Rapport livrable 5
Quebec

LIS CNRS

N. Blaukat, A. Grosmartial, H. Glotin
in collaboration with : A. Simard, J.-F. Jetté

Table des matières

1 Introduction

The following study dives into the downsides and gains of different active learning techniques when applied to a data set composed of chirps and vocalizations from more than two hundred species of birds and frogs. The labels of the studied files are based on expert annotations from J.F Jetté. Thanks to his intensive work and dedication, rendering this study possible, the effects of data augmentation as well as the influence of different hyperparameters will be tested. The main objective being to shed light on the importance of data augmentation and the quality of samples when building the best model possible to classify animal vocalizations.

There are multiple possibilities to obtain the representation of a signal in the time-frequency domain, used by convolutional neural networks. In this study the Wigner-Ville's distribution (WVD) will be used. While the short-time fourier transfer/spectrogram uses a fixed basis to analyze sounds, the WVD has an adaptive basis, which gives great representations of chirps. A chirp is a signal whose frequency varies through time, and is often found in bird vocalizations.

In order to achieve the objectives mentioned above, multiple experiments were conducted with different Wigner-Ville representations of the samples, and varying noises (type and level) added to the signal before hand.

2 Representation

From the expert annotations, multiple different data sets can be created. The main factor that comes to play while building these sets is the quality of the audio recording. These were labelled as either good, mid or bad. With the help of this criterion, species were ranked according to their harmonic mean :

$$I = \left(\frac{1}{\text{number of good samples} + 1e^{-3}} + \frac{1}{\text{number of samples} + 1e^{-3}} \right)^{-1}$$

This indicator looks at the relative importance of a class according to two important values, the number of different samples, necessary to get a broad overview of the class, and the quality of these samples, important to learn well.

rank	specie	number of samples	number of good samples	harmonic mean
1	wtsp	443	235	307.1
2	pscr	465	213	292.2
3	swsp	285	148	194.8
4	licl	266	133	177.3
5	oven	252	130	171.5
6	mawa	263	119	163.9
7	heth	231	101	140.5
8	yrwa	228	101	140.0
9	coye	225	100	138.5
10	nawa	217	97	134.1
11	alfl	200	95	128.8
12	veer	200	88	122.2
13	amre	193	87	119.9
14	phre	187	84	115.9
15	resq	165	83	110.4

TABLE 1 – Fifteen first species ranked according to their harmonic mean

Once the species were ranked according to their potential in being well classified by the model, different sets were created.

$$\begin{cases} \textbf{Set 1 :} \text{ First 5 classes [1...5] with the highest I, 1711 samples (859 good)} \\ \textbf{Set 2 :} \text{ Next 5 classes [6...10] with the highest I, 1164 samples (518 good)} \\ \textbf{Set 3 :} \text{ Next 5 classes [11...15] with the highest I, 945 samples (437 good)} \end{cases} \quad (1)$$

As explained in the introduction, the sound files are then switched from the time domain to the time-frequency domain, obtaining a 2D representation of the signal adapted to convolution. This switch to the time-frequency domain is done with the WVD. This distribution has many parameters, however the one that will be experimented upon is the value of L, which changes the

frequency precision. For L = 1, the transform will be a simple spectrogram. When L increases, chirps will gain in precision while constant frequency time periods will be blurred.

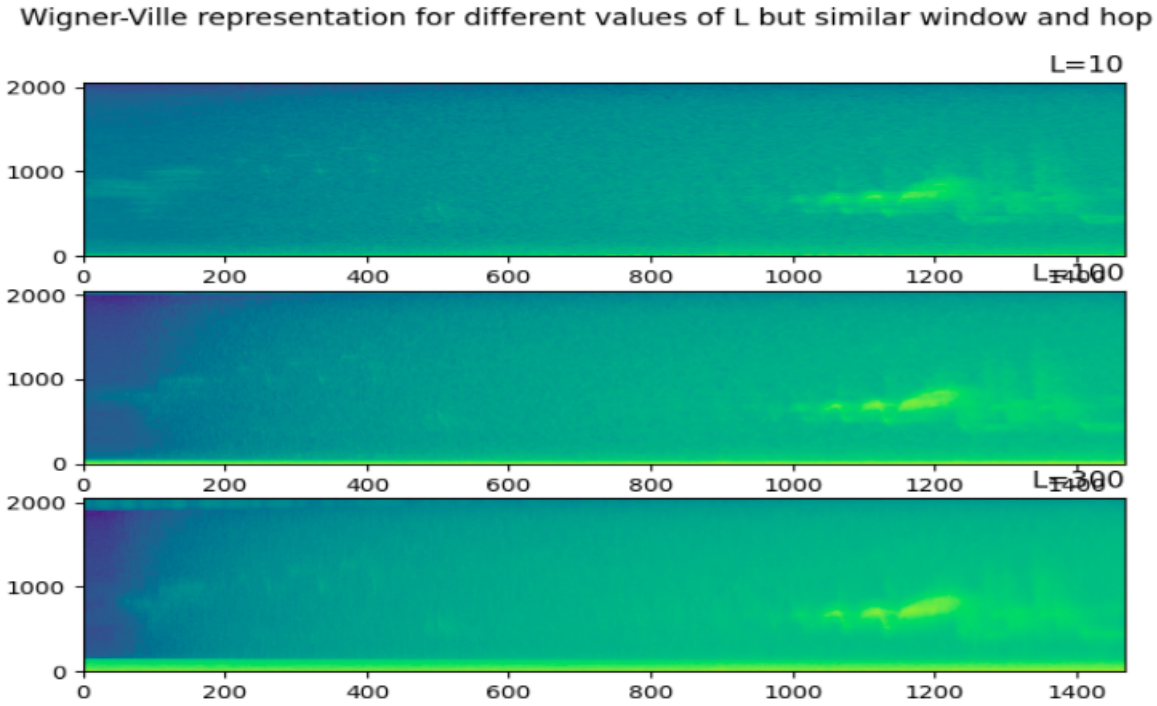


FIGURE 1 – Evolution of the representation of a signal for increasing values of L

Representations similar to the ones pictured above are fed to a neural network, the difference being noise was added before the switch to the time-frequency domain. The WVD of a 10 second signal is an image of shape (1024 x 930). The neural network chosen for all experiments is the AlexNet model represented below. It model was tested beforehand on the AudioMNIST dataset, consisting of 30 000 samples of 60 different people saying numbers from 0 to 10 and predicted 100% accurately on the testset.

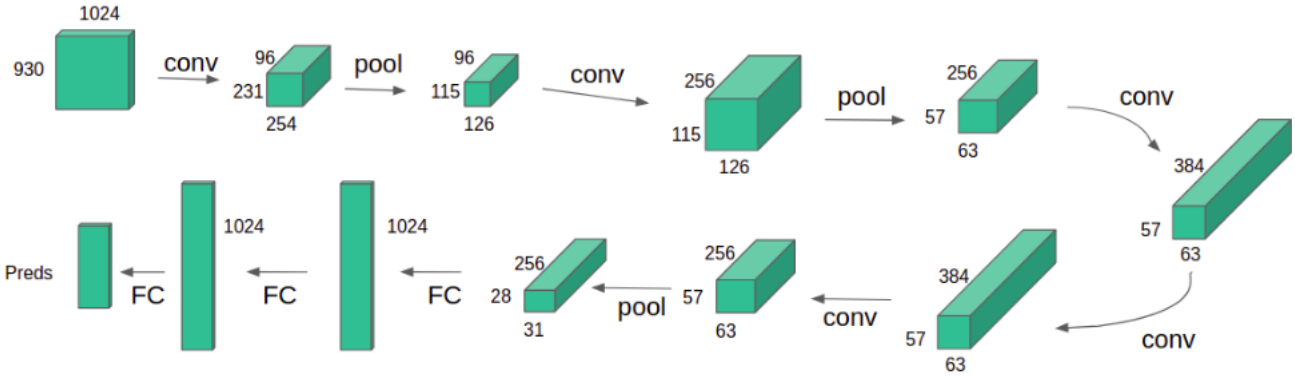


FIGURE 2 – AlexNet model applied to a 1024x930 input

3 Parameters

3.1 Data augmentation

Annotating sound samples for active learning proving to be costly, the use of data augmentation leads to an automatic and fast extension of the annotated corpus, that leads to better generalization results. This study is based on different types of data augmentation, all tripling the size of the data set. The samples will be fed to the neural network with three distinct representations. These three representations are created by juxtaposing noise to the original sample at different levels or Sound to Noise Ration. The different values of SNR used for this study are $SNR \in \{\infty, 6, 12\}$. The following noises were used in independent experiments :

White noise :

$$x_{noise}[n] = x_{norm}[n] + \vec{\epsilon} \quad \text{with} \quad \begin{cases} \vec{\epsilon} \text{ with random values in} \left[0, 10^{-6/20}\right] & \text{for a SNR of 6} \\ \vec{\epsilon} \text{ with random values in} \left[0, 10^{-12/20}\right] & \text{for a SNR of 12} \end{cases} \tag{2}$$

Brown noise :

$$x_{noise}[n] = x_{norm}[n] + \text{cumulative sum}(\epsilon) \quad \text{with} \quad \begin{cases} \vec{\epsilon} \text{ w/ rnd values in} \left[0, 10^{-6/20}\right] & \text{for SNR=6} \\ \vec{\epsilon} \text{ w/ rnd values in} \left[0, 10^{-12/20}\right] & \text{for SNR=12} \end{cases} \tag{3}$$

Pink noise :

```
def pink_noise(size, rng, ncols=16, axis=-1):
    """Generates pink noise using the Voss-McCartney algorithm.

    size: either a tuple of int or an int. If an int : number of sample to generate.
    ncols: number of random sources to add.
    axis: axis which contains the sound samples. Generate white noise otherwise.

    returns: NumPy array of shape size
    """
    if type(size) is not tuple:
        size = (size,)
    array = rng.rand(*size)
    assert -len(size) <= axis < len(size)
    axis  %= len(size)
    axis +=1
    # the total number of changes is nrows
    cols = rng.geometric(0.5, size)
    cols[cols >= ncols] = 0
    cols = (1.*(np.arange(1,ncols).reshape((-1,) + len(size)*(1,)) == cols)).swapaxes(axis,-1)
    cols[...,0] = 1.
    cols = np.cumsum(cols).reshape(cols.shape).astype(int).swapaxes(axis,-1)
    array = np.concatenate([array[np.newaxis],rng.rand(cols.max()+1)[cols]],axis=0).sum(0)
    return array
```

Wind noise :

$$x_{noise}[n] = x_{norm}[n] + 10^{-SNR/20} \cdot \text{wind}_{norm} \tag{4}$$

with wind noise from https ://www.youtube.com/watch ?v=jSQ2sLhTcvY&t=57s

While these four added noises were the only ones used, it is completely possible to add anthropogenic noises such as traffic, or rain and storm noises. Both could be found in certain samples down the road and are thus relevant data augmentation techniques for this study.

3.2 Hyperparameters

Data augmentation being the main parameter, many different experiments were done with the same added noise, by changing the learning rate and the weight decay loss. While the ratio of the batch size to the learning rate affects generalization, the batch size was fixed to 8 samples for all experiments, and the learning rate varied in the interval $[0.001, 0.1]$. The weight decay loss varied from $[0.0002, 0.01]$. All experiments were done relative to a control group with the following parameters :

$$\begin{cases} \text{Learning rate :} \;\; LR = 0.005 \\ \text{Weight decay loss :} \;\; WDL = 0.0002 \\ \text{Batch size :} \;\; BS = 8 \end{cases} \tag{5}$$

When one parameter (LR or WDL) was experimented upon, the other one was fixed.

4 Results with the wind noise

Down below are the results for the experiments with the data augmentation based on the wind sound taken from YouTube, with the same sampling frequency as the samples. As we can see, some parameters lead to strong overfitting and no model was able to generalize. For the experiments, only the samples that were mentioned to be of good quality from the first five classes were considered. For each sample, the three SNRs were fed to the neural network in the training part, and only the original signals' spectrograms were kept for the testing.

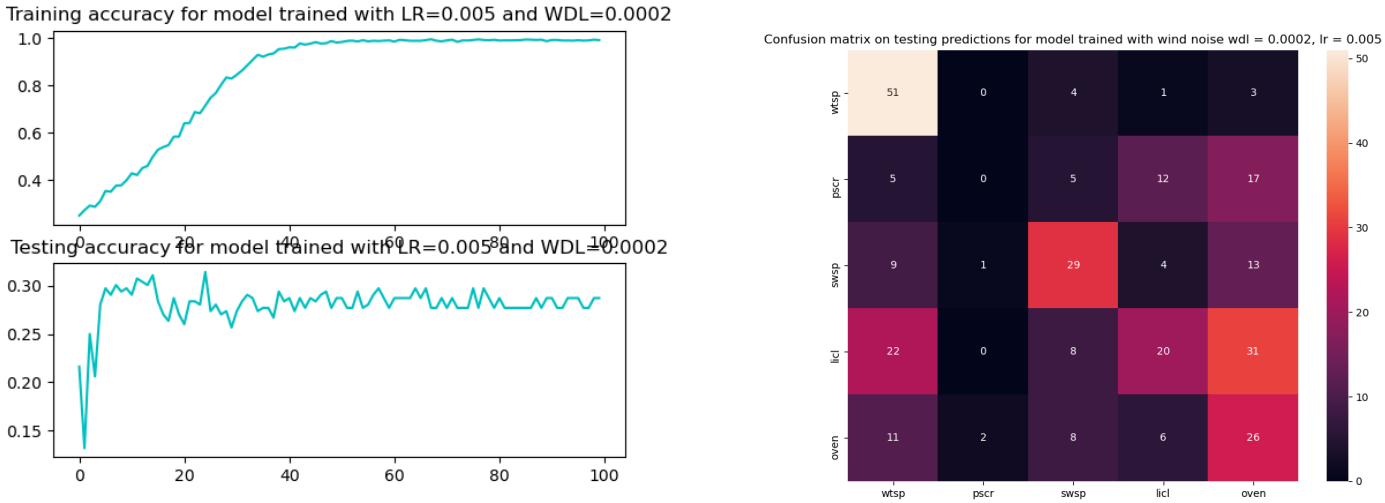


FIGURE 3 – Training/testing accuracy and confusion matrix for model trained with wdl = 0.0002 and lr = 0.005

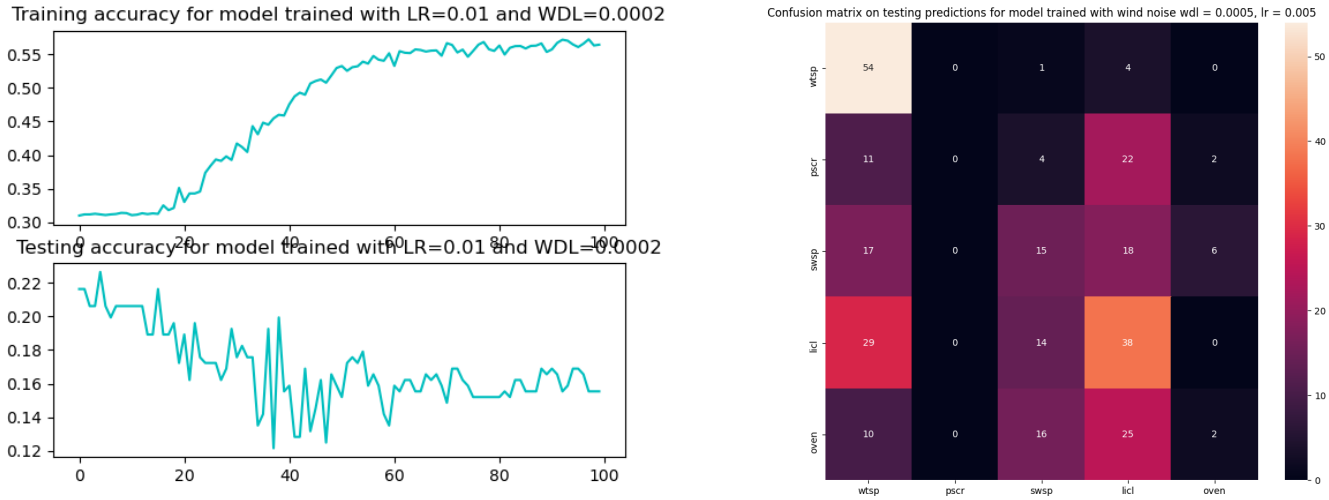


FIGURE 4 – Training/testing accuracy and confusion matrix for model trained with wdl = 0.0005 and lr = 0.005

