Underwater Classification

Giuseppe D'Avino, Mario Lezzi and Daniele Dello Russo

ABSTRACT

This study focuses on distinguishing between sounds generated by humans(Target) and those produced by fish(Non Target), with the aim of improving monitoring of marine ecosystems and mitigate the impact of anthropogenic activities. Using advanced machine learning and signal analysis techniques, an automatic classification system was developed that can accurately identify and separate anthropogenic sounds from marine fauna sounds. The methodology includes collection of a large dataset of si sample underwater recordings, extraction of distinguishing features such as frequency spectra and MFCC, and training models of Machine Learning. The training was divided into two main phases: in the first phase, it was implemented a binary classification, in which the model had to determine whether each recording of the samples contained sounds generated by humans or fish. In the second phase, a multiclass classification for sounds generated by fish. In this phase, the system was trained to recognize to which specific subclass each sample belonged.

1. Introduction

In the information age, the amount of audio data generated and collected is continuously growing, with audio representing a rich and complex source of information. However, managing and analyzing this data poses a significant challenge, especially when the volume becomes substantial. Several domains require the management and analysis of large amounts of audio data: for example, the ocean is an acoustically rich and complex environment, characterized by the continuous interaction between natural and man-made sounds. With the development of advanced underwater recording technologies, the collection of marine acoustic data has become a common practice for numerous purposes, such as scientific research, marine wildlife conservation, and monitoring of human activities. In this context, it is crucial to deepen the understanding and analysis of marine acoustic signals, in order to classify them and distinguish between sounds generated by human activities (Targets) and sounds produced by marine fauna, particularly fish (Non Targets). This distinction is of fundamental importance for a variety of reasons: from the protection and conservation of marine ecosystems, to the improvement of scientific research, to the development of innovative technologies and environmental monitoring tools. The use of advanced machine learning and acoustic signal analysis techniques provides a promising solution for more sustainable and conscious management of marine ecosystems.

The main goals of this study are as follows:

(i) Conduct a thorough review of the existing literature on acoustic signal recognition in marine environments in order to identify the essential characteristics of audio signals in this context. This step will allow the creation of a template of the essential features for the identification of sounds produced by marine fauna and anthropogenic sounds.

ORCID(s):

(ii) To develop a classification model that can distinguish marine bioacoustic sounds from audio recordings, effectively separating biological signals from those of anthropogenic origin. Next, the goal will be to improve the model to distinguish not only at the intraclass level, but also at the interclass level, with the aim of specifically identifying animal species detected in the analyzed signals.

(iii) Optimize the numerical features extracted from an audio signal to describe it in a clear and detailed manner. In this regard, various acoustic features and their informational potential in the context of scientific research and sound identification will be discussed.

These goals represent a crucial step toward developing effective tools for understanding marine acoustic signals, with significant implications for both environmental conservation and scientific progress.

2. State of the Art

Bioacoustic analysis is a rapidly expanding field, supported by the integration of new technologies,as artificial intelligence, which is revolutionizing every aspect of data analysis. Before the advent of AI, bioacoustic analysis was performed by a combination of manual techniques and traditional acoustic instruments. These methods were often laborious and required significant human intervention to collect, analyze, and interpret sound data. Clearly, prior to the introduction of the new technologies, each bioacoustic data had to be analyzed individually, treated as a case study, and submitted for review by experts in the field, significantly slowing down the signal analysis and classification steps.(6)Blumstein2011(Gibb et al., 2019; Blumstein et al., 2011). In the context of analysis and classification, the basic steps can be summarized as follows: textbfData collection: This step was carried out by audio recording and manual sampling, requiring the physical presence of experts to accurately distinguish the bioacoustic signals to be collected. **Data pre-processing** : During this phase, filtering and source cleaning was performed to reduce background noise and improve the

quality of the recorded sound, using low-pass (LowPass) and high-pass (HighPass) filters. Segmentation was performed manually to select only those elements relevant to the application context. **Sound analysis**: The collected signals were analyzed one by one, transforming them into spectrograms and oscillograms, analyzing frequency and making manual measurements. **Classification and Identification** : At this stage, dichotomous keys were used to identify and distinguish different animal species based on vocalizations, which have specific characteristics for each species. In addition, direct comparison was made with known samples belonging to certain species.
Documentation and archiving: Finally, cataloguing was carried out, associating the recordings with the various known animal species; if there was no match, the signals were classified as inconsistent bioacoustic components or suspected to belong to existing subspecies.
Limitations of pre-IA analyses included the time taken and the substantial resources needed for each case study. In addition, there was a component of subjectivity and personal interpretation by experts, as well as limited processing capacity, as experts in a given field could only work on a portion of the research at a time.(6)Blumstein2011 (Gibb et al., 2019; Blumstein et al., 2011).
Today, with the introduction of AI, several solutions have been developed to automate the previously described steps. In the field of bioacoustic signal analysis, a new approach based on the use of AI is used to automate and empirically analyze audio sources. In this context, several projects are described on which our work is based to research and develop a new system capable of performing interclass and intraclass distinction of bioacoustic signals in the underwater environment.(9) (Kahl et al., 2021).

3. Case Study: BirdNet

QutEcoacoustics/audio-analysis:QUT Ecoacoustics Analysis Programs is a software package that can perform a variety of analyses on environmental audio recordings. Although these analyses are designed for recordings of long duration (1-24 hours), they can be performed on any sample in a format supported by the software. The software is capable of:

- Calculate spectral acoustic indices and summaries at varying resolutions.
- Produce multi-index, false-color and long-period spectrograms.
- Compute critical statistics of annotations downloaded from an Acoustic Workbench.
- Perform various acoustic event recognizers.

Although this project cannot distinguish between bioacoustic and non-bioacoustic signals, it provides an important guideline on how to structure the analysis of samples related to this specific class. The AI model used is a linear classifier based on neural networks. Specifically, the code allows the construction of a linear classifier with one or two hidden layers, controlled by the parameter `hidden_units`. The model is implemented using the Keras library (imported through keras), and the training process is enhanced by a Bayesian optimization algorithm for automatic tuning of hyperparameters (using `keras_tuner.BayesianOptimization`). Our goal was to use the knowledge gained from this project as a starting point for our application. The fundamental limitation of this project is definitely the use of long duration recording to be able to clearly distinguish bioacoustic signals, moreover the dataset used for the analysis turns out to be vertical by going to consider only volatiles, finally the application does not take into account the distinction of anthropogenic sounds. Subsequent to the BirdNET project, which highlighted the effectiveness of AI in bird song recognition, an innovative approach has emerged in the field of bioacoustic analysis applied to marine monitoring. The article (11) (Malfante et al 2018).presents an advanced method for passive monitoring of ocean viability, with a particular focus on fish populations. The study develops a discriminative model based on supervised machine learning techniques, specifically Random Forest (RF) and Support Vector Machines (SVM), to classify fish sounds. The model is distinguished by the use of features extracted from the temporal, frequency and cepstral domains of acoustic signals. Tested on real sounds recorded in different marine areas, the system achieved a classification accuracy of 96.9 percent, significantly exceeding the results obtained by traditional methods.
A particularly relevant aspect of the study is the detailed approach to feature extraction. The authors explore the impact of features from the different domains, demonstrating that although each domain contains useful discriminative information, the combination of features extracted from the three domains offers superior performance. To optimize classification, a feature selection method based on feature weights in the RF model was adopted. Two subsets of features were identified: Most Valuable Features (MVF) and Valuable Features (VF). The MVF includes three key features-energy kurtosis from the frequency domain (F28), average kurtosis from the time domain (T7), and threshold crossing rate from the time domain (T15)-and achieved an average accuracy of 91.5% (RF) and 91.3% (SVM). The VF set, which includes the MVF and 16 additional features, improved the overall accuracy to 95.6% (RF) and 94.7% (SVM). These results suggest that although the use of all features may not be necessary to obtain high-quality results, the targeted selection of features allows the effectiveness of the classification system to be maintained. This is particularly useful for real-time applications with limited computational resources. In addition, the analysis shows that the importance of features can vary by class, underscoring the need for careful selection to optimize classification based

