A Parallel Classification Model for Marine Mammal Sounds Based on Multi-Dimensional Feature Extraction and Data Augmentation

Wenyu Cai 1,*, Jifeng Zhu 1,*, Meiyuan Zhang 2 and Yong Yang 1

1 College of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China
2 College of Electrical Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China
* Correspondence: caiwy@hdu.edu.cn (W.C.); jifengzhu@hdu.edu.cn (J.Z.)

Abstract: Due to the poor visibility of the deep-sea environment, acoustic signals are often collected and analyzed to explore the behavior of marine species. With the progress of underwater signal-acquisition technology, the amount of acoustic data obtained from the ocean has exceeded the limit that human can process manually, so designing efficient marine-mammal classification algorithms has become a research hotspot. In this paper, we design a classification model based on a multi-channel parallel structure, which can process multi-dimensional acoustic features extracted from audio samples, and fuse the prediction results of different channels through a trainable full connection layer. It uses transfer learning to obtain faster convergence speed, and introduces data augmentation to improve the classification accuracy. The k-fold cross-validation method was used to segment the data set to comprehensively evaluate the prediction accuracy and robustness of the model. The evaluation results showed that the model can achieve a mean accuracy of 95.21% while maintaining a standard deviation of 0.65%. There was excellent consistency in performance over multiple tests.

Keywords: feature extraction; convolutional neural network; transfer learning; data augmentation; marine mammal classification

1. Introduction

Human activities in the oceans have a huge impact on marine ecological environments. In particular, noises generated by oil-exploration platforms, large offshore wind-power equipment, and ships are serious threats to the survival of marine animals. Therefore, the study of the quantitative distribution and migration behavior of these animals is of great significance in maintaining biodiversity and ecological balance [1–4]. As one of the most important species in the ocean, marine mammals are highly sensitive to changes in the environmental in which they live, so they are often regarded as the most suitable research objects to evaluate the quality of the ecological environment [5]. Marine mammals can generally be divided into five categories in the field of biology, namely cetaceans, pinnipeds, sirenians, marine otters, and polar bears [6].

Marine mammals spend most of their life under the water surface and rely on acoustic signals to communicate with each other. The most effective way to collect their acoustic signals is to use passive acoustic monitoring (PAM) that can cover a wide range of the target sea areas [7]. Since this type of system does not emit sonar signals, it can protect the living habits of marine animals to the maximum extent. Since the PAM system is located underwater, detection of behavior is not affected by bad weather, so it has a strong stability. The PAM system generally consists of an array of hydrophones with different depths. It can be mounted on a moving ship by cables to form a mobile PAM, which can provide continuous and real-time monitoring data, or it can form a fixed PAM by being connected to buoys or anchored at the sea bottom, which can provide long-term automatic
offline data acquisition [8]. A typical PAM equipment is the underwater acoustic signal acquisition system composed of 160 hydrophones arranged in the Norwegian Sea, which has a coverage of 100,000 km$^2$ [9]. Figure 1 shows an audio signal of about 6 s collected from Antarctica by PAM equipment [10]. The bottom half of the figure is the unprocessed time-domain representation of the audio signal, which is sufficient to distinguish the quiet state from the target state. In order to further analyze the audio sample, the signal can be converted from the time domain to the time–frequency (T–F) domain through short-time Fourier transform, where the target marine mammal sound pattern can be seen. However, the spectrogram needs to be further processed before the next classification algorithm, because the audio is often filled with long-term quiet data and the simultaneous existence of multiple species. As the audio represented in this figure shows, most of the time there is no target signal and only ambient noise, and the signal came from the call of a Weddell seal that lives in Antarctica. We need to use an algorithm to remove the quiet state and filter out the desired target from the noisy background in the preprocessing stage.

![Figure 1. Time domain and T-F domain representation of an audio recording from Weddell seal.](image)

After collecting a large amount of acoustic data, target samples can be identified by a special analysis method. Traditionally, the task is performed manually by experts in the field of underwater acoustics who listen carefully to the audio or observe the spectrum to identify whether there are acoustic signals of target animals. However, this method has some limitations. One limitation is that humans can only hear sounds between 20 Hz and 20 kHz, but the sound of many marine mammals exceeds this range. For example, narwhals captured by researchers in Northwest Greenland had a call signal of more than 200 kHz, hence part of their sounds cannot be heard by human ears [11].

A natural idea is to design an end-to-end algorithm without any manual processing. This has the advantage of not using domain-knowledge-based feature engineering. The goal of some related research is to enable the algorithm to automatically extract features from raw waves, and this technical path is called feature learning. Its representative methods include dictionary learning [12], sparse representation learning [13], and deep learning [14]. Some researchers believe that features based on feature engineering are still more efficient than end-to-end models based on feature learning in some cases [15]. Typical examples of extraction from the time domain are zero crossing rate [16], autocorrelation coefficients [17], and the time-domain envelope [18]. Compared with the time-domain features, there are more optional features in the frequency domain, such as spectral envelope [19], spectral flux [20], and Mel spectrogram [21]. By decomposing these features, we
can get cepstral features, Mel frequency cepstral coefficients (MFCCs) [22], linear-frequency cepstral coefficients (LFCCs) [23], gammatone-feature cepstral coefficients [24], and linear predictive coding (LPC) [25], as well as some feature histograms of oriented gradients [26], sub-band power distribution [27], and local binary patterns [28] developed from the field of image recognition.

Several studies have proved that in most cases, the performance of frequency-domain features is better than that of time-domain features in target detection and classification tasks [29]. There is some interesting evidence about this in the evolutionary history of biology. In the human cochlear structure, which has been formed through evolution, the auditory system responds to the external sound with loudness and pitch, which corresponds to the power and frequency of the sound, respectively, and the frequency-domain feature spectrogram contains all this information. The equal loudness contours and critical-band theory discovered by Harvey Fletcher in 1933 [30] proved that the relationship between critical band and sound frequency is a nonlinear relationship and further simplified the spectrogram. Through the matrix operation of the spectrogram and nonlinear Mel-filter banks, the widely used acoustic features Mel spectrogram is obtained.

