contributes to the model’s predictions. Higher kurtosis tends to push predictions towards the positive class, indicating that sharper peaks in data distributions are crucial for certain outcomes. In contrast, lower kurtosis pulls predictions towards the negative class. Standard deviation primarily influences predictions negatively when the values are low, but its effect is more nuanced at higher values, demonstrating both positive and negative impacts. The mean shows a broad spread of SHAP values across both positive and negative sides, highlighting its pivotal role in driving predictions. Skewness, mostly clustering near zero, has a relatively minor but occasionally significant impact. By observing the absolute size of the SHAP value (distance from the zero point), it can be seen that the contributions of kurtosis and Std are relatively large.

The analyses show that the standard deviation (Std) remains stable in both datasets. This suggests that changes in amplitude of audio segments are a reliable indicator for detecting cetacean sounds/Kurtosis performs slightly less well, implying that asymmetry in the amplitude distribution and tail coarseness may not be sufficiently reliable.

The kurtosis feature is poorly detected after noise reduction using smoothed spectral subtraction and EMD modal decomposition and Wiener filtering. Whereas, both EEMD and VMD have good detection results. This shows that EEMD and VMD methods are more resilient in noise reduction.

The experiments were completed on the same computer. The LAPTOP-0NCUK0A6 (Huawei, Shenzhen, China) was equipped with an Intel Core i7-10750H processor, 16 GB of RAM, and a 64-bit Windows operating system. The detection algorithm was implemented in the Python (version: Python 3.7.0) environment. The timing results are shown in Table 1. CNN has the longest execution time and the most memory consumption. The running time and memory usage of shape statistical features (SSFs) are second only to band-limited energy detection (BLED), and the detection effect is better than the BLED, which indicates that the detection method of shape statistical features is more rapid and reliable.

**Table 1.** The memory consumption, total algorithm time, and F1 score of convolutional neural network (CNN), MFCC, band-limited energy detection (BLED), and shape statistical feature (SSF).

	Memory Consumption	Total Time	F1 Score
CNN	580.54 MB	293,940 ms	0.93
MFCC	58.43 MB	40,870 ms	0.88
BLED	8.35 MB	1330 ms	0.81
SSF	10.39 MB	17,470 ms	0.97

## 6. Conclusions and Discussion

Obviously, when the Std feature is used for automatic detection, the detection algorithm performs well, and an F1 score of more than 90% is obtained. The performance of kurtosis is slightly worse, and the performance of the standard deviation and skewness is the worst, with an F1 score of 40%. When using the XGBoost model to combine the four features for automatic detection, the automatic detection effect is the best. This study also highlights the advantages of mixing numerous characteristics, particularly when employing machine learning models like XGBoost. The integration of mean, standard deviation, skewness, and kurtosis significantly improves the detection accuracy, which is better than the performance of a single feature. This highlights the complementarity of these features in capturing different aspects of cetacean sounds, thereby improving the robustness of the overall detection.

The findings of this study have significant implications for the development of autonomous detection systems in marine biology and conservation activities. These algorithms can effectively recognize and classify cetacean sounds in a variety of auditory situations utilizing shape statistical properties, promoting study of marine species' behavior, distribution, and population dynamics.

Furthermore, the application of the detection methods in dataset 2 highlights their utility in practical applications, as environmental noise is often a major obstacle to existing detection methods. By incorporating these features into autonomous monitoring systems [39], we can provide continuous, non-invasive monitoring of marine habitats, thereby helping to protect and maintain biodiversity.

Although the findings are promising, future research should focus on numerous areas: Currently, this study uses a simple way of directly extracting statistical features from the original audio, which is ideal for situations that do not require comprehensive time–frequency information. Spectrum analysis, on the other hand, is extremely useful for applications that require detailed time–frequency information and frequency-related properties. Such analysis can provide essential frequency domain information, such as the distribution of sound frequencies and spectral energy. This makes it easier to extract time–frequency characteristics related to sound features, such as spectrum average, standard deviation, skewness, and kurtosis. These features help distinguish between distinct sounds, analyze sound frequency distributions, and aid in tasks such as sound categorization, recognition, and analysis. Future study could also benefit from investigating advanced feature engineering approaches, such as wavelet analysis or higher-order statistics, in order to discover more distinguishing features and improve detection accuracy. Continuous refinement and optimization of machine learning models, including the investigation of alternative classifiers and ensemble approaches [40], is critical for increasing detection accuracy and generalizability. Field testing and verification study in a variety of marine conditions are required to determine the practicality and scalability of the automatic detection system [41]. Extending the method to include a broader variety of marine species and sound categories may increase the use of automatic detection systems in ecological research and conservation monitoring. Furthermore, combining these systems with existing acoustic monitoring networks, such as the Ocean Observing Initiative (OOI) or the Integrated Ocean Observing System (IMOS), will allow for large-scale, collaborative research initiatives and improve our understanding of marine ecosystems.

Finally, using shape statistical features in conjunction with machine learning technology to recognize marine organism sounds is quite important. Researchers can obtain new insights into the underwater world by using data-driven methodologies, which will help to maintain and manage marine biodiversity in a sustainable manner.