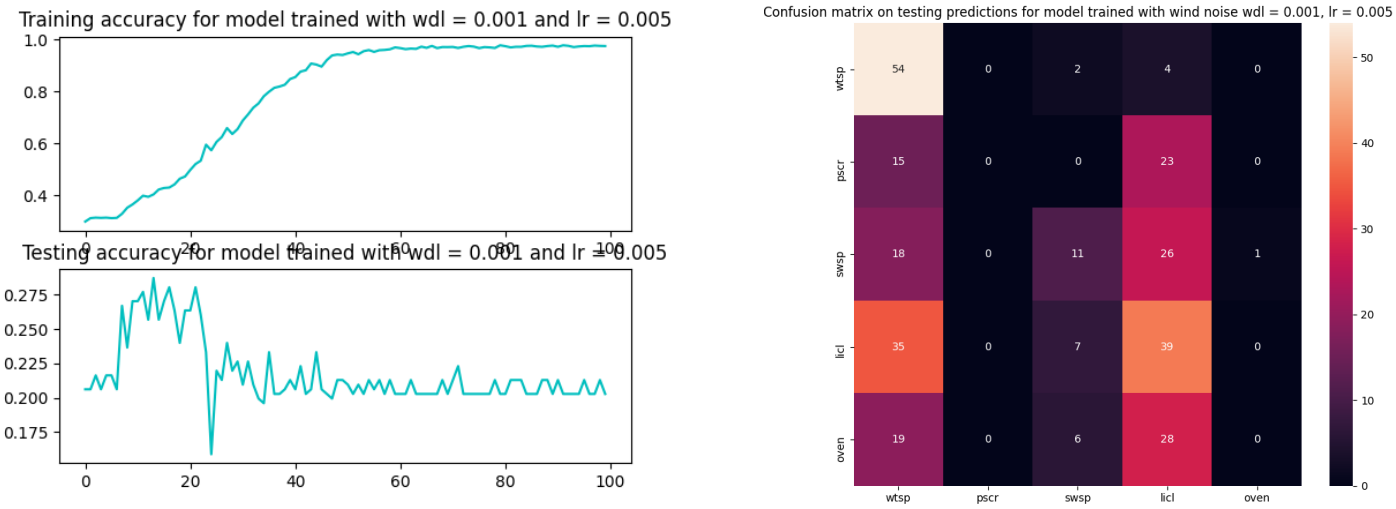


FIGURE 5 – Training/testing accuracy and confusion matrix for model trained with wdl = 0.001 and lr = 0.005

Presented above are the best three experiments with wind noise, with poor results, reaching about 30% accuracy on the testing set.

5 Results with white noise

Down below are the results for the experiments with the data augmentation with white noise added to the signal. In most experiments, two classes are over predicted to the detriment of others. This result is shown here were two classes (columns) are almost never outputs of the model.

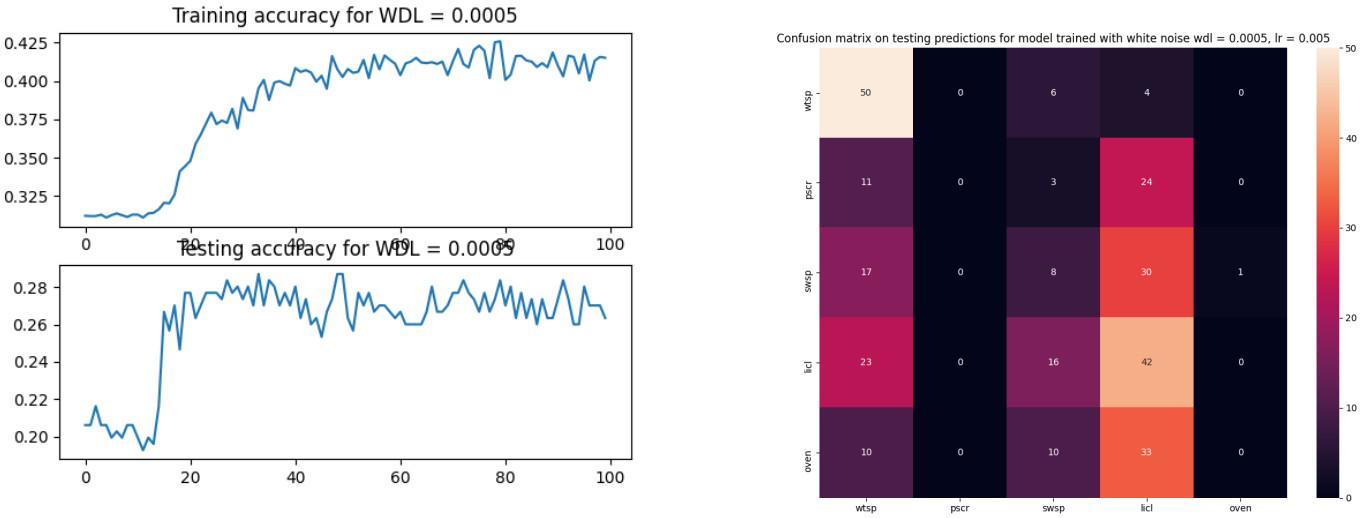


FIGURE 6 – Training/testing accuracy and confusion matrix for model trained with wdl = 0.0005 and lr = 0.005

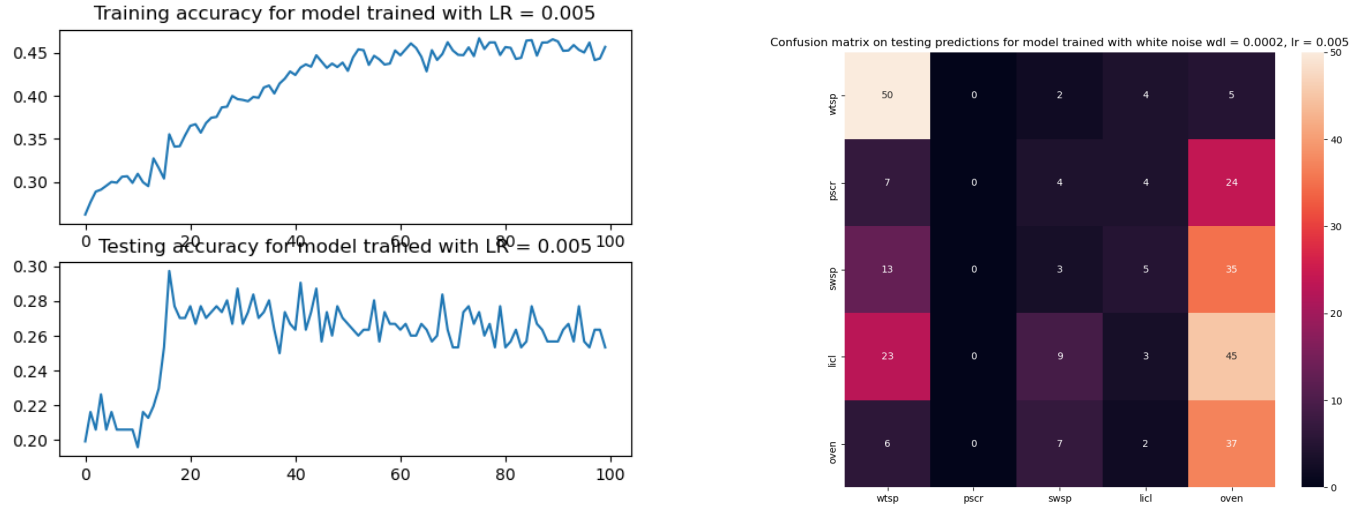


FIGURE 7 – Training/testing accuracy and confusion matrix for model trained with wdl = 0.0002 and lr = 0.005

These two experiments are the best of five experiments with different hyperparameters. However, once again, they did not do classify the test samples well. Some experiments here, as opposed to the ones with the wind noise, did not lead to overfitting and the model did not learn the train set as well.

6 Results after removing potentially overlapping samples

The presence of multiple species in the same 10s windows could explain the poor results. In this section, we removed samples with overlapping bird calls/songs from the original 11,051 samples, shrinking the training set to 7,601 samples. 5,018 samples present species that are found both in xeno-canto and Quebec datasets.

In order to assess the quality of Quebec samples, a model was trained on xeno-canto and alternatively tested on : Quebec 5,018 samples, and a set of previously unseen xeno-canto data. Figure 9 and 10 display the two confusion matrices associated to these tests. The model training, validation and testing curve is presented in Figure 8.

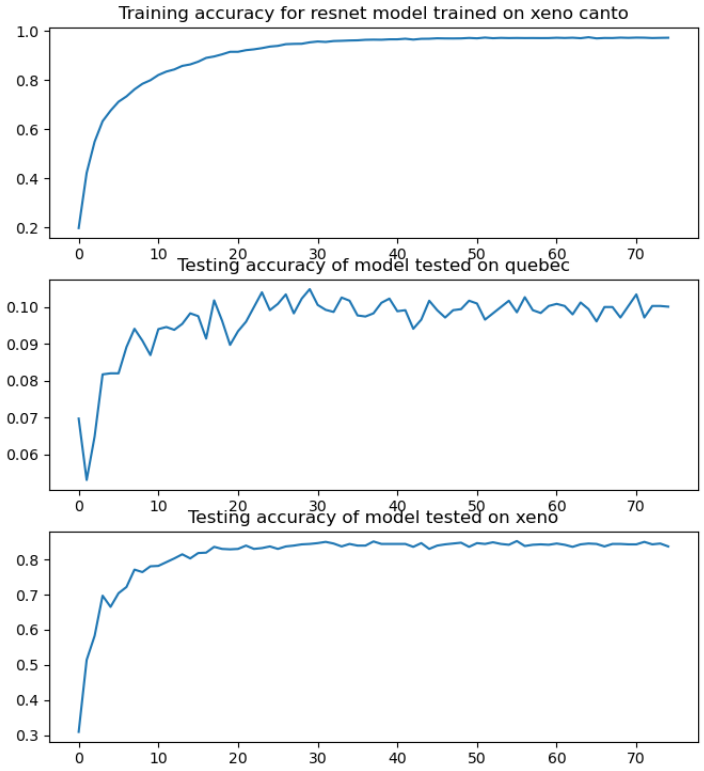


FIGURE 8 – Resnet 18 accuracies while training on xeno canto

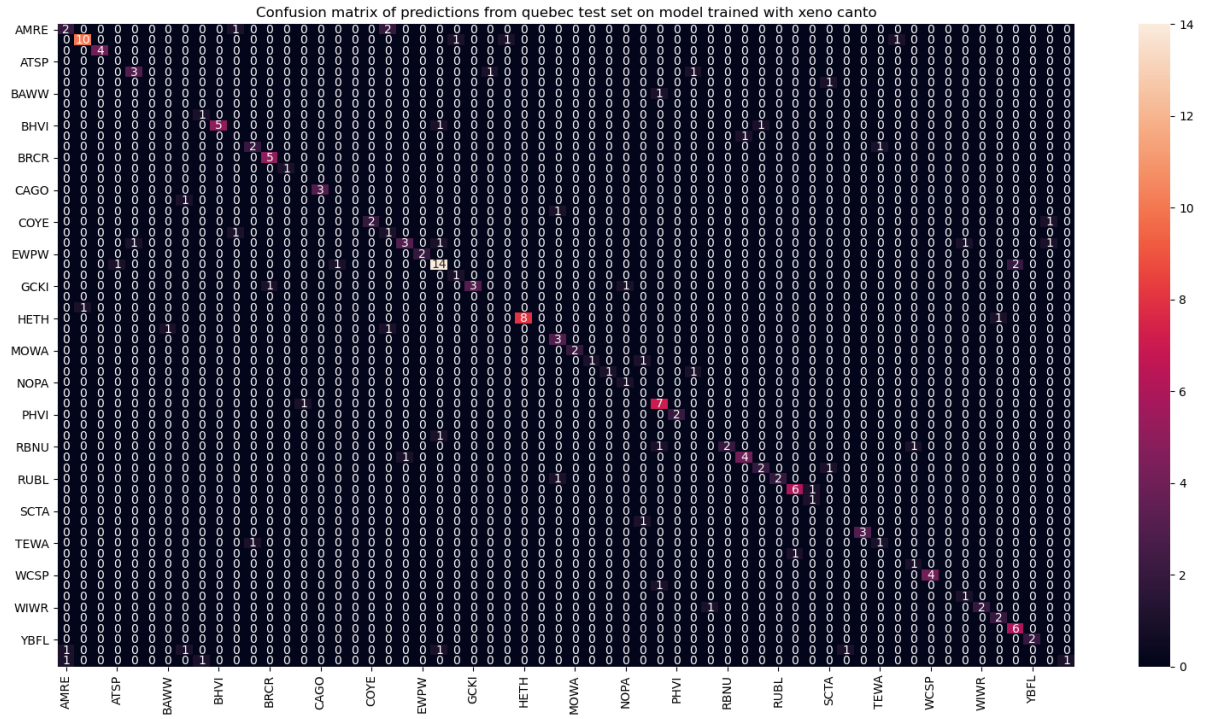


FIGURE 9 – Confusion matrix of xeno-canto testset. X and Y-axis are species labels.