After obtaining the features, we need to find an efficient classification algorithm to automatically identify the audio. Before machine learning became popular, most of the research was based on traditional pattern-recognition algorithms, such as spectrogram correlation [31], edge detection [32], dynamic time-warping [33] and energy-ratio mapping [34]. Compared to these traditional recognition algorithms, machine learning based on a large number of data samples can fit a very complex and accurate prediction function, so it is increasingly used in the field of acoustic detection. Typical machine learning based audio recognition algorithms comprises decision trees [35], linear discriminant analysis (LDA) [36], support vector machines (SVMs) [37], the Gaussian mixture model (GMM) [38], self-organizing maps (SOMs) [39], long short-term memory (LSTM) [40], the hidden Markov model (HMM) [41], and convolutional neural networks (CNNs) [42–46].

As computers become more powerful, researchers are processing larger amounts of data. Therefore, CNN-based algorithms, due to their data-driven attributes, can find the complex relationship between features and labels from a large number of samples, thus showing more potential than other algorithms in the application field [47]. The mechanism of CNNs is to extract local features from the low-level structure, and then the high-level structure generates global judgment. This process can be understood through visualization of the receptive field. Figure 2 is a marine mammal audio selected at random from the Watkins Marine Mammal Sound Database [10], and fed into the pre-trained CNN model Xception after extracting the spectrogram [48]. The receptive field is visualized at layers 6, 12, and 48. It can be seen that some local features similar to the input spectrogram can be distinguished at the lower level, but the meaning of the receptive field cannot be understood by human eyes at the higher level.

In the field of marine-mammal sound classification, some research has been carried out based on machine-learning algorithms. For example, González-Hernández et al. [49] used the octave analysis method to extract the features of underwater acoustic signals, and constructed a fully connected neural network to classify 11 types of underwater sounds, including several marine mammal species, and achieved a classification accuracy of about 90% under the condition of low computational load. Kirsebom et al. [36] achieved a recall ratio of 80% and a precision of 90% for the North Atlantic right whale under diverse environmental conditions through two-dimensional CNN, and proved that dataset with higher variance was beneficial to the model’s training. Luet et al. [43] solved the common overfitting problem of deep models by modifying the last layers of a pre-trained model, and performed excellent detection and classification of marine-mammal calls with extensive overlapping living areas. The model achieved 97.42% classification accuracy in a dataset with a high signal-to-noise ratio. Frasier et al. [50] proposed an algorithm to classify dolphin click signals, it used time–frequency profiles and click-interval distributions as features, and developed an automatic classification method based on unsupervised networks.
majority of current studies are based on single-dimensional acoustic features [51], but the complexity of acoustic signals means that a single acoustic feature cannot describe the signal well. Better performance can be achieved by selecting and fusing acoustic features from different dimensions.

Research of this paper was focused on cetaceans. These can be divided into odontocetes and mysticetes [52,53]. In order to improve the accuracy of marine-mammal sound classification, this paper proposes a classification model named MDF-PNet; MDF refers to multi-dimensional features, respectively, and P refers to parallel. Its innovation is in the design of a multi-channel parallel model, which can fuse four different and complementary acoustic features, and use a trainable full-connection layer to get the final prediction result. By using transfer-learning techniques, it can achieve faster convergence speed. At the same time, through data augmentation of the data set, the mean accuracy is further improved. Using the k-fold cross validation strategy to cut the data set, we can evaluate the accuracy and robustness of the model.

2. Materials and Methods

2.1. Data Description

The quality of the acoustic dataset plays an important role on machine-learning models. A high signal-to-noise ratio (SNR) and accurately labeled dataset are beneficial in ensuring more accurate parameters and higher prediction accuracy of unknown samples. The data used in this paper is from the Watkins Marine Mammal Sound Database [10,54] and contains more than 15,000 labeled digital sound clips of 51 marine mammal species from the 1940s to the 2000s. The location of the sampling area extends from Bering Strait to the coast of Australia, and the original sampling rate ranged from 5.12k to 192k.

This paper is dedicated to classifying six species in ‘all cuts’ to evaluate the proposed model. Considering that too few samples will lead to bias in test results, the species were selected based on sample size to better measure the performance of the model. The specific distribution and characteristics of these samples are shown in Table 1.
Table 1. Distribution and characteristics of samples.

| Species Name             | Abbreviation | Sample Size (No Augmentation) | Sample Size (Augmentation) | Sampling Rate (Hz) |
|--------------------------|--------------|------------------------------|----------------------------|--------------------|
| Humpback whale           | HW           | 604                          | 1208                       | 5.12k–30k          |
| Long-finned pilot whale  | LFPW         | 1213                         | 1213                       | 23.5k–166.6k       |
| Pantropical spotted dolphin| PSD         | 1034                         | 1034                       | 40.96k–192k        |
| Sperm whale              | SW           | 1422                         | 1422                       | 10k–166.6k         |
| Striped dolphin          | SD           | 681                          | 1362                       | 47.6k–192k         |
| White-sided dolphin      | WSD          | 560                          | 1120                       | 60.6k–166.6k       |
| Total                    |              | 5514                         | 7359                       | 5.12k–192k         |

By randomly selecting the audio sample for each species and performing STFT (short-time Fourier transform) on the sample, the corresponding relationship between the original audio waveform and the spectrogram could be obtained as Figure 3. It can be seen that although each audio looks very similar in the time domain, the difference between them becomes very obvious after the signals are transformed from the time domain to the frequency domain. Therefore, the features used in this paper mainly come from the frequency domain.

Figure 3. Comparison of acoustic signals from different species of marine mammals in time and frequency domains.