on the specific requirements of different features.(11)(Malfante et al. 2018) In summary, the integration of artificial intelligence has revolutionized bioacoustic analysis, improving efficiency and accuracy in the classification of sound signals. Projects such as BirdNET and the approach described by Malfante et al. demonstrate that the combined use of advanced techniques and targeted feature selection enables superior results in the detection and classification of sounds in both terrestrial and marine contexts. These advances highlight the potential of AI in transforming the bioacoustic analysis of sounds.

4. Dataset Structure

The dataset used in this study was divided into two main categories: Target and Non-Target(96 sub-directories). The Target(16 sub-directories) directory includes audio signals of anthropogenic origin, while the Non Target directory contains BioAcoustic signals. The records for the creation of this dataset were extracted from sources available online, with the goal of building a composite dataset to be used as the basis for the development of the artificial intelligence model. However, it is our intention in the future to use an ad hoc created dataset to improve the quality and accuracy of the model. Being a heterogeneous dataset, with a number of samples of 2663 , the technical characteristics of the audio signals vary considerably: the sampling has a frequency range from 600 Hz to 384,000 Hz, with signals in both mono and stereo formats. The signal amplitude ranges from 0 to 2, while the bit depth (bit depth) ranges from 8 bits PCM up to 32 bits. The duration of the audio signals is also highly variable, with recordings lasting from a few seconds to over 30 minutes. In addition, the dataset is unbalanced among signal categories. The manipulations and interventions required to balance and prepare the data will be described in detail in the following sections.

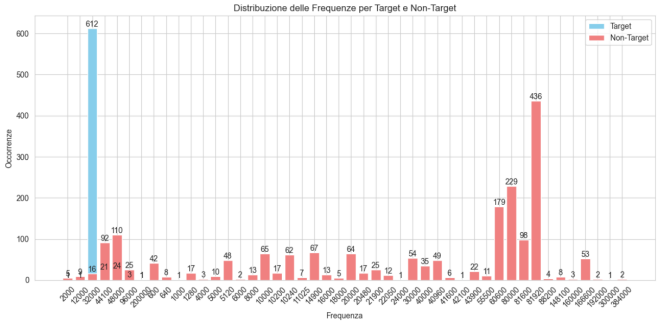

Figure 1: Target and Non-Target frequency distribution

5. Analysis

The first step in our study involved a thorough preliminary analysis of the dataset. At this stage, a thorough search and removal of duplicate files was conducted, ensuring the uniqueness of each audio signal. Then, for each sample, the following key properties were computed and recorded:

- **Amplitude** : used to evaluate the loudness of the audio signal in order to better understand the sound dynamics present in the recordings.
- **Number of Channels** : employed to determine whether the audio signals were in mono or stereo format.
- **Bit Depth**: indicating the resolution of the audio data, that is, the precision with which each sample of the signal is represented.
- **Sampling Rate**: representing the number of samples acquired per second, used to evaluate the temporal quality of the signal.
- **Duration**: measured in terms of the temporal length of each audio signal, as it varies significantly within the dataset.
- **Phase**: refers to the relative position of the starting point of an oscillation or vibration with respect to a time reference. It indicates at what point in the cycle the wave is at a given instant.

These parameters provided us with a detailed overview of the characteristics of the dataset and formed the basis for the subsequent data preprocessing steps.

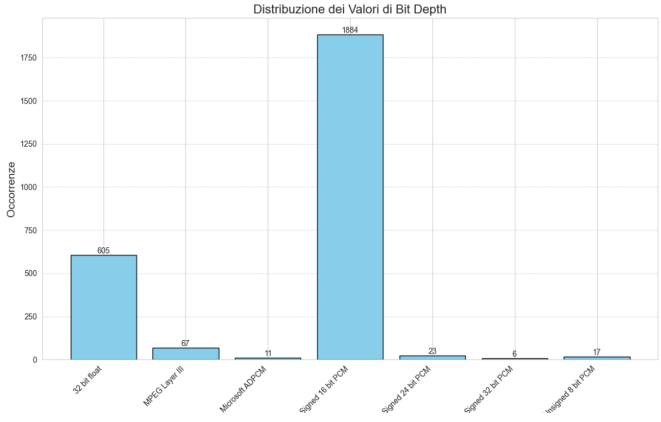

Figure 2: Bit Depth value distribution in Target

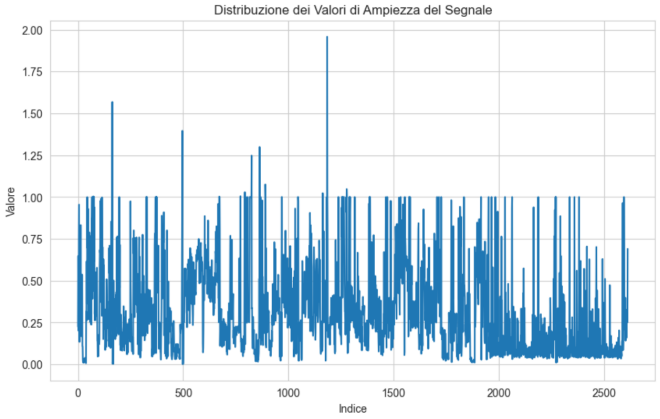

Figure 3: Signal amplitude

6. Pre-Processing

After completing the preliminary analysis, pre-processing of the audio data was performed to prepare it for model training. First, the audio signals were normalized to equalize the amplitude (between -1 and 1) and reduce extreme variations that could have compromised the performance of the model. Normalization made the signals more homogeneous and comparable. Next, all recordings were converted to a 16-bit bit depth in single-channel to standardize audio quality, ensuring uniform resolution across all data and facilitating the model learning process.

