Bird sound recognition

L. Ceccomancini 1667798, L. Loretucci 1903794, A. Melnic 1692625, M. Meo 1599032, D. Russo 1714011

27 February 2021

ABSTRACT

The aim of the project is to build an algorithm that is able to classify bird species from sounds and other complementary informations such as location, time, date and altitude of the observation. Given that birds are high up in the food chain, the construction of such an algorithm can be useful not only for monitoring bird populations but also for detecting deterioration of habitat quality. Given the large number of existing species and the great variety of environments in which they can be found, we decided to start from an easier problem in which we consider 10 of the most common species: this allows us to have a greater amount of data per species and to select cleaner audio since the data collection phase is from the internet. We collected audio tracks of 5 seconds, in particular we worked mainly on the *frequency spectrum* of each audio. The best classification was obtained with the combination of 3 different models (*SVM*, *Logistic*, *Random Forest*) and, afterwards, we tried to visualize the reasons behind the wrong classifications. We hope to continue developing this project, even increasing the degree of difficulty, in order to create an app for recognizing birds from audio recordings.

Key words: bird sound – classification – audio signals – bioacoustics – ornithology

1 DATA RETRIEVING

The database for this paper has been scraped from **xeno-canto** [1], which is a website dedicated to sharing bird sounds from all over the world. Its pros are the large number of recordings and the well-organized database. As a starting point, we decided to build our classifier only on the **top 10 species** with the highest number of recordings. To do this, we took advantage of the database structure, that groups birds by the taxonomy method: *order - family - genus - species*, giving for each the number of recordings. After having got the species in which we are interested in, the database is created by gathering together information from two different pages: the former contains a table with generic information about the recording and the latter allows us to retrieve GPS information. Not all the database rows are gathered, we use some filters to obtain only the better ones, filtering by audio quality and other information that makes us sure to take only birds that are correctly associated with the corresponding audio. For what concerns the audio recording, it is directly processed during the scraping thanks to the library **librosa** [2], we keep at maximum 5 seconds of audio: starting from one of the highest in amplitude peak (assuming that the highest peaks correspond to the bird corresponding to the label) we keep two and half seconds before the peak and two and half after it. Our starting database contains *1366* rows and *110261* columns. Among the columns, the first 11 contains information about the birds (*id*, *common name*, *scientific name*, *date*, *time*, *country*, *latitude*, *longitude*, *elevation*, *background*, *type*), the other ones represent the **wave form**s. Due to the high dimension of the database, we opt for the parquet format, which is lighter and faster to load.

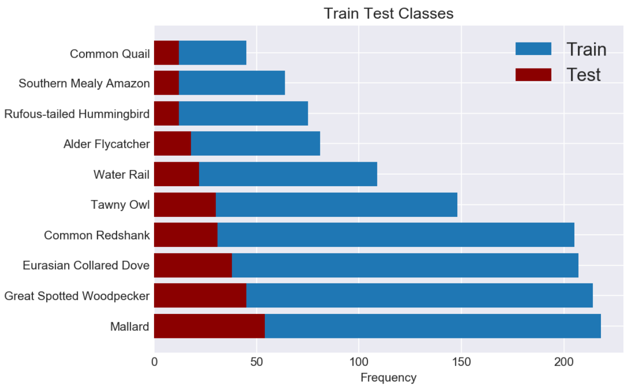


Figure 1. Frequency of the labels in both train and test.

2 DATA EXPLORATION

The dataset is composed by audio features (derived from the wave forms) and other type of features like country, hour, date of localization and more specific features like the type of call and sex of the bird. These two type of data will be processed separately and the description of the process is in the next sections.