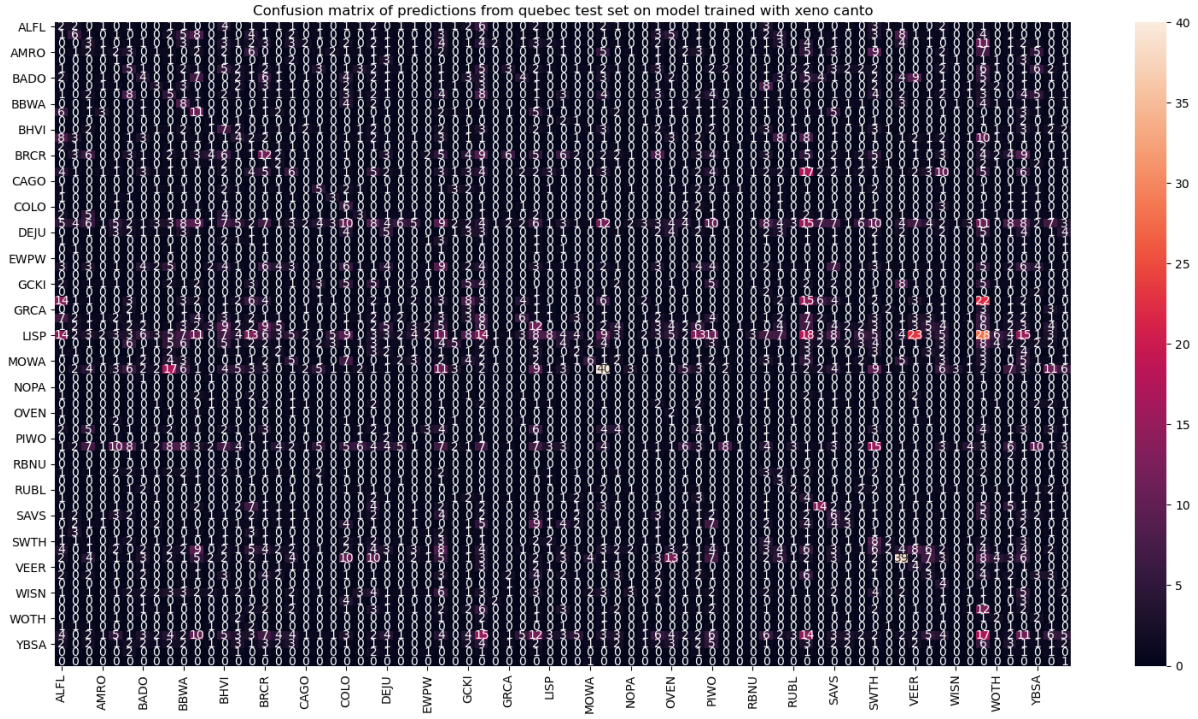


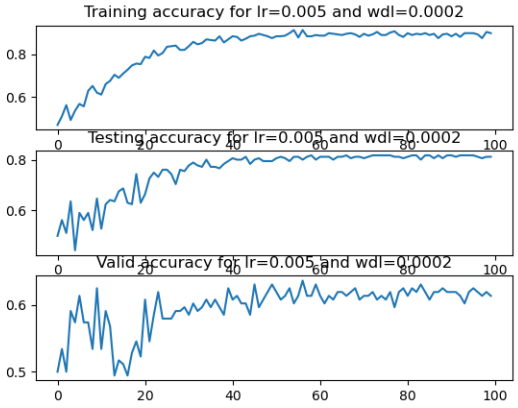
FIGURE 10 – Confusion matrix of Quebec testset. X and Y-axis are species labels.

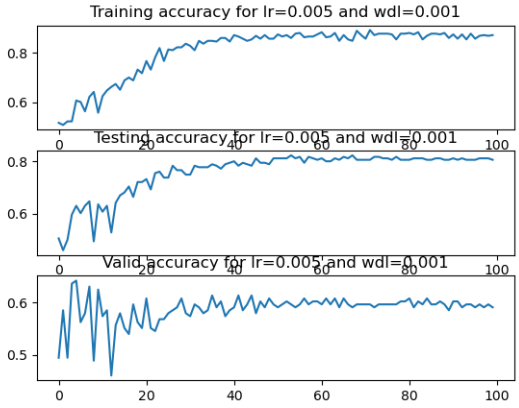
The model stopped training on xeno-canto when it reached 78% accuracy on the training set. When testing on this model, it performed well on xeno-canto (72%) and poorly on Quebec (6.9 %). Two classes of the model are over-predicted. However, the problem lies in the model and not in these two species because removing them from the testing set does not improve the overall score.

7 Learning the model classification ability

From the 5,018 samples from Quebec tested on the xeno-canto model, 344 were well classified. This following section tries to understand whether the model can distinguish classifiable and not-classifiable samples. 344 correctly classified samples and 344 incorrectly classified samples were taken to form a new dataset to train, test and validate a new model, with the model ability to classify the sample as new labels. In total, 4 different tests were performed on 10s samples and 1.5s samples : with and without data augmentation (white noise and wind noise), and with two different values of weight decay loss (wdl). The first 1.5 seconds record the background noise of a sample.

7.1 Training on full length samples





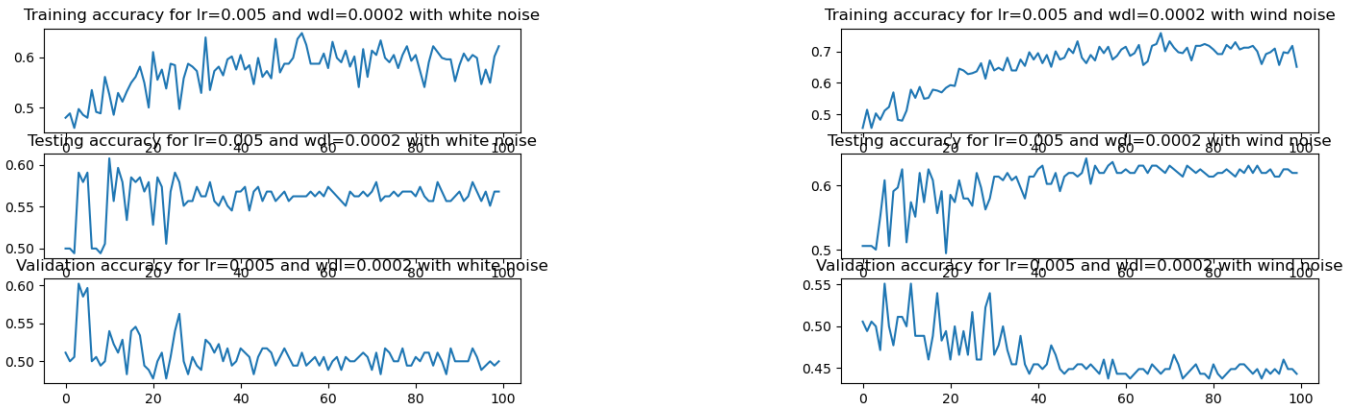
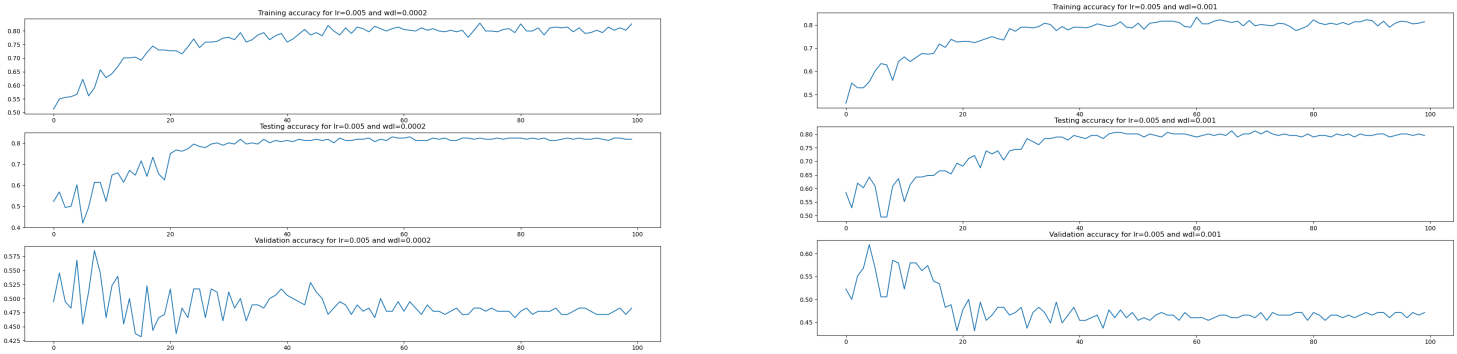
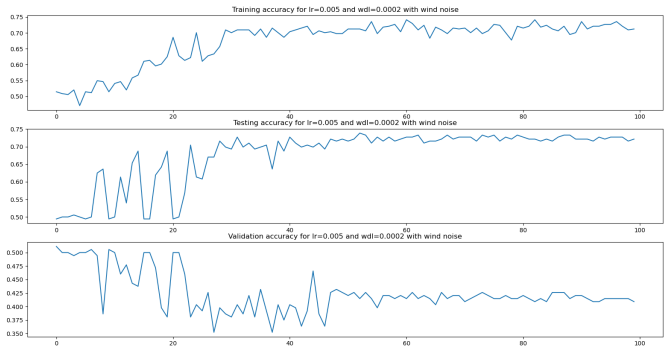


FIGURE 11 – Training, testing and validation accuracy for AlexNet model trained with different parameters

Background noise between classifiable and not-classifiable signals are different (Figure 11). Models without data augmentation differentiate good and bad samples for classification, reaching 80% on the testset and above 60% on the validation set for balanced classes. The two experiments with data augmentation did not perform as well as the experiments without data augmentation. The added noise seems to hide differences between samples.

7.2 Training on background noise only





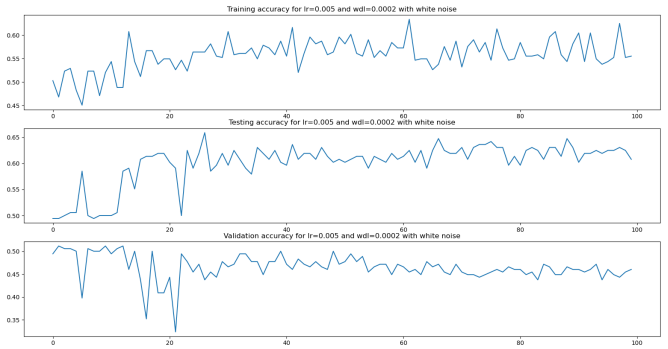


FIGURE 12 – Training, testing and validation accuracy for AlexNet model trained with different parameters on first 1.5 seconds of signal

Similarly as in the previous experiment, this model is able to sort classifiable and not-classifiable samples (Figure 12). However, the model did not perform well on the validation set. Although background noise is a decisive factor to distinguish classifiable and not-classifiable samples, it is not the only one. Other differences, if not in the background noise, are related to bird vocalizations, and are not present in this experiment.

8 Effect of weather on model performance

The previous section established that the model can distinguish classifiable from not-classifiable samples. Therefore, finding out the conditions that enable the model to correctly classify is essential. The conditions leading to mostly wrong predictions will be excluded from Quebec dataset to feed classifiable samples to the model. Here, we investigated the correlation between weather conditions and the model performance.

Weather conditions are extracted from hourly meteorological data originating from the closest weather station to each sampling site (Table 2).

TABLE 2 – Corresponding weather stations and sampling sites (and their latitude and longitude).

Sampling sites	Latitude	Longitude	Weather station
057_156_H01	58.51665	-77.99432	Inukjuak (ECCC)
057_156_T01	58.49209	-78.08743	Inukjuak (ECCC)
057_182_T01	61.31273	-73.66665	Parc national des Pingualuit (ECCC)
073_137_F02	55.29361	-77.69222	Kuujjuarapik
080_175_T01	58.646671	-69.997047	Aux Feuilles
086_180_H01	58.05207	-68.53668	Kuujjuaq A (NavCAN)
086_180_H02	58.34573	-68.35947	Kuujjuaq A (NavCAN)
086_180_T01	58.05353	-68.5117	Kuujjuaq A (NavCAN)
086_180_T02	58.20953	-68.3735	Kuujjuaq A (NavCAN)
090_189_H01	58.73125	-66.01998	Kangiqsualujjuaq A (NavCAN)

090_189_T01	58.72503	-66.00285	Kangiqsualujjuaq A (NavCAN)
099_088_F01	48.14553	-79.27893	-
099_088_H01	48.10778	-79.51675	-
105_101_F01	49.10239	-76.99387	Quévillon (SOPFEU)
105_101_H01	48.8313	-77.12239	Quévillon (SOPFEU)
111_115_F01	49.83417	-74.37779	Chibougamau-Chapais (ECCC et NavCAN)
111_115_H01	49.843	-74.3822	Chibougamau-Chapais (ECCC et NavCAN)
122_092_F01	46.77018	-75.46088	-
122_092_H01	46.87738	-75.56468	Mont-Saint-Miichel
122_092_H02	46.78965	-75.47688	Mont-Saint-Miichel
124_086_F01	46.08892	-75.83353	Barrage Rapides-des-Cèdres
124_086_H01	45.91049	-76.04464	Lac-Sainte-Marie (FADQ)
124_086_H02	45.97708	-75.93081	Lac-Sainte-Marie (FADQ)
127_116_H02	48.75341	-72.05228	Chute-du-Diable (Rio Tinto)
128_089_F01	45.98357	-75.16566	-
128_089_H01	46.00082	-75.17989	-
128_089_H02	45.96686	-75.16729	-
129_094_H01	46.41719	-74.40726	-
129_123_F01	49.11631	-70.60315	Onatchiway (ECCC)
129_123_H01	49.00715	-70.66965	Onatchiway (ECCC)
130_086_F01	45.60086	-75.12686	-
131_120_F01	48.58144	-70.90059	Falardeau
131_120_F02	48.60349	-70.82977	Falardeau
131_120_H01	48.57225	-70.86144	Falardeau
132_116_F01	48.21545	-71.26847	Chicoutimi
132_116_H01	48.21511	-71.27799	Chicoutimi
135_104_F01	46.79467	-72.30312	Lac-aux-sables
135_104_H01	46.80876	-72.29919	Lac-aux-sables
136_095_F01	45.94968	-73.43805	L'Assomption (ECCC)
136_095_H01	45.99033	-73.29996	L'Assomption (ECCC)
136_116_T01	47.66656	-70.7793	La-Galette
137_107_F01	46.95814	-71.69524	Dunford (SOPFEU)
137_107_H01	46.87663	-71.66492	Sainte-Catherine-de-la-Jacques-Cartier
137_107_H02	46.86802	-71.67593	Sainte-Catherine-de-la-Jacques-Cartier
137_107_H03	46.73119	-71.43624	Québec\Jean-Lesage Intl (ECCC)
137_110_H01	47.11555	-71.36092	Parc national de la Jacques-Cartier
137_110_H02	47.45812	-71.24557	L'Étape (ECCC)
137_111_F01	47.30817	-71.16354	Forêt Montmorency
137_111_H01	47.25727	-71.16294	Forêt Montmorency
137_144_F01	50.20795	-66.67467	Pointe-Noire-CS (ECCC)
137_144_H01	50.19985	-66.56328	Pointe-Noire-CS (ECCC)

138_093_F01	45.537	-73.31895	-
138_093_H01	45.54361	-73.31133	-
139_087_F01	45.03273	-73.78258	Hemmingford-Four-Winds
139_087_H01	45.00642	-73.81944	Hemmingford-Four-Winds
139_087_H02	45.02227	-73.89027	Hemmingford-Four-Winds
139_103_F01	46.2486	-71.95004	Lemieux (ECCC)
139_103_H01	46.39063	-71.81608	Fortierville
139_103_H02	46.65885	-71.83463	Deschambault
141_108_F01	46.67234	-71.03031	Beauséjour
141_108_H01	46.78227	-71.03661	Lauzon
142_091_F01	45.02226	-73.06106	-
142_091_H02	45.02898	-73.07446	-
142_111_F01	46.71082	-70.70377	Armagh-2
142_111_H01	46.8895	-70.43522	Notre-Dame-du-Rosaire
142_111_H02	46.8705	-70.47272	Notre-Dame-du-Rosaire
145_102_F01	45.859	-71.18744	Barrage Jules-Allard
145_102_H01	45.96444	-71.13774	Barrage Jules-Allard
145_102_H02	45.84158	-71.17387	Barrage Jules-Allard
145_141_F01	48.93628	-66.04865	Mont-Ernest-Laflamme (?)
145_141_H01	49.09083	-66.03606	Petit-Mont-Saint-Anne (?)
146_133_F01	48.36166	-67.0126	Marguerite
146_133_H01	48.48845	-67.04545	Marguerite
146_133_H02	48.48683	-67.11034	Marguerite
148_101_F01	45.45978	-70.99964	La Patrie
148_101_H01	45.447226	-70.886072	La Patrie
149_142_F01	48.98756	-65.49925	-
149_142_H01	48.93254	-65.32671	-
149_142_H02	48.94138	-65.38064	-

8.1 weather and overall model performance

Air water content, extreme temperature and rain can impact the recording quality long before the measurement. Therefore averaged data on the last 12 and 24 hours before the recording are processed along with weather data at the recording time. Using recordings made in the first week of May, June, July and August of each year in every site for which weather data are available, the correlation between prediction accuracy and weather data was computed. A standard PCA was performed on the data (Figure 14 and 15). Bad prediction accuracy on Quebec dataset mostly occurs for medium to high wind speed, medium temperatures, low rainfall values, and at low latitudes. However, the prediction success arrow does not seem to strongly correlate with presented axes, indicating other explaining parameters for the poor prediction accuracy of the model on Quebec dataset.