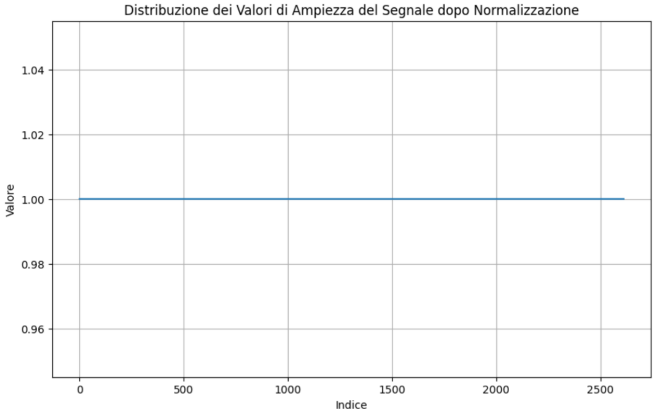

Figure 4: Signal amplitude after normalization

In addition, the sampling rate of all samples was equalized to a standard value of 96 kHz. This step was critical to ensure comparability between the different recordings and to ensure that the model received consistent inputs. The recordings were then divided into segments of uniform 4-second duration, a step that simplifies data processing and analysis. In the case of shorter recordings, periods of silence were added to achieve the desired length. The result produced is a dataset containing 50993 samples.
These pre-processing steps were essential to ensure that the data were consistent and of high quality, laying a solid foundation for the effective training of machine learning models.

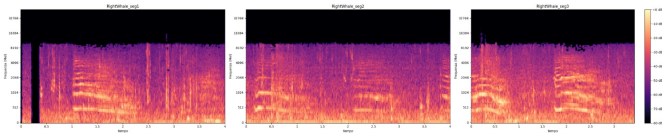

Figure 5: Spectrograms segmented audio sample

7. Methodologies

Feature extraction from audio signals is a key step in building artificial intelligence models for bioacoustic analysis. Each selected numerical feature allows describing specific properties of the signal, facilitating the distinction between bioacoustic and anthropogenic sounds, as well as the classification of sound sources. In this context, spectro-temporal features provide detailed information about the spectral content and temporal structure of signals, which are crucial for proper identification and classification.

- **Spectral Centroid Mean** The spectral centroid represents the weighted average of the frequencies present in a signal, and can be regarded as an indicator of the "brightness" of the sound. In mathematical terms, the spectral centroid is defined as:

$$\frac{1}{E}\sum_{i} i \cdot E_i$$

where i represents the position or index of an element (e.g., a frame or frequency bin), Ei represents the energy associated with that particular position i, E is the total energy of the signal, calculated as the sum of all Ei. In signal and audio applications, this can be seen as a measure of where the "energy center of gravity" is in a given domain (such as the frequency domain, for example in the audio spectrum).

- **Spectral Bandwidth RMS** The RMS spectral bandwidth measures the dispersion of frequencies around the spectral centroid and reflects the complexity of the signal:

$$RMS_i = \sqrt{\frac{1}{E}\sum_{i} i^2 E_i - i^{-2}}$$

This feature is crucial for distinguishing between simple and complex sounds, where anthropogenic sounds tend to have greater bandwidth than bioacoustic (**?**)ignalsLerch2012(Lerch, 2012).

- **Standard Deviation** The standard deviation of the spectrum describes the variation of frequencies from

the spectral centroid:

$$\sigma_s = \sqrt{\frac{1}{n-1}\sum_i (s[i] - \mu_s)^2}$$

This feature makes it possible to assess the spectral stability of the signal, which is useful for identifying sounds with frequent fluctuations typical of some anthropogenic noiseSharma2020 (Sharma et al., 2020).

- **Skewness** Skewness (Skewness) of the spectrum quantifies the symmetry of the frequency distribution around the centroid:

$$\frac{1}{n}\sum_i \left(\frac{s[i] - \mu_s}{\sigma_s}\right)^3$$

Positive values indicate a prevalence of higher than average frequencies, while negative values indicate the opposite. This feature is essential for distinguishing sounds with atypical spectral distributions, such as those produced by some marine (**?**)peciesBishop2006(Bishop, 2006).

- **Kurtosis** Kurtosis measures the "pointiness" of the frequency distribution:

$$\frac{1}{n}\sum_i \left(\frac{s[i] - \mu_s}{\sigma_s}\right)^4$$

High kurtosis indicates the presence of accentuated spectral peaks, typical of impulsive sounds such as cetacean clicks, which are useful for identifying specific sound events (5)(Figueroa et al., 2015).

- **Mean skewness** Mean skewness is the average skewness calculated over several segments or windows of the signal, as in time series or audio data. One divides the signal into several parts (windows) and calculates the skewness for each window, finally obtaining the average of these values.The formula is as follows:

$$\sqrt{\frac{\sum_i (i - \bar{i})^3 \cdot E_i}{E \cdot \mathrm{rms}_i^3}}$$

In the case of audio signals, mean skewness can be used to understand whether the signal has a tendency to concentrate more energy on one side of the distribution, such as in the higher or lower frequencies, and how this tendency changes over time. Mean skewness could provide a general indication of the distribution of energy in the signal.

- **Mean kurtosis** Mean Kurtosis is a statistical measure that indicates the "smallness" of a signal's distribution by assessing how concentrated the data are around the mean. In practice, high values of kurtosis may signal rare events or defects in the signal, while low values indicate a more uniform distribution.

$$\sqrt{\frac{\sum_i (i - \bar{i})^4 \cdot E_i}{E \cdot \mathrm{rms}_i^4}}$$

- **Shennon Entropy** Shannon entropy measures the amount of information or complexity of the signal:

$$-\sum_j p(s_j)\log_2(p(s_j))$$

This feature is indicative of signal complexity, with higher values suggesting greater variability, which is useful for differentiating between complex and simple soundsCover2006(Cover & Thomas, 2006).

- **Renyi Entropy** Renyi entropy makes it possible to analyze signals that have a complex or nonuniform probability distribution, which is common in many applications, such as signal compression and pattern recognition. Unlike Shannon entropy, which considers only the probabilities of events, Renyi entropy introduces a parameter (α) that allows events to be weighted differently. Renyi entropy is defined as follows:

$$\frac{1}{1-\alpha}\log_2\left(\sum_j p(s_j)^\alpha\right)$$

Using Renyi entropy, signals of different complexity can be compared and their information content determined. For example, high values of α emphasize rare events, while lower values focus on more likely events, making this measure useful for applications where significant anomalies or variations in the signal need to be detected.

- **Rate of Attack** Rate of attack is a measure that quantifies the rate at which a signal increases in amplitude over a specific time interval. In audio system design and signal processing, this parameter is particularly important for understanding how a signal, or sound, evolves over time.

$$\max_i \left(\frac{s[i] - s[i-1]}{n}\right)$$

- **Rate of decay** Rate of Decay is an important parameter in the analysis of audio signals and their perception, indicating the rate at which a sound loses its intensity after reaching its peak.

$$\min \quad \frac{s[i] - s[i+1]}{n}$$

When a sound is produced, its initial intensity is maximum. However, over time, the sound pressure

begins to decrease due to various factors, such as sound absorption in the environment and dispersion. The Rate of Decay describes how quickly this process occurs.

- **Silence Ratio** The silence ratio quantifies the proportion of "silence" in the signal, defined as the percentage of frames below a certain threshold:

$$\frac{\#(s \text{ where } s < \text{threshold})}{\sum_i^n s[i]}$$

This feature is useful for discriminating between continuous and intermittent sounds, often associated with natural and anthropogenic phenomena respectively (7)(Giannakopoulos & Pikrakis, 2014).

