Journal of *Marine Science and Engineering*

MDPI

*Article*

# On the Importance of Passive Acoustic Monitoring Filters

**Rafael Aguiar [1,*], Gianluca Maguolo [2], Loris Nanni [2], Yandre Costa [3] and Carlos Silla, Jr. [1]**

[1] Programa de Pós-Graduação em Informática, Escola Politécnica, Pontifícia Universidade Católica do Paraná, Curitiba 80215-901, Brazil; carlos.sillajr@gmail.com
[2] Department of Information Engineering, University of Padua, 35131 Padua, Italy; gianluca.maguolo@unipd.it (G.M.); loris.nanni@unipd.it (L.N.)
[3] Departamento de Informática, Universidade Estadual de Maringá, Maringá 87200-000, Brazil; yandre@din.uem.br
* Correspondence: aguiar.pr@gmail.com

**Abstract:** Passive acoustic monitoring (PAM) is a noninvasive technique to supervise wildlife. Acoustic surveillance is preferable in some situations such as in the case of marine mammals, when the animals spend most of their time underwater, making it hard to obtain their images. Machine learning is very useful for PAM, for example to identify species based on audio recordings. However, some care should be taken to evaluate the capability of a system. We defined PAM filters as the creation of the experimental protocols according to the dates and locations of the recordings, aiming to avoid the use of the same individuals, noise patterns, and recording devices in both the training and test sets. It is important to remark that the filters proposed here were not intended to improve the accuracy rates. Indeed, these filters tended to make it harder to obtain better rates, but at the same time, they tended to provide more reliable results. In our experiments, a random division of a database presented accuracies much higher than accuracies obtained with protocols generated with PAM filters, which indicates that the classification system learned other components presented in the audio. Although we used the animal vocalizations, in our method, we converted the audio into spectrogram images, and after that, we described the images using the texture. These are well-known techniques for audio classification, and they have already been used for species classification. Furthermore, we performed statistical tests to demonstrate the significant difference between the accuracies generated with and without PAM filters with several well-known classifiers. The configuration of our experimental protocols and the database were made available online.

**Keywords:** passive acoustic monitoring (PAM); audio classification; texture classification; PAM-filters; experimental protocols for audio classification; statistical tests



**Citation:** Aguiar, R.; Maguolo, G.; Nannis, L.; Costa, Y.; Silla, C., Jr. On the Importance of Passive Acoustic Monitoring Filters. *J. Mar. Sci. Eng.* **2021**, *9*, 685. https://doi.org/10.3390/jmse9070685

Academic Editors: Marta Belchior Lopes and Pedro Reis Costa

Received: 28 April 2021
Accepted: 15 June 2021
Published: 22 June 2021



## 1. Introduction

The techniques of passive acoustic monitoring (PAM) are tools to automatically detect, localize, and monitor animals [1]. Passive refers to the fact that the system is noninvasive, as it does not interfere with the environment. It is an acoustic system because the surveillance is performed through audio signals. For example, a recording device connected to the Internet could acquire data from an environment and send the captured data to a classification system that identifies which species are nearby.

In the case of marine animals, the use of audio data might be preferred over image data [2]. The reasoning for that is because visual survey methods for some marine animals, such as whales, may detect only a fraction of the animals present in the area. This happens because visual observers can only see them during the very short period when they are on the surface, and also because visual surveys can be undertaken only during daylight hours and in relatively good weather.

Research based on audio of marine mammals has been conducted for different purposes. For example, Sayigh et al. [3] found that bottlenose dolphins are likely to use

signature whistles to identify each other. Furthermore, Sayigh et al. [4] indicated that identification does not happen with other kinds of vocalizations. Still considering the same species, King et al. [5] verified the role of vocalizations in complex social abilities and relationships. For another species of dolphins, the Indo-Pacific humpback dolphin, Caruso et al. [6] recorded and classified echolocation vocalizations to investigate the distribution and acoustic behavior of the species.

Examples of recent research related to PAM in the aquatic environment is the segmentation of dolphin vocalizations [7], aiming at finding regions of interest in long-time audio files, a task that is costly and time consuming when performed by humans. The use of acoustic indicators to monitor coral reefs [8] is another example of research on PAM in the aquatic environment. In another vein, Rountree et al. [9] highlighted the importance of acoustic monitoring in freshwater, an area in which there are many research gaps to be explored.

Global warming, industrial fishing, oil spills, and other factors cause much damage to and many changes in the environment of marine mammals. By virtue of this, it is very important to supervise marine mammals. The practical use of species identification was applied to the North Atlantic right whales, focusing on environmental conservation. Collisions between ships and these animals are one of the main threats to this species [10]. In order to avoid these collisions, in 2013, there was an international challenge [11] to automatically identify if a given audio contained or did not the vocalizations of such species. The data from the challenge were collected by floating buoys. This kind of recognition can help ships change their route to avoid a possible collision. Beyond this challenge, there are other research works on North Atlantic right whales' identification in the scientific literature [12–14]. In fact, the situation of this species is so critical, that another challenge was proposed in 2015 [15]. In this challenge, the database was composed of aerial images from these animals, and the classification task was to identify each individual, to help researchers track the health and general status of the individuals, focusing on conservation efforts.

Although PAM techniques and machine learning are useful in marine research, there are two potential hazards concerning machine learning that we may note in systems related to species classification. The first one is the risk of using vocalizations of the same individuals both in the training and test sets simultaneously; this may lead the system to be able to recognize the individuals instead of the species. Not only the vocalization of the same individual can bias the system, but even a characteristic noise presented in several samples of a class can be distinguished from noises of other classes, which is the second hazard. This can happen when samples of the same class are recorded in the same location with the same devices; the environment creates noise, and the recording devices may distinguish the audio files. Based on these issues and aiming at more reliable results for PAM systems, we proposed the concept of PAM filters, which means trying to use the same individual always in the same set, whether it be the training or test set.

In the database we used, it was possible to separate the individuals from the same class by location and date of record, trying our best to avoid the recognition of individuals, devices, and noise. Experiments were also conducted with a randomized version of the database, and the results were fairly disparate: as we expected, the accuracy rates decreased when the filters proposed here were applied, indicating the influence of other aspects of the audio in the learning system. The main result of our investigation was to collect a database where a very reliable testing protocol was created, and we made it all available, so other researchers can validate their own classification systems.

The rest of this document is organized as follows: in Section 2, we present the database, the novel idea, the concepts related to this research, and the methodology. The results are presented in Section 3 and discussed in detail in Section 4. Final remarks are presented in Section 5.

## 2. Materials and Methods

In this section, we describe the database used for the experimentation (Section 2.1), the novelty (Section 2.2), and the creation of the protocols and their peculiarities (Section 2.3). Furthermore, we present the theoretical concepts of the research in Section 2.4, and with these all these discussed, the methodologies of the experiments are presented in Section 2.5.

### 2.1. Watkins Marine Mammal Sound Database

Marine mammals are an informal group of animals that rely on a marine ecosystem for their existence. According to the taxonomy committee from The Society for Marine Mammalogy [16], marine mammals are classified into three orders and several families, genera, species, and subspecies; some of them may already be extinct. In this work, we used the Watkins Marine Mammal Sound Database (WMMSD) [17]. The audio files were recorded from the 1940s to the 2000s, and the species in the audio files were identified by biologists. Recently, this database has been gaining attention from the PAM community [18,19].

The database is composed of almost 1600 entire tapes. Each tape is composed of several minutes of recording, and they may contain vocalizations of several species. Smaller cuts of these long-length audio files are available on the website, usually with vocalizations of only one species. They are divided into two sections on the website, "all cuts" and "best cuts". "Best cuts" represents high-quality and low-noise cuts. "All cuts" contains all the audio files from the "best cuts" plus other ones, which are lower in quality and noisier.