2.2. MDF-PNet Structure

The detailed structure of MDF-PNet model is shown in Figure 4. The MDF-PNet model consists of four neural network branches based on transfer learning: Mel spectrogram branch, MFCC branch, LFCC branch, and mean MFCC branch. Each branch processes different kinds of features, and four special acoustic features were chosen since they can provide complementary feature descriptions from different perspectives. Specifically, the Mel spectrogram describes the samples from an image visual perspective, while MFCC and
LFCC use statistical methods to extract the essential information of the samples. Among these branches, the Mel spectrogram branch is a pre-trained DenseNet [55], and the shape of its input feature is $128 \times 345 \times 3$ (frequency $\times$ time $\times$ channel). The MFCC branch and LFCC branch are pre-trained MobileNet [56] and Xception [48], respectively, and the input tensors are also processed in the same way as the Mel spectrogram branch. In addition, the input of the mean MFCC branch is a one-dimensional tensor. The last layer of four branches is the softmax layer with six nodes, which corresponds to six kinds of marine-mammal species, and its output tensor represents the probability that the sample belongs to each species. The calculation formula of softmax layer output is described in Equation (1):

$$y_j = \frac{e^{x_j}}{\sum_{i=1}^{N} e^{x_i}}, \quad 1 \leq j \leq 6$$

where $x_i$ is the output value of the $i$th node of fully connected layer, and $y_j$ is the output value of the $j$th node of softmax layer. The parameter $N$ is 6, which means that each branch will eventually divides all samples into six categories and it predicts a vector with the shape of $[6 \times 1]$.

![Figure 4. The network structure of MDF-PNet.](image)

The MDF-PNet uses stochastic gradient descent (SGD) as the optimization method of parameters. Compared to other optimization methods, SGD has a higher parameter-update frequency so as to reduce the probability of jumping into local optimal solutions. The cross-entropy (CE) loss is used as the optimization target for the proposed model, which is easier to train than mean-square-error (MSE) loss [57]. In this paper, the cross-entropy loss is applied to evaluate the difference between predicted and actual labels. The calculation method of cross-entropy loss is shown as Equation (2):

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij})$$

where $L$ represents the cross-entropy loss of a batch of samples, $N$ is the number of samples in this batch, and $M$ is the number of species in classification. $y_{ij}$ refers to the $j$th elements of one-hot encoded label of the $i$th sample and $p_{ij}$ refers to the prediction probability that the $i$th sample belongs to category $j$. $L$ is the optimization target and by observing the loss curve in the process of model training, we can see that the value of $L$ will gradually decrease, which means that the model has gradually fitted the correct mapping curve.

The predicted results of the four branches are denoted as Y1, Y2, Y3, and Y4, respectively. The four independent prediction results are connected to form a $[24 \times 1]$ vector, and
then a final \([6 \times 1]\) vector is obtained through a trainable fully connected layer and softmax layer as the prediction result. Compared with the fusion method of averaging each branch in some existing studies [58], the learnable weight parameters have better flexibility. As shown in Figure 5, this fusion layer is named the dense layer, which has better adjustability for the weight of each branch compared with the common average calculation method.

**Figure 5.** The fusion block of MDF-PNet at the connection layer.

### 2.3. Data Augmentation Method

For the classification task in deep learning, one basic requirement is that the number of samples in different categories is similar since the data imbalance will affect the generalization ability of the classifier with fewer samples [59]. However, there were too few samples of the three species in the original dataset, so data augmentation was used to keep their sample sizes roughly the same. Traditional data augmentation methods include flipping, rotating, scaling, cropping, and color-jittering. Previous studies have proved that these methods can improve the classification accuracy [60]. For sound data, the samples obtained after flipping the waveform were opposite to the original samples on the time axis in the T–F domain, and belonged to different samples from the perspective of the classification model. Such data augmentation method is often used in sound classification scenes [61]. In this paper, the raw waveform was horizontally flipped for data augmentation and Figure 6 is an example of sample-flipping. The total number of samples after augmentation increased from 5514 to 7359, and the number of samples of all species was basically at the same level.

**Figure 6.** Data augmentation example.
2.4. Feature Extraction Algorithm

Several studies have proposed that MFCC has stronger representation ability in the low-frequency region, while LFCC is better in the high-frequency region [62]. Therefore, we can improve the expressive ability by fusing different types of features and their unique advantages. Four frequency-domain features were selected in this paper: including Mel spectrogram, MFCC, LFCC, and mean MFCC. MFCC is the most widely used feature in the field of audio recognition. The specific calculation can be divided into the following two steps:

Step (1): Perform STFT on the original audio and filter through Mel filter banks. Since the acquisition equipment when collecting data has different sampling rates, this parameter represents the number of times the equipment samples the signal every second. Therefore, all samples should be resampled to a certain fixed frequency, which is 10 kHz in this paper, before the short-time Fourier transform. Although the shape of input tensor about the MDF-PNet model is fixed, the original length of each sample is inconsistent. Therefore, the audio samples are processed through the strategy of random cropping to get the same length. For samples that are too long, the cropping is carried out, and for the short sample, the value zero is randomly added on both sides. Therefore, the length of 30,000 sampling points was unified equivalent to 3 s of audio. After preprocessing, 1024 points were selected as the frame length and 512 points as the step size, and then STFT operation was performed on the audio sample to obtain a complex matrix \( S_t(k) \) with size 59 \( \times \) 513. Finally, according to the number of Mel filters selected, the Mel filter bank matrix \( H_m(k) \) with size of 513 \( \times \) 128 was generated and multiplied by the results of STFT to obtain a power spectrum of Mel scale, i.e., a 59 \( \times \) 128 real matrix \( X_m(t) \). The calculation process is shown in Equations (3) and (4) [63]:

\[
X_m(t) = \sum_{k=0}^{N-1} [S_t(k)]^2 H_m(k), \quad 0 \leq t < L, \quad 0 \leq m < M
\]  

(3)

\[
H_m(k) = \begin{cases} 
\frac{k - o(m)}{c(m) - o(m)} & 0 \leq k \leq c(m) \\
\frac{h(m) - k}{h(m) - c(m)} & c(m) \leq k \leq h(m) 
\end{cases}
\]  

(4)

where \( X_m(t) \) denotes the power spectrum matrix, \( L \) denotes the number of data frames in the direction of time axis, \( M \) denotes the number of Mel filter banks, \( S_t(k) \) denotes the spectrum obtained by STFT, \( H_m(k) \) is the expression of Mel filter banks, where \( o(m), c(m) \), and \( h(m) \) are frequency points, and their interval determines the type of filter banks.