- **Threshold Crossing Rate** The Threshold Crossing Rate measures how often the signal crosses a certain amplitude threshold:

$$\frac{\#(\text{Threshold Crossing})}{n}$$

This feature is useful for identifying distinct sound events such as clicks or spikes, which are frequent in bioacoustic sounds such as those emitted by marine mammals (14)(Popescu et al., 2009).

- **Mean** The average provides an indicator of the general level of the signal.

$$\frac{1}{n}\sum_{i=1}^{n} s[i]$$

- **Max over mean** Max over Mean is a statistical measure used in signal analysis to describe the relationship between the maximum value of a signal and its mean. This measure is particularly useful for assessing the variability and distribution of a signal's values over time. The mathematical formula is defined as follows:

$$\text{Max over Mean} = \frac{\max(s[i])}{\frac{1}{n}\sum_i s[i]}$$

- **Min over mean** Min over Mean is a statistical measure used in signal and audio analysis. It represents the ratio of the minimum value of a signal to its mean. This measure can provide useful information about the distribution of values in the signal and its general characteristics. The formula for calculating Min over Mean is:

$$\text{Min over Mean} = \frac{\min(s[i])}{\frac{1}{n}\sum_i s[i]}$$

- **Energy measurements** Energy measurements refer to techniques used to quantify the energy contained in an audio signal or sound wave. The formula for calculating Energy measurements is:

Energy for a discrete signal

$$E = \sum_{n=0}^{N-1} |x(n)|^2$$

Energy for a continuous signal

$$E = \int_{-\infty}^{\infty} |x(t)|^2 \, dt$$

- **MFCC: Mel-Frequency Cepstral Coefficients** These coefficients are widely used in audio analysis to represent the waveform in a way that better matches human perception of frequencies. MFCCs capture the spectral characteristics of the signal, making them particularly useful for identifying timbres and sound textures. The formula for calculating the n-th MFCC is:

$$c_n = \sum_{k=1}^{K} \log(S(k)) \cdot \cos\left(\frac{n\pi(k-\frac{1}{2})}{K}\right)$$

- **ZCR: Zero-Crossing Rate** This measure indicates the frequency with which a signal crosses zero, providing useful information about the time complexity of the sound. It is calculated as:

$$ZCR = \frac{1}{N}\sum_{n=1}^{N-1} |\text{sgn}(x[n]) - \text{sgn}(x[n-1])|$$

- **Spectral Centroid** This characteristic represents the center of mass of the power spectrum, indicative of the predominant tone in the audio signal. The formula for calculating it is:

$$C = \frac{\sum_f f \cdot |X(f)|^2}{\sum_f |X(f)|^2}$$

- **Spectral Bandwidth** This feature measures the dispersion of frequencies with respect to the spectral centroid and is calculated as:

$$BW = \frac{\sum_f (f-C)^2 \cdot |X(f)|^2}{\sum_f |X(f)|^2}$$

- **Chroma Features** Chroma Features represent the intensities of the twelve musical notes played in a song, enabling analysis of harmony and pitch. These features are particularly valuable for understanding the harmonic relationships between notes in the context of marine sounds, thus contributing to a richer and more layered analysis of the dataset.The formula is:

$$C_k = \sum_f |X(f)|^2 \cdot \text{h}(f,k)$$

- **Spectral Contrast** Spectral contrast is a measure that describes the intensity differences between frequency bands in an audio spectrum. It captures the interaction between adjacent frequencies and provides a representation of the harmonic structure of a signal.It is calculated using the following formula:

$$SC_k = \frac{1}{N} \sum_{n=1}^{N} \max(0, |X(f_n)|^2 - |X(f_k)|^2)$$

8. Oversampling & Classification

Initially, a binary classification was conducted to distinguish between anthropogenic and bioacoustic sounds. This first stage allowed the development of a simple model that could identify the two main groups of sounds with good accuracy. Subsequently, the problem was extended to a multiclass classification, with the goal of further categorizing bioacoustic sounds into different subclasses, such as species or types of vocalizations. This approach provided a deeper and more detailed understanding of the audio features present in the dataset.
To train and test the model, the dataset was divided into three parts: 80% of the data was devoted to training, 10% to validation, and 10% to testing. This division ensured that the model was trained on a sufficiently large portion of the data, while maintaining an independent set to evaluate performance during training and to test generalization on never-before-seen data.
Before the application of SMOTE, the training set for binary classification had 34,635 samples for the Target class and 5,791 for the Non-Target class. After the use of SMOTE, both classes were balanced, bringing the number of samples to 9,181 for each Target subclass and 1,087 for each Non-Target subclass. Later, SMOTE was applied between the Target class and the Non-Target class, reaching a total of 146,896 samples for both, thus improving the representation of the minority class.
Regarding multiclass classification, SMOTE was used to balance all subclasses of the Target class, bringing them to 9,181 samples each. At this point, the Target class consists of a training set with 16 subclasses (all balanced), a validation set with 6 subclasses, and a test set with 7 subclasses. During the application of SMOTE, the parameter `k_neighbors` was tested with different values to determine the appropriate number of neighbors to be considered in the generation of synthetic samples. A relatively low value of `k_neighbors` was chosen because many of the subclasses were underrepresented in the dataset. In addition, classes with only one sample were removed, as they did not allow SMOTE to function properly. This approach better preserved the characteristics of each minority subclass, avoiding the creation of synthetic samples that were too distant from the actual data. To further improve the representation of minority classes and reduce the risk of overfitting, SMOTE was applied in both classifications. This helped maintain the diversity of bioacoustic signals and allowed the model to learn distinctive features for each category (3)(Chawla et al., 2002). Their results indicate a significant increase in accuracy and recall for minority classes, contributing to a more robust overall performance.

9. Features Integration

The combination of features described above has been shown to be effective in improving the classification accuracy of bioacoustic signals compared with the use of single features. Recent studies have shown that the selection and optimization of features such as MFCCs (Mel-frequency cepstral coefficients), combined with spectral centroids and skewness, can significantly increase the performance of machine learning models such as Support Vector Machines and Random Forests (Malfante et al., 2018). By integrating these features, a robust model can be built that can not only distinguish between anthropogenic and bioacoustic sounds, but also accurately classify the marine species involved. This approach provides an advanced and highly reliable acoustic monitoring system.
Several numerical features from BirdNet were also used, including MFCCs, Zero Crossing Rate, Spectral Centroid, and Chroma. These features were selected for their ability to capture important acoustic aspects, proving effective in signal classification.
Subsequently, another feature set proposed by Dhamodaran et al. in their study “A Survey on Audio Feature Extraction for Automatic Music Genre Classification” was also considered, further expanding the range of features analyzed to improve the performance of the model.

10. Training

For model training, three machine learning algorithms were used: Random Forest, Support Vector Machine (SVM) and LightGBM. Each of these models was chosen for its distinctive features and ability to deal with classification problems in complex contexts such as bioacoustic signals.

- **Random Forest**: This model is an ensemble of decision trees that uses the bagging method to improve accuracy and reduce the risk of overfitting. The Random Forest is particularly effective in handling datasets with many variables, allowing it to capture nonlinear relationships in the data. In this specific case, the model was implemented with 100 decision trees, setting a value of *random seed* of 42 to ensure reproducibility of the results. The training was performed iteratively thanks to the parameter *warm_start*texttt=True, which allowed a tree to be added progressively for each iteration while maintaining the previously trained models. This strategy allowed the model to be optimized by

gradually increasing the number of trees from 1 to 100.