We chose to use the "best cuts" for the classification task, as noise reduction and segmentation were not the focus of this work. "Best cuts" contains 1694 audio files from 32 species of marine mammals. There are 25 samples that contained vocalizations of more than one species, and those were removed as we did not intend to handle multilabel classification, nor audio segmentation.

The website also provides a metadata file for each audio cut. The data contain additional information, such as the date and the location of the record. However, most metadata files do not present information for all their fields. Furthermore, several classes have most of their samples recorded for just a few locations and dates. This led us to suspect that they may contain samples of the same individual. Furthermore, the noise pattern in such samples was homogeneous, so using the metadata files, it was possible to see that the cuts were extracted from the same long-length tapes.

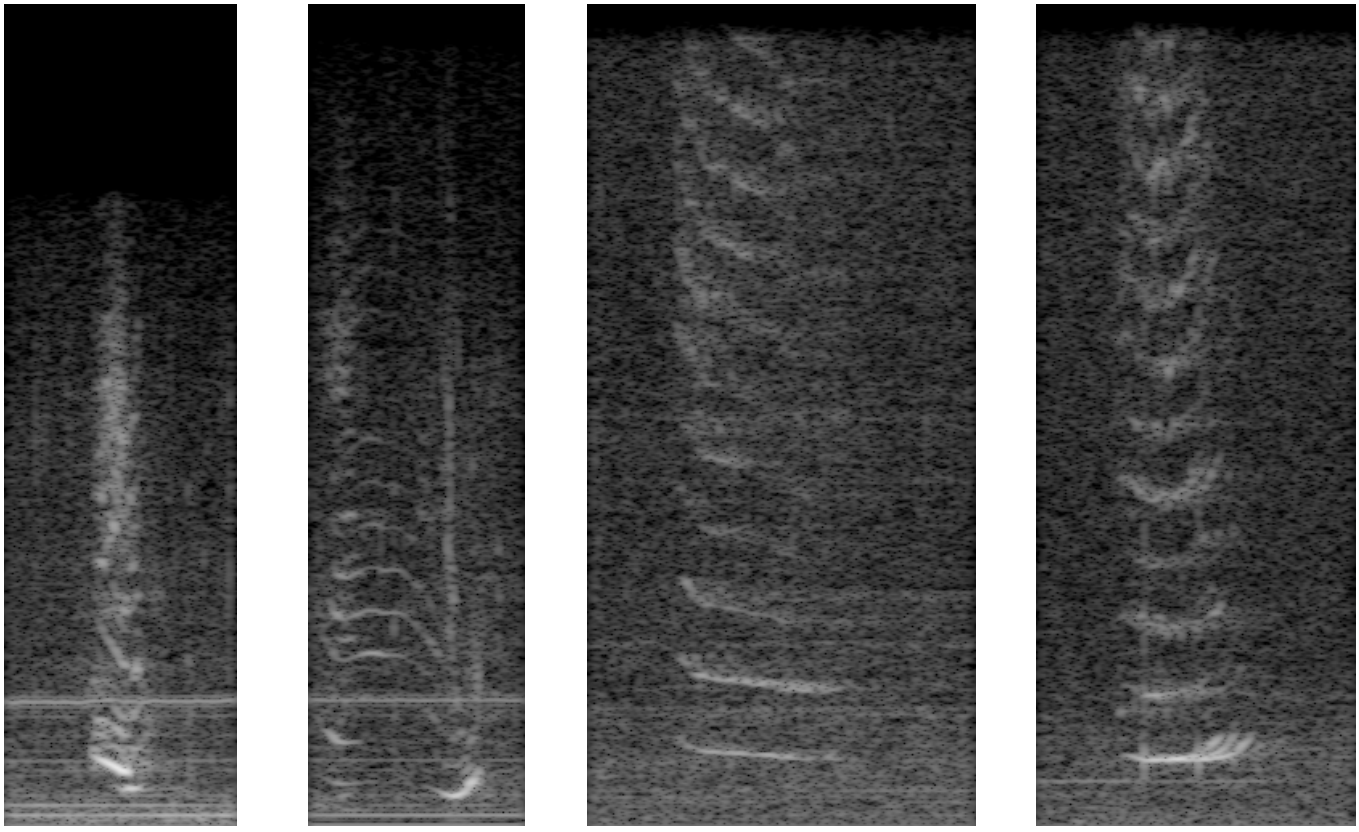
Other research found that the recent scientific literature used the same database for species classification [18,19]. Trawicki [18] to discriminate species of the infraorder Cetacea using Mel-frequency cepstral coefficients and hidden Markov models. The database used in that work was composed of 521 samples from 11 classes, and the highest accuracy rate obtained was 84.11%. Lu et al. [19] performed the classification and identification of marine mammals using deep and transfer learning. The database used was composed of only three classes, and the number of samples was uncertain, as they used segments obtained from large audio files. The highest accuracy obtained was 97.42%. Although some high-accuracy rates were achieved in those works reported in the literature, it is worth mentioning that none of those research works demonstrated any concern about the division of the database aiming to avoid, for example, the occurrence of samples taken from the same individuals, both in the training and test sets. In fact, this concern was not even expected in those works, since they did not take into account the PAM filter analysis, which is introduced in this work in Section 2.2.

### 2.2. Pam Filter

In another audio classification task, music genre classification, Flexer [20] defined the concept of the "artist-filter", which means to have all samples of the same artist either in the training or test set. The author noticed that experiments with samples from the same artists in the training and test sets presented higher accuracies and lower standard deviations, which suggested that music genre classification systems were learning the artist instead of the genre.

We considered that experiments with randomized sets for training and testing may produce overestimated results of species classification, because the classifiers may be making decisions based on underlying patterns. This is based on the information presented in the metadata files of the WMMSD, which indicates that there are many samples of the same individuals and, also, the same pattern of noise was present in several samples of the same class. The concept of "artist-filter" led us to conceive of "PAM filters", where we tried to use the same individual's vocalization either in the training or test set, never in both them. In our experiments, this was achieved by separating the files by location and date of recording. As a consequence, we also separated the noise pattern of the environment, which can be easily distinguished. Furthermore, we probably separated the devices used for recording.

Figure 1 presents four different vocalizations of the species *Eubalaena glacialis*. All the spectrograms were generated with the same parameters. According to the metadata files, Figure 1a,b were extracted from the same long-length tape, recorded on the same day, with the same devices. Vertical lines that represent the same noise in the lower area of Figure 1a,b are discernible. The vocalization in Figure 1a is described as "grunt", and in Figure 1b, it is described as "one long groan".



(**a**) Recorded in 1956 (**b**) Recorded in 1956 (**c**) Recorded in 1981 (**d**) Recorded in 1981

**Figure 1.** Different vocalizations of *Eubalaena glacialis*.

Figure 1c,d also share the same tape, date, and devices. Both vocalizations were described as "moan". Although noise was not visible in the vertical lines, nor reported in the metadata, a texture pattern resembling "salt and pepper" was present all over the spectrogram, probably generated by the sound of the ocean.

### 2.3. Watkins Experimental Protocols

To investigate the impact of PAM filters, we created three different protocols of the same database, with and without PAM filters. They were called Watkins Experimental

Protocols (WEPs). WEP#1 did not have PAM filters applied; on the other hand, WEP#2 and WEP#3 did. They are described in Sections 2.3.1–2.3.3, and their specifications are available online [21]. It is important to remark that we tried to use as many samples and classes of the database as possible, but due to the limitations of the locations and dates, the protocols had different numbers of samples and classes.