The classes are unbalanced as seen in figure 1 and this can possibly cause some problems in the train phase as described in the section 5. In the splitting of train and test the proportion of the classes is maintained. In figure 3 are represented the average binned spectra

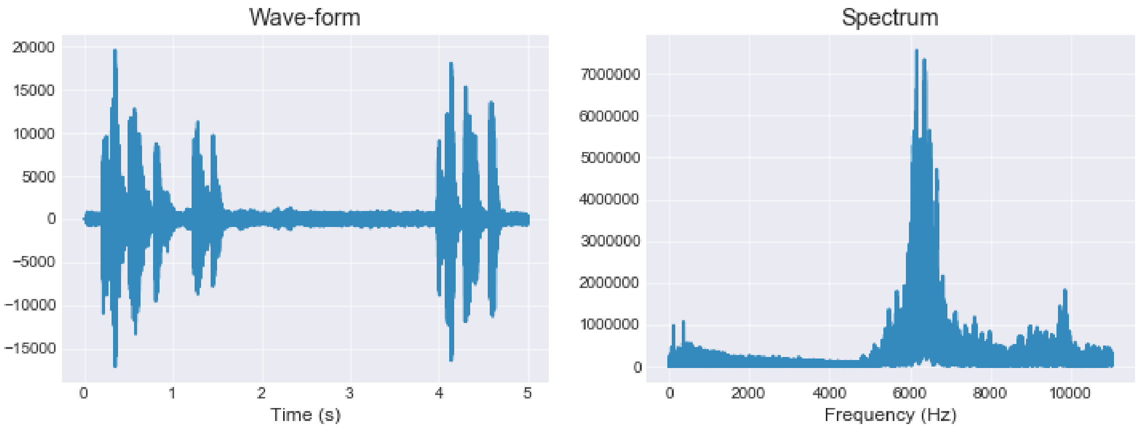


Figure 2. A wave-form and its spectrum.

for all the classes, i.e., the type of audio feature that we used in our model.

3 FEATURE ENGINEERING

In the literature of audio classification two type of audio features are mainly analysed : **temporal features** and **spectral features**. As examples from the first category there are: ZCR, amplitude envelope, RMS, etc; from the second category there are: MFCCs, spectral centroid, spectral roll off, etc. As it will be shown in the next sections, our winning strategy consisted in choosing only spectral features, since from our point of view, in the classical ML setup they are more effective.

3.1 Audio features

The first feature selected is the *spectrum* in the Mel scale. Given a discrete signal x_n, with $n = 0,..., M - 1$, we get the Discrete Fourier Transform (DFT) as:

$$\hat{x}_k = \mathrm{DFT}(x_n) = \sum_{n=0}^{M-1} x_n e^{-ik2\pi n/M} \qquad k = 0, 1, ..., M - 1 \qquad (1)$$

This transformation gives us the spectrum of the signal. From a computational point of view this can be achieved by using the Fast Fourier Transform (FFT) algorithm implemented in the `numpy` package. Additionally, since the DFT returns complex numbers we only took the modulus and discarded the phase.

After obtaining the signal in the frequency domain $\hat{x}_k$ it was converted into the Mel scale according to the following formula: $f_{Mel} = 2595 \log(1 + \frac{f}{700})$ [3] (more precisely this transformation should not be applied in the 1 kHz range, but through cross validation we evaluated that there was no big difference with this, so we decided to convert all frequencies in this scale). Furthermore we applied a band-pass filter by cutting all the frequencies below 200 Hz and over 10000 Hz, where the low and high cut were selected through cross validation. This first part is implemented in `Audio_Processing.return_ffts` in the `Preprocessing` class.

Since from (1) we are getting a vector of same dimension as the wave-form we decided to bin it and take averages of the elements inside each bin, where the number of bins M' was selected through cross validation. The final vector has the form:

$$\hat{x}_i = \frac{1}{n_i} \sum_{k \in B_i} \hat{x}_k, \qquad i = 0, 1, ..., M' - 1 \qquad (2)$$

where $\hat{x}_i$ represents the i-th element of the vector associated to the bin B_i containing n_i number of elements. The code for this function is contained in `Audio_Processing.bin_data`.

As second audio feature we used the *centroid*, defined as the center of mass of the frequency spectrum:

$$\mathrm{Centroid} = \frac{\sum_k k\hat{x}_k}{\sum_k \hat{x}_k} \qquad (3)$$

This is implemented in `Audio_Processing.eval_spectral_centroid`.

Each audio feature of the first type was scaled separately from each other to have mean 0 and standard deviation 1 (it was scaled along rows), meanwhile the second feature was scaled together with all the other features of this type (along column axis).