- **Support Vector Machine (SVM)**: Support Vector Machines (SVMs) are a powerful classification tool designed to find the optimal hyperplane separating classes in the multidimensional plane. They are particularly useful in high-dimensional scenarios, where different kernel functions can be used to handle nonlinearity. The SVM was configured to use the loss function *'hinge'*, which is particularly effective for binary classification problems. The parameter *random_state*=42 is set to ensure reproducibility of the results.
- **LightGBM**: This boosting algorithm is designed to be time and space efficient. It uses a tree-based gradient boosting approach, and is capable of handling large datasets and achieving high performance due to its ability to reduce the number of data elements required for training. LightGBM's model is initialized with the parameter *random_state* set to a fixed value (42) to ensure reproducibility of the results, ensuring that each training run produces the same results, regardless of the inherent variability in the data or execution conditions.

  The training approach used involves dividing the dataset into incremental blocks, determined by the argument *n_steps*=10, which specifies the number of training steps. This method allows the model to learn incrementally, adapting to the information contained in each block of data in a systematic manner. During each iteration, the model is trained using an increasing subset of data, which helps to improve training efficiency and optimise performance, especially in scenarios with large volumes of data.

These models were trained using the balanced dataset through SMOTE, allowing a fair evaluation of their performance in classifying bioacoustic signals.

11. Results

This section reports the results obtained from the three implemented classification models. The objective of this analysis is to evaluate the effectiveness of each model in correctly classifying the dataset, using three different feature sets selected from the literature. These feature sets were constructed with the intention of capturing different relevant aspects of the data, in order to compare the performance of the models on different representations of the problem. The performance metrics reported consider both the results obtained during the validation phase and those observed on the test set, allowing for a comparison of the models' behaviour on data never seen before. This ensures that performance is robust and generalisable. To measure the performance of each model, the following metrics were used: -textbfPrecision: indicates the proportion of correct predictions among those assigned to the positive class. -textbfRecall: measures the model's ability to correctly identify all samples of the positive class. -**F1-score**: a balanced combination of precision and recall, useful for evaluating performance when there is an imbalance between classes. -**Accuracy**: represents the overall percentage of correct predictions out of the total number of samples. In addition, both the **macro avg** (the simple average of the metrics for each class, regardless of the class distribution) and the **weighted avg** (which considers the proportion of classes) were considered. This approach makes it possible to assess how the models handle any imbalances between classes. In the evaluation process, both a validation set and a test set were used to estimate the generalisation capabilities of the models. **Validation set**: used during the training process to optimise the models and select the best parameters. The metrics calculated on this set help monitor the fit of the models to the data and avoid overfitting. **Test set**: used exclusively after training and parameter selection. It serves to evaluate the final effectiveness of the model on completely new data, providing a realistic estimate of its performance. The results obtained by the models on each feature set are shown in the tables below, with the metrics calculated on both the validation and test sets. Differences between the two assessments may indicate the model's ability to generalise correctly or signal overfitting or underfitting. The identified features were divided into 3 datasets:

Table 1
Comparative Data Sets

First Set (11)	
Spectral Centroid Mean	Spectral Bandwidth RMS
Standard Deviation	Skewness
Kurtosis	Shannon Entropy
Renyi Entropy	Rate of Attack
Rate of Decay	Threshold Crossings
Silence Ratio	Mean
Max Over Mean	Min Over Mean
Energy-Measurements	
Second Set (BirdNet)	
MFCC 1	MFCC 2
MFCC 3	MFCC 4
MFCC 5	MFCC 6
MFCC 7	MFCC 8
MFCC 9	MFCC 10
MFCC 11	MFCC 12
MFCC 13	ZCR
Spectral Centroid	Spectral Bandwidth
Chroma 1	Chroma 2
Chroma 3	Chroma 4
Chroma 5	Chroma 6
Chroma 7	Chroma 8
Chroma 9	Chroma 10
Chroma 11	Chroma 12
Third Set (19)	
MFCC 1	MFCC 2
MFCC 3	MFCC 4
MFCC 5	MFCC 6
MFCC 7	MFCC 8
MFCC 9	MFCC 10
MFCC 11	MFCC 12
MFCC 13	ZCR
Spectral Contrast	Tonnetz 1
Tonnetz 2	Tonnetz 3
Tonnetz 4	Tonnetz 5
Tonnetz 6	Chroma 1
Chroma 2	Chroma 3
Chroma 4	Chroma 5
Chroma 6	Chroma 7
Chroma 8	Chroma 9
Chroma 10	Chroma 11
Chroma 12	Tempo

	Precision	Recall	F1-score	Support
macro avg(val)	0.99	0.99	0.99	4333
weighted avg(val)	0.99	0.99	0.99	4333
macro avg(test)	0.94	0.79	0.84	6234
weighted avg(test)	0.93	0.92	0.92	6234

Accuracy on the validation set 0.9914
Accuracy on test set: 0.9231

Table 2
Binary training results with Random Forest (first set). The Random Forest performed excellently during validation, with precision, recall and F1-score close to 1, indicating high accuracy in classifying positive and negative instances. The accuracy on the validation set is 99.14%, confirming the effectiveness of the model. On the test set, the model achieved an accuracy of 92.3%, showing a slight decrease in performance, especially in recall (0.79 for the macro mean). However, the accuracy and F1-score values remain high, suggesting that the model maintains good performance, despite the test set presenting some additional challenges

.

	Precision	Recall	F1-score	Support
macro avg(val)	0.81	0.87	0.83	4333
weighted avg(val)	0.89	0.87	0.88	4333
macro avg(test)	0.82	0.79	0.80	6234
weighted avg(test)	0.88	0.89	0.88	6234

Accuracy on the validation set: 0.8786
Accuracy on the test set: 0.8928

Table 3
Binary training results with SVM (first set). SVM performed well on the validation set, with an accuracy of 87.86%. However, the recall in the macro average (0.87) is lower than the Random Forest, indicating that the model may have difficulty capturing some minority classes. On the test set, accuracy is slightly improved (89.28%), with balanced performance between accuracy and recall. This suggests that the SVM is able to generalise well on unseen data, but may still have room for improvement in dealing with variability in the dataset

.

	Precision	Recall	F1-score	Support
macro avg(val)	0.98	0.99	0.98	4333
weighted avg(val)	0.99	0.99	0.99	4333
macro avg(test)	0.98	0.98	0.98	6234
weighted avg(test)	0.99	0.99	0.99	6234

Accuracy on validation set: 0.9892
Accuracy on validation set: 0.9885

Table 4
Binary training results with LightGBM (first set). The LightGBM model performed excellently, with a very high F1-score (0.98-0.99), indicating an excellent balance between accuracy and recall. The accuracy on the validation set is 98.92 %, confirming the model's ability to handle different classes effectively. The results on the test set, with an accuracy of 98.85%, confirm that the model generalises very well, maintaining performance almost identical to that obtained during validation, demonstrating high robustness

.