#### 2.3.1. Watkins Experimental Protocol #1: Ten-Fold Cross-Validation

The first protocol was defined without concerning the PAM filter. We used the database randomly divided into ten folds for cross-validation, and classes with fewer than ten samples were removed. Table 1 presents the species, the number of samples used in WEP#1, the number of locations where the samples were recorded, and the number of samples present at each location.

**Table 1.** Composition of WEP#1: 31 species. Columns also present the number of locations were the samples were recorded, the number of samples recorded at each location, and the number of samples per class.

| Species | Number of Locations | Samples by Location | Number of Samples in WEP#1 |
|---|---|---|---|
| *Balaena mysticetus* | 2 | 1; 49 | 50 |
| *Balaenoptera acutorostrata* | 1 | 17 | 17 |
| *Balaenoptera physalus* | 2 | 5; 45 | 50 |
| *Delphinapterus leucas* | 4 | 1; 6; 16; 27 | 50 |
| *Delphinus delphis* | 3 | 2; 15; 35 | 52 |
| *Erignathus barbatus* | 3 | 3; 9; 15 | 27 |
| *Eubalaena australis* | 2 | 7; 18 | 25 |
| *Eubalaena glacialis* | 4 | 3; 12; 19; 20 | 54 |
| *Globicephala macrorhynchus* | 4 | 5; 16; 18; 26 | 65 |
| *Globicephala melas* | 4 | 11; 12; 14; 28 | 65 |
| *Grampus griseus* | 3 | 1; 21; 45 | 67 |
| *Hydrurga leptonyx* | 1 | 10 | 10 |
| *Lagenodelphis hosei* | 1 | 87 | 87 |
| *Lagenorhynchus acutus* | 3 | 12; 12; 31 | 55 |
| *Lagenorhynchus albirostris* | 2 | 20; 37 | 57 |
| *Megaptera novaeangliae* | 3 | 1; 17; 46 | 64 |
| *Monodon monoceros* | 3 | 4; 10; 36 | 50 |
| *Odobenus rosmarus* | 3 | 1; 16; 21 | 38 |
| *Ommatophoca rossi* | 3 | 11; 19; 20 | 50 |
| *Orcinus orca* | 5 | 1; 2; 5; 8; 19 | 35 |
| *Pagophilus groenlandicus* | 1 | 47 | 47 |
| *Peponocephala electra* | 1 | 56 | 56 |
| *Physeter macrocephalus* | 6 | 2; 2; 2; 9; 12; 33 | 60 |
| *Pseudorca crassidens* | 2 | 11; 48 | 59 |
| *Stenella attenuata* | 2 | 11; 54 | 65 |
| *Stenella clymene* | 2 | 14; 49 | 63 |
| *Stenella coeruleoalba* | 4 | 8; 12; 27; 34 | 81 |
| *Stenella frontalis* | 1 | 58 | 58 |
| *Stenella longirostris* | 2 | 1; 113 | 114 |
| *Steno bredanensis* | 1 | 50 | 50 |
| *Tursiops truncatus* | 3 | 1; 10; 13 | 24 |
| Number of samples | | | 1645 |

#### 2.3.2. Watkins Experimental Protocol #2: Training/Test Protocol

As listed in Table 1, some species, such as *Balaenoptera acutorostrata*, had all its records in the same place. This indicated that the samples might contain the vocalizations of the same individual, and also, the noise pattern generated by the environment and the devices were usually similar. Such species were removed.

To create an experimental protocol using the PAM filters, we scanned the metadata files to separate the database according to the locations, and each location of each class was allocated exclusively into the training or the test set.

Furthermore, in other cases, all samples from the same class were recorded at only a few different locations. For species where the samples belonged to only two locations, the location with more samples went to the training set and the second location to the test set. Classes with three or more locations were distributed, trying to achieve a 70% for training and 30% for testing split.

Table 2 presents the data used in WEP#2 and WEP#3. Classes with an unfilled number of samples were not in the protocol. WEP#2 is composed of 24 classes, with 908 samples for training and 412 for testing.

**Table 2.** Data used in WEP#2 and WEP#3. Classes and the number of samples of the training and test set of WEP#2 and the two folds of WEP#3. Unfilled cells mean the class was not used, while 24 classes were used in WEP#2 and 20 in WEP#3.

| Species | Number of Samples | | | |
|---|---|---|---|---|
| | WEP#2 | | WEP#3 | |
| | Train | Test | Fold 1 | Fold 2 |
| *Balaena mysticetus* | 49 | 1 | - | - |
| *Balaenoptera acutorostrata* | - | - | - | - |
| *Balaenoptera physalus* | 45 | 5 | - | - |
| *Delphinapterus leucas* | 27 | 23 | 23 | 27 |
| *Delphinus delphis* | 35 | 17 | 35 | 17 |
| *Erignathus barbatus* | 15 | 12 | 15 | 12 |
| *Eubalaena australis* | 18 | 7 | - | - |
| *Eubalaena glacialis* | 31 | 23 | 31 | 23 |
| *Globicephala macrorhynchus* | 34 | 31 | 31 | 34 |
| *Globicephala melas* | 37 | 28 | 28 | 37 |
| *Grampus griseus* | 45 | 22 | 45 | 22 |
| *Hydrurga leptonyx* | - | - | - | - |
| *Lagenodelphis hosei* | - | - | - | - |
| *Lagenorhynchus acutus* | 31 | 24 | 31 | 24 |
| *Lagenorhynchus albirostris* | 37 | 20 | 37 | 20 |
| *Megaptera novaeangliae* | 46 | 18 | 46 | 18 |
| *Monodon monoceros* | 36 | 14 | 36 | 14 |
| *Odobenus rosmarus* | 21 | 17 | 21 | 17 |
| *Ommatophoca rossi* | 30 | 20 | 20 | 30 |
| *Orcinus orca* | 19 | 16 | 19 | 16 |
| *Pagophilus groenlandicus* | - | - | - | - |
| *Peponocephala electra* | - | - | - | - |
| *Physeter macrocephalus* | 33 | 27 | 33 | 27 |
| *Pseudorca crassidens* | 48 | 11 | 11 | 48 |
| *Stenella attenuata* | 54 | 11 | 11 | 54 |
| *Stenella clymene* | 49 | 14 | 14 | 49 |
| *Stenella coeruleoalba* | 42 | 39 | 42 | 39 |
| *Stenella frontalis* | - | - | - | - |
| *Stenella longirostris* | 113 | 1 | - | - |
| *Steno bredanensis* | - | - | - | - |
| *Tursiops truncatus* | 13 | 11 | 13 | 11 |
| Sums | 908 | 412 | 542 | 539 |

#### 2.3.3. Watkins Experimental Protocol #3: Two-Fold Cross-Validation

The third and last protocol was two-fold cross-validation. It was created to use all the samples either for training or testing, improving the reliability of the results. We could not increase the number of folds due to the number of locations, since eight classes had all their samples recorded at only two locations (see Table 1), so increasing the number of folds to three would result in a major cut of the classes.

However, a minor cut of the classes was still necessary. For example, the class *Stenella longirostris* had recordings taken from two locations, one of them with one sample and the other one with one-hundred thirteen. If it was used in the cross-validation, there would be one-hundred thirteen samples of a class tested in a model trained with just one sample of the class, an unfair task. Empirically, we decided to remove classes that held fewer than ten samples in either one of the folds. Table 2 presents the details of WEP#3. It used 20 class. These two folds held 542 and 539 samples each.

### 2.4. Theoretical Framework

In this section, we present the theoretical information about the experimental protocol used in our experiments. First, we describe how the audio signal was manipulated and the spectrograms generated (Section 2.4.1). Then, we detail the feature extraction (Section 2.4.2) and the classifiers with which we experimented (Section 2.4.3). We also discuss the deep learning architecture used (Section 2.4.4) and, lastly, the evaluation metric calculated to compare the results (Section 2.4.5).