Step (2): Perform discrete cosine transform (DCT) on logarithm of power spectrum.

Firstly, the power spectrum is converted into decibels (dB unit), and then DCT transformation is carried out to obtain the MFCC matrix. The dimension of MFCC matrix is finally obtained by the calculation of Equation (5).

\[
mfcc(t,c) = \sum_{m=1}^{M} \frac{c\pi(m - 0.5)}{M} \cos \left( \frac{c\pi(m - 0.5)}{M} \right)
\]  

(5)

where \( mfcc(t,c) \) is the calculated MFCC matrix, \( t \) denotes the index in the time axis, \( c \) denotes the index of cepstral coefficient, \( M \) denotes the number of Mel filter banks, and \( X_m(t) \) denotes the input power spectrum matrix. In some applications, the dynamic features of MFCC are also calculated, i.e., the first-order difference and the second-order difference form a high-dimensional feature together with the static features [64].

The calculation process of the Mel spectrogram is similar, except that it only uses Step (1). The difference between LFCC and MFCC is that the Mel filter bank in Step (1) is replaced by a linear filter bank for LFCC.
To summarize, the difference between these features can be intuitively observed through visualization. An audio sample was randomly selected from the data set in Table 1, and a high-resolution spectrogram was obtained through STFT to compare. After feature extraction processing through the above steps, four real number matrices could be obtained. The visualization results are shown in Figure 7, where the X-axis represents time with second unit and Y-axis represents frequency with Hz unit. It can be seen that although the resolution of Figure 7b is much lower than Figure 7a, the signal contour can still be seen. Furthermore, signals cannot be directly distinguished from Figure 7c,d after DCT transformation. In conclusion, it is promising to combine these four features to improve audio recognition accuracy.

![Figure 7](image)

**Figure 7.** Frequency-domain features of the same audio sample. (a) Spectrogram (513 × 938), (b) Mel spectrogram (32 × 59), (c) MFCC (32 × 59), and (d) LFCC (32 × 59).

2.5. Transfer-Learning Method

Three pre-training models including DenseNet-201, MobileNet, and Xception were used to initialize parameters in the proposed MDF-PNet, which were based on the subset of ImageNet and fully trained. The number of parameters of these networks were 20.2 M, 4.2 M, and 22.9 M, and the depths of these models were 201, 88, and 126 layers, respectively. Since the feature tensor generated in the feature extraction stage was different from the shape of ImageNet, each pre-training model needed to be modified slightly before usage. The method used was to modify the input layer of all models to match the corresponding feature tensor size and then remove the weight of full connection layer, and finally add the full connection layer consistent with the sample label dimension and retrain it.

Compared with the huge computational workload required by two-dimensional CNN, one-dimensional CNN could achieve the same excellent effect in a simpler way. As a part of the proposed MDF-PNet, it can improve the overall performance of classification. So far, in addition to these three pre-training models, a one-dimensional CNN as the fourth parallel structure was constructed to match the 1D Mean MFCC vector. After training 3000 epochs, the obtained weights were saved as the initial weights of the fourth model. The structure of this one-dimensional CNN is shown in Table 2.
Table 2. The detailed parameters of used 1D-CNN.

| Layers                          | Description                                      |
|---------------------------------|--------------------------------------------------|
| Input layer                     | Input tensor: 128 × 1                            |
| | Conv + dropout + maxPooling1D | Convolution: 16 kernels (3 × 1, stride = 1)      |
| |                             | Pooling: (2 × 1, stride = 1)                     |
| | Flatten + dropout             | Activation function: ReLu × 4                    |
| fully-connected                 | Convert the shape of tensor from 8 × 16 to 128 × 1|
| Fully-connected                 | Units = 128                                      |
| Softmax Layer                   | Output tensor: 6 × 1                              |

2.6. Data Segmentation

In the field of statistics, it is of great significance to reasonably divide the data set so as to obtain a more balanced sample distribution. Especially for small-scale samples, simply dividing the data set into training sets and test sets will often lead to the inability to accurately evaluate the performance of the model. Compared with this segmentation method, a more reasonable way is to use k-fold cross validation to test the model repeatedly, and use the mean value to characterize the performance of the model. As shown in Figure 8 below, this method firstly cuts out 20% of the data as the test set, which is not involved in subsequent parameter adjustment and model training, and then divides the remaining 80% of the data into five parts for cycle training and testing to obtain five validation accuracy values. Finally, the prediction accuracy of the model is obtained by calculating the mean value. Accuracy value can be calculated by Equation (6) below [65]:

\[
\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}
\]

where \(N\) refers to the number of species, the value is 6 in this paper. The metrics in multi-classification are derived from binary classification. \(TP, FP, TN,\) and \(FN\) refer to true positive, false positive, true negative and false negative, respectively. Accuracy is an indicator designed to evaluate the consistency between the prediction result of the model and the label, so a more intuitive understanding can be considered as the ratio between the number of correctly classified samples and the total number of samples. The following test results for baseline and MDF-PNet were obtained based on this data segmentation method.

![Diagram](image)

Figure 8. Segmentation method of data set.

3. Results

3.1. Performance Comparison between Baseline and MDF-PNet

In order to compare the performance of different models in similar fields, it is necessary to choose recent research in the field of ecological informatics as the baseline. In reference [43], the authors proposed a modified AlexNet based on transfer learning to detect and classify marine-mammal sounds. The data used by the author come from the ‘best of cuts’ part of the marine mammal call dataset. Compared to the part marked ‘all
cuts’, this part of dataset has a higher signal-to-noise ratio, so it can often obtain excellent classification accuracy. In this paper, the model proposed in this reference is used as the baseline and compared with the proposed MDF-PNet.