	Precision	Recall	F1-score	Support
macro avg	0.98	0.99	0.99	4333
weighted avg	0.99	0.99	0.99	4333
macro avg(test)	0.98	0.96	0.97	6234
weighted avg(test)	0.98	0.98	0.98	6234

Accuracy on validation set: 0.9903
Accuracy on test set: 0.9807

Table 5
Binary training results with Random Forest (second set). The performance of the Random Forest model on the second set is high, with F1-score and accuracy values close to 1 for both the macro and weighted mean. The accuracy on the validation set is 99.03%, indicating an excellent ability to correctly classify instances. On the test set, the accuracy is slightly lower (98.07%), but the metrics remain consistent with those of the validation set, suggesting that the model is highly reliable and stable

	Precision	Recall	F1-score	Support
macro avg(val)	0.84	0.84	0.84	4333
weighted avg(val)	0.89	0.89	0.89	4333
macro avg(test)	0.78	0.73	0.75	6234
weighted avg(test)	0.85	0.86	0.86	6234

Accuracy on validation set: 0.8975
Accuracy on test set: 0.8681

Table 6
Binary training results with SVM (second set). The SVM model showed good overall performance with an accuracy of 89.75% on the validation set. However, the recall in the macro average (0.84) indicates that some classes may not have been captured optimally. The results on the test set (accuracy 86.81%) confirm that the model performs slightly worse on unseen data, suggesting that it may be more sensitive to dataset variability than other models.

	Precision	Recall	F1-score	Support
macro avg(val)	0.98	0.97	0.98	4333
weighted avg(val)	0.98	0.98	0.98	4333
macro avg(test)	0.98	0.97	0.97	6234
weighted avg(test)	0.99	0.99	0.99	6234

Accuracy on validation set: 0.9836
Accuracy on test set: 0.9854

Table 7
Binary training results with Light GBM (second set). The LightGBM performed very well with an accuracy of 98.36% on the validation set and high metrics for precision, recall and F1-score. These results indicate that the model has a high ability to generalise, maintaining high accuracy (98.54%) even on the test set. The consistent performance between the two sets confirms the model's ability to fit the data well, without overfitting

	Precision	Recall	F1-score	Support
macro avg(val)	0.98	0.99	0.98	4333
weighted avg(val)	0.99	0.99	0.99	4333
macro avg(test)	0.97	0.94	0.96	6234
weighted avg(test)	0.97	0.97	0.97	6234

Accuracy on validation set: 0.9898
Accuracy on test set: 0.9748

Table 8
Binary training results with Random Forest (third set).
The Random Forest demonstrated an excellent ability to classify instances of the third set, with very high accuracy, recall and F1-score. The accuracy on the validation set is 98.98%, confirming the robustness of the model. On the test set, the model achieved slightly lower accuracy (97.48%), but with similar metrics, demonstrating good generalisation even on unseen data

	Precision	Recall	F1-score	Support
macro avg(val)	0.82	0.80	0.81	4333
weighted avg(val)	0.87	0.87	0.87	4333
macro avg(test)	0.76	0.73	0.74	6234
weighted avg(test)	0.85	0.85	0.85	6234

Accuracy on validation set: 0.8792
Accuracy on test set: 0.8591

Table 9
Binary training results with SVM (third set). SVM model showed a lower performance than the other models, with an accuracy of 87.92% on the validation set. The recall and F1-score metrics (0.73 and 0.74) on the test set indicate a slight decrease in performance compared to the validation, with an accuracy of 85.91%. The model may have more difficulty in dealing with variability in the dataset than solutions such as Random Forest or LightGBM

	Precision	Recall	F1-score	Support
macro avg(val)	0.98	0.99	0.99	4333
weighted avg(val)	0.99	0.99	0.99	4333
macro avg(test)	0.98	0.96	0.97	6234
weighted avg(test)	0.98	0.98	0.98	6234

Accuracy on validation set: 0.9915
Accuracy on test set: 0.9809

Table 10
Binary training results with Light GBM (third set).
The results obtained indicate that the LightGBM model demonstrated an excellent ability to correctly classify instances of the third data set. Its high accuracy, recall and F1-score suggest that the model is able to distinguish between the two classes with high reliability. The results on the test set were the same as those on the validation set

	Precision	Recall	F1-score	Support
macro avg(val)	0.29	0.22	0.24	3427
weighted avg(val)	0.37	0.36	0.36	3427
macro avg(test)	0.39	0.43	0.41	5119
weighted avg(test)	0.42	0.39	0.40	5119

Accuracy on validation set: 0.3580
Accuracy on test set: 0.3970

Table 11
Multiclass training results with Random Forest (first set).
The Random Forest model for multiclass classification did not perform satisfactorily, with low metrics for accuracy, recall and F1-score. The accuracy of 35.80% on the validation set suggests a poor ability to distinguish between classes. On the test set, the model also reported similar results (accuracy 39.70%), confirming the difficulty in generalising in complex multiclass contexts

.

	Precision	Recall	F1-score	Support
macro avg(val)	0.10	0.09	0.04	3427
weighted avg(val)	0.28	0.18	0.08	3427
macro avg(test)	0.10	0.15	0.11	5119
weighted avg(test)	0.07	0.20	0.08	5119

Accuracy on validation set: 0.1806
Accuracy on test set: 0.2026

Table 12
Multi-class training results with SVM (first set). The performance of the SVM was particularly disappointing, with very low values for all metrics. Accuracy on the validation set was only 18.06%, and on the test set it improved slightly to 20.26%. This indicates that the model struggles to correctly identify the different classes, with performance remaining very limited on both sets

.

	Precision	Recall	F1-score	Support
macro avg(val)	0.38	0.61	0.28	3427
weighted avg(val)	0.36	0.36	0.36	3427
macro avg(test)	0.31	0.68	0.31	5119
weighted avg(test)	0.41	0.39	0.39	5119

Accuracy on validation set: 0.3645
Accuracy on test set: 0.3862

Table 13
Multiclass training results with Light GBM (first set).
The LightGBM model showed a slight improvement over previous models, with an accuracy of 36.45% on the validation set. Recall metrics are particularly high compared to precision and F1-score, suggesting a greater ability to capture classes compared to models such as Random Forest and SVM. The results on the test set (accuracy 38.62%) are in line with those of the validation, but still remain insufficient to consider the model fully reliable.

	Precision	Recall	F1-score	Support
macro avg(val)	0.50	0.54	0.52	3427
weighted avg(val)	0.53	0.54	0.53	3427
macro avg(test)	0.60	0.56	0.57	5119
weighted avg(test)	0.56	0.55	0.55	5119

Accuracy on validation set: 0.5404
Accuracy on test set: 0.5458

Table 14
Multi-class training results with Random Forest (second set). The Random Forest model showed significant improvements over the other multiclass models, with an accuracy of 54.04% on the validation set. The accuracy, recall and F1-score metrics indicate that the model is able to distinguish classes better than previous versions. On the test set, accuracy is also similar (54.58%), demonstrating a slightly better ability to generalise than the other models

.