Other features were also tested, like MFCCs, ZCR, RMS,spectral roll-off, spectral bandwidth and also different type of binnings, but these always decreased the cross validation score and for this reason we discarded them.

3.2 Other features

We have many features on the observations such as country, date, hour and elevation.

These were processed separately from the audio features,in particular some of them like date and hour were completely transformed, meanwhile others like country and coordinates remained the same.

We felt that date and hour were too specific to trace the characteristics of the birds so we decided to use them to extrapolate the daily period of observation as day or night, in order to understand if the bird was nocturnal or diurnal and the annual period to get in which season these birds were observed in order to better understand their behavior.

Day

We use the date, latitude and longitude of the observation of a bird, to calculate the time of sunrise and sunset in those specific coordinates

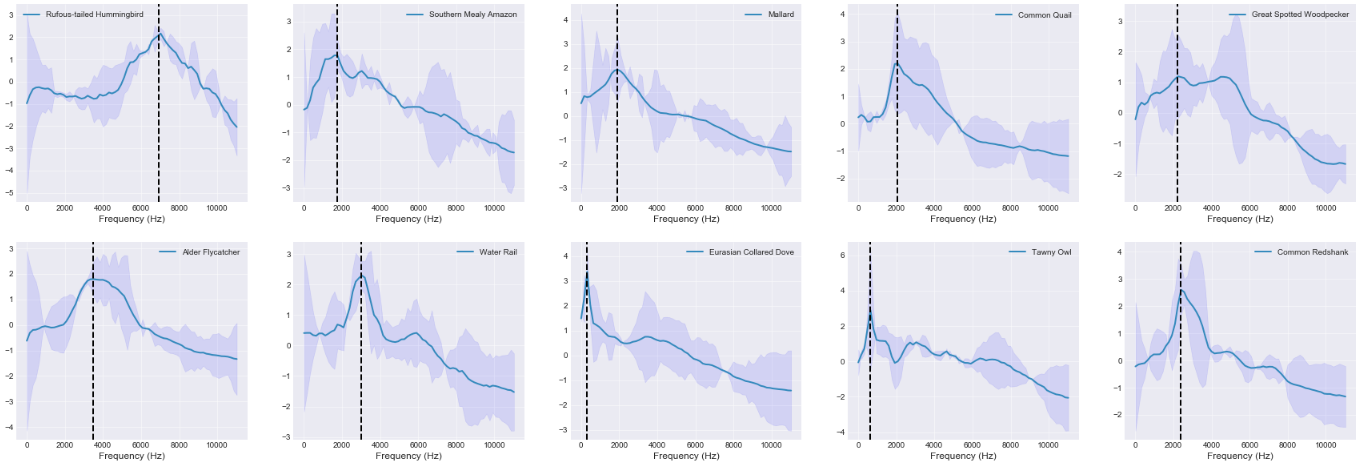


Figure 3. Average binned spectra for all classes (the data processing procedure is explaned in the Feature Engineering section) highlighting the peaks that almost all classes share and representing the standard deviation around all points to give an idea of the variability of the signal.

on that specific day and then we use hour to check if the observation time is between the two. The feature takes the value

(i) *day* if sighting hour between sunrise and sunset
(ii) *night* otherwise

Season

We use latitude to calculate the hemisphere and then with this one, the day and the month of the observation we calculate the season.
If hemisphere is equal to north we use this formula to calculate *se*(season):

$$md = month * 100 + day \tag{4}$$

where *se* takes the value

(i) 0 (spring) if md is between 320 and 620
(ii) 1 (summer) if md is between 621 and 922
(iii) 2 (fall) if md is between 622 and 1222
(iv) 3 (winter) otherwise

If hemisphere is equal to south we use this formula to change and take the correct season *se*:

$$se = (se + 2)\%3 \tag{5}$$

Elevation

Originally the elevation was an integer number, but for our purposes this information was too much detailed, and to reduce the noise of it we discretized it in 3 bins:

(i) *low altitude* ranging from 0 to 240 m
(ii) *medium altitude* ranging from 240 to 500 m
(iii) *high altitude* from 500 m to above

Type

This column contains the tags associated with the recordings. Each of them belongs to one of the following categories: *call*, *sex*, *stage*, *special*. This information has a strong impact on the model, because birds emit different sound depending on the situation, on the sex and on the live stage, as matter of fact, the new columns increase the accuracy of our model. When for a category of given recordings are present more than one tag, we assume that the information is missing, because usually two tags are at odds with each other.