#### 2.4.1. Signal

Several research works that addressed audio classification performed the feature extraction in the visual domain, typically using spectrogram images. Investigations have already been developed to handle tasks such as infant crying motivation [22], music genre classification [23], and music mood classification [24]. The visual domain has also been used with animal vocalizations, in tasks such as species identification and detection [13,25].

Spectrograms are time–frequency representations of a signal, and they can be plotted into an image. From a digital audio, a spectrogram can be generated using the Discrete Fourier Transform (DFT). It shows the intensity of the frequency values as time varies. An example of a spectrogram image of a vocalization of a marine mammal is presented in Figure 2. The $X$-axis represents time, the $Y$-axis information about the frequency, and the $Z$-axis (i.e., the color intensity of the image pixels) the intensity of the signal.

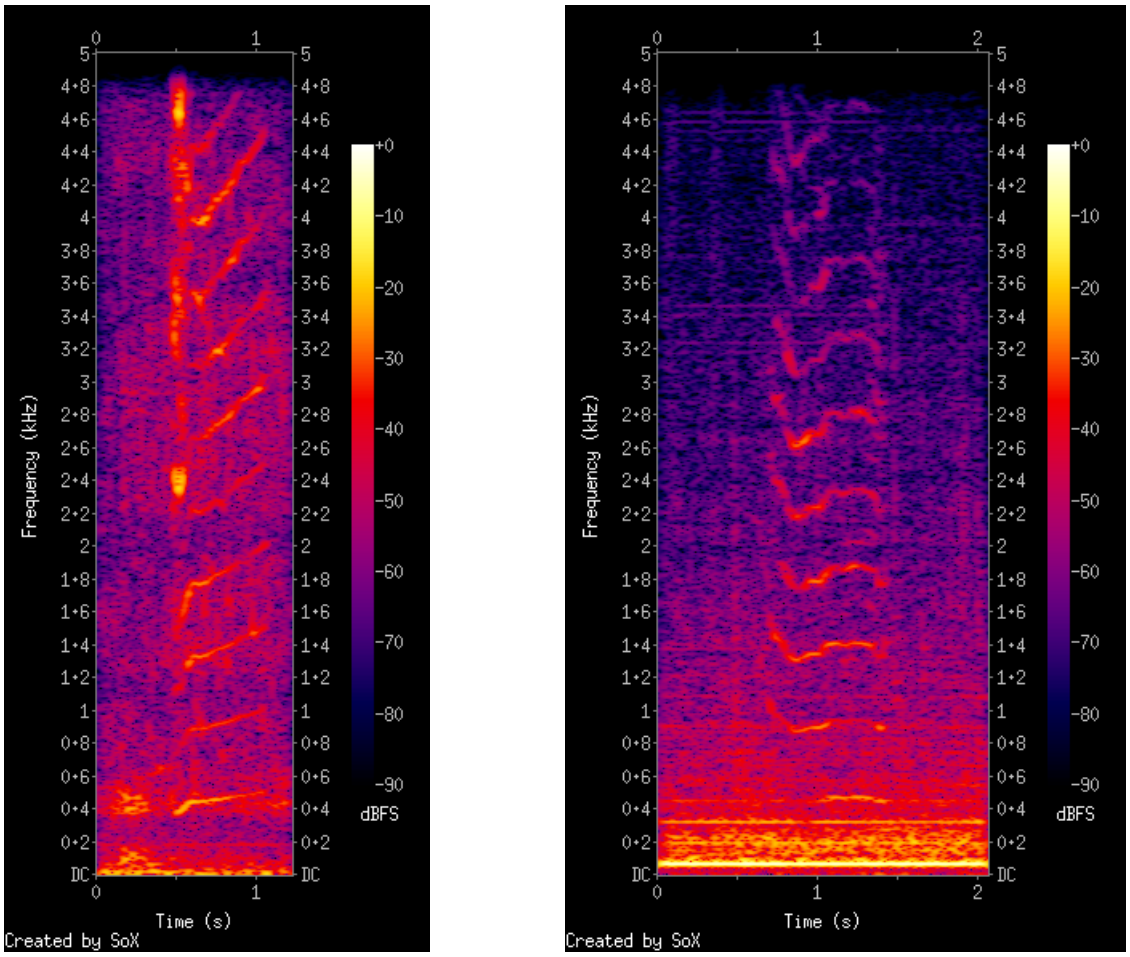
The spectrogram of Figure 2a represents a vocalization of a *Delphinapterus leucas* individual. The audio was a bit more than one second long. Its metadata file described it as a "moan". It was recorded in 1965, from Coudres Island, Canada. Figure 2b represents a vocalization of a *Eubalaena glacialis* individual. The audio was a bit longer than two seconds. The metadata also described the vocalizations as "moan". It was recorded on the coast of Massachusetts, USA, in 1959.

#### 2.4.2. Features

Texture is an important visual attribute in digital images. In the case of spectrograms in particular, texture is a very prominent visual property. In this vein, the textural content of spectrograms has been used in several audio classification tasks, such as music genre classification [26], voice classification [27], bird species classification, and whale recognition [13].

In [28], the authors proposed the Local Binary Pattern (LBP). The texture of an image is described using a histogram. Each cell of the histogram holds the number of occurrences of a binary pattern. One binary pattern is calculated for each pixel and is based on its neighborhood. Equation (1) is computed for each pixel for the extraction of the LBP histogram.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad (1)$$




(**a**) *Delphinapterus leucas* (**b**) *Eubalaena glacialis*

**Figure 2.** Examples of the spectrograms of the WMMSD.

The parameter $P$ represents the number of neighbor pixels to be taken into account, and $R$ stands for the radius, the distance between the pixel and its neighbors. $g_c$ and $g_p$ stand for the gray level of the central pixel (i.e., the pixel for which the LPB has been calculated) and the gray level of one neighbor, respectively. The function $s$ is defined in Equation (2).

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2}$$

An example of the computation of the LBP in one pixel of an image is illustrated in Figure 3, considering the parameters $P = 8$ and $R = 1$.

The binary pattern begins in the top left and goes clockwise. The pattern of the central pixel of Figure 3 is defined as the sequence of bits:

$$(01011011)_2.$$

It is possible to conceptualize the binary pattern as a decimal number. This is computed by applying the sum of binary digits times their power of two ($2^n$):

$$(0 \times 128) + (1 \times 64) + (0 \times 32) + (1 \times 16) + (1 \times 8) + (0 \times 4) + (1 \times 2) + (1 \times 1) = 91_{10};$$

therefore, the decimal number of the pattern from Figure 3 is 91.

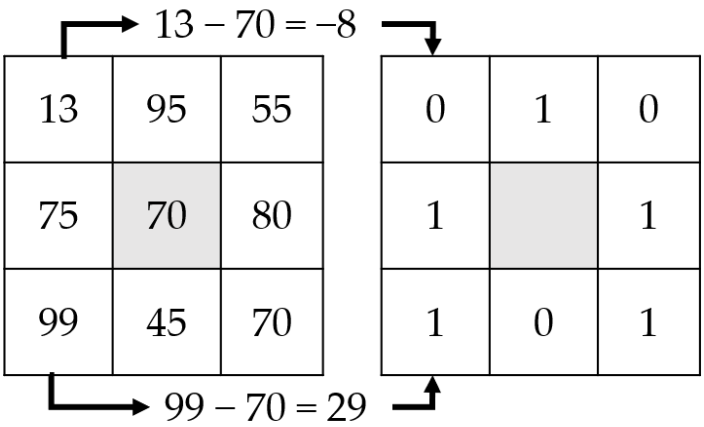



**Figure 3.** Example of the computation of the LBP in a grayscale image. Adapted from [29].

Changes in the parameter $P$ imply changes in the number of features: eight neighbors' binary described generate 256 possible patterns ($2^8$). LBP with $P = 8$ presents several nonuniform patterns, which are binary sequences that present more than two bitwise changes, for example: the binary pattern in Figure 3 is nonuniform; it presents six bitwise changes. Usually, LBP is used with $P = 8$, $R = 2$, and gathering together all the nonuniform patterns into just one feature, it results in 59 features. However, Ojala et al. [30] observed that nonuniform patterns do not contain the fundamental properties of texture, and they suggested to sum up all the nonuniform patterns in the same feature.

