Applying Feature Selection Combination in Audios of Whale for Improving Classification

Cephas A S Barreto [2], Victor V Targino [1], Tales V de M Alves [1], Lucas V Bazante [1], Rafael V R de Oliveira [1], Ricardo A R do A Junior [1], João C. Xavier-Júnior [1], Anne Magály de P. Canuto [2]

[1] Instituto Metrópole Digital (IMD)
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

[2]Departamento de Informática e Matemática Aplicada (DIMAp)
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

cephasax@gmail.com,victorvieira.rn@gmail.com,

talesvinicius1998@gmail.com,lucasbazante1@gmail.com,

rafaelvini1999@gmail.com,ric.alex55@yahoo.com,

jcxavier@imd.ufrn.br,anne@dimap.ufrn.br

Abstract. *Audio signal processing has been under investigation for the last decades. The majority of the works found in literature focus on signal analysis and classification. Most of them integrate Machine Learning (ML) algorithms with the audio signal processing techniques. As the performance of any ML algorithm depends on the features of a dataset used for training and testing purposes, using a dataset derived from the extraction of features from an audio is not trivial due to the fact that the correct combination of extraction techniques with the selection of the most relevant attributes needs to take place. In this sense, this paper proposes an empirical analysis on different audio extraction techniques combined with feature selection for improving Whale audio classification. Usually, the application of audio extraction techniques results in poor classification performance. However, the combination of feature selection can achieve better results. The experimental results have been promising, indicating that the idea of combining different audio extraction techniques with feature selection can improve the performance of ML classification algorithms over whales' audios by 22 percentage points.*

1. Introduction

Humans are used to hear different types of audios from speech to Environmental sounds everyday [Sharma et al. 2020]. However, there are other variation of sounds which we are not used, for instance, those produced by aquatic mammals. Their sound production is

different from that of humans, mainly because some of them (e.g., whales) produce sound through the phonic lips, which function like the human nasal cavity [Xian 2016].

Additionally, as we are not used to that sort of sound, in open sea, a vessel could collide with a large mammal, such as a whale, causing its death. In this sense, much needs to be done in order to prevent the extermination of such important animals to the marine environment. In fact, The Marinexplore and Cornell University have launched a worldwide challenge for creating an ML algorithm for detecting North Atlantic right whale calls from audio recordings, aiming to prevent collisions with shipping traffic [Karpištšenko 2013].

In fact, audio signal processing has been under investigation for the last decades, and the majority of the works have been focusing on signal analysis and classification [Halkias et al. 2013], [Al-Shoshan 2006]. Most of them integrate Machine Learning (ML) algorithms with the audio signal processing techniques. As the frequency of such type of sound (i.e., whale) is quite high, different audio signal processing techniques need to be applied in order to extract the best suitable feature [Oppenheim 1978]. Moreover, the classification performance of ML algorithms can be deteriorated with dataset originated from poor audio signal processing [Costa et al. 2020]. Finally, even when using the more appropriated audio signal processing tools, the generated features need to go through a selection process so that the most related ones are selected to a specific domain [Jović et al. 2015].

In this sense, this paper proposes an empirical analysis on different audio extraction techniques combined with feature selection for improving whale audio classification. For audio extraction we used the following techniques: (1) Linear Predictive Coding (LPC); (2) Method of Moments; (3) Mel Frequency Cepstral Coefficients (MFCC); (4) Power Spectrum; (5) Strongest Frequncy via FFT Maximum; (6) Strongest Frequency via Spectral Centroid; and (7) Strongest Frequency via Zero Crossing. These techniques were chosen after some preliminary experiments due to their best results when compared with other techniques. Regarding ML algorithms, we used some well-known classifiers, which are: Multiple Layer Perceptron (MLP) ; k-NN ; Decision Tree; Naive Bayes; Support Vector Machine (SVM); and classifier ensembles (e.g.,Boosting on Decision Tree [Wu et al. 2008], Random Forest [Cutler et al. 2012], Bagging on Decision Tree [Bühlmann and Yu 2002] and a custom classifier ensemble composed by all already mentioned base classifiers.