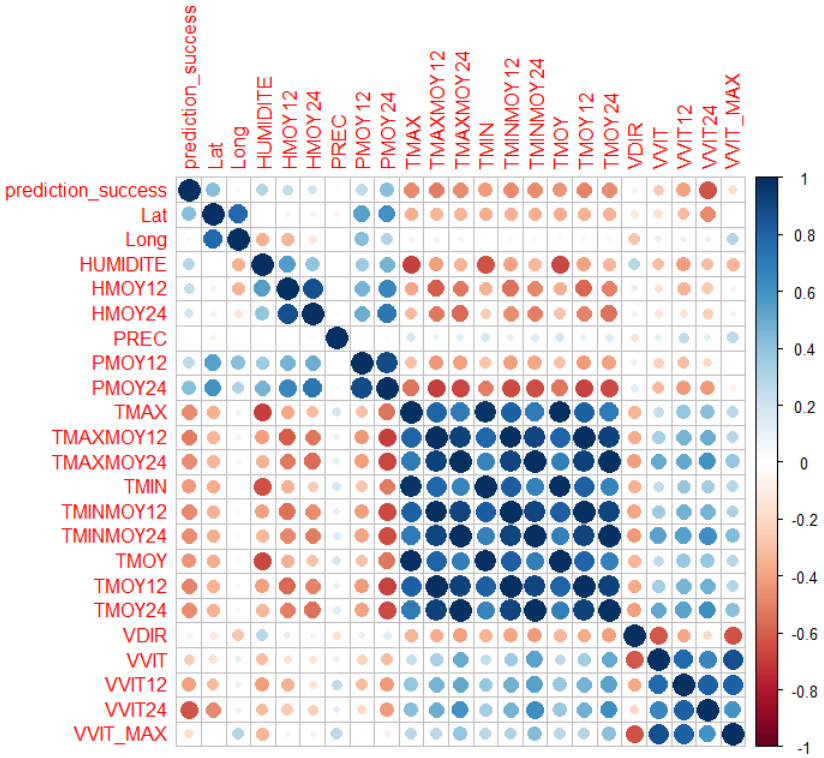


FIGURE 13 – Correlation matrix between prediction success and weather data

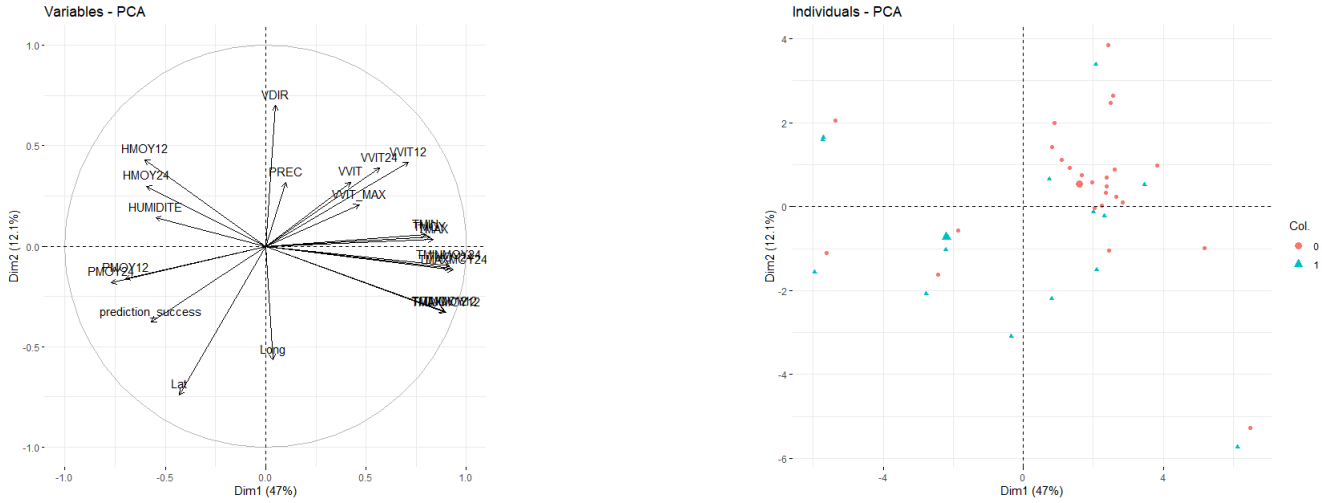


FIGURE 14 – Main axes of data variation in a two-dimensional plane and data projection with correct (1) and wrong (0) prediction with weather data 4 h before recording

wind	average wind velocity (m/s) over the hour
temp	average temperature (°C) over the span of the hour
prec	average precipitation (mm) over the span of the hour
result	average score of the model on all samples of a recording (1 - 100% correct, 0 - 0% correct)

TABLE 3 – Explicit legend for the following graphs.

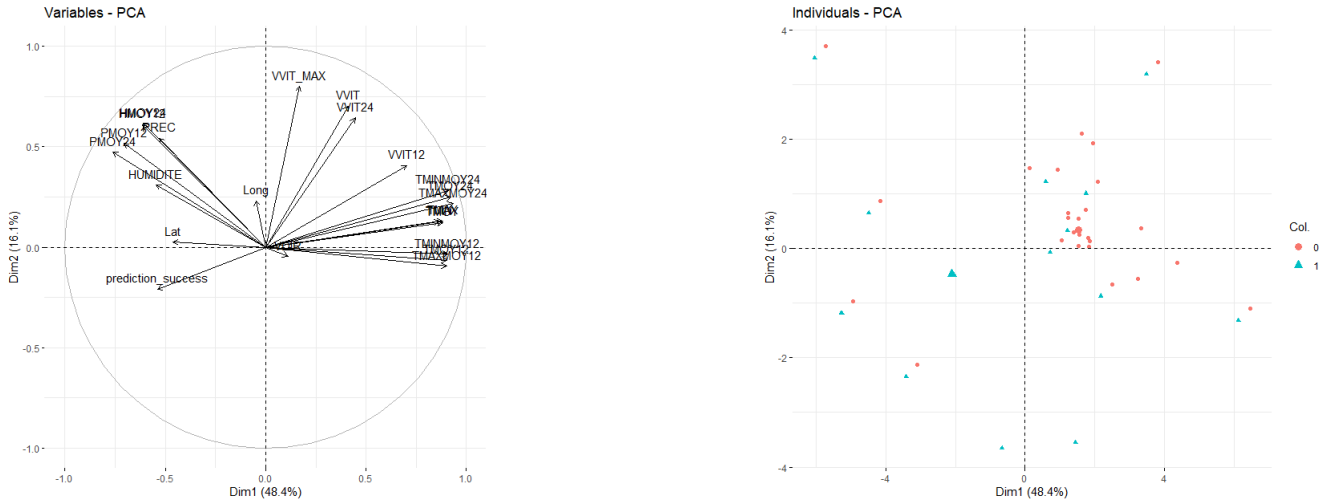


FIGURE 15 – Main axes of data variation in a two-dimensional plane and data projection with correct (1) and wrong (0) prediction with weather data 8 h before recording

These global results show no clear correlation between weather and predictions. The following subsections take a closer look over a 24h time period at the model performance and the weather, with the y-axis using a symmetrical log scale. Here is the legend for every graph :

8.2 Wind speed

In Figure 16, the result is the log of the mean of predictions (1 for correct, 0 for incorrect) for every hour in a recording site. A 24h span is too narrow to see any tendency.

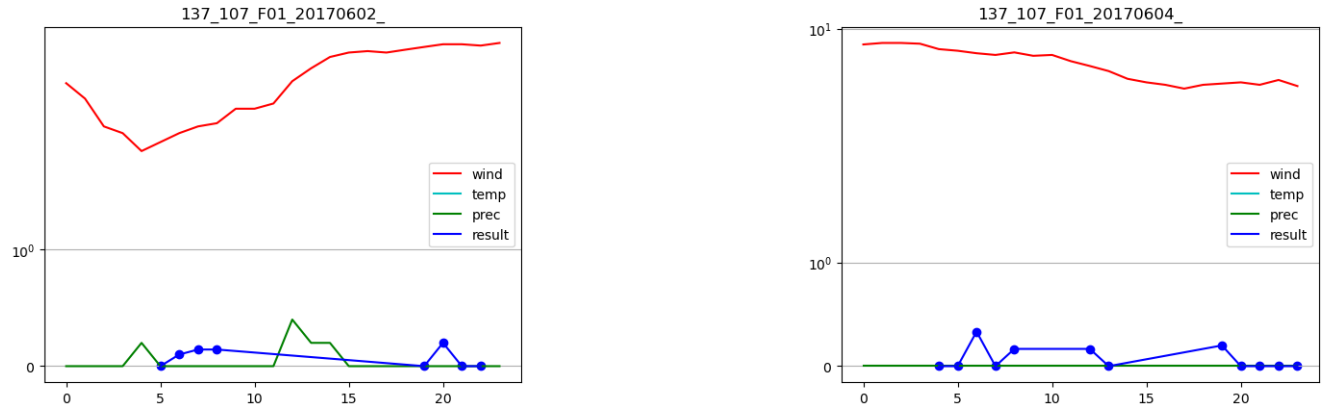
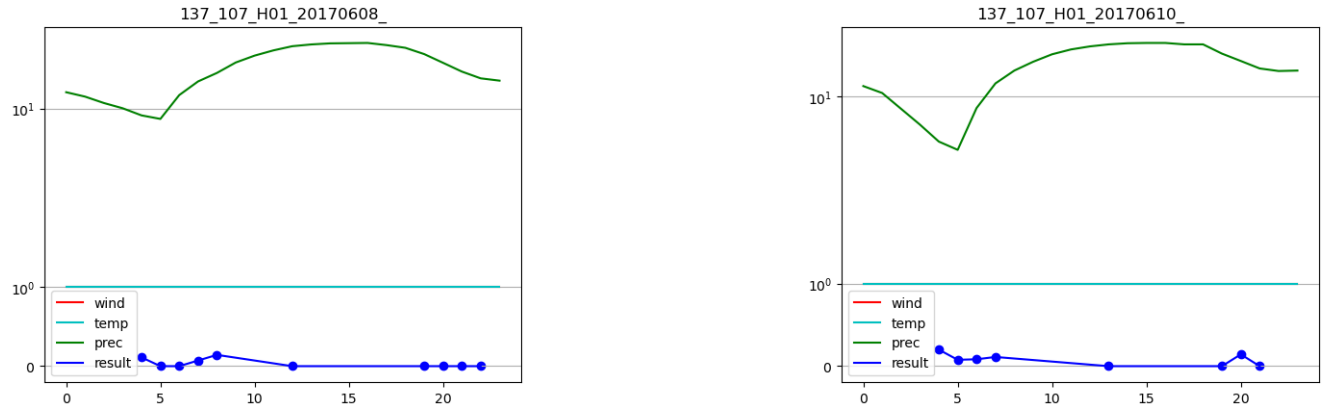


FIGURE 16 – Wind speed (km/h) and model performance for two specific dates in one site

8.3 Precipitation



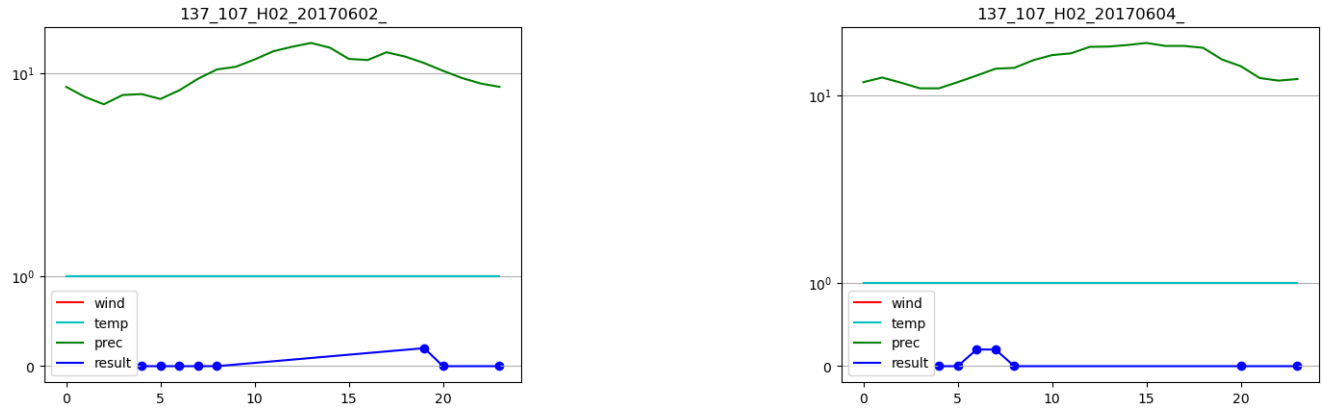
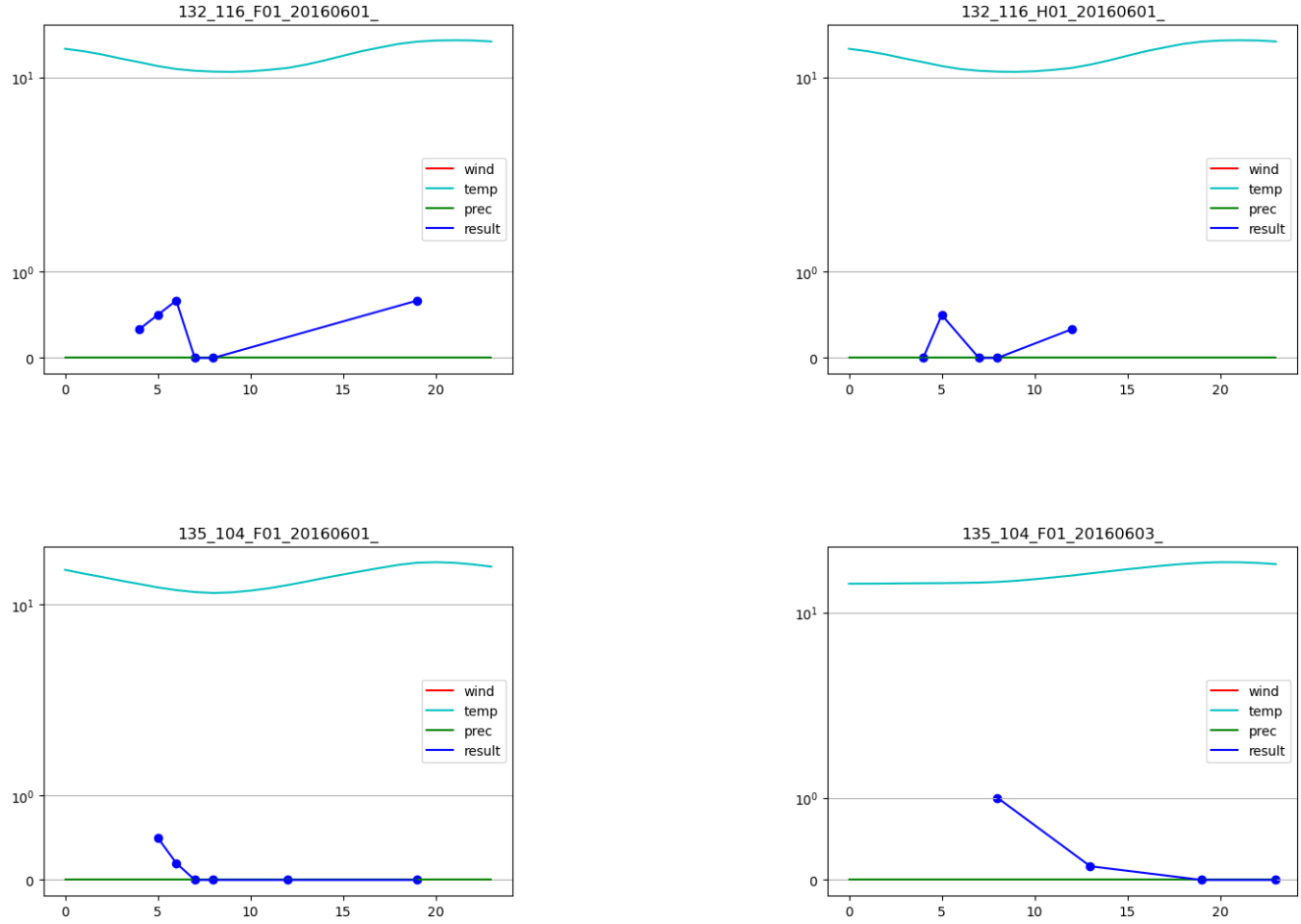