After the data set was segmented using the above data segmentation method, the performance of baseline and MDF-PNet on train set and validation set was tested. In order to maintain strict consistency in data distribution, the same random seed was used to segment the data sets. Loss curve and accuracy curve can accurately visualize this process. The former represents the output of loss function, while the latter reflects the accuracy of prediction. Considering that the generalization ability of the model on validation set is the most important index pursued by the algorithm, Figures 9 and 10, respectively, represent the iterative process of baseline and MDF-PNet on the validation set. The light-colored lines in the figure represent the process of five cycles of testing, and the dark lines represent their mean values.

![Figure 9](image.png) **Figure 9.** Five test results of baseline. (a) The accuracy curve and (b) the loss curve.

![Figure 10](image.png) **Figure 10.** Five test results of MDF-PNet. (a) The accuracy curve and (b) the loss curve.

In addition to the mean, the standard deviation is also an important indicator in statistics. It can measure the dispersion of the results in the case of multiple measurements. A model with better performance should have a lower standard deviation. Table 3 shows the mean and standard deviation of baseline and MDF-PNet under the five tests.

| Model/Data          | Mean Accuracy of Five Tests | Standard Deviation |
|---------------------|----------------------------|-------------------|
| Baseline model      | 88.40%                     | 1.36%             |
| MDF-PNet model      | 95.21%                     | 0.65%             |

In order to more intuitively compare the accuracy of the two models, the two models were used to predict the same batch of data for evaluation. This batch of data is the part of the test set obtained by the k-fold segmentation strategy above. Since it has never
participated in the previous training, it is a batch of completely independent data. The predicted results are represented using the confusion matrix in Figure 11. The labels on the axes correspond to the abbreviation of species name, and their full names are shown in Table 1 above.

![Confusion matrix of baseline and MDF-PNet on the same batch of samples. (a) Baseline and (b) MDF-PNet.](image)

When evaluating the performance of binary classification models, precision, recall and F1 score can be used to comprehensively characterize the classification ability of the model. The specific steps are as follows: First, it is assumed that the labels of the samples can be divided into positive and negative. Then prediction results are compared with the real labels to obtain the four categories of true positive, true negative, false positive, and false negative. Finally, according to the different focus of the research objectives, the following statistical metrics can be obtained in Equations (7)–(9) [66]:

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \\
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

The focus of precision and recall in the above formulae is different. The former focuses on evaluating the accuracy of prediction results, while the latter focuses on measuring the coverage of prediction results. To some extent, these two indicators have certain contradictory attributes, that is, when only one indicator is considered to be highest, the other indicator will perform worse. So F1 score, as a weighted statistical index of precision and recall, can be used to evaluate the model performance more comprehensively. In the multi-label classification task, the above statistical metrics are still effective, so the confusion matrix can be further calculated by the above formula to obtain the statistical results shown in Table 4.

Table 4. Comparison of statistical metrics performance between baseline and MDF-PNet.

| Species/Metrics | Precision (Baseline/MDF-PNet) (%) | Recall (Baseline/MDF-PNet) (%) | F1 Score (Baseline/MDF-PNet) (%) |
|-----------------|----------------------------------|------------------------------|---------------------------------|
| HW              | 94.19/99.61                     | 95.67/99.61                  | 94.92/99.61                     |
| LFPW            | 86.61/91.57                     | 89.43/92.68                  | 88.00/92.12                     |
| PSD             | 93.88/98.96                     | 89.76/93.17                  | 91.77/95.88                     |
| SW              | 78.24/88.84                     | 78.90/90.72                  | 78.57/89.77                     |
| SD              | 94.95/96.01                     | 89.15/97.97                  | 91.96/96.97                     |
| WSD             | 93.95/99.57                     | 99.15/98.72                  | 96.48/99.14                     |
3.2. Effect of Data Augmentation on Performance

Finally, in order to study the effect of the data augmentation method on the performance of baseline and MDF-PNet, the following cross-test was conducted based on the model used and data augmentation was performed. In order to keep the consistency of the input data under different conditions, this step is also based on the above data segmentation method. The mean accuracy curve obtained by multiple tests on the validation set is shown in Figure 12. Table 5 lists the mean accuracy and standard deviation of the cross-test, where the better results are highlighted in red.

![Accuracy vs Epoch](image)

**Figure 12.** Cross-test the combination of two conditions. (a) baseline and (b) MDF-PNet.

**Table 5.** Comparison of accuracy after cross-test.

| Model                          | Mean Accuracy of Five Tests | Standard Deviation |
|-------------------------------|-----------------------------|-------------------|
| Baseline + no data augmentation | 81.89%                      | 3.29%             |
| Baseline + data augmentation   | 88.40%                      | 1.36%             |
| MDF-PNet + no data augmentation| 93.67%                      | 1.07%             |
| MDF-PNet + data augmentation   | 95.21%                      | 0.65%             |

4. Discussion

By comparing Figures 9 and 10, we can clearly see three advantages of MDF-PNet compared with baseline: (1) The convergence speed is faster. MDF-PNet only needed 20 epochs to make the loss curve become stable, while the baseline required about 200 epochs to achieve the same effect. (2) The mean accuracy of prediction was higher. When the results of five tests were averaged, the mean prediction accuracy of MDF-PNet for the validation set was 95.21%, while the mean prediction accuracy of baseline under the same test conditions was 88.40%. (3) The standard deviation of the prediction results was smaller. The standard deviation of MDF-PNet was 0.65% for five tests, while the baseline was 1.36%. This means that MDF-PNet has less dispersion and better consistency in its prediction performance. Table 3 records these statistical results in detail.