#### 2.4.3. Classifiers

To analyze the impact of the PAM filters, we tried several classifiers to observe their performances. The first classifier considered in this work was the well-known K-Nearest Neighbors (KNNs). This is an instance-based algorithm, which means it does not produce a model; it only stores the instances of training. During testing, the algorithm computes the distance between each test sample to each training sample. The prediction of a test sample is based on the classes of the nearest neighbors. The parameters of KNN are the distance metric and $K$, the number of neighbors. Other simpler classifiers used in our experiments were Naive Bayes (NB), a probabilistic classifier that is known to assume the independence of the features, and Decision Trees (DTs), a tree composed of conditional statements created using information gain or impurity metrics.

Another classifier selected for the research was the Support Vector Machine (SVM), a binary classifier proposed by Cortes and Vapnik [31]. It can be easily applied to multi-class problems using one of the following strategies: one-vs.-one or one-vs.-all. Three considerable benefits of the SVM are the kernel method, the maximum-margin hyperplane, and the soft margin. The kernel method consists of mapping the feature space into another feature space, dimensionally higher, where the data can be linearly separable. The maximum-margin hyperplane divides the data by taking into account that the nearest points on each side are equally distant from the hyperplane. The soft margin is a technique to reduce overfitting, and it treats the $C$ samples nearest to the margin as outliers and, with that, increases the distance between the classes. The parameters of SVM are $C$ and the kernel function (which may have its own parameters).

Ensembles of classifiers based on decision trees were also used here, as they were recently considered the state-of-the-art in many machine learning tasks [32]. These are bagging, Random Forest (RF), Extremely Randomized Trees (ERTs), AdaBoost (AB), and Gradient Boosting (GB). The classifier bagging builds several instances of classifiers from random subsets of the training database. RF combines the concept of bagging, but with the idea of random subsets of features per classifier. ERTs are similar to random forest, the differences being in the choice of the attributes and in the definition of cut-points, which

are fully random. AB generates new classifiers by increasing the weight of samples that were wrongly classified by the previous models. Then, the outcome is obtained by the weighted predictions of all created models. GB is similar to AB, but the new classifiers are created using only the residual error from the previous classifier.

#### 2.4.4. Deep Learning

To diversify the experiments of this work, we also executed deep learning tests. We used a pretrained neural network fine-tuned with the training samples, similar to what was implemented by Lu et al. [19]. As is common in convolutional neural networks, we used the spectrograms as the input. The deep learning model was used here in such a way that it provided features, in a non-handcrafted fashion, and it also performed the classification. Therefore, in the deep learning experiments, the classifiers described in Section 2.4.3 and the features described in Section 2.4.2 were not applicable.

Residual Networks-50 layers (ResNet-50) was proposed by He et al. [33] as a prize-winning network in competitions on detection and localization in images. The authors introduced the concept of the shortcut connection to neural networks, where the output of a layer is connected to the input of more than one of its following layers. ResNets are easier to train because the jump connections between two layers help the gradient to flow and reduce the probability of exploding or vanishing gradients. In this investigation, we used ResNet-50; however, He et al. [33] tested up to ResNet-152, and it outperformed ResNet-50; yet, the gain in accuracy was small when compared to the additional computational costs.

#### 2.4.5. Evaluation Metric

To evaluate the results of our experiments, we used the accuracy metric. It is described in Equation (3).

$$Acc = \frac{\#CP}{\#samples}, \tag{3}$$

where $\#CP$ stands for the number of samples correctly predicted and $\#samples$ indicates the total number of tested samples, the ones that were correctly and wrongly predicted.

### 2.5. Experimental Methodology

The database, the novelty, the experimental protocols, and the theoretical concepts of this investigation were previously presented in Sections 2.1–2.4, so now, we describe each step of our research based on such explanations. First, we present the non-deep learning experiments of this work, and they are illustrated in Figure 4.

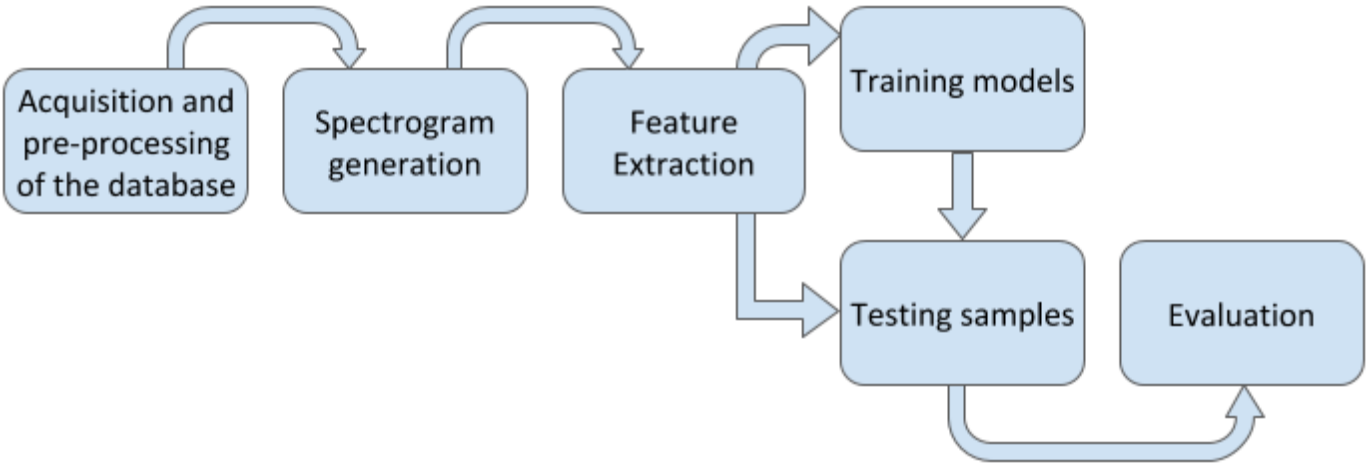



**Figure 4.** Illustration of the general methodology of this work.

The first step was to acquire the database. WMMSD is available online, but the download is exclusively sample-by-sample. A crawler script was developed to download

all the samples of the database and the metadata files. After that, as a preprocessing task, all the samples were converted to the same sample rate, 22,050 Hz. To accomplish this task, we used LibRosa [34] (version 0.6.3; python package for music and audio analysis, McFee, Brian et al.; Austin, TX, USA). The audio files were converted to spectrogram images using the software SoX (version 14.4.2, maintained by Chris Bagwell et al., Dallas, TX, USA) [35]. The features of the LBP were extracted with the software library Scikit-Image (version 0.16, created by van der Walt et al., Berkeley, CA, USA) [36]. The features from the training sets were used to create a model with the library Scikit-Learn (version 0.22, Pedregosa et al., Montreal, QC, Canada) [37]. The features of the testing samples were, then, predicted by the model. Finally, the accuracy rates were calculated.