Background

This information has been included in the database, with the idea of using it later. It reports the name of other birds present in the recordings, if they are present. From this column, we generate a new one that tell us just whether other birds are present or not in the recordings, but it doesn't improve our model. Anyway, this variable can be better exploit in future version, for instance using directly the name of the birds or using the names to get the frequencies to remove in the recordings.
All these features contained null values, we decided to substitute them with the mode for categorical variables and mean for continuous ones (only for coordinates and elevation). Subsequently, after the transformations described above, we encoded all the variables, expect coordinates, in a dummy way. The transformation and processing of these variables is contained in the `Preprocessing.` and `model.fill_na`, `model.scaling`, `model.prepare_df`.

4 MODELS

Different models were tested, the best score on the test set, according to accuracy and our personalized score was obtained with a soft Voting of 3 models: SVM (rbf, C=24), Logistic (l2 penalty, C=0.45) and Random Forest (5000 estimators). The train and test set were processed separately and the parameters were tuned through cross validation with 10 folds. The Voting model uses the three models that were tuned to achieve individually their highest CV score. In the following table the results and comparison between the different models on the test set.

Model	our score	accuracy
Voting	0.967	0.876
Logistic l2 penalty	0.959	0.839
Random Forest trees=5000	0.958	0.825
SVM rbf	0.961	0.821
LDA	0.961	0.814
K-NN k=5	0.918	0.723
QDA	0.777	0.474

As it will be described in the next section, our model is predicting the right label 87% of the times, but in those 13% where the model is wrong, the right class is predicted, most of the times, as the second most probable. Since there are 10 classes, from our point of view, the accuracy is not an ideal score, and for this reason we defined our personalized score.

From the model we output a distribution of classes and then we rank the classes from the one with highest probability to the one with the lowest one. From this output, in our personalized score, we considered where, in the ranking list, the correct label is placed. We did this by sorting the output according to their probability, and then assigning a score going evenly from 1 to 0 to the prediction depending on the place of the correct label in the rank. For example if the right label is predicted as first then it is assigned a score of 1, if it is predicted as second it is assigned a score of 0.88, if third 0.77, and so on. If it is predicted as last in the rank then it is assigned a score of 0.

From our results, for the best model, we can see that our score is 0.967, that means that most of the times the label is predicted at least between the first 2 positions. The code for the training and test evaluation is in the `model` class.

5 RESULTS

As anticipated in the previous section, most of the times the true label is in the first place in our prediction and when it is not then it is in the second one (most of the times). From the bar plot in figure 4 we can see in which position the true label is placed in the prediction rank. It is possible to notice that 97% of the times the prediction is within the first 3 places and never in the last 2 places.

From the confusion matrix in figure 5 we can notice that almost every class is predicted correctly except Common Quaill and Water Rail. In particular they are often miss-classified with the Common Redshank class. The Common Quail as shown in figure 1 , does not have many observations in the test set and most important it does not have many observation in the train. For the first reason is not possible to establish if there is an exchange between the two classes or is just a coincidence and for the second reason the predictions on it may be poor because the observations are not enough to learn all the important features of the class. Meanwhile the Water Rail has enough observations in the test and train (more than the Alder Flycatcher that has a much higher score on its class) but still performs poorly. We then analysed the binned spectra for the Water Rail and Common Redshank to understand if similarities can be found. Our main idea was the one to found differences between the spectra of the correctly predicted and the wrongly predicted Water Rails, but this is not the case. From the figure 6 it is possible to notice that the two spectra doesn't have big differences between them, or at least not visible from these features. Furthermore we thought that maybe this mix between the classes can be caused by the non-audio features and for this we conducted the test evaluation without those features also, but again the result of the confusion matrix is identical (for the Water Rail class) as before and is shown in figure 7.