Figure 11 illustrates the performance of the baseline and MDF-PNet on the same batch of samples. It is obvious that MDF-PNet is better than baseline in the classification ability of most marine mammal species. Compared with the confusion matrix, by introducing statistical indicators to quantify the classification ability of them, we can get a more accurate perception of the improvement of the model performance. All indicators obtained according to Equations (6)–(8) have been listed in detail in Table 4. The left side of the slash in each cell belongs to baseline, and the corresponding statistics results on the right belong to MDF-PNet. The better results among them are marked with red. As can be seen, the performance of MDF-PNet was better than baseline in most metrics, which fully reflects the effective performance improvement of the model. There are many factors that affect the performance of a model. One of the most important is the quality and size of the dataset used. Therefore, the comparison of metrics is only valuable when dealing with the same dataset. Trawickiet et al. [41] used a hidden Markov model to classify marine mammals.
in the subset of the dataset, and obtained a classification accuracy of 84.11%, but too few samples were used, so the obtained results had a high variance. There is also work based on different dataset of marine-mammal calls. For example, Kirsebom et al. [36] used a deep neural network with skip structure to detect North Atlantic right whale upcalls. After the evaluation on the test set, their method could achieve a recall up to 80% and a precision of 90% on the dataset, both of which seem to be slightly worse than the model proposed in this paper.

Figure 12 shows the effect of the data augmentation step for baseline and MDF-PNet, in evaluating the significance of data equalization for the model performance, and comparison was performed under completely consistent data segmentation criteria. Since the convergence speed of baseline was slower than MDF-PNet, it had more iterations on coordinates than the latter.

The numerical results of Figure 12 are shown in Table 5. It can be seen that for baseline that the data-augmentation step increased its mean accuracy from 81.89% to 88.40%, an increase of 6.51%. For MDF-PNet, this step increased mean accuracy from 93.67% to 95.21%, an increase of 1.54%. For any model, this step makes the standard deviation of the five test results smaller and achieves better consistency, which fully demonstrates that this step is effective and necessary. The highest mean classification accuracy was achieved under MDF-PNet with data augmentation. Compared with baseline, its mean accuracy was increased by 6.81%.

5. Conclusions

In this paper, we propose a classification model for marine mammal sounds based on multi-dimensional feature extraction. The data augmentation step was applied to the unbalanced data to make the distribution more balanced among the species, thus improving the mean classification accuracy of the model. At the same time, MDF-PNet used transfer learning to obtain the initial weights and thus achieved faster convergence. In the stage of acoustic feature selection, the proposed MDF-PNet model used multiple complementary features in the frequency domain to combine their respective advantages, and a full connection layer was retrained at the back end to connect the prediction results of different branches.

After the data set was uniformly segmented by k-fold cross validation, the MDF-PNet and similar research in the same field were repeatedly compared and tested. It can be seen that the mean classification accuracy of MDF-PNet increased by 6.81% compared with the baseline, and there was a lower standard deviation between different test results, which indicates that the performance had been greatly improved.

In our future work, we will explore whether the proposed model has a better solution in the fusion stage of prediction results, such as retraining a deep model to replace the current trainable full connection layer.

Author Contributions: Conceptualization, W.C.; methodology, W.C. and J.Z.; software, J.Z.; validation, J.Z.; formal analysis, M.Z.; investigation, M.Z.; writing—original draft preparation, W.C. and J.Z.; writing—review and editing, W.C., J.Z. and Y.Y.; visualization, Y.Y.; supervision, W.C.; project administration, W.C.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Natural Science Foundation of Zhejiang Province (No.LZ22F010004 and LZJWY22E090001), the Fundamental Research Funds for the Provincial Universities of Zhejiang (No.GK209907299001-001), the National Natural Science Foundation of China (No.61871163 and No.61801431), and the Stable Supporting Fund of Acoustics Science and Technology Laboratory.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Williams, R.; Wright, A.J.; Ashe, E.; Blight, L.K.; Bruintjies, R.; Canessa, R.; Clark, C.W.; Cullis-Suzuki, S.; Dakin, D.T.; Erbe, C.; et al. Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management. Ocean. Coast. Manag. 2015, 115, 17–24. [CrossRef]

2. Merchant, N.D.; Brookes, K.L.; Faulkner, R.C.; Bicknell, A.W.; Godley, B.J.; Witt, M.J. Underwater noise levels in UK waters. Sci. Rep. 2016, 6, 36942. [CrossRef]

3. Blair, H.B.; Merchant, N.D.; Friedlaender, A.S.; Wiley, D.N.; Parks, S.E. Evidence for ship noise impacts on humpback whale foraging behaviour. Biol. Lett. 2016, 12, 20160005. [CrossRef] [PubMed]

4. McDonald, M.A.; Hildebrand, J.A.; Wiggins, S.M. Increases in deep ocean ambient noise in the Northeast Pacific west of San Nicolas Island, California. J. Acoust. Soc. Am. 2006, 120, 711–718. [CrossRef] [PubMed]

5. Weilgart, L.S. A brief review of known effects of noise on marine mammals. Int. J. Comp. Psychol. 2007, 20, 159–168.

6. Reeves, R.R. Marine Mammals: History of Exploitation. In Encyclopedia of Ocean Sciences; Elsevier: Amsterdam, The Netherlands, 2019; pp. 601–609. [CrossRef]

7. Gervaise, C.; Simard, Y.; Aulanier, F.; Roy, N. Performance Study of Passive Acoustic Systems for Detecting North Atlantic Right Whales in Seaways: The Honguendo Strait in the Gulf of St. Lawrence; Department of Fisheries and Oceans, Government of Canada: Ottawa, ON, Canada, 2019.

8. Mellinger, D.; Stafford, K.; Moore, S.; Dziak, R.; Matsumoto, H. An overview of fixed passive acoustic observation methods for cetaceans. Oceanography 2007, 20, 36–45. [CrossRef]

9. Wang, D.; Garcia, H.; Huang, W.; Tran, D.D.; Jain, A.D.; Yi, D.H.; Gong, Z.; Jech, J.M.; Godø, O.R.; Makris, N.C.; et al. Vast assembly of vocal marine mammals from diverse species on fish spawning ground. Nature 2016, 531, 366–370. [CrossRef]

10. Woods Hole Oceanographic Institution. Watkins Marine Mammal Sound Database. Available online: https://cis.whoi.edu/science/B/whalesounds/index.cfm (accessed on 1 March 2022).