The deep learning methodology is slightly different. A specific representation of it is presented in Figure 5. A pretrained ResNet-50 was fine-tuned with the spectrograms of the training samples (the same generated in Figure 4). After that, the spectrograms of the testing samples were predicted with the neural network. The deep learning experiments were carried out using MATLAB (version 2019b, The MathWorks, Inc., Natick, MA, USA) [38].

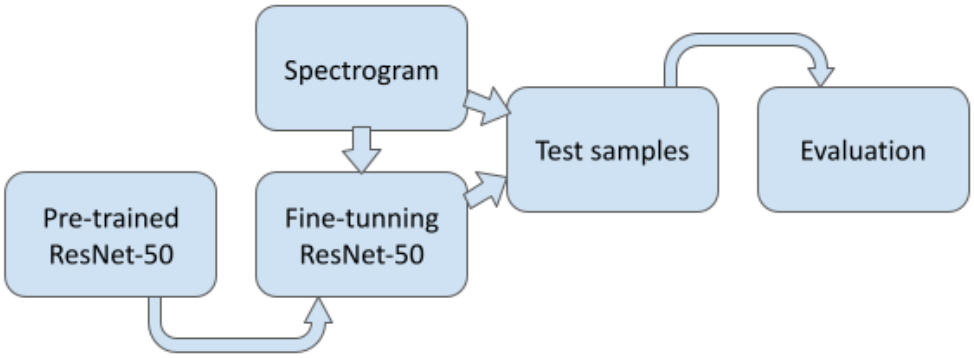



**Figure 5.** Illustration of the methodology of the deep learning experiments.

All the experiments were conducted in the three Watkins experimental protocols. For WEP#1, the protocol composed of 31 classes, samples were randomly divided into ten folds for cross-validation, without using the same individuals or the same noise pattern in the training and test sets. Both WEP#2 and WEP#3 involved PAM filters. WEP#2 had 24 classes and was a training and test protocol, and WEP#3 had 20 classes and was a two-fold cross-validation protocol. The number of classes and of samples were not the same due to the limitations observed in Table 1.

## 3. Results

Table 3 presents the results obtained with the three protocols presented in Section 2.3, WEP#1 without PAM filters and WEP#2 and #3 with PAM filters. We tried to use as much of the database as possible, but the lack of variety in the locations and dates of the samples imposed different limitations on each protocol, as shown in Tables 1 and 2.

The best results found using the WEP#1 were with deep learning and the classifier SVM, 78.10%$\sigma$2.73 and 64.62%$\sigma$3.49, respectively. The protocol had 1645 samples from 31 classes. It was randomly divided into ten folds for cross-validation and did not apply the PAM filter.

On the other hand, the protocols with the PAM filters presented much lower results, with or without deep learning. The best results with WEP#2 were 24.27% and 19.17%, and it was a training/testing protocol. Although it had fewer samples than WEP#1, 1320, it also had fewer classes, 24. These results were achieved with deep learning and the KNN classifier.

The two-fold cross-validation protocol with PAM filters, WEP#3, also obtained its best results with deep learning and the KNN (but with a different number of neighbors). The best accuracies were 21.38%$\sigma$1.82 and 18.86%$\sigma$5.16. WEP#3 contained fewer samples than WEP#1, 1081, and it also dealt with fewer classes, 20.

**Table 3.** Experimental results in the WMMSD. Features extracted with the LBP and several different classifiers. Experiments with a deep learning architecture.

| Classifiers | Parameters | Acc. and Standard Deviation (When Applicable) | | |
|---|---|---|---|---|
| | | WEP#1 (31 Classes) | WEP#2 (24 Classes) | WEP#3 (20 Classes) |
| NB | Gaussian naive Bayes | 38.78%$\sigma$3.76 | 18.45% | 16.56%$\sigma$2.29 |
| DT | Splitting with entropy | 39.89%$\sigma$3.82 | 11.65% | 12.39%$\sigma$0.47 |
| | Splitting with Gini impurity | 43.25%$\sigma$2.80 | 10.44% | 10.36%$\sigma$0.74 |
| KNN | Manhattan distance, K = 1 | 59.53%$\sigma$4.51 | 18.45% | 17.02%$\sigma$1.76 |
| | Manhattan distance, K = 3 | 61.03%$\sigma$3.64 | 16.02% | 17.29%$\sigma$3.46 |
| | Manhattan distance, K = 5 | 61.08%$\sigma$3.13 | 17.48% | 17.38%$\sigma$4.38 |
| | Manhattan distance, K = 7 | 59.54%$\sigma$3.36 | 16.50% | 17.38%$\sigma$4.90 |
| | Manhattan distance, K = 11 | 59.13%$\sigma$4.06 | 17.48% | 16.09%$\sigma$4.12 |
| | Euclidean distance, K = 1 | 57.16%$\sigma$4.32 | 19.17% | 17.20%$\sigma$2.03 |
| | Euclidean distance, K = 3 | 57.83%$\sigma$3.56 | 17.72% | 17.20%$\sigma$3.33 |
| | Euclidean distance, K = 5 | 58.20%$\sigma$3.60 | 18.20% | 18.86%$\sigma$5.16 |
| | Euclidean distance, K = 7 | 57.53%$\sigma$3.46 | 17.96% | 17.94%$\sigma$5.42 |
| | Euclidean distance, K = 11 | 56.29%$\sigma$3.22 | 17.72% | 17.38%$\sigma$4.39 |
| SVM | Grid search | 64.62%$\sigma$3.49 | 13.35% | 13.69%$\sigma$1.36 |
| Ensembles | AdaBoost | 12.09%$\sigma$3.55 | 01.46% | 09.62%$\sigma$0.22 |
| | Bagging | 50.75%$\sigma$3.72 | 13.11% | 13.13%$\sigma$1.26 |
| | Extremely randomized trees | 46.91%$\sigma$4.66 | 16.02% | 14.24%$\sigma$1.51 |
| | Gradient boosting | 54.54%$\sigma$3.78 | 12.38% | 11.84%$\sigma$0.31 |
| | Random forest | 47.29%$\sigma$4.51 | 13.11% | 15.07%$\sigma$2.43 |
| Deep learning | ResNet-50 | **78.10% $\sigma$2.73** | **24.27%** | **21.38% $\sigma$1.82** |

The results indicated that it was more likely to achieve better accuracies without PAM filters, since WEP#1 presented accuracies much higher than the other two protocols, WEP#2 and WEP#3, which applied PAM filters. Therefore, the outcome corroborated the hypothesis that PAM based on machine learning can be biased by individuals, the devices used in the recording, and the noise pattern. In fact, we could easily notice the same noise pattern in the samples recorded on the same date and at the same location.

In Table 4, we present the best results obtained in Table 3, but also experimenting with the same classes of WEP#3 in the arrangements of the protocols WEP#1 and WEP#2. Therefore, the columns designated with "20 classes" present results with the same classes and samples, the difference between them being that WEP#1 was a 10-fold cross-validation protocol without PAM filters and WEP#2 was a training and testing protocol with PAM filters.

The differences between the accuracy rates within WEP#1 were more expressive with deep leaning and the KNN, when $K = 5$. For WEP#1, the experiments with a fewer number of classes presented lower accuracy rates, instead of what was expected. Probably, this happened specifically due to the 11 classes that were removed. Most of them were not suitable for the PAM filters due to the lack of variety, and this made them easier to classify. For example, one of these classes was *Steno bredanensis*: it had 50 samples, and all of them were recorded on 4 September 1985, in the Mediterranean Sea; the metadata files describe the locations of the recordings, either close to Sicily or Malta, which are close to each other.