The rest of the paper is divided into six section and organised as follows. Section 2 presents the background concepts related to this paper. Section 3 presents the studies related to this paper, while Section 4 presents the experimental methodology used in this empirical analysis. The obtained results are presented in Section 5. Finally, Section 6 describes the final remarks as well as the future works of this paper.

2. Background

This section describes some important concepts related to this paper, more specifically: Machine Learning Classification algorithms and Classifier Ensembles algorithms.

2.1. Machine Learning

According to [Mitchell 1997], Machine Learning (ML) is defined as the field of study that gives computers the ability to learn without being explicitly programmed. ML has been widely applied in different application and there are several ML methods that can be broadly divided into supervised, semi-supervised and unsupervised methods.

In this paper, we will focus in the supervised learning category. There are several supervised algorithms in the literature and the main ones are:

- Multiple Layer Perceptron (MLP)[Gardner and Dorling 1998];
- k-NN [Silverman and Jones 1989];
- Decision Tree [Quinlan 1986];
- Naive Bayes [Domingos and Pazzani 1997];
- Support Vector Machine (SVM) [Cortes and Vapnik 1995];
- among others [Wu et al. 2008].

2.2. Classifier Ensemble

An ensemble of classifiers is a model built using two or more classifiers (base classifiers) to solve a classification task, reducing the variance of a single model [Gharroudi 2017]. Usually, the base classifiers of an ensemble are weak classifiers (such as Naïve Bayes, Decision Tree, *k*-Nearest Neighbour). However, sometimes more powerful classifiers also can be used as a base classifier of an ensemble (Support Vector Machine, Multilayer Perceptron, Bayes Network, and others). Another characteristic of an ensemble is that the type of all base classifiers determines if the structure ensemble is homogeneous or heterogeneous. The first one occurs when all base classifiers of an ensemble are the same type, and the last one occurs when at least one of the base classifiers is a different type from others [Gharroudi 2017].

The idea of creating an ensemble of classifiers to combine their different outputs is to improve the performance of a task is a promising possibility. In addition, when several classifiers outputs are combined, an additional information is generated by each instance (different label for the same instance). As the selection of the classifier that always correctly classifies an instance is an impossible task, it is necessary to select a label for this instance, combining the outputs of base classifiers through majority weighted vote. The majority vote selects the most voted label among those presented by the base classifiers, while the weighted vote computes the vote using weight criteria to weigh the classifiers with the most effectiveness.

3. Related Work

This section reviews related work on audio extraction techniques and ML classification algorithms. The reason being is due to the fact that the worldwide challenge launched by the Marinexplore and Cornell University has provided a whale audio dataset with more than 5,000 sounds, and also that there are other datasets used for whale sound classification. In this sense, we can discuss the similarities between related works and our proposition.

In [Mazhar et al. 2007], the authors presented results on recognition of individual humpback whales based on their vocalisation data. Cepstral coefficients from song units extracted from audio records of seven humpback whales, re-sampled at 8KHz, were used

as a pre-processed dataset to train a multi-class SVM classifier. According to them, the test phase results indicated classification accuracy as high as 99%.

In another work [Ness et al. 2013], the authors have used audios from the Orchive, which is a large collection of over 20,000 hours of audio recordings from the OrcaLab research facility located off the northern tip of Vancouver Island. The features of 197 orcas' calls (audios) were extracted with the application of MFCC coefficients, Centroid, Rolloff, Flux and Zero crossings, then the mean and standard deviation for each of these features were used as an output file for training a SVM classification algorithm. The authors reported an accuracy of 98.5% on this small set of orcas' calls.