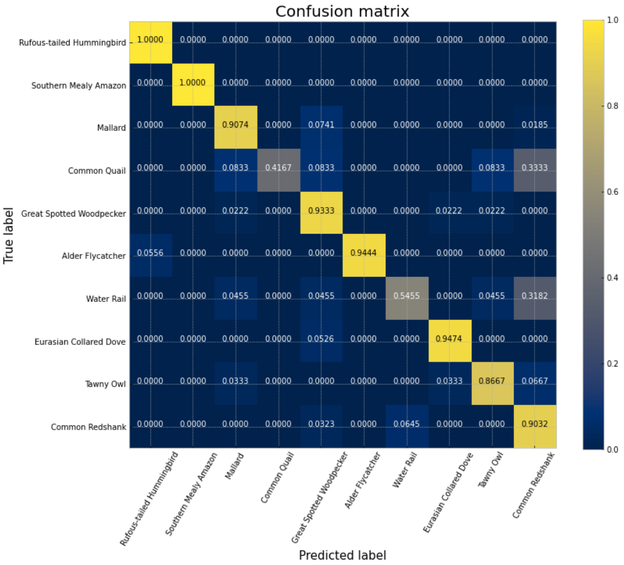


Figure 4. Confusion matrix indicating, given the correct label on the y-axis, the fraction with which a certain label on the x-axis is predicted.

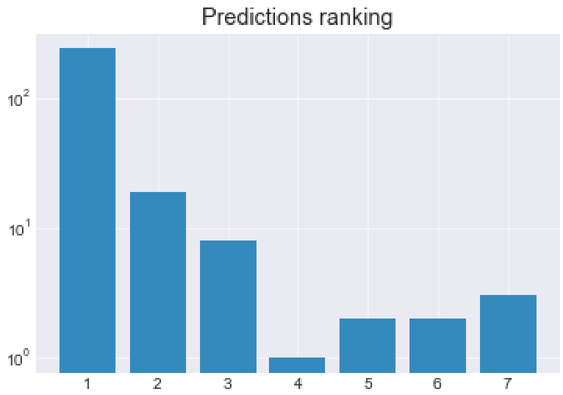


Figure 5. Counts of the position of the test true label in the ranking output.

In the end, we decided to compare our results with the **BirdNET** application that identify 984 birds species by sound. We random sampled 10 audio recordings among our corrected predictions, showed in the first 10 rows of the Table1, and 10 audio recordings among our wrong predictions, showed in the second 10 rows of the table, in order to check the the robustness of our model compared to the BirdnNET one. As it's possible to see, we manage to predict one more correct label respect to BirdNET in the correct sample; instead, in the wrong sample BirdNET misses the label 4 times out of 10.

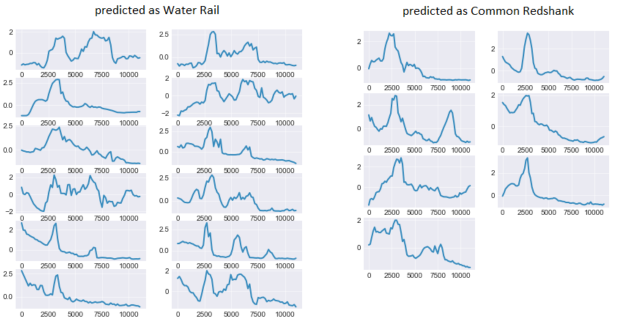
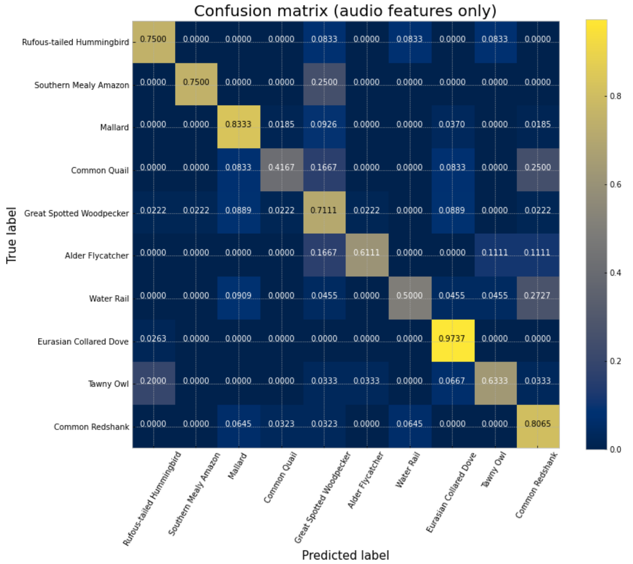