**Table 4.** Best results obtained in Table 3, also testing exactly the same subset from WEP#3 in the protocols WEP#1 and WEP#2. The number of classes and the number of samples for training and testing also presented as $(\#training; \#testing)$. The values are approximated where there is cross-validation.

| Classifiers | Parameters | Acc. and Standard Deviation (When Applicable) | | | | |
|---|---|---|---|---|---|---|
| | | WEP#1 | | WEP#2 | | WEP#3 |
| | | 31 Classes ≈ (1481; 164) | 20 Classes ≈ (973; 108) | 24 Classes (908; 412) | 20 Classes (683; 398) | 20 Classes ≈ (541; 540) |
| KNN | Euclidean distance, K = 1 | 57.16%$\sigma$4.32 | 56.28%$\sigma$4.94 | 19.17% | 19.60% | 17.20%$\sigma$2.03 |
| | Euclidean distance, K = 5 | 58.20%$\sigma$3.60 | 52.05%$\sigma$3.08 | 18.20% | 17.84% | 18.86%$\sigma$5.16 |
| SVM | Grid search | 64.62%$\sigma$3.49 | 62.99%$\sigma$4.80 | 13.35% | 15.08% | 13.69%$\sigma$1.36 |
| Deep learning | ResNet-50 | **78.10% $\sigma$2.73** | **73.44% $\sigma$3.21** | **24.27%** | **25.38%** | **21.38% $\sigma$1.82** |

Within WEP#2, the differences between the rates were smaller, since there was more diversity in both sets of classes (20 and 24). One of the four classes that was used in the original definition of WEP#2, but not in WEP#3, was *Stenella longirostris*, which had one-hundred thirteen samples recorded from Hawaii in 1971 and just one other sample recorded in 1981 from the Atlantic Ocean, making it impracticable to use with cross-validation and PAM filters.

Yet, in Table 4, the difference between the results obtained with the same samples and classes remained similar to what was observed in Table 3. Randomizing a database of vocalizations of marine mammals can lead to overestimated results, if the same individuals, noise patterns, and recording devices are presented in both the training and test sets. This can even make a 10-fold cross-validation protocol less reliable.

The low accuracy rates obtained with all protocols, especially WEP#2 and WEP#3, were a consequence of a real-world database, composed of samples that were recorded over several different decades. Furthermore, some species of the database can share the same biological family or even genus, and similar species can produce similar vocalizations.

## 4. Discussion

Deep learning presented the best results for all protocols. While the SVM presented the second best accuracy in WEP#1, the KNN, a much simpler classifier, outperformed SVM in the two PAM filter protocols, WEP#1 and WEP#2. The ensemble of classifiers is considered the state-of-the-art for several problems [32], but in our experiments, they did not perform very well, especially AdaBoost. The decision tree performed poorly. Naive Bayes presented better results than the SVM in the PAM filter protocols, but also performed poorly for WEP#1. There was a huge difference between the results with and without PAM filters.

It is important to reiterate the potential of ResNet-50. One deep learning architecture outperformed all the other classifiers, and this architecture was not trained from scratch. The initial weights were generated with a general-propose image database [39] and the fine-tuning was only carried out with the vocalizations of marine mammals in the spectrogram format, similar to the idea of Lu et al. [19].

The negative impact caused by PAM filters on the accuracy rates be due to more than one reason. First, PAM filters aim to put the same individuals either in the training or test sets. Furthermore, the samples recorded on the same dates and at the same locations are likely to share the same recording devices and ambient sounds. Therefore, for WEP#1, different folds might share individuals, pattern noises, and recording devices. The application of PAM filters removed such underlying features and decreased the accuracy rates, as we expected. In other words, if the particularities of the samples are not taken into account

during the arrangement of experimental protocols, the evaluation of a PAM system might be overestimated.

Further, we performed the Friedman statistical test [40,41] on the accuracy rates obtained with each combination of the protocols and classifiers from Table 3. The test compared the means of at least three samples; it is similar to Analysis Of Variance (ANOVA), but nonparametric. In our case, the samples were the protocols of the database, two with PAM filters and the other one without them. The null hypothesis of the Friedman test was that there was no difference between the samples. The result of the Friedman test indicated that at least one of the samples (protocols) was significantly different from another one, with $\alpha < 0.05$.

At this point, to find which one was different from the others, we executed the Wilcoxon signed-rank test, in pairs of protocols. Since there were three samples, we also applied Bonferroni correction to avoid the statistical error of Type 1. Therefore, now, $\alpha_b < 0.167$. The null hypothesis of the Wilcoxon test argues that there is no significant difference between the pairs. The hypothesis was retained only with the pair WEP#2 and WEP#3, and in both comparisons with WEP#1, it was rejected.

## 5. Conclusions

PAM systems based on machine learning can be used to support several different application tasks. However, the evaluation protocol is a critical point that must be carefully crafted, so as not to perpetrate the wrong assumptions, which could compromise the system as a whole.

The novelty of this research, PAM filters, aimed to avoid the same individual, noise pattern, and record device in both the training and test sets, in order to better analyze the performance of a species classification system. In our results, the use of PAM filters decreased the accuracy rates with several different classifiers, indicating that a PAM system may learn other aspects of the audio, and this can lead to an overestimation of the system.

Unfortunately, audio wildlife databases with information such as dates, locations, individuals, and devices used in the recordings are not easily found. However, our results suggested that they must be used to create appropriated experimental protocols. The main outcome of this research was the definition of a reliable experimental protocol, which is available online with the database [21]. With that, other PAM investigations can validate their own classification systems.

In general, the accuracy rates we found with all protocols were unsatisfactory, especially the protocols with the PAM filters, and that provided us room for improvement. In future works, we will try to explore the similarity between species of the same genus with hierarchical classification, and we intend to apply noise reduction and resampling techniques to balance the database.

**Author Contributions:** Conceptualization, L.N., Y.C. and C.S.J.; methodology, R.A., G.M., L.N. and C.S.J.; software, R.A. and G.M.; validation, R.A.; formal analysis, R.A.; investigation, R.A., G.M., L.N., Y.C. and C.S.J.; resources, R.A., G.M., L.N., Y.C. and C.S.J.; data curation, R.A. and G.M.; writing—original draft preparation, R.A.; writing—review and editing, R.A., L.N., Y.C. and C.S.J.; visualization, R.A, L.N., Y.C. and C.S.J.; supervision, L.N., Y.C. and C.S.J.; project administration, L.N., Y.C. and C.S.J.; funding acquisition, R.A., L.N., and C.S.J. All authors read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by Araucaria Foundation, the National Council for Scientific and Technological Development (CNPq), the Coordination of Superior Level Staff Improvement (CAPES), and the National Council of State Research Support Foundations (CONFAP).

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available on FigShare at https://doi.org/10.6084/m9.figshare.14068106.