On the other hand, in [Frasier 2015] the authors used a non-parametric classification tree (CART) and a Random Forest algorithms to analyse 1019 calls of the eastern Beaufort Sea beluga recorded over 14 days off Icy Cape, Alaska. The frequencies and duration measurements of those 1019 calls resulted in 34 categories with 83% agreement in classification for both CART and Random Forest algorithms.

None of these related works used datasets larger than 1,100 whales' audios which indicates that the bigger the dataset is the more difficult becomes the feature extraction and the training processes. Moreover, the dataset provided in [Karpištšenko 2013] has more than 5,000 sounds of whales plus background noise. In this sense, it important to address this change as a whole by analysing the audio extraction tools combined with the feature selection process aiming to improve classification performance. Finally, differently from other works, we perform a deeper analysis on combining audio extraction techniques and feature selection aiming to classify whale's audio and background noise.

4. Experimental Methodology

As previously mentioned, this paper performs an exploratory study on feature selection combination to classify Whale sounds. In order to do this, Figure 1 presents the general structure of the experimental methodology. As it can be observed in this figure, the whale audios are extracted and feature selectors are applied to obtain seven audio features datasets. All feature extracted in one combined dataset. Then, the individual classification methods as well as the classifier ensemble are applied to the combined dataset. Finally, the obtained resulted are analysed by a statistical test.

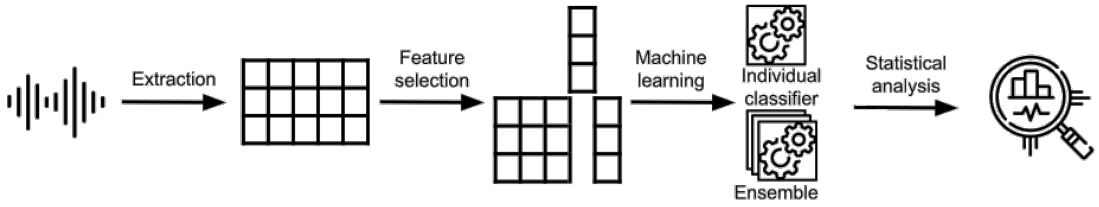

Figure 1. The general structure of the experimental methodology

4.1. The original dataset

The original dataset was extracted from the Kaggle repository, and it is was used in the ICML 2013 Whale Challenge[1]. Due to this fact, some works have been carried out aiming

[1]https://www.kaggle.com/competitions/the-icml-2013-whale-challenge-right-whale-redux/data

to increase the accuracy performance over this dataset. Moreover, as the original proposal, this dataset is a challenging one for classifcation tasks.

The sound recordings were acquired of single species (one call per hour), Right Whale (3 types of calls). The used equipment was hydrophones of 16 bit, 2-10kHz MARU, bit depth = 11, Sensitivity -167.5 dB re: 1uPa/V. The audio set consists of 47149 files in ".aif" format, with two classes (whale and noise), and their duration vary between 1 and 2 seconds. 41895 of them are considered as "noise" and 5254 are considered as "whale sound".

4.2. The feature extractors

The original audioset configuration was preserved since no pre-processing method,(noise reduction, audio correction, etc.) was applied in the audioset. The JAudio software [McKay 2010] was applied to extract features from audio data and the following features were extracted: (1) *LPC; (2) Method of Moments; (3) MFCC; (4) Power Spectrum; (5) Strongest Frequncy via FFT Maximum; (6) Strongest Frequency via Spectral Centroid;* and (7) *Strongest Frequency via Zero Crossing* [McKay 2010]. These features were extracted using 44.1 kHz as sampling rate and all other parameters with default values (e.g. window size equal to 512 samples, and no window overlap. Table 1 presents the number of features for each feature extractor.

Table 1. Number of attributes for each feature after feature selection

Extractor	Numb. Features
LPC	45
Method of Moments	23
MFCC	36
Power Spectrum	45
SF FFT Maximum	9
SF Spectral Centroid	6
SF Zero Crossing	9
Sum	166

In terms of instances, an under-sampling technique was applied, decreasing the size of the majority class (noise sound) to make it the same as the minority one. In the original dataset, the majority class had 9 times more data than the minority one. As a consequence, the under-sampled dataset contains 10.508 instances, a major reduction compared to the 47.149 instances without under-sampling. In addition, all attributes were also normalised and the missing values were replaced by the mean value of the corresponding attribute.