11. Rasmussen, M.H.; Koblitz, J.C.; Laidre, K.L. Buzzes and high-frequency clicks recorded from narwhals (Monodon monoceros) at their wintering ground. AquatMamm 2015, 41, 256–264. [CrossRef]

12. Zubair, S.; Yan, F.; Wang, W. Dictionary learning based sparse coefficients for audio classification with max and average pooling. Digit. Signal Process. 2013, 23, 960–970. [CrossRef]

13. Henaff, M.; Jarrett, K.; Kavukcuoglu, K.; LeCun, Y. Unsupervised Learning of Sparse Features for Scalable Audio Classification. In ISMIR; Citeeseer: Princeton, NJ, USA, 2011; pp. 681–686.

14. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204. [CrossRef]

15. Jiang, Y.; Bosch, N.; Baker, R.S.; Paquete, L.; Oculiugh, J.; Andres, J.M.A.L.; Moore, A.L.; Biswas, G. Expert Feature-Engineering vs. Deep Neural Networks: Which Is Better for Sensor-Free Affect Detection? In Artificial Intelligence in Education; Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 198–211. [CrossRef]

16. Ramaiah, V.S.; Yan, F.; Wang, W. Dictionary learning based sparse coefficients for audio classification with max and average pooling. Digit. Signal Process. 2013, 23, 960–970. [CrossRef]

17. Shannon, B.J.; Paiwali, K.K. Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. Speech Commun. 2006, 48, 1458–1485. [CrossRef]

18. Caetano, M.; Rodet, X. Improved Estimation of the Amplitude Envelope of Time-Domain Signals Using True Envelope Cepstral Smoothing. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4244–4247. [CrossRef]

19. Wisniewski, M.; Zielinski, T. Joint application of audio spectral envelope and tonality index in e-asthma monitoring system. IEEE J. Biomed. Health Inform. 2015, 19, 1009–1018. [CrossRef]

20. Tzanetakis, G.; Cook, P. Musical genre classification of audio signals. IEEE Trans. Audio Speech Process. 2002, 10, 293–302. [CrossRef]

21. Sharan, R.V.; Moir, T.J. Acoustic event recognition using cochleagram image and convolutional neural networks. Appl. Acoust. 2019, 148, 62–66. [CrossRef]

22. Ogundile, O.O.; Usman, A.M.; Babalola, O.P.; Versfeld, D.J.J. Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of Mysticetes’ vocalisations. Ecol. Inform. 2021, 63, 101306. [CrossRef]

23. Dewi, S.P.; Prasasti, A.L.; Irawan, B. Analysis of LFCC Feature Extraction in Baby Crying Classification Using KNN. In Proceedings of the 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Bali, Indonesia, 5–7 November 2019; pp. 86–91. [CrossRef]

24. Valero, X.; Alias, F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Trans. Multimedia. 2012, 14, 1684–1689. [CrossRef]