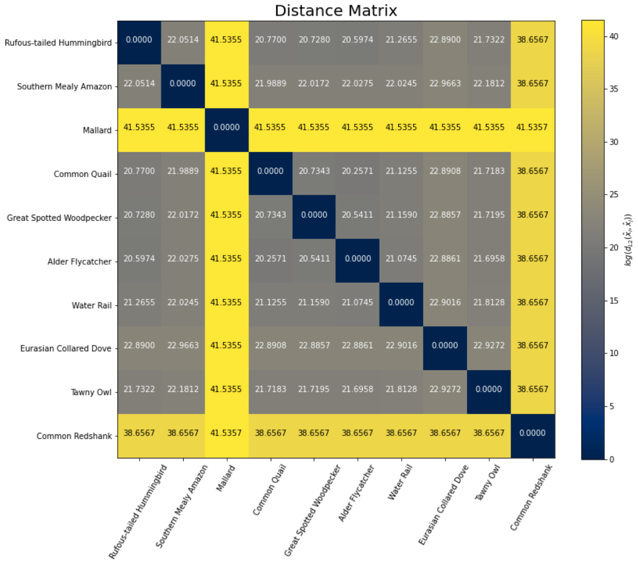
Figure 6. Confront between the binned spectra for the correct and wrong prediction for the Water Rail class.



	Rufous-tailed Hummingbird	Southern Mealy Amazon	Mallard	Common Quail	Great Spotted Woodpecker	Alder Flycatcher	Water Rail	Eurasian Collared Dove	Tawny Owl	Common Redshank
Rufous-tailed Hummingbird	0.7500	0.0000	0.0000	0.0000	0.0833	0.0000	0.0833	0.0000	0.0833	0.0000
Southern Mealy Amazon	0.0000	0.7500	0.0000	0.0000	0.2500	0.0000	0.0000	0.0000	0.0000	0.0000
Mallard	0.0000	0.0000	0.8333	0.0185	0.0926	0.0000	0.0000	0.0370	0.0000	0.0185
Common Quail	0.0000	0.0000	0.0833	0.4167	0.1667	0.0000	0.0000	0.0833	0.0000	0.2500
Great Spotted Woodpecker	0.0222	0.0222	0.0889	0.0222	0.7111	0.0222	0.0000	0.0889	0.0000	0.0222
Alder Flycatcher	0.0000	0.0000	0.0000	0.0000	0.1667	0.6111	0.0000	0.0000	0.1111	0.1111
Water Rail	0.0000	0.0000	0.0909	0.0000	0.0455	0.0000	0.5000	0.0455	0.0455	0.2727
Eurasian Collared Dove	0.0263	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9737	0.0000	0.0000
Tawny Owl	0.2000	0.0000	0.0000	0.0000	0.0333	0.0333	0.0000	0.0667	0.6333	0.0333
Common Redshank	0.0000	0.0000	0.0645	0.0323	0.0323	0.0000	0.0645	0.0000	0.0000	0.8065

Figure 7. Confusion matrix for the model tested without the non-audio features.

Somehow this confirms that some audio in the wrong audio sample is actually hard to predict: listening to the audio, we realized that they are mainly recordings in which the songs of 2 different birds are present, in most of the cases one in the background and one dominant but sudden.