**Author Contributions:** Methodology, Z.Z.; Software, B.Z. (Boqing Zhu); Investigation, B.Z. (Bingbing Zhang); Writing—original draft, Y.Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is based on the research supported by the National Natural Science Foundation of China (grant number: 52101391).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Dataset 1 was provided by Jacques van Wyk, Jaco Versfeld, and Johan du Preez.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Simpson, S.; Radford, A.; Nedelec, S.; Ferrari, M.C.O.; Chivers, D.P.; McCormick, M.I.; Meekan, M.G. Anthropogenic noise increases fish mortality by predation. *Nat. Commun.* **2016**, *7*, 10544. [[CrossRef](#)] [[PubMed](#)]
2. Davenport, A.M.; Erbe, C.; Jenner, M.-N.M.; Jenner, K.C.S.; Saunders, B.J.; McCauley, R.D. Pygmy Blue Whale Diving Behaviour Reflects Song Structure. *J. Mar. Sci. Eng.* **2022**, *10*, 1227. [[CrossRef](#)]
3. Acar, G.; Adams, A.E. ACMENet: An underwater acoustic sensor network protocol for real-time environmental monitoring in coastal areas. *IEE Proc.-Radar Sonar Navig.* **2006**, *153*, 365–380. [[CrossRef](#)]
4. Simard, Y.; Roy, N.; Giard, S.; Aulanier, F. North Atlantic right whale shift to the Gulf of St. Lawrence in 2015, revealed by long-term passive acoustics. *Endanger. Species Res.* **2019**, *40*, 271–284. [[CrossRef](#)]
5. Kowarski, K.A.; Martin, S.B.; Maxner, E.E.; Lawrence, C.B.; Delarue, J.J.Y.; Miksis-Olds, J.L. Cetacean acoustic occurrence on the US Atlantic Outer Continental Shelf from 2017 to 2020. *Mar. Mammal Sci.* **2023**, *39*, 175–199. [[CrossRef](#)]
6. Rako-Gospic, N.; Picciulin, M. Underwater noise: Sources and effects on marine life. In *World Seas: An Environmental Evaluation*; Academic Press: Cambridge, MA, USA, 2019; pp. 367–389. [[CrossRef](#)]
7. Bittle, M.; Duncan, A. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. In *Proceedings of Acoustics*; Australian Acoustical Society: Victor Harbor, SA, USA, 2013; Volume 2013, pp. 1–8.
8. Landeira, V.A.R.; Santos, J.O.; Nagano, H. Comparing and Combining Audio Processing and Deep Learning Features for Classification of Heartbeat Sounds. In *Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, 14–19 April 2024; pp. 7220–7224. [[CrossRef](#)]
9. Wang, X.; Liu, A.; Zhang, Y.; Xue, F. Underwater Acoustic Target Recognition: A Combination of Multi-Dimensional Fusion Features and Modified Deep Neural Network. *Remote Sens.* **2019**, *11*, 1888. [[CrossRef](#)]
10. Abeßer, J. A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Appl. Sci.* **2020**, *10*, 2020. [[CrossRef](#)]
11. Hu, S.; Hou, R.; Liao, Z.; Chen, P. Recognition and location of marine animal sounds using two-stream ConvNet with attention. *Front. Mar. Sci.* **2023**, *10*, 1059622. [[CrossRef](#)]
12. Bergler, C.; Schröter, H.; Cheng, R.X.; Barth, V.; Weber, M.; Nöth, E.; Hofer, H.; Maier, A. ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Sci. Rep.* **2019**, *9*, 10997. [[CrossRef](#)]
13. Cai, W.; Zhu, J.; Zhang, M.; Yang, Y. A Parallel Classification Model for Marine Mammal Sounds Based on Multi-Dimensional Feature Extraction and Data Augmentation. *Sensors* **2022**, *22*, 7443. [[CrossRef](#)]
14. Allen, A.N.; Harvey, M.; Harrell, L.; Jansen, A.; Merckens, K.P.; Wall, C.C.; Cattiau, J.; Oleson, E.M. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Front. Mar. Sci.* **2021**, *8*, 607321. [[CrossRef](#)]
15. Talaei Khoei, T.; Ould Slimane, H.; Kaabouch, N. Deep learning: Systematic review, models, challenges, and research directions. *Neural Comput. Appl.* **2023**, *35*, 23103–23124. [[CrossRef](#)]
16. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [[CrossRef](#)]
17. Yektaeian, M.; Amirfattahi, R. Comparison of spectral subtraction methods used in noise suppression algorithms. In *Proceedings of the 2007 6th International Conference on Information, Communications & Signal Processing*, Singapore, 10–13 December 2007; IEEE: New York, NY, USA, 2007; pp. 1–4. [[CrossRef](#)]
18. Sun, W.; Zhou, J.; Meng, X.; Yang, G.; Ren, K.; Peng, J. Coupled Temporal Variation Information Estimation and Resolution Enhancement for Remote Sensing Spatial–Temporal–Spectral Fusion. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18. [[CrossRef](#)]
19. Ribas, D.; Miguel, A.; Ortega, A.; Lleida, E. Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement. *Appl. Sci.* **2022**, *12*, 9000. [[CrossRef](#)]
20. Srinivasarao, V.; Ghanekar, U. Speech intelligibility enhancement: A hybrid wiener approach. *Int. J. Speech Technol.* **2020**, *23*, 517–525. [[CrossRef](#)]
21. Niu, X.D.; Lu, L.R.; Wang, J.; Han, X.C.; Li, X.; Wang, L.M. An improved empirical mode decomposition based on local integral mean and its application in signal processing. *Math. Probl. Eng.* **2021**, *2021*, 8891217. [[CrossRef](#)]
22. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [[CrossRef](#)]