25. Ogundile, O.O.; Usman, A.M.; Babalola, O.P.; Versfeld, D.J.J. A hidden Markov model with selective time domain feature extraction to detect inshore Bryde’s whale short pulse calls. Ecol. Inform. 2020, 57, 101087. [CrossRef]
26. Rakotomamonjy, A.; Gasso, G. Histogram of gradients of time-frequency representations for audio scene detection. *IEEE/ACM Trans. Audio Speech Lang. Process*. 2014, 23, 142–153. [CrossRef]
27. Dennis, J.; Tran, H.D.; Chng, E.S. Image feature representation of the subband power distribution for robust sound event classification. *IEEE Trans. Audio Speech Lang. Process*. 2013, 21, 367–377. [CrossRef]
28. Kobayashi, T.; Ye, J. Acoustic Feature Extraction by Statistics Based Local Binary Pattern for Environmental Sound Classification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3052–3056. [CrossRef]
29. Padovese, B.; Frazao, F.; Kirsebom, O.S.; Matwin, S. Data augmentation for the classification of North Atlantic right whales upcalls. *J. Acoust. Soc. Am.* 2021, 149, 2520–2530. [CrossRef]
30. Fletcher, H. Auditory patterns. *Rev. Mod. Phys.* 1940, 12, 47–65. [CrossRef]
31. Mellinger, D.K.; Clark, C.W. Recognizing transient low-frequency whale sounds by spectrogram correlation. *J. Acoust. Soc. Am.* 2000, 107, 3518–3529. [CrossRef] [PubMed]
32. Gillespie, D. Detection and classification of right whale calls using an ‘edge’ detector operating on a smoothed spectrogram. *Can. Acoust.* 2004, 32, 39–47.
33. Brown, J.C.; Miller, P.J.O. Automatic classification of killer whale vocalizations using dynamic time warping. *J. Acoust. Soc. Am.* 2007, 122, 1201–1207. [CrossRef] [PubMed]
34. Klinck, H.; Mellinger, D.K. The energy ratio mapping algorithm: A tool to improve the energy-based detection of odontocete echolocation clicks. *J. Acoust. Soc. Am.* 2011, 129, 1807–1812. [CrossRef] [PubMed]
35. Esfahanian, M.; Erdol, N.; Gerstein, E.; Zhuang, H. Two-stage detection of north Atlantic right whale upcalls using local binary patterns and machine learning algorithms. *Appl. Acoust.* 2017, 120, 158–166. [CrossRef]
36. Kirsebom, O.S.; Frazao, F.; Simard, Y.; Roy, N.; Matwin, S.; Giard, S. Performance of a deep neural network at detecting North Atlantic right whale upcalls. *J. Acoust. Soc. Am.* 2020, 147, 2636–2646. [CrossRef]
37. Shen, W.; Tu, D.; Yin, Y.; Bao, J. A new fusion feature based on convolutional neural network for pig cough recognition in field situations. *Inf. Process. Agric.* 2021, 8, 573–580. [CrossRef]
38. Pentapati, H.; Vasamsetti, S.; Tenneti, M. MFCC for voiced part using VAD and GMM based gender recognition. *AMA_B* 2017, 60, 581–592. [CrossRef]
39. Allen, A.N.; Harvey, M.; Harrell, L.; Jansen, A.; Merkens, K.P.; Wall, C.C.; Cattiau, J.; Oleson, E.M. A convolutional neural network for pig cough recognition in field situations. *Inf. Process. Agric.* 2021, 8, 573–580. [CrossRef]
40. Ibrahim, A.K.; Zhuang, H.; Chollet, F. Xception: Deep Learning with Depth wise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]
41. Trawicki, M.B. Multispecies discrimination of whales (Cetaceans) using Hidden Markov Models (HMMS). *Ecol. Inform.* 2021, 81, 101223. [CrossRef]
42. Mishachandar, B.; Vairamuthu, S. Diverse ocean noise classification using deep learning. *Appl. Acoust.* 2021, 181, 108141. [CrossRef]
43. Tu, L.; Han, B.; Yu, F. Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecol. Inform.* 2021, 62, 101277. [CrossRef]
44. Dugan, P.J.; Clark, C.W.; LeCun, Y.A.; Van Parijs, S.M. DCL System Using Deep Learning Approaches for Land-Based or Ship-Based Real Time Recognition and Localization of Marine Mammals; Bioacoustics Research Program, Cornell University: Ithaca, NY, USA, 2015.
45. Shiu, Y.; Palmer, K.J.; Roch, M.A.; Fleishman, E.; Liu, X.; Nosal, E.-M.; Helble, T.; Cholewiak, D.; Gillespie, D.; Klinck, H. Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* 2020, 10, 607. [CrossRef] [PubMed]
46. Allen, A.N.; Harvey, M.; Harrell, L.; Jansen, A.; Merkens, 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]
47. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.-A. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* 2019, 146, 3590–3628. [CrossRef] [PubMed]
48. Chollet, F. Xception: Deep Learning with Depth wise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
49. Gonzalez-Hernandez, F.R.; Sanchez-Fernandez, L.P.; Suarez-Guerra, S.; Sanchez-Perez, E.A. Marine mammal sound classification based on a parallel recognition model and octave analysis. *Appl. Acoust.* 2017, 119, 17–28. [CrossRef]
50. Frasier, K.E.; Roch, M.A.; Soldevilla, M.S.; Wiggins, S.M.; Garrison, L.P.; Hildebrand, J.A. Automated classification of dolphin echolocation click types from the Gulf of Mexico. *PloS Comput. Biol.* 2017, 13, e1005823. [CrossRef]
51. Usman, A.M.; Ogundile, O.O.; Versfeld, D.J.J. Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access* 2020, 8, 105181–105206. [CrossRef]
52. Reichenbach, W.J.; Greene, J.; Malme, C.I.; Thomson, D.H. Marine Mammals and Noise; Elsevier: Amsterdam, The Netherlands, 1995. [CrossRef]
53. Jefferson, T.A.; Webber, M.A.; Pitman, R.L.; Gagnon, S.; Pitman, R.L.; Tavolacci, L.P.; Hildebrand, J.A. Marine Mammals of the World: A Comprehensive Guide to Their Identification, 2nd ed.; Elsevier: London, UK; San Diego, CA, USA, 2015. [CrossRef] [PubMed]
54. Sayigh, L.; Daher, M.A.; Allen, J.; Gordon, H.; Joyce, K.; Stuhlmann, C.; Tyack, P. The Watkins Marine Mammal Sound Database: An Online, Freely Accessible Resource. *Proc. Mtgs. Acoust.* 2016, 27, 040013. [CrossRef]
55. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]

56. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications; NASA/ADS: Cambridge, MA, USA, 2017.

57. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv. Neural Inf. Process. Syst. 2018, 31, 8792–8802.

58. Zhong, M.; Castellote, M.; Dodiia, R.; Lavista Ferres, J.; Keogh, M.; Brewer, A. Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. 2020, 147, 1834–1841. [CrossRef] [PubMed]

59. Prati, R.C.; Batista, G.E.A.P.; Silva, D.F. Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. Knowl. Inf. Syst. 2015, 45, 247–270. [CrossRef]

60. Taylor, L.; Nitschke, G. Improving Deep Learning with Generic Data Augmentation. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1542–1547. [CrossRef]

61. Mushtaq, Z.; Su, S.-F.; Tran, Q.-V. Spectral images based environmental sound classification using CNN with meaningful data augmentation. Appl. Acoust. 2021, 172, 107581. [CrossRef]

62. Zhou, X.; Garcia-Romero, D.; Duraiswami, R.; Espy-Wilson, C.; Shamma, S. Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 559–564. [CrossRef]

63. Ye, Y.; Lao, L.; Yan, D.; Wang, R. Identification of weakly pitch-shifted voice based on convolutional neural network. Int. J. Digit. Multimed. Broadcast. 2020, 2020, 8927031. [CrossRef]

64. Noda, J.J.; Travieso, C.M.; Sanchez-Rodriguez, D.; Dutta, M.K.; Singh, A. Using Bioacoustic Signals and Support Vector Machine for Automatic Classification of Insects. In Proceedings of the 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 11–12 February 2016; pp. 656–659. [CrossRef]

65. Qayyum, A.; Anwar, S.M.; Awais, M.; Majid, M. Medical image retrieval using deep convolutional neural network. Neurocomputing 2017, 266, 8–20. [CrossRef]

66. Salman, A.; Maqbool, S.; Khan, A.H.; Jalal, A.; Shafait, F. Real-time fish detection in complex backgrounds using probabilistic background modelling. Ecol. Inform. 2019, 51, 44–51. [CrossRef]