4.3. The analysed feature datasets

In order to perform an exploratory study of the different feature extractors, three different scenarios will be assessed, which are:

1. Individual feature extractors: In this scenario, the classification methods will be analysed for each feature extractor individually;

2. Data combination: In this scenario, the features of all extractors are combined in a combined feature dataset and this dataset will be used by the classification methods;
3. Decision combination: In this scenario, the classification methods will be trained by the individual feature extractors and their decision will be combined.

For the data combination scenario, four versions of the combined dataset were created: the whole combined dataset with all the generated attributes; a combined dataset with a feature selection made through Random Forest; and two datasets with a selection of attributes eliminating randomly 30% and 70% of the attributes, respectively.

4.4. Methods and Materials

All datasets showed and discussed in the previous section are presented to 5 individual classification algorithms (*k*-NN, Decision Tree, Naive Bayesian, Support Vector Machine and Multilayer Perceptron) and four classifier ensemble methods (Bagging, Boosting, Random Forest and Stacking), three homogeneous classifier ensemble methods, all of them using Decision Tree as base classifier (Bagging, Boosting and Random Forest) and a custom Ensemble Classifier, which uses voting scheme to predict the classes and has *k*-NN, Decision Tree, Multilayer Perceptron, Naive Bayes and Support Vector Machine as base classifiers .

All classifications methods were executed in Python using the scikit-learn library. The classification algorithms were executed using the default parameter values and the classifier ensembles used 10 base classifiers. The only exception is Random Forest which uses 100 random trees.

In this paper, in order to obtain a better estimation of the accuracy rates, a 10-fold cross validation method is applied to all classification algorithms. Therefore, the values presented in this papers are the average of 20 cases (10 folds of 2 runs).

Finally, for the statistical analysis, we will apply the Friedman test [Demšar 2006] with the Nemenyi Post-hoc test [Hollander et al. 2013]. These tests will be applied in order to validate these the obtained values, from a statistical point of view. In order to provide a better visualization of the post-hoc results, the Critical Difference Diagram (CD Diagram) will be used [Demšar 2006]. This diagram is an interpretable visualization tool for displaying the results of a set of methods.

5. The obtained results

This section presents the results obtained in the empirical analysis. As previously mentioned, three different scenario will be assessed, which are: individual feature extractors, data combination and decision combination. The next three subsections will present the results of these three scenarios. Finally, the last subsection describes the results of the statistical analysis.

5.1. Individual feature extractors

As described in Section 4.2, seven different feature extractors are applied in the Whale audios. Table 2 presents the results of all eight classification methods for all feature extractors. The results of this table represent the average of 20 accuracy values (10 folds

Table 2. Accuracy results for Individual feature extractors

Audio feature	Decision Tree	K-NN	MLP	NB	SVM	Bagging	Boosting	RF	Avg
LPC	0.542	0.576	**0.621**	0.543	0.617	0.579	0.539	0.612	0.579
Method of Moments	0.544	0.558	**0.624**	0.563	0.592	0.574	0.547	0.606	0.576
MFCC	0.628	0.682	**0.764**	0.670	0.754	0.693	0.632	0.725	0.694
Power Spectrum	0.639	0.589	0.632	0.499	0.580	0.695	0.637	**0.710**	0.623
Strongest Freq. - FFT	0.525	0.533	0.528	0.507	0.523	0.539	0.532	**0.543**	0.529
Strongest Freq. - Spectral Centroid	0.521	0.491	0.502	0.520	0.494	0.543	0.516	**0.552**	0.517
Strongest Freq. - Zero Cross	0.536	0.540	0.571	0.515	0.569	0.562	0.564	**0.575**	0.554

and 2 runs). Finally, in this table, the bold numbers represent the highest accuracy value for each feature extractor.