FIGURE 17 – Precipitation (mm) and model performance on four specific dates in one site

8.4 Temperature



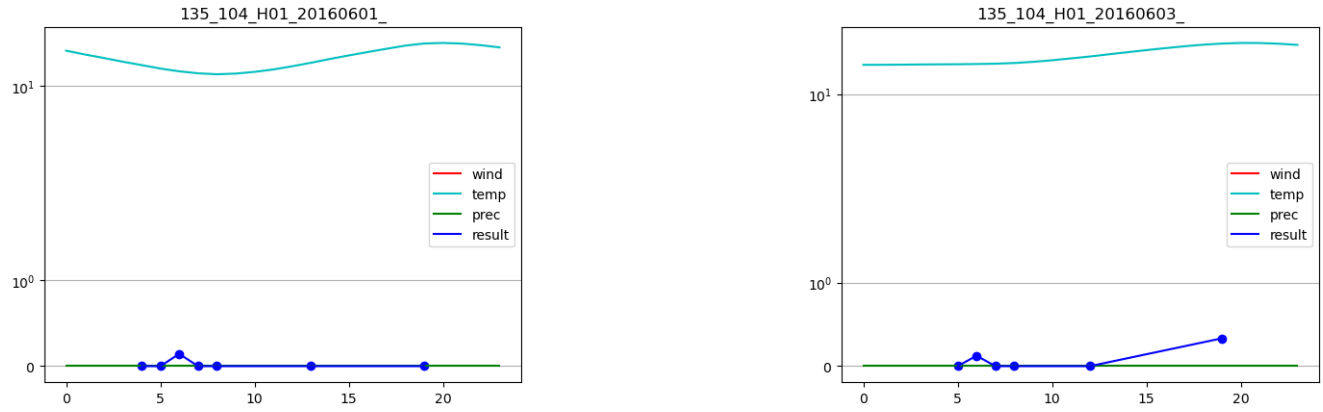


FIGURE 18 – Temperature (°C) and model performance on six specific dates in one site

The model performance does not seem to correlate to weather data. However, it looks like rain lowers a sample probability of being correctly classified. Wind does not seem to have a major impact on the AlexNet model, but the direction of the wind is not known and might be an important factor. The temperature does not vary enough throughout a given day to see a clear influence.

9 Influence of recording sites on the prediction accuracy

The model performs differently between sites (Figure 19). However, this distribution can be approximated by a Gaussian curve. Therefore, the site does not seem to play a role on the model performance. The model fails to correctly identify samples from several sites, namely 122_092_H02, 105_101_F01, 111_115_F01, 057_182_T01, and 136_095_H01. No correlation between these sites and the number of samples available for each of them has been noticed.

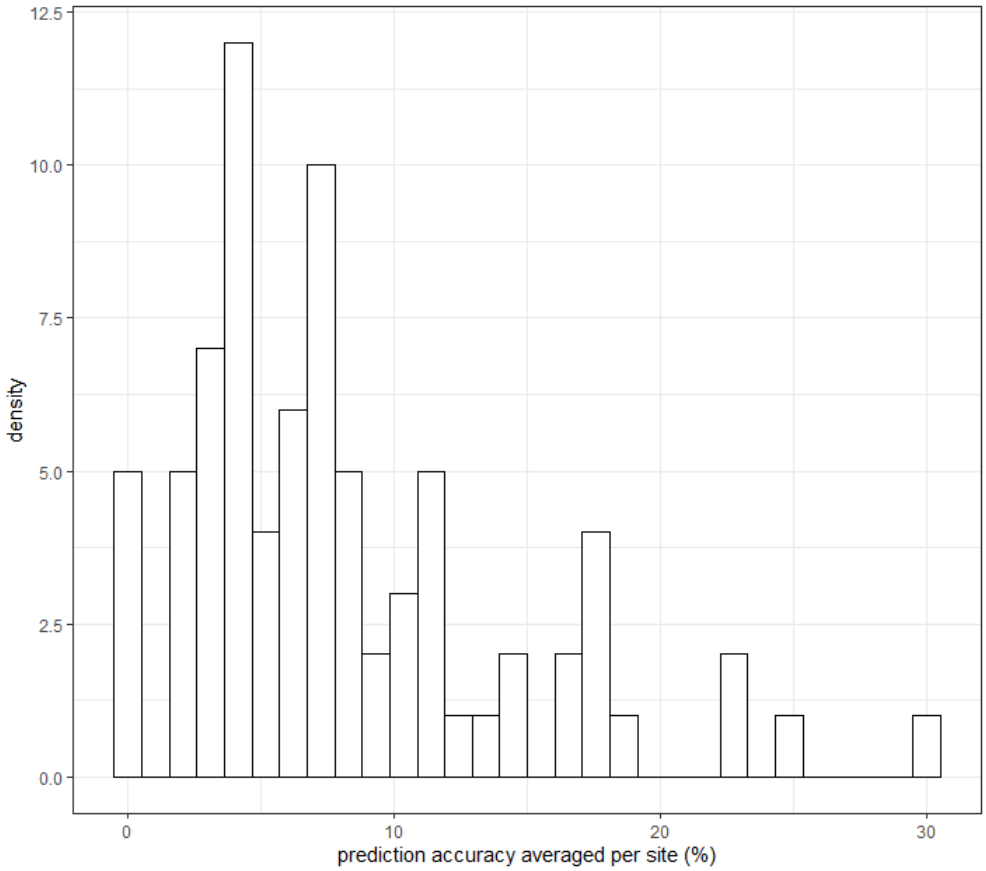


FIGURE 19 – Distribution of prediction accuracy per site (average in %)

We investigated the potential link between geographic coordinates and prediction accuracy (Figure 20, 21, 22).

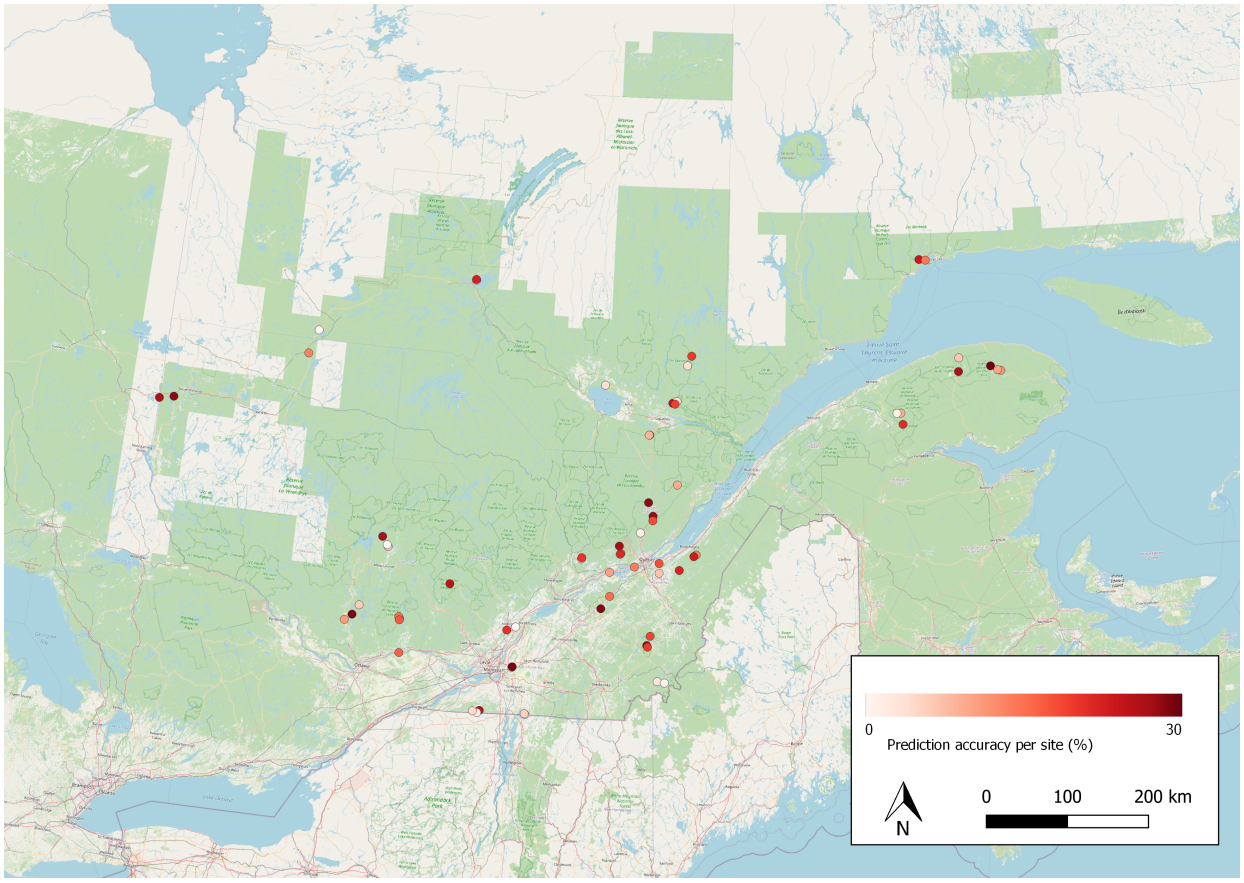


FIGURE 20 – Prediction accuracy per localization

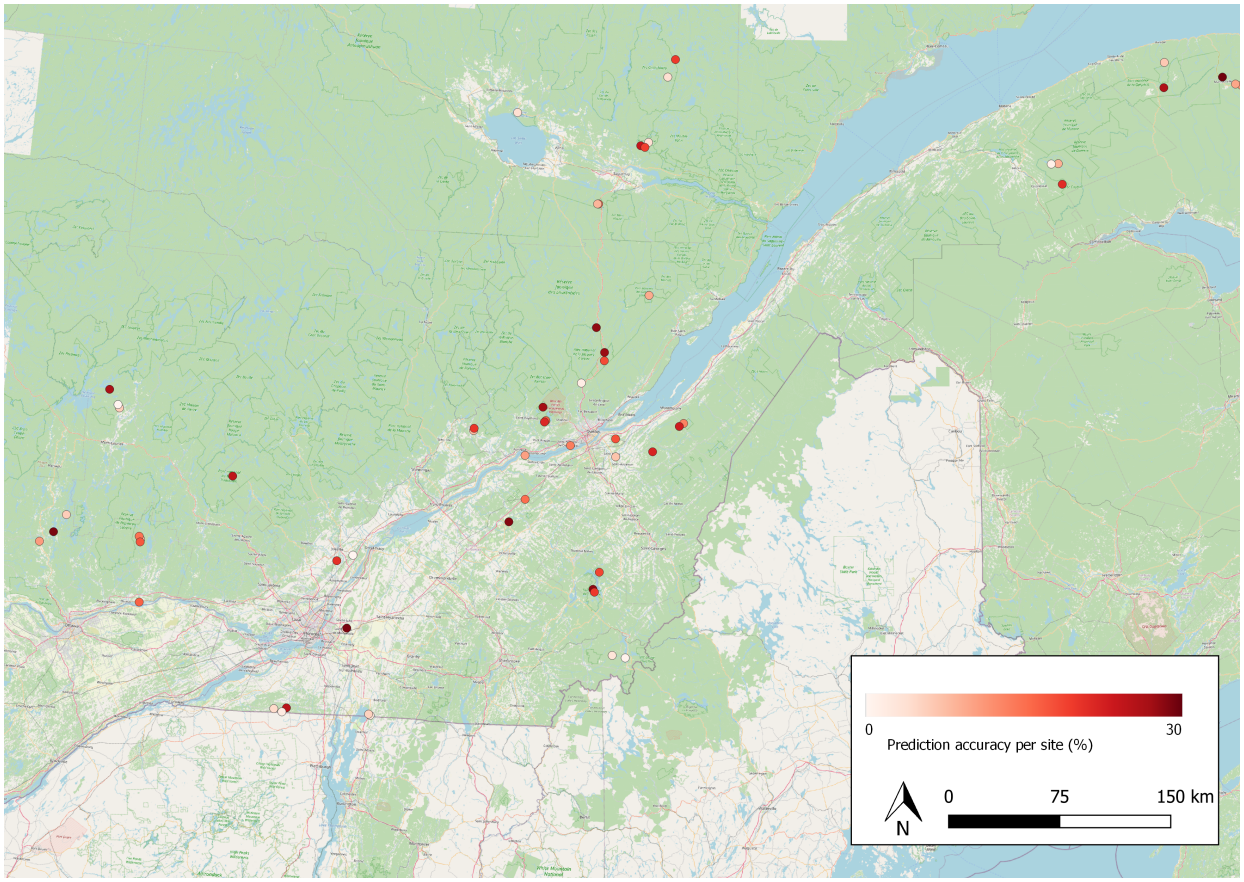


FIGURE 21 – Prediction accuracy per localization

FIGURE 22 – Prediction accuracy per localization

The maps on Figure 20, 21, 22 show that recorders located in the same site but in different habitats present contrasted results. Well-predicted sites in one habitat are often next to poorly-predicted sites. In a further section, confusion matrices per site will be presented.

10 Addition of abiotic sounds

The train set is modified by adding different wind noises extracted from the Quebec recordings to the xeno-canto set. This addition of abiotic sounds could help mitigate the differences in recording and weather conditions between xeno-canto and Quebec data. The 5 most present classes (wtsp, swsp, oven, mawa, heth) in the Quebec dataset were used for this experiment. The model is trained on xeno-canto recordings of these species that last less than 2 minutes, and tested on non overlapping samples from the Quebec dataset, then on xeno-canto recordings of these species that last in between 2 and 4 minutes (Table 1). Wind noise is classified in three categories of intensity following the expert's annotations. Categories and wind samples are randomly added to xeno-canto data.