23. Wu, X.; Ai, Q.; Chen, H.; Liu, L. A Complete EEMD with Adaptive Noise Analysis and Improved LSSVM for Transformer Winding Looseness Fault Diagnosis. In Proceedings of the 2024 9th Asia Conference on Power and Electrical Engineering (ACPEE), Shanghai, China, 11–13 April 2024; pp. 1155–1159. [\[CrossRef\]](#)
24. Sun, Z.; Xi, X.; Yuan, C.; Yang, Y.; Hua, X. Surface electromyography signal denoising via EEMD and improved wavelet thresholds. *Math. Biosci. Eng.* **2020**, *17*, 6945–6962. [\[CrossRef\]](#)
25. Hu, M.; Zhang, S.; Dong, W.; Xu, F.; Liu, H. Adaptive denoising algorithm using peak statistics-based thresholding and novel adaptive complementary ensemble empirical mode decomposition. *Inf. Sci.* **2021**, *563*, 269–289. [\[CrossRef\]](#)
26. Li, P.; Fu, Y.; Pan, L.; Li, J.; Tang, J.; Qin, Y. Multi-Model Short-Term Load Forecasting Based on VMD-XGBoost-BiGRU. In Proceedings of the 2024 3rd International Conference on Energy, Power and Electrical Technology (ICEPET), Chengdu, China, 17–19 May 2024; pp. 1282–1285. [\[CrossRef\]](#)
27. Zhao, N.; Mao, Z.; Wei, D.; Zhao, H.; Zhang, J.; Jiang, Z. Fault diagnosis of diesel engine valve clearance based on variational mode decomposition and random forest. *Appl. Sci.* **2020**, *10*, 1124. [\[CrossRef\]](#)
28. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N. An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition. *Acoustics* **2019**, *1*, 410–422. [\[CrossRef\]](#)
29. Chittaragi, N.B.; Koolagudi, S.G. Dialect identification using chroma-spectral shape features with ensemble technique. *Comput. Speech Lang.* **2021**, *70*, 101230. [\[CrossRef\]](#)
30. Dwyer, R. Use of the kurtosis statistic in the frequency domain as an aid in detecting random signals. *IEEE J. Ocean. Eng.* **1984**, *9*, 85–92. [\[CrossRef\]](#)
31. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.
32. Subramani, S.; Hari, S.; Asha, K.; Raj, V.F.D.; Shaji, L.P. Enhanced Short-Term Photovoltaic Power Prediction using a Hybrid Improved Whale Optimization Algorithm with XGBoost. In Proceedings of the 2024 Second International Conference on Smart Technologies for Power and Renewable Energy (SPECon), Ernakulam, India, 2–4 April 2024; pp. 1–4. [\[CrossRef\]](#)
33. Ben Jabeur, S.; Stef, N.; Carmona, P. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Comput. Econ.* **2023**, *61*, 715–741. [\[CrossRef\]](#)
34. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
35. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: New York, NY, USA, 2017; pp. 131–135. [\[CrossRef\]](#)
36. Palanisamy, K.; Singhanian, D.; Yao, A. Rethinking CNN models for audio classification. *arXiv* **2020**, arXiv:2007.11154. [\[CrossRef\]](#)
37. Jiang, H.; Hu, B.; Liu, Z.; Wang, G.; Zhang, L.; Li, X.; Kang, H. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Methods Med.* **2018**, *2018*, 6508319. [\[CrossRef\]](#)
38. Gourisaria, M.K.; Agrawal, R.; Sahni, M.; Singh, P.K. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discov. Internet Things* **2024**, *4*, 1. [\[CrossRef\]](#)
39. Spatial-Temporal Ship Pollution Distribution Exploitation and Harbor Environmental Impact Analysis via Large-Scale AIS Data. *J. Mar. Sci. Eng.* **2024**, *12*, 960. [\[CrossRef\]](#)
40. Alamir, M.A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers. *Appl. Acoust.* **2021**, *175*, 107829. [\[CrossRef\]](#)
41. Glaviano, F.; Esposito, R.; Cosmo, A.D.; Esposito, F.; Gerevini, L.; Ria, A.; Molinara, M.; Bruschi, P.; Costantini, M.; Zupo, V. Management and Sustainable Exploitation of Marine Environments through Smart Monitoring and Automation. *J. Mar. Sci. Eng.* **2022**, *10*, 297. [\[CrossRef\]](#)

**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.