In terms of feature extractors, the last line represent the average accuracy of the feature extractors. As it can be observed, MFCC obtained the highest accuracy values, followed by Power Spectrum, LPC, Methods of Moments, Zero Cross, FFT and Spectral Centroid. This is an expected result since MFCC has presented impressive values for audio data.

In terms of classification methods, the bold numbers represent best method for each feature extractor. As it can be observed in Table 2, the best methods are MLP for the individual classification methods and Random Forest (RF) for the ensemble classifiers. The performance superiority of MLP was expected since it is the most elaborated method, along with SVM. The RF superiority is also expected since this ensemble has shown better results than Bagging, Boosting and Stacking in the literature. Nevertheless, it is important to highlight that RF is composed of 100 classifiers, while the remaining ensemble methods are composed of 10 classifiers. This might be an important fact that improves the performance of Random Forest.

5.2. Data combination

The second scenario represents the combination of all feature extractors. As mentioned previously, four datasets were created in this scenario. In terms of dataset dimensionality, the whole dataset (Whole) contains 166 attributes. The dataset reduced by Random Forest (RF-reduction) contains 70 attributes. Finally, the randomly reduced data sets have 50 and 116 attributes for 70% of reduction (rand-reduction-70) and 30% of reduction (rand-reduction-70), respectively. Table 3 presents the results of all classification methods for these four datasets of the data combination scenario. Once again, the bold numbers represent the best classification method for each dataset.

Table 3. Average results - Data Combination - accuracy

Dataset	Dec. Tree	K-NN	MLP	NB	SVM	Bagging	Boosting	Rand. Forest	Ensemble
Whole	0.862	0.497	0.560	0.586	0.497	0.942	0.861	**0.972**	0.778
RF-reduction	0.867	0.892	**0.987**	0.630	0.908	0.946	0.868	0.970	0.954
rand-reduction-30	0.846	0.498	0.524	0.602	0.497	0.927	0.843	**0.960**	0.778
rand-reduction-30	0.850	0.497	0.557	0.581	0.497	0.938	0.849	**0.971**	0.758

In terms of improvement of the individual feature extractors (comparing Tables 2 and 3), the data combination datasets improved the performance of the individual feature extractors. This improvement is stronger for the classifier ensembles. For Bagging, for instance, this improvement is higher than 30 percentage points. On the other hand, for the

individual classifiers, the improvement in performance is stronger for Decision Tree. For the remaining methods, the improvement in performance is slight, mainly for k-NN and Naive Bayes.

In terms of the data combination datasets, the RF-reduction dataset presented the best results. In fact, the improvement of this dataset is very impressive, in which all methods delivered strong improvements in performance. The results of this table show that the use of an efficient feature reduction method, leaving only the important attributes can produce efficient classification methods.

Finally, in terms of classification methods, unlike the previous section, the best individual classification method is Decision Tree. However, the MLP network delivered the highest accuracy value (for the RF-reduction dataset). For the classifier ensemble, once again, the best method is Random Forest.

5.3. Decision combination

This section presents the results of third scenario, which is decision combination. In order to do this, a classifier ensemble is used, in which each classifier uses a feature extractor. In this case, for a fair comparison, an ensemble with 10 classifiers is used. As we have only 7 feature extractors, 3 of them were used in 2 classifiers. In these case, different classifiers were used (DT and MLP). As we have only one method, this result is presented in Table 4.

Table 4. Average results - Decision Combination - accuracy

	Ensemble
Average	0.936

When comparing data combination (Table 3) with decision combination (Table 4), we can observe that decision combination delivered better performance the majority of the data combination methods. The only exception is Random Forest and some cases for Bagging. This results show that the use of decision combination is also a promising attempt to combine features for audio whales.