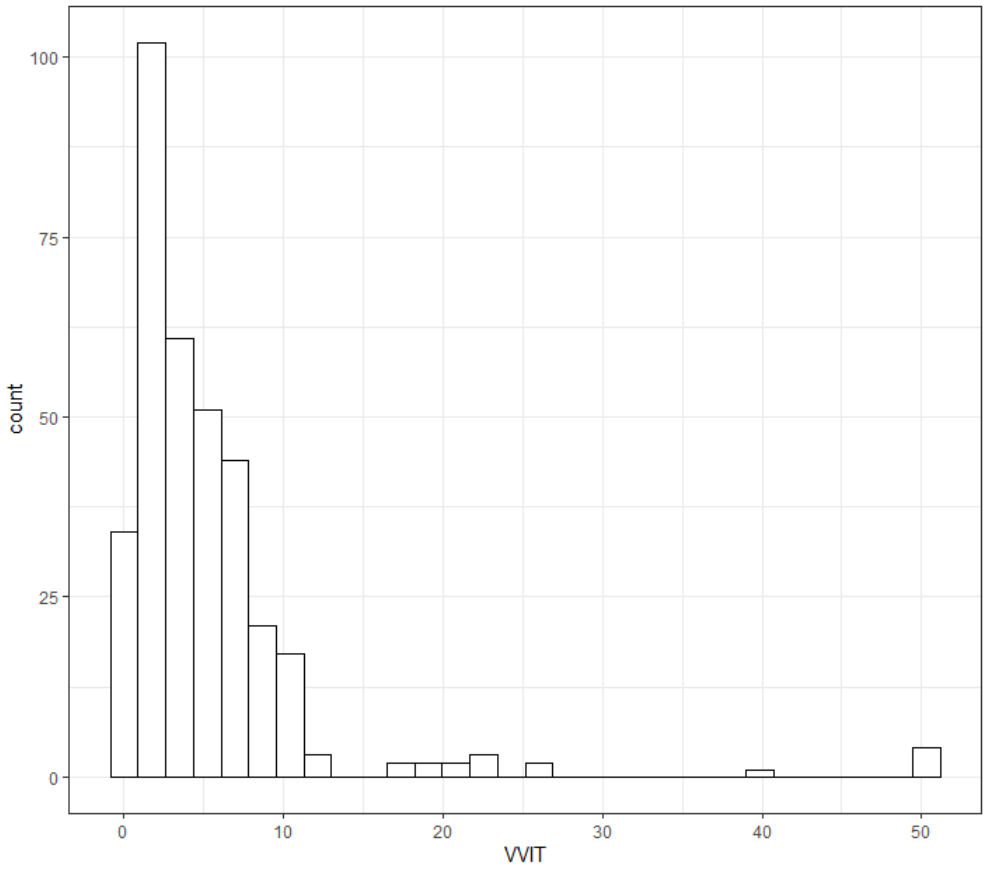


FIGURE 23 – Distribution of windspeed (km/h) in top 5 Quebec data

In the top 5 Quebec dataset, recordings are made mainly in low-intensity wind (<10 km/h) conditions (Figure 23). Sites with more than 10 samples are selected for further analysis (Table 4).

Down below are the results of the experiment with abiotic wind added to the xeno-canto train set (right) compared to the experiment where the model was trained on untouched samples. There was no data augmentation so the train set and the two test sets were identical. The results show that the model performed slightly better when it had seen abiotic wind in the train set.

site	count
073_137_F02	10
136_095_H01	10
137_144_H01	10
145_141_H01	10
146_133_H01	10
142_111_H01	11
146_133_H02	11
122_092_H01	13
142_111_H02	16
148_101_F01	30
137_107_H01	41
137_107_F01	42
137_107_H02	44
148_101_H01	50

TABLE 4 – number of samples per selected sites in top 5 Quebec dataset

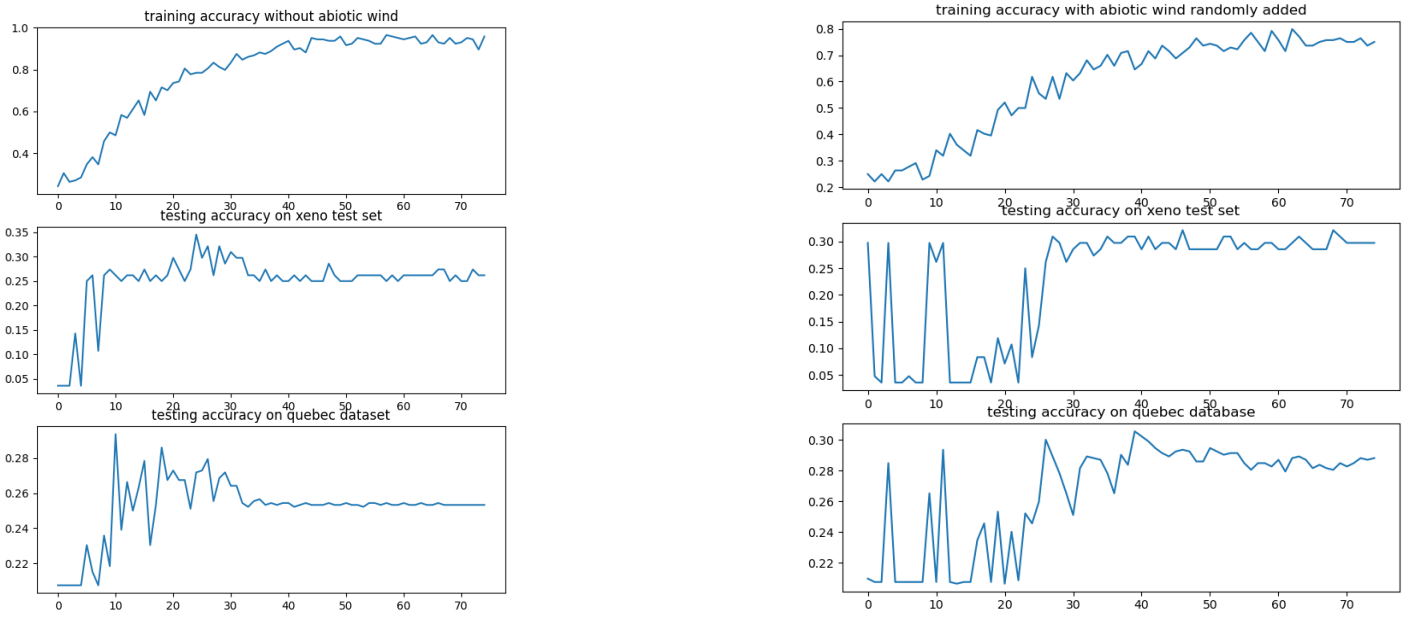


FIGURE 24 – AlexNet default model trained on xeno dataset and tested on quebec and xeno test sets

Here is the confusion matrix for the quebec test set :

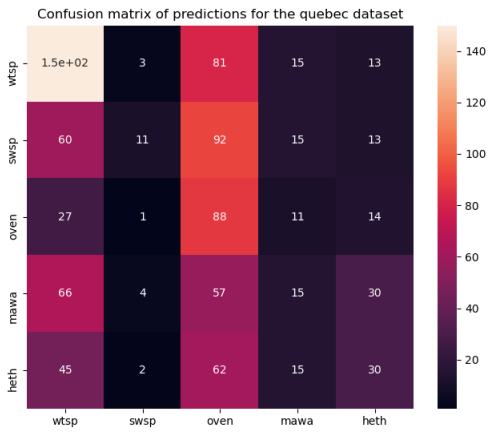


FIGURE 25 – Confusion matrix for the quebec test set

Although weather and background noise in general seem to account for a part of the poor model accuracy, this section experiments show that these sound characteristics taken as such are not enough to improve prediction accuracy. At this stage, we are unable to take into account weather data in the model but as the model improves, we will come back to these experiments to eventually remove bad quality recordings from the testing dataset.

11 Octave analysis

In order to differentiate background noise in different sites, we performed an octave analysis on all recorded samples. This yields a measure of acoustic energy per band of frequencies (Figure 26). Octave analysis divides a recording into increasing band of frequencies in which the central frequency is doubled is each octave (i.e. : the first band goes from f1 to 2*f1, the second from 2*f1 to 4*f1, and so on). The first band starts around 100 Hz and the last band ends around 16 000 Hz.

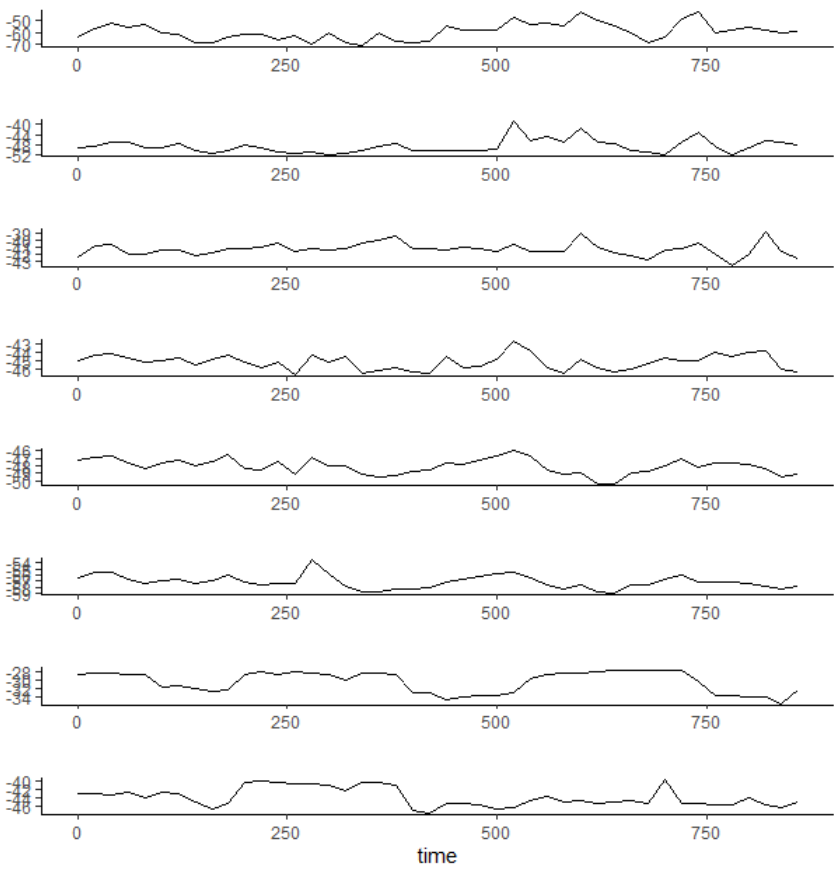


FIGURE 26 – Confusion matrix for the quebec test set

However, no correlation between energy in specific octaves and prediction accuracy is found in 309 samples from different sites of the year 2018 (Figure 27).

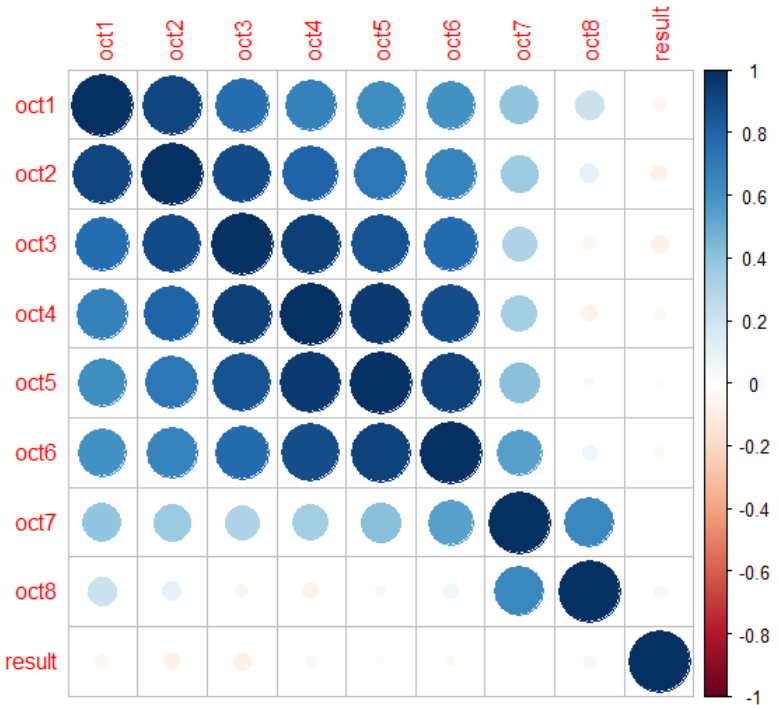
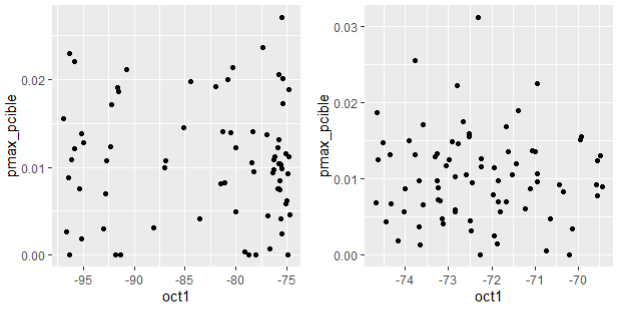
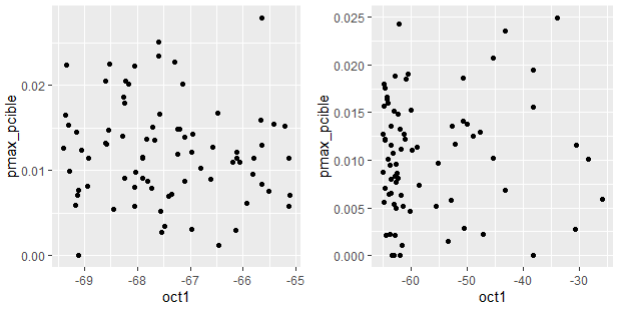
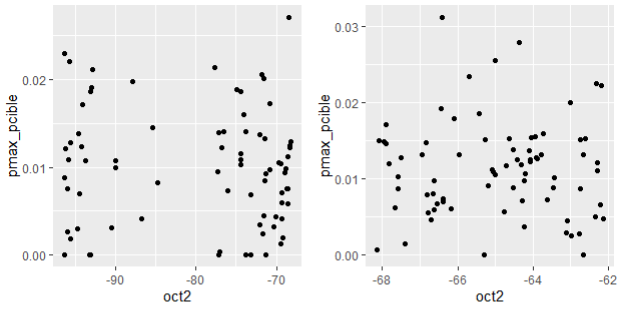


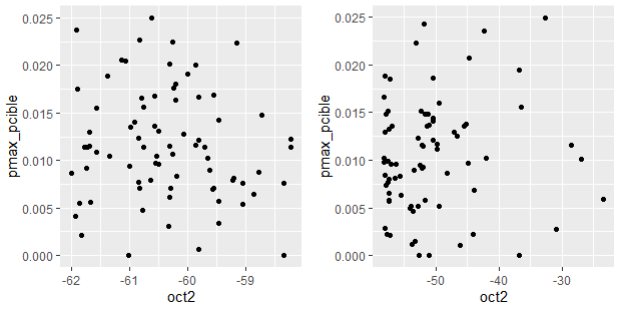
FIGURE 27 – Confusion matrix for the quebec test set

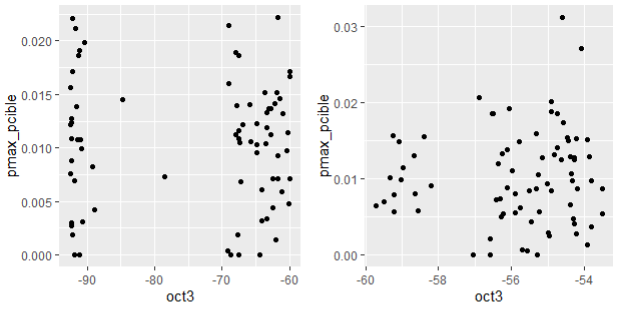
The difference between the probability attributed by the model to a species and the maximum probability shows the distance between good and bad predictions. We represented this result for each of the 309 samples in 4 categories of energy per octave band (Figure 28). There is no clear tendency.