	Rufous-tailed Hummingbird	Southern Mealy Amazon	Mallard	Common Quail	Great Spotted Woodpecker	Alder Flycatcher	Water Rail	Eurasian Collared Dove	Tawny Owl	Common Redshank
Rufous-tailed Hummingbird	0.0000	22.0514	41.5355	20.7700	20.7280	20.5974	21.2655	22.8900	21.7322	38.6567
Southern Mealy Amazon	22.0514	0.0000	41.5355	21.9889	22.0172	22.0275	22.0245	22.9663	22.1812	38.6567
Mallard	41.5355	41.5355	0.0000	41.5355	41.5355	41.5355	41.5355	41.5355	41.5355	41.5357
Common Quail	20.7700	21.9889	41.5355	0.0000	20.7343	20.2571	21.1255	22.8908	21.7183	38.6567
Great Spotted Woodpecker	20.7280	22.0172	41.5355	20.7343	0.0000	20.5411	21.1590	22.8857	21.7195	38.6567
Alder Flycatcher	20.5974	22.0275	41.5355	20.2571	20.5411	0.0000	21.0745	22.8861	21.6958	38.6567
Water Rail	21.2655	22.0245	41.5355	21.1255	21.1590	21.0745	0.0000	22.9016	21.8128	38.6567
Eurasian Collared Dove	22.8900	22.9663	41.5355	22.8908	22.8857	22.8861	22.9016	0.0000	22.9272	38.6567
Tawny Owl	21.7322	22.1812	41.5355	21.7183	21.7195	21.6958	21.8128	22.9272	0.0000	38.6567
Common Redshank	38.6567	38.6567	41.5357	38.6567	38.6567	38.6567	38.6567	38.6567	38.6567	0.0000

Figure 8. Distance matrix that at each point is equal to the l2 distance between the average spectra of the classes. The Common Redshank has one of the highest distance between all the other classes and it means that is not possible to explain through the similarity between the spectra the reason of mixing between the Water Rail and Common Redshank.

True label	Our prediction	BirdNET prediction
Alder Flycatcher	Alder Flycatcher	Alder Flycatcher
Eurasian Collared Dove	Eurasian Collared Dove	Eurasian Collared Dove
Eurasian Collared Dove	Eurasian Collared Dove	Eurasian Collared Dove
Alder Flycatcher	Alder Flycatcher	Alder Flycatcher
Mallard	Mallard	Mallard
Alder Flycatcher	Alder Flycatcher	Alder Flycatcher
Water Rail	Water Rail	Water Rail
Mallard	Mallard	Mallard
Great Spotted Woodpecker	Great Spotted Woodpecker	Great Spotted Woodpecker
Great Spotted Woodpecker	Great Spotted Woodpecker	Fieldfare
Mallard	Common Redshank	Gadwall
Water Rail	Mallard	Water Rail
Tawny Owl	Common Redshank	Tawny Owl
Water Rail	Great Spotted Woodpecker	Water Rail
Water Rail	Tawny Owl	Cetti Warbler
Eurasian Collared Dove	Great Spotted Woodpecker	Eurasian Collared Dove
Tawny Owl	Common Redshank	Icterine Warbler
Water Rail	Common Redshank	Water Rail
Common Redshank	Water Rail	Wood Sandpiper
Great Spotted Woodpecker	Eurasian Collared Dove	Great Spotted Woodpecker

Table1. Comparison table between some of our predictions and BirdNET predictions on the same audio. The first 10 rows are sampled among our correct predictions, the second 10 rows are sampled among our wrong predictions.

6 CONCLUSIONS

The prediction of our model for us is satisfying. Overall we reached an accuracy of 87 % on the test and an accuracy of 97% if we consider the correct label to be within the 3 top ranked classes. Some interesting facts also emerged from the post-model analysis, i.e. the confusion of the 2 classes like discussed in the Results section, issue that we couldn't precisely explain with our means. Lastly, even if our model was trained on only 10 classes, in many predictions it gave the same result of BirdNET, and in those where our model was having trouble, many times, also the BirdNET had the same trouble.

7 REFERENCES

[1] xeno-canto
[2] librosa
[3] BirdNET
[4] Douglas O'Shaughnessy (1987). Speech communication: human and machine.
[5] Mrinmoy Bhattacharjee, S.R.M. Prasanna, Prithwijit Guha, Time-Frequency Audio Features for Speech-Music Classification
[6] Xin Zhang and Zbigniew W. Ras (2007). Analysis of Sound Features for Music Timbre Recognition.
[7] Dan Stowell and Mark D. Plumbley (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning
[8] Stefan Kahl1, Thomas Wilhelm-Stein1, Holger Klinck3, Danny Kowerko2 and Maximilian Eibl1 (2018). Recognizing Birds from Sound - The 2018 BirdCLEF Baseline System

This paper has been typeset from a TEX/LATEX file prepared by the author.