5.4. Statistical Analysis

In order to assess the obtained results in a more robust way, a statistical analysis is performed. In this analysis, the Friedman test is applied. In case of rejection of the null hypothesis, the post-hoc Nemenyi test is applied. As already mentioned, for a better visualization of the obtained result are presented in a CD (Critical Difference) diagram.

Figure 2 presents the CD diagram of all eight classification methods for the Whole dataset. As it can be observed from Figure 2, the first four methods (all of them are classifier ensembles) have similar performance (linked by a horizontal line) and they provided superior accuracy than the individual classification methods, from a statistical point of view.

Figures 3, 4 and 5 present the CD diagram for the remaining three datasets (RF-reduction, rand-reduction-30 and rand-reduction-70). All these diagrams are similar, with most of the classifier ensemble methods with better performance than most of the individual classification methods, from a statistical point of view.

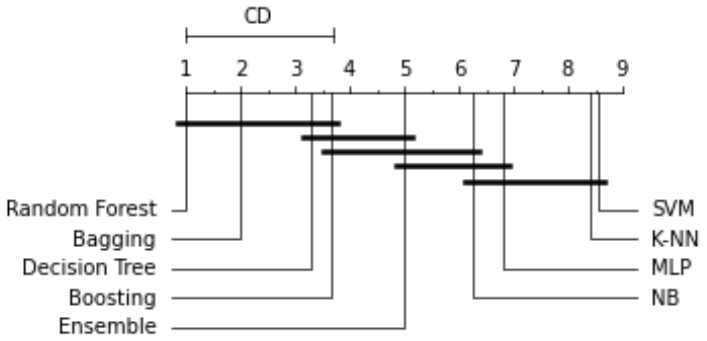

Figure 2. Critical Difference for methods - whole dataset

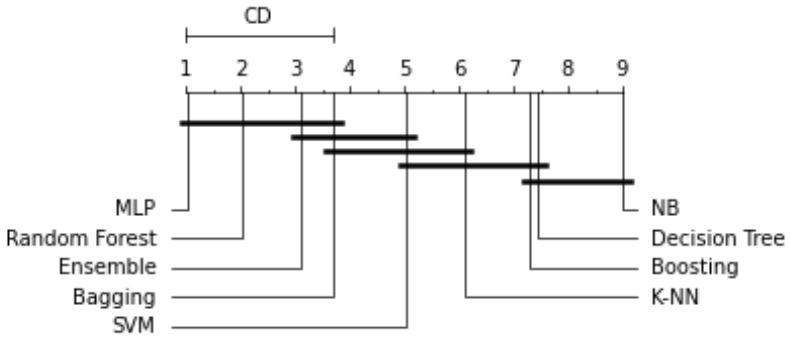

Figure 3. Critical Difference for methods - RF-reduction dataset

6. Conclusions and Future Work

This work presented an empirical analysis on seven different feature extractors, being: LPC, Method of Moments, MFCC, Power Spectrum, Strongest Frequency via FFT Maximum, Strongest Frequency via Spectral Centroid, and Strongest Frequency via Zero Crossing. In addition, we also performed a combination of all feature extractor, resulting in four datasets.

The experimental results have shown that the classification accuracy of all classification methods (i.e., base classifiers and classifier ensembles) was not higher than 76.4% when using the aforementioned individual feature extractors. However, when we combined all feature extractors (i.e., second scenario) the overall classification accuracy has increased, reaching 98.7% for MLP over RF-reduction dataset. Moreover, the improvement was really strong for the classifier ensembles. Random Forest, for instance, was the best classifier ensemble achieving 97% in three of four datasets.

As a direction for future work, it would be interesting to extend the experiments for more feature extractors as well as more classification methods. Additionally, a feature distribution method for ensembles could also be used, increasing the diversity of the classifier ensemble and, as a consequence, improving its performance.

References

Al-Shoshan, A. I. (2006). Speech and music classification and separation: A review. *Journal of King Saud University - Engineering Sciences*, 19(1):95–132.