**Acknowledgments:** The authors are grateful to the NVIDIA Corporation for supporting this research with the donation of a Titan XP GPU.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AB | AdaBoost |
| ANOVA | Analysis of Variance |
| DT | Decision Trees |
| DTF | Discrete Fourier Transform |
| ERT | Extreme Random Forest |
| GB | Gradient Boosting |
| KNNs | K-Nearest Neighbors |
| LBP | Local Binary Pattern |
| NB | Naive Bayes |
| PAM | Passive Acoustic Monitoring |
| ResNet-50 | Residual Network-50 layers |
| RF | Random Forest |
| SoX | Sound eXchange |
| SVM | Support Vector Machine |
| WEP | Watkins Experimental Protocol |
| WMMSD | Watkins Marine Mammal Sound Database |

## References

1. Bittle, M.; Duncan, A. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. In Proceedings of the Annual Conference of the Australian Acoustical Society 2013, Acoustics 2013: Science, Technology and Amenity, Victor Harbor, Australia, 17–20 November 2013; pp. 208–215.
2. Mellinger, D.; Barlow, J. *Future Directions for Acoustic Marine Mammal Surveys: Stock Assessment and Habitat Use*; Technical Report; National Oceanic and Atmospheric Administration: Washington, DC, USA, 2003.
3. Sayigh, L.S.; Tyack, P.L.; Wells, R.S.; Solow, A.R.; Scott, M.D.; Irvine, A. Individual recognition in wild bottlenose dolphins: A field test using playback experiments. *Anim. Behav.* **1999**, *57*, 41–50. [CrossRef]
4. Sayigh, L.S.; Wells, R.S.; Janik, V.M. What's in a voice? Dolphins do not use voice cues for individual recognition. *Anim. Cogn.* **2017**, *20*, 1067–1079. [CrossRef] [PubMed]
5. King, S.L.; Friedman, W.R.; Allen, S.J.; Gerber, L.; Jensen, F.H.; Wittwer, S.; Connor, R.C.; Krützen, M. Bottlenose Dolphins Retain Individual Vocal Labels in Multi-level Alliances. *Curr. Biol.* **2018**, *28*, 1993–1999.e3. [CrossRef] [PubMed]
6. Caruso, F.; Dong, L.; Lin, M.; Liu, M.; Gong, Z.; Xu, W.; Alonge, G.; Li, S. Monitoring of a Nearshore Small Dolphin Species Using Passive Acoustic Platforms and Supervised Machine Learning Techniques. *Front. Mar. Sci.* **2020**, *7*, 267. [CrossRef]
7. Kohlsdorf, D.; Herzing, D.; Starner, T. An Auto Encoder For Audio Dolphin Communication. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7. [CrossRef]
8. Dimoff, S.A.; Halliday, W.D.; Pine, M.K.; Tietjen, K.L.; Juanes, F.; Baum, J.K. The utility of different acoustic indicators to describe biological sounds of a coral reef soundscape. *Ecol. Indic.* **2021**, *124*, 107435. [CrossRef]
9. Rountree, R.A.; Bolgan, M.; Juanes, F. How Can We Understand Freshwater Soundscapes Without Fish Sound Descriptions? *Fisheries* **2019**, *44*, 137–143. [CrossRef]
10. Smithsonian Ocean Portal. North Atlantic Right Whale. Available online: https://ocean.si.edu/ocean-life/marine-mammals/north-atlantic-right-whale (accessed on 1 February 2021).
11. Kaggle Inc. The ICML 2013 Whale Challenge: Develop Recognition Solutions to Detect and Classify Right Whales for BIG Data Mining and Exploration Studies. Available online: https://www.kaggle.com/c/the-icml-2013-whale-challenge-right-whale-redux (accessed on 1 February 2021).
12. Freitas, G.K.; Costa, Y.M.G.; Aguiar, R.L. Using spectrogram to detect North Atlantic right whale calls from audio recordings. In Proceedings of the 2016 35th International Conference of the Chilean Computer Science Society (SCCC), Valparaiso, Chile, 10 October 2016–10 February 2017; pp. 1–6. [CrossRef]
13. Nanni, L.; Aguiar, R.L.; Costa, Y.M.G.; Brahnam, S.; Silla, C.N., Jr.; Brattin, R.L.; Zhao, Z. Bird and whale species identification using sound images. *IET Comput. Vis.* **2018**, *12*, 178–184. [CrossRef]
14. Nanni, L.; Costa, Y.M.G.; Aguiar, R.L.; Mangolin, R.B.; Brahnam, S.; Silla, C.N., Jr. Ensemble of convolutional neural networks to improve animal audio classification. *EURASIP J. Audio Speech Music. Process.* **2020**, *2020*. [CrossRef]
15. Kaggle Inc. Right Whale Recognition: Identify Endangered Right Whales in Aerial Photographs. Available online: https://www.kaggle.com/c/noaa-right-whale-recognition (accessed on 1 February 2021).

16. The Society for Marine Mammalogy. List of Marine Mammal Species and Subspecies. Available online: https://marinemammalscience.org/science-and-publications/list-marine-mammal-species-subspecies/ (accessed on 1 February 2021).
17. Woods Hole Oceanographic Institution. Watkins Marine Mammal Sound Database. Available online: https://cis.whoi.edu/science/B/whalesounds/index.cfm (accessed on 1 February 2021).
18. Trawicki, M.B. Multispecies discrimination of whales (cetaceans) using Hidden Markov Models (HMMS). *Ecol. Inform.* **2021**, *61*, 101223. [CrossRef]
19. Lu, T.; Han, B.; Yu, F. Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecol. Inform.* **2021**, *62*, 101277. [CrossRef]
20. Flexer, A. A Closer Look on Artist Filters for Musical Genre Classification. In Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, 23–27 September 2007; pp. 341–344.
21. Aguiar, R.; Maguolo, G.; Nanni, L.; Costa, Y.; Silla , C., Jr. Vocalization of Marine Mammals. Available online: https://figshare.com/articles/dataset/Database_of_spectrograms_of_marine_mammals/14068106 (accessed on 1 April 2021). [CrossRef]
22. Felipe, G.Z.; Aguiar, R.L.; Costa, Y.M.G.; Silla, C.N., Jr.; Brahnam, S.; Nanni, L.; McMurtrey, S. Identification of Infants' Cry Motivation Using Spectrograms. In Proceedings of the 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), Osijek, Croatia, 5–7 June 2019; pp. 181–186.
23. Nanni, L.; Costa, Y.M.G.; Aguiar, R.L.; Silla, C.N., Jr.; Brahnam, S. Ensemble of deep learning, visual and acoustic features for music genre classification. *J. New Music. Res.* **2018**, *47*, 383–397. [CrossRef]
24. Tavares, J.C.C.; Costa, Y.M.G. Music mood classification using visual and acoustic features. In Proceedings of the 2017 XLIII Latin American Computer Conference (CLEI), Cordoba, Argentina, 4–8 September 2017; pp. 1–10.
25. Merchan, F.; Guerra, A.; Poveda, H.; Guzmán, H.M.; Sanchez-Galan, J.E. Bioacoustic Classification of Antillean Manatee Vocalization Spectrograms Using Deep Convolutional Neural Networks. *Appl. Sci.* **2020**, *10*, 3286. [CrossRef]
26. Costa, Y.M.G.; Oliveira, L.; Koerich, A.; Gouyon, F. Music Genre Recognition Using Spectrograms. In Proceedings of the International Conference on Systems, Signals and Image Processing, Sarajevo, Bosnia and Herzegovina, 16–18 June 2011.
27. Montalvo, A.; Costa, Y.M.G.; Calvo, J.R. Language identification using spectrogram texture. In *Iberoamerican Congress on Pattern Recognition*; Springer: Montevideo, Uruguay, 2015; pp. 543–550.
28. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. *Pattern Recognit.* **1994**, *1*, 582–585.
29. Lakshmiprabha, N.S. Face Image Analysis using AAM, Gabor, LBP and WD features for Gender, Age, Expression and Ethnicity Classification. *arXiv* **2016**, arXiv:1604.01684.
30. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
31. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
32. Sagi, O.; Rokach, L. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
34. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conference, SciPy, Austin, TX, USA, 6–12 July 2015. [CrossRef]
35. Bagwell, C.; Sykes, R.; Giard, P. SoX—Sound eXchange, Version 14.4.2; The Swiss Army knife of Sound Processing Programs. Available online: http://sox.sourceforge.net/ (accessed on 1 February 2021).
36. van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. scikit-image: image processing in Python. *PeerJ* **2014**, *2*, e453. [CrossRef]
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
38. MATLAB, Version 2019b. *Programming and Numeric Computing Platform*; The MathWorks, Inc.: Natick, MA, USA, 2019.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
40. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
41. Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.* **2011**, *1*, 3–18. [CrossRef]