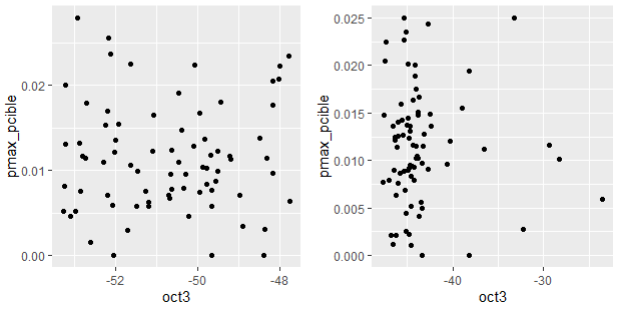

pmax_pcible
oct1

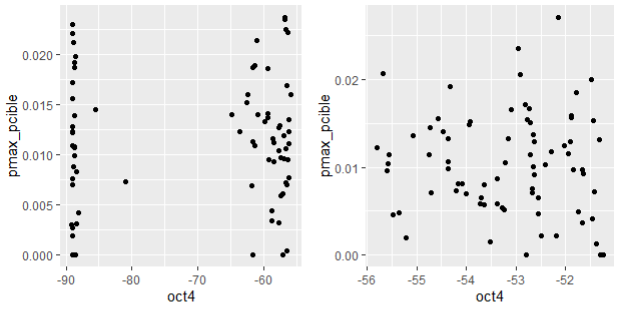

pmax_pcible
oct1

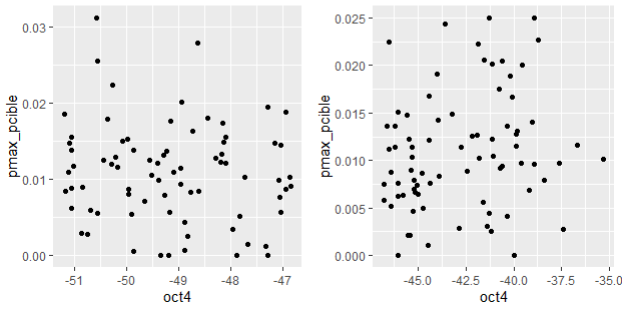

pmax_pcible
oct2


pmax_pcible
oct2


pmax_pcible
oct3


pmax_pcible
oct3


pmax_pcible
oct4


pmax_pcible
oct4

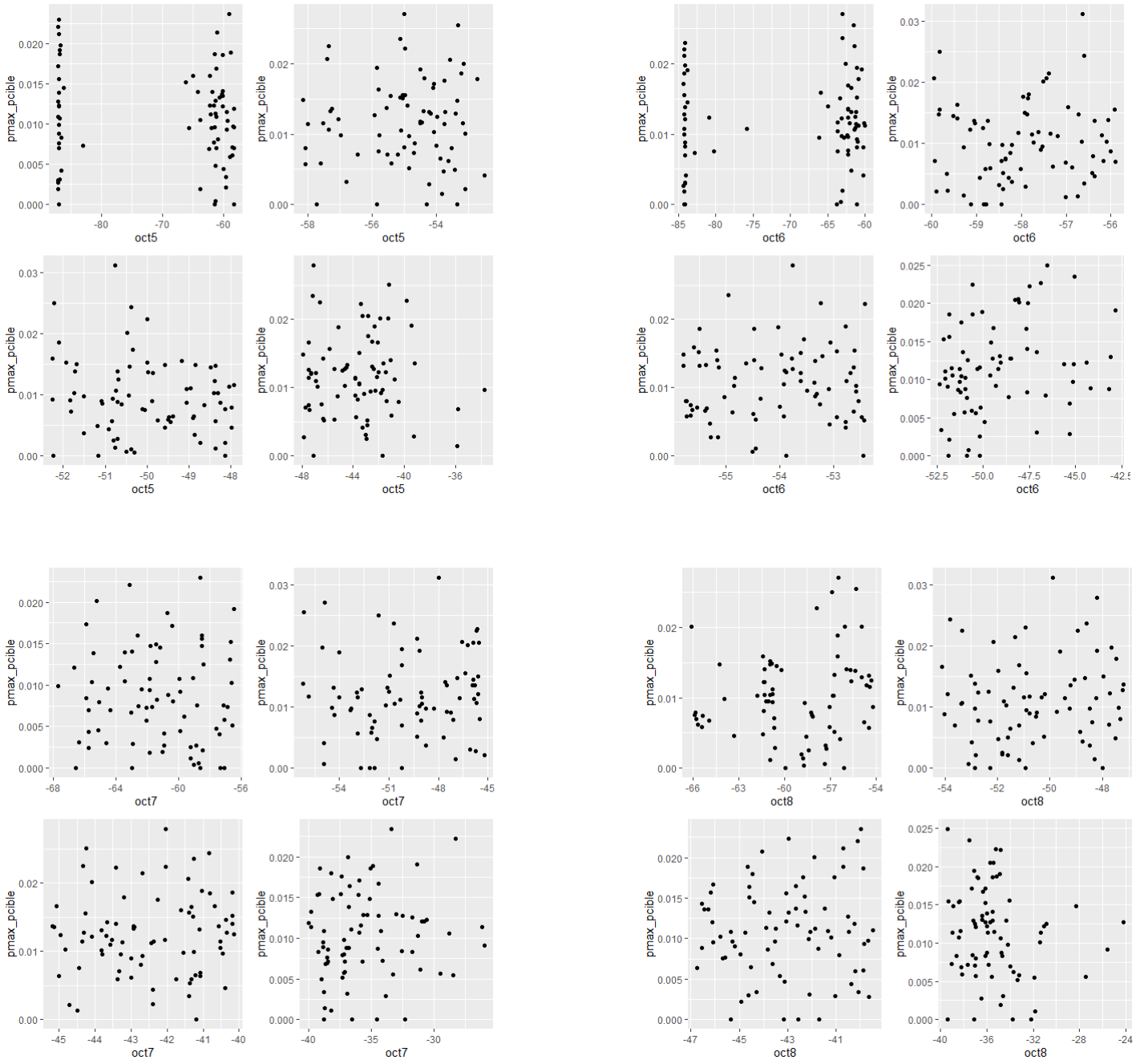


FIGURE 28 – Pmax-Ptarget for each octave band (oct 1 to oct8). The closer the probability is to 0, the closer is the prediction to the correct species

12 distance from the correct prediction per site and per specie

Beside results from octave analysis, we also looked at the difference between good and bad predictions per habitat for 43 species out of 75, a choice made by experts' recommendations based on the ecological relevance of these species (Figure 29, 30, 35).

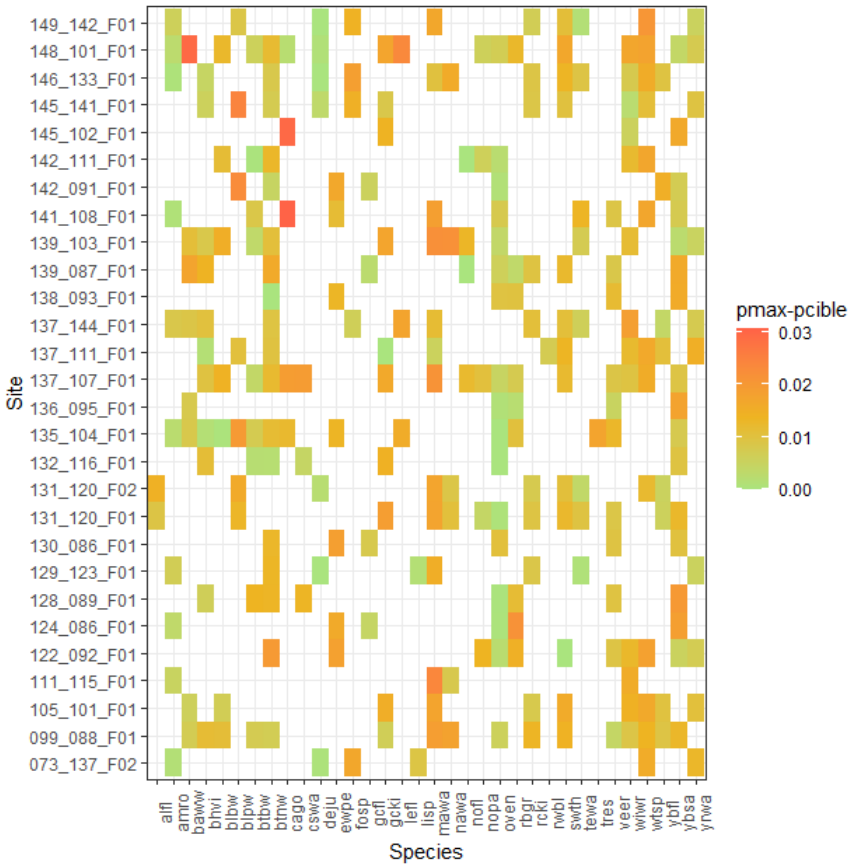


FIGURE 29 – Average difference between Pmax and Ptarget per forest sites

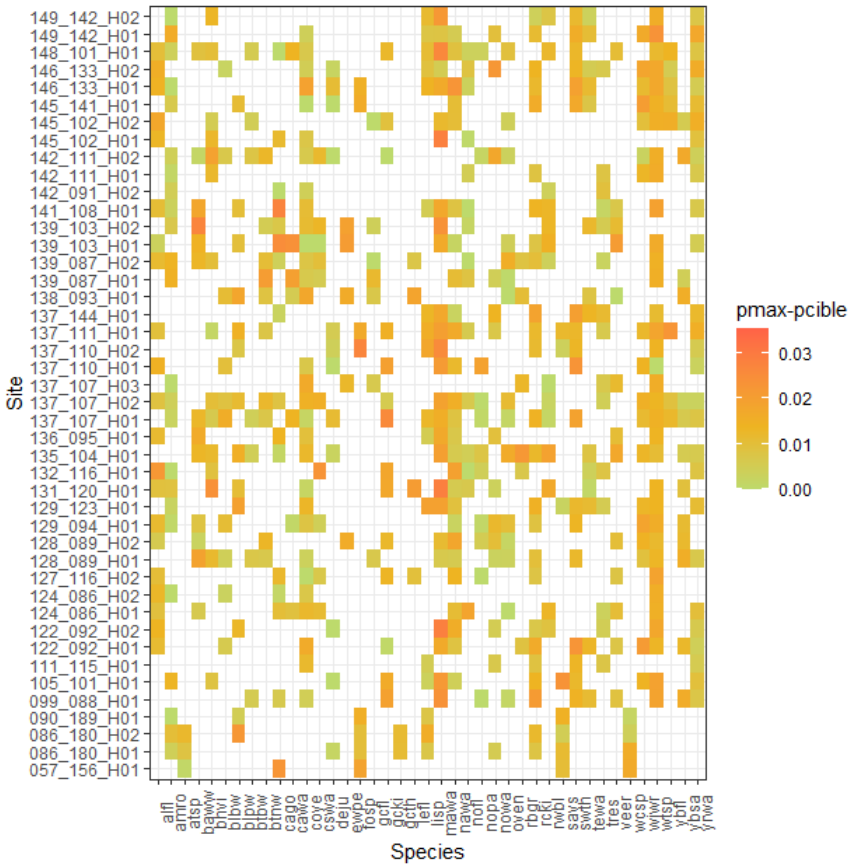


FIGURE 30 – Average difference between Pmax and Ptarget per wetlands sites

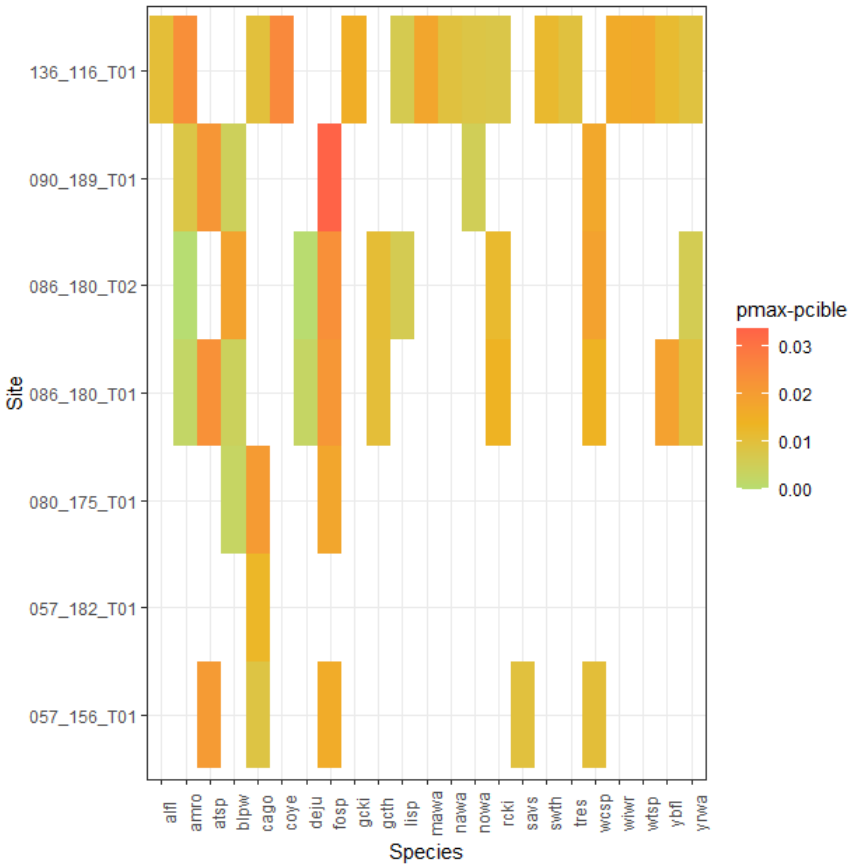
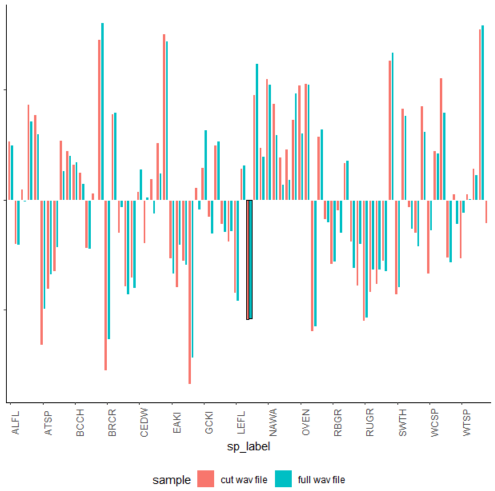


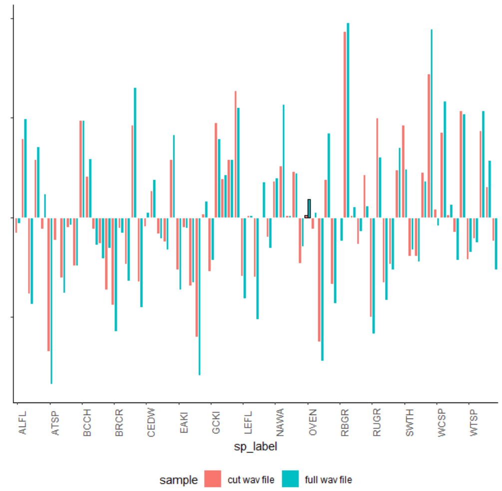
FIGURE 31 – Average difference between Pmax and Ptarget per tundra sites

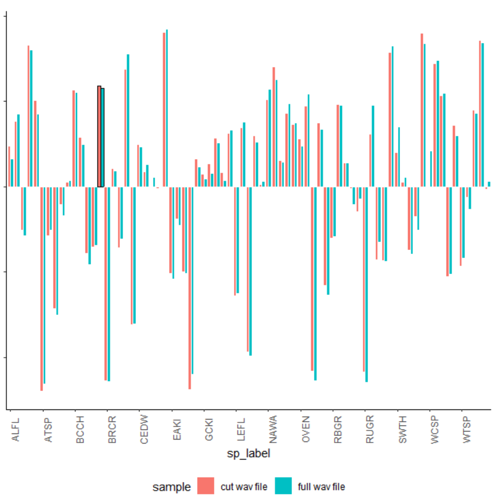
There seems to be no species or site that recurrently yields poor predictions. The number of species in tundric sites is reduced compared to other environments. These sites were thought to be more prone to bad predictions because of the adverse weather conditions and the openness of the habitat.