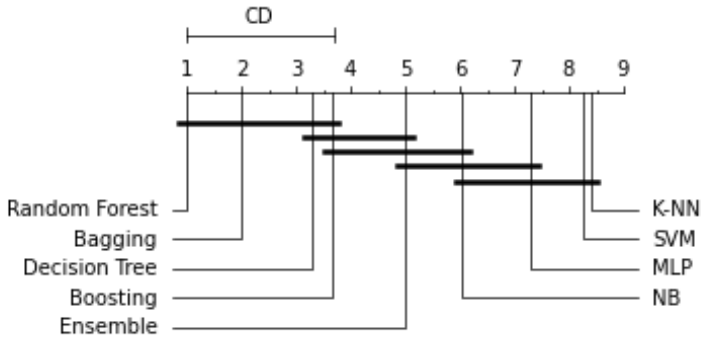

Figure 4. Critical Difference for methods - rand-reduction-30 dataset

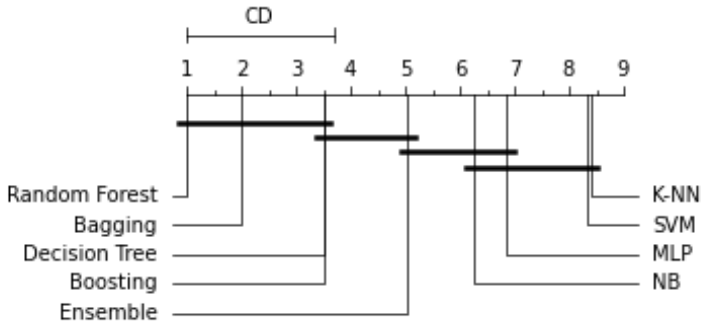

Figure 5. Critical Difference for methods - rand-reduction-70 dataset

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The annals of Statistics*, 30(4):927–961.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning - ML*, 20:273–297.

Costa, R. R., Gorgonio, A. C., da S Barreto, C. A., Lima, D. F., de P Canuto, A. M., and Xavier-Junior, J. C. (2020). Detection of respiratory problems through lung audios using machine learning. *Anais do Encontro de Computacao do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, (4).

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, pages 157–175. Springer.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-oneloss. *Machine Learning - ML*, 29:103–130.

Frasier, K. E. (2015). Beluga whale (delphinapterus leucas) vocalizations and call classification from the eastern beaufort sea population. *The Journal of the Acoustical Society of America*, 137(6):3054—-3067.

Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

Gharroudi, O. (2017). *Ensemble multi-label learning in supervised and semi-supervised settings*. Theses, Université de Lyon.

Halkias, X. C., Paris, S., and Glotin, H. (2013). Classification of mysticete sounds using machine learning techniques. *The Journal of the Acoustical Society of America*, 134(5):3496–3505.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons.

Jović, A., Brkić, K., and Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205.

Karpištšenko, A. (2013). The marinexplore and cornell university whale detection challenge. `https://www.kaggle.com/competitions/whale-detection-challenge/discussion/4472`.

Mazhar, S., Ura, T., and Bahl, R. (2007). Vocalization based individual classification of humpback whales using support vector machine. In *OCEANS 2007*, pages 1–9. IEEE.

McKay, C. (2010). *Automatic music classification with jMIR*. PhD thesis, McGill University, Montreal, Quebec, Canada.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, London.

Ness, S. R., Symonds, H., Spong, P., and Tzanetakis, G. (2013). The orchive : Data mining a massive bioacoustic archive. *CoRR*, abs/1307.0589.

Oppenheim, A. (1978). *Applications of Digital Signal Processing*. Prentice-Hall signal processing series. Prentice-Hall.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(3):81—106.

Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020.

Silverman, B. and Jones, M. C. (1989). E. fix and j.l. hodges(1951): an important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, 57(3):233–247.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

Xian, Y. (2016). *Detection and classification of whale acoustic signals*. PhD thesis, Duke University.