	Precision	Recall	F1-score	Support
macro avg(val)	0.26	0.21	0.19	3427
weighted avg(val)	0.43	0.33	0.29	3427
macro avg(test)	0.21	0.25	0.19	5119
weighted avg(test)	0.36	0.24	0.20	5119

Accuracy on validation set: 0.3309
Accuracy on test set: 0.2352

Table 15
Multiclass SVM training results (second set).
The results obtained by training the SVM model on the second multiclass dataset show a modest performance. The average precision, recall and F1-score are low, with a weighted F1-score of 0.08 on the validation set, and similar values on the test set. This suggests that the model has great difficulty in correctly distinguishing the different classes. The overall accuracy on the validation set is 0.18, with a slight increase on the test set (0.20). These results confirm that the SVM model is not suitable for this specific multiclass task, presenting a very low accuracy both in validation and testing.

	Precision	Recall	F1-score	Support
macro avg(val)	0.60	0.60	0.49	3427
weighted avg(val)	0.56	0.56	0.55	3427
macro avg(test)	0.63	0.63	0.63	5119
weighted avg(test)	0.60	0.58	0.59	5119

Accuracy on validation set: 0.5579
Accuracy on test set: 0.5814

Table 16
Multiclass training results with Light GBM (second set).
The Light GBM model applied to the second multiclass dataset performed better than the previous models. The average F1-score, precision and recall suggest a balance between classes, with a weighted F1-score of 0.39. However, individual precision and recall indicate that the model may struggle with some specific classes. The overall accuracy on the validation set is 0.3645, while on the test set the value is very similar, 0.3862. Although these results are still far from optimal performance, the model seems to be slightly more robust than SVM and Random Forest in the multiclass context, confirming its ability to handle complex datasets.

	Precision	Recall	F1-score	Support
macro avg(val)	0.35	0.30	0.31	3427
weighted avg(val)	0.52	0.52	0.51	3427
macro avg(test)	0.50	0.42	0.42	5119
weighted avg(test)	0.57	0.56	0.56	5119

Accuracy on validation set: 0.5182
Accuracy on test set: 0.5589

Table 17
Multiclass Random Forest training results (third set).
The results of the Random Forest model for the third multiclass dataset indicate a significant improvement in performance compared to previous models. The average precision, recall and F1-score are significantly higher, with a weighted F1-score of 0.57. These results suggest that the model has gained a better ability to distinguish between different classes compared to the other sets. The accuracy on the validation set is 0.5404, with a slight improvement on the test set (0.5458). This shows that the model is able to generalize better, although there may still be room for improvement for more accurate classification across all classes.

	Precision	Recall	F1-score	Support
macro avg(val)	0.16	0.15	0.14	3427
weighted avg(val)	0.35	0.35	0.33	3427
macro avg(test)	0.32	0.38	0.34	5119
weighted avg(test)	0.48	0.46	0.46	5119

Accuracy on validation set: 0.3510
Accuracy on test set: 0.4604

Table 18
Multiclass SVM training results (third set).
The results obtained with the SVM on the third dataset are rather disappointing. The overall accuracy is only 35.10%, indicating that the model correctly classified less than a third of the instances. Better results for the test set.

	Precision	Recall	F1-score	Support
macro avg(val)	0.59	0.67	0.55	3427
weighted avg(val)	0.53	0.53	0.52	3427
macro avg(test)	0.55	0.68	0.54	5119
weighted avg(test)	0.61	0.59	0.60	5119

Accuracy on validation set: 0.5279
Accuracy on test set: 0.5945

Table 19
Multiclass training results with Light GBM (third set).
The table shows that the Light GBM model, applied to the third multiclass dataset, has achieved the best results so far for this task. The average precision, recall and F1-score are relatively high, with a weighted F1-score of 0.55 and a good ability to balance between classes. The accuracy on the validation set is 0.5450, while on the test set it is slightly lower, 0.5407. This suggests that the model is able to generalize well to new data, although there are still some difficulties in correctly identifying instances belonging to all classes

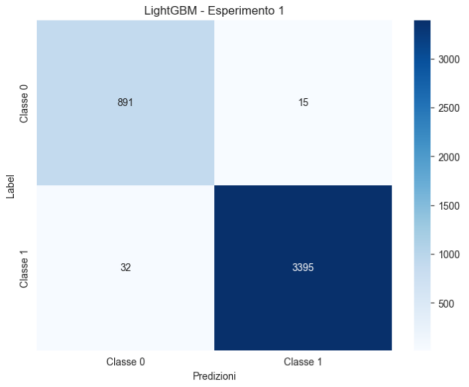

Figure 6: Confusion matrix first set binary classification with LightGBM

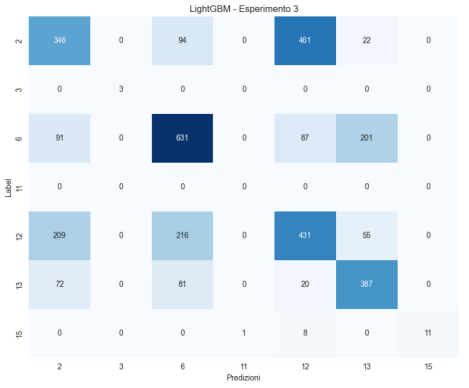

Figure 7: Third set confusion matrix multiclass classification with LightGBM

12. Experiments on the best models

After analyzing the performance of the models, the two best ones were identified: LightGBM for binary classification (first set), with an accuracy value on the test set equal to 0.9885, and LightGBM for multiclass classification (third set), with an accuracy value of 0.5945. To further optimize the performance, several experiments were conducted by changing the sample rate and feeding new datasets to the models. The experiments regarding binary classification are shown below.
The first experiment, performed on LightGBM using the first dataset with a sampling rate of 44100 Hz, produced the following results:

	Precision	Recall	F1-score	Support
macro avg(val)	0.96	0.95	0.95	4333
weighted avg(val)	0.98	0.98	0.98	4333
macro avg(test)	0.94	0.82	0.86	6234
weighted avg(test)	0.93	0.93	0.92	6234

Accuracy on validation set: 0.9811
Accuracy on test set: 0.9301

Subsequently, a second experiment was performed always for the binary classification with LightGBM, using the first data set but with a sampling rate of 192000 Hz. The results obtained are the following:

	Precision	Recall	F1-score	Support
macro avg(val)	0.97	0.97	0.97	4333
weighted avg(val)	0.98	0.98	0.98	4333
macro avg(test)	0.92	0.78	0.83	6234
weighted avg(test)	0.92	0.91	0.91	6234

Accuracy on validation set: 0.9785
Accuracy on test set: 0.9143

Finally, a third experiment for binary classification with LightGBM was conducted on the first dataset, this time using a sampling rate of 384000 Hz. The results were as follows:

	Precision	Recall	F1-score	Support
macro avg(val)	0.97	0.97	0.97	4333
weighted avg(val)	0.98	0.98	0.98	4333
macro avg(test)	0.93	0.79	0.84	6234
weighted avg(test)	0.92	0.92	0.91	6234

Accuracy on validation set: 0.9815
Accuracy on test set: 0.9211

Below are the results of the multiclass classification experiments using the third feature set on LightGBM.

Esperimento con sample rate a 44100 Hz:

	Precision	Recall	F1-score	Support
macro avg(val)	0.51	0.69	0.46	3427
weighted avg(val)	0.83	0.82	0.82	3427
macro avg(test)	0.52	0.81	0.43	5119
weighted avg(test)	0.78	0.77	0.77	5119

Accuracy on validation set: 0.8240
Accuracy on test set: 0.7675

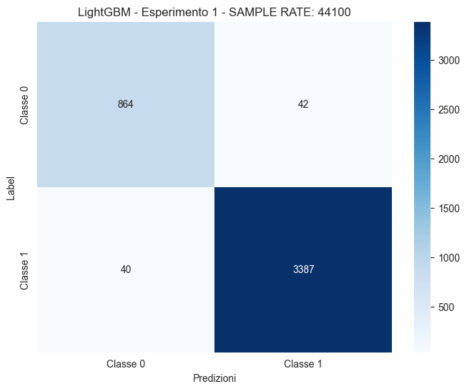

Figure 8: Confusion matrix experiment 1 (Malfante features) binary classification 44100Hz with LightGBM

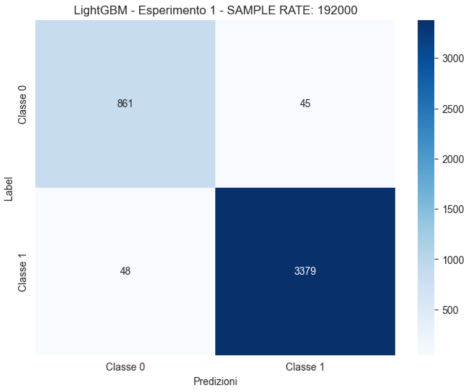

Figure 9: Confusion matrix experiment 1 (Malfante features) binary classification 192000Hz with LightGBM

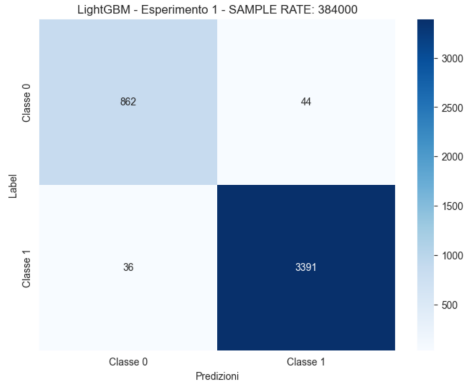

Figure 10: Confusion matrix experiment 1 (Malfante features) binary classification 384000Hz with LightGBM

13. Conclusions

In conclusion, the analysis of the results obtained highlights that for the binary classification task, the LightGBM and Random Forest models proved to be the most performing, reaching accuracies higher than 98
The results of the LightGBM binary classification experiments, performed by varying the sample rate of the dataset, demonstrate that resampling at frequencies other than 96,000 Hz does not significantly improve the model

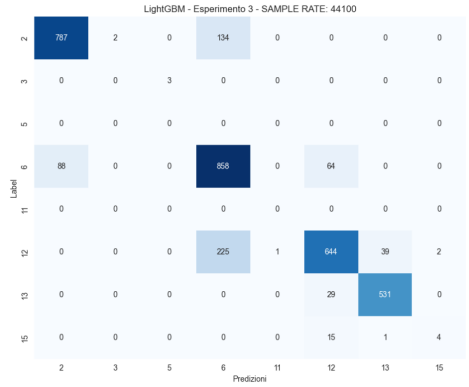

Figure 11: Confusion matrix experiment 3 (Dhamodaran features) 44100Hz multiclass classification with LightGBM

performance. In particular, reducing the sample rate to 44,100 Hz caused a decrease in accuracy on the test set, from 0.9885 to 0.9301. A similar behavior was also observed with the increase in sample rate to 192,000 Hz and 384,000 Hz, with an accuracy further reduced to 0.9143 and 0.9211 respectively.
In the multiclass classification, the use of a sample rate of 44,100 Hz led to significant improvements in the model performance. The accuracy on the test set increased from 0.5945 to 0.7675, while on the validation set it went from 0.5279 to 0.8240.
The results obtained highlight the importance of the choice of sampling frequency in performance optimization.

14. Future developments

Future developments of this project could focus on several key aspects to improve the accuracy and effectiveness of the classification model. First, it would be appropriate to use a larger and more diverse dataset, including a greater variety of bioacoustic and anthropogenic sounds, collected in different environmental conditions. A more representative dataset would improve the model's ability to generalize and adapt to a wider range of real-world scenarios. Furthermore, the integration of advanced combinations of acoustic features could allow a more detailed representation of audio signals, allowing to better capture the peculiarities of bioacoustic and anthropogenic sounds. The use of feature selection and extraction techniques, such as principal component analysis (PCA) or deep learning-based methods, could optimize the representation of signals, improving the accuracy of the model. Finally, the use of more advanced and complex machine learning models, such as convolutional neural networks (CNN) or deep learning-based models, could increase the system performance. An ensemble approach, combining different classification algorithms, could further refine the results, exploiting the complementarity between the various models to achieve greater robustness in distinguishing between anthropogenic and bioacoustic sounds. These future developments will offer significant

potential to improve the system's ability to support sustainable management of marine ecosystems through more accurate identification of acoustic signals.

15. Bibliography

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Blumstein, D. T., Mennill, D. J., Clemins, P., Girod, L., Yao, K., Patricelli, G., ... & Kirschel, A. N. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology, 48*(3), 758-767.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

[4] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

[5] Figueroa, L., Belloni, M., & Abascal-Mena, R. (2015). Statistical modeling of cetacean click properties for automatic classification. *Journal of the Acoustical Society of America, 137*(3), 1025-1034.

[6] Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution, 10*(2), 169-185.

[7] Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press.

[8] Hildebrand, J. A. (2009). Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series, 395*, 5-20.

[9] Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics, 61*, 101236.

[10] Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. John Wiley & Sons.

[11] Malfante, M., Mars, J. I., Dalla Mura, M., & Gervaise, C. (2018). Automatic fish sounds classification in real-life marine environments: Performance evaluation and perspectives. *Journal of the Acoustical Society of America, 143*(5), 2834-2845.

[12] Mellinger, D. K., & Clark, C. W. (2000). Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America, 107*(6), 3518-3529.

[13] Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence, 116*, 374-388.

[14] Popescu, M., Ehrlich, R., & Xu, W. (2009). Detection and classification of acoustic events for in-home monitoring of an elder person living alone. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 314-319).

[15] Salamon, J., Bello, J. P., Farnsworth, A., & Rosenheim, K. (2017). Feature learning with convolutional neural networks for bioacoustics. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2662-2666). IEEE.

[16] Sharma, R., Agarwal, S., & Jain, R. (2020). Acoustic signal classification using hybrid features and machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing, 11*(3), 1171-1180.

[17] Sueur, J., Aubin, T., & Simonis, C. (2008). Equipment review: Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics, 18*(2), 213-226.

[18] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293-302.

[19] Dhamodaran, P., Balakrishnan, V., & Dharmavaram, A. (2023). A Survey on Audio Feature Extraction for Automatic Music Genre Classification. *International Journal of Recent Technology and Engineering, 12*(3), 45-52.