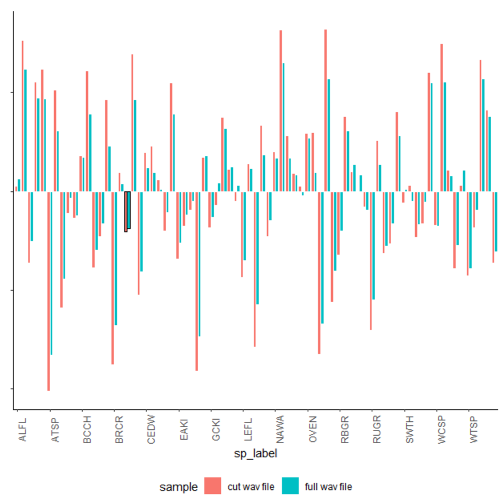
13 Modification of the pooling function

If a 10 s recording sample is mostly made up of background noise, the resulting maximum probability will be the sum of randomly attributed species because the focus is not on the actual animal vocalization. We randomly sampled 6 recordings to see whether the cutting of samples to the exact bird vocalization would improve the probability to predict the correct species (Figure 32). There seems to be no significant effect on these random samples. However, when listening to these samples, they are either very low or seem to present multiple birds singing at the same time. This requires further investigation.


ALFL ATSP BCCH BRCR CEDW EAKI GCKI LEFL NAWA OVEN RBGR RUGR SWTH WCSP WTSP
sp_label
sample
cut wav file
full wav file


ALFL ATSP BCCH BRCR CEDW EAKI GCKI LEFL NAWA OVEN RBGR RUGR SWTH WCSP WTSP
sp_label
sample
cut wav file
full wav file


ALFL ATSP BCCH BRCR CEDW EAKI GCKI LEFL NAWA OVEN RBGR RUGR SWTH WCSP WTSP
sp_label
sample
cut wav file
full wav file


ALFL ATSP BCCH BRCR CEDW EAKI GCKI LEFL NAWA OVEN RBGR RUGR SWTH WCSP WTSP
sp_label
sample
cut wav file
full wav file

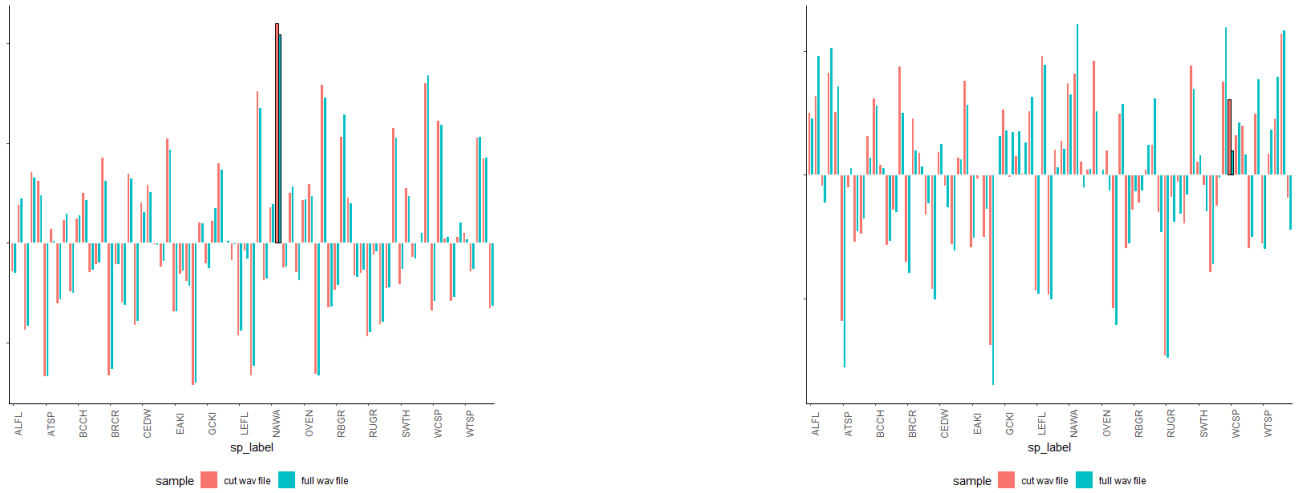


FIGURE 32 – Outputs of the model for six random samples. The highest positive value is the species attributed to a sample by the model.

14 Entropy Analysis

In this following subsection, all experiments are conducted on the output of the last layer of the resnet model, a vector assigning to each class a probability that the sample belongs to that specific class. Additionally, instead of forwarding the 10 second sample in the model, the sample was divided into four 3s sections with a 1s overlap (0s-3s, 2s-5s, 4s-7s, 6s-9s). This results from the fact that several birds sing on a 10sec interval. Smaller interval enable to isolate single species of birds. The result was for each sample a (4x43) matrix.

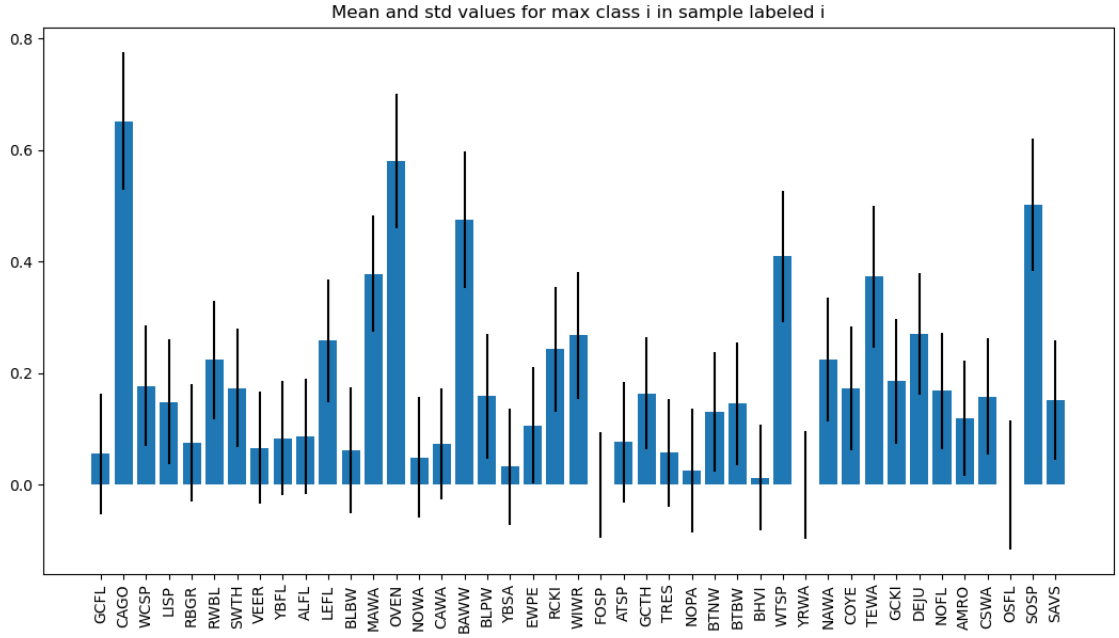


FIGURE 33 – Mean value of the correct label in the 4 sections (blue), and mean of the standard deviations for each section (bars)

The correct label has a high probability in the last layer of the resnet model (Figure 33).

The predicted class is the class with the highest probability in the time section and with the lowest value of entropy (Figure 34).

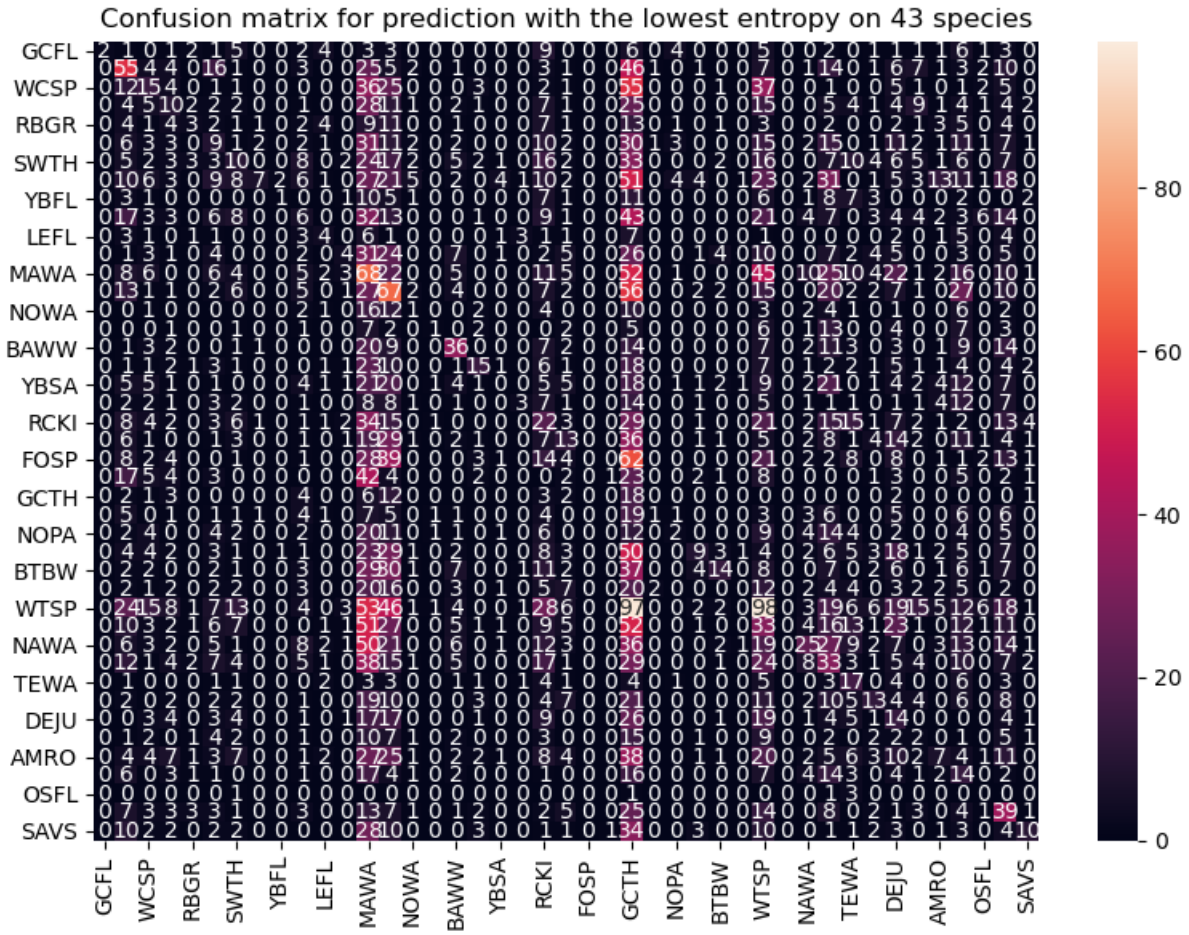


FIGURE 34 – Confusion matrix for predictions based on the section with the lowest entropy

The strongest classes are selected (Figure 33). All classes averaging over 0.15 (potential values are between 0 and 1) for the correct classes were chosen, creating a dataset of (4x43) matrices belonging to 21 classes. A small model with 3 fully connected layers and some dropout was trained on 75% of this new dataset.

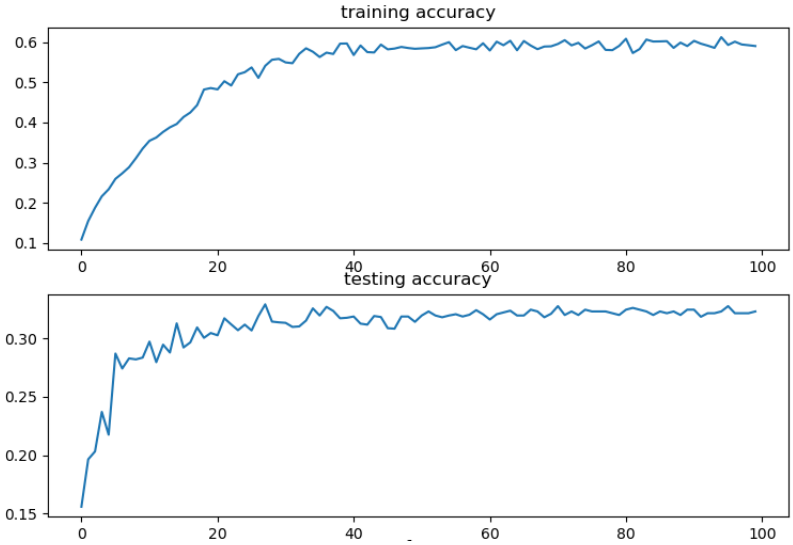


FIGURE 35 – Accuracies for model trained on the output of the resnet model and on the strongest 21 classes

Accuracy was also computed on each 3 sec interval. This analysis reveals that accuracy is higher for tests made on the 4-7 sec interval of 10 sec samples.

interval (s)	accuracy	best species selected from this interval	accuracy for best species
0-3	7%	CAGO, MAWA, OVEN	25%
2-5	11%	BAWW, CAGO, MAWA,OVEN	26%
4-7	25%	BAWW, CAGO, OVEN,SOSP,TEWA,WTSP	42%
6-9	14%	CAGO, MAWA,OVEN,SOSP,WTSP	26%

TABLE 5 – accuracy for 4 intervals in 10sec samples

A pooling function taking into account the hierarchy in intervals accuracy could be applied. However, this effect likely stems from the window selection around the time given by J.-F. Jetté at which the identified species produces a vocalization.

15 Transfer Learning

Result vectors from the resnet model are fed into a new model composed of fully connected layers. This experiment was inspired by the promising and interesting analysis of the probability vectors in the previous subsection. Bird vocalizations last on average 3.7 seconds. The new matrix of results was composed of 4s signals every 0.5 seconds in the 10s window. Encouraging results on Quebec dataset are observed (Figure 36).

learning rate	batch size				
	16	32	64	128	256
0.1	0.089				0.08
0.01	0.35	0.37	0.07		0.32
0.001	0.62	0.57	0.56		
0.0001		0.58	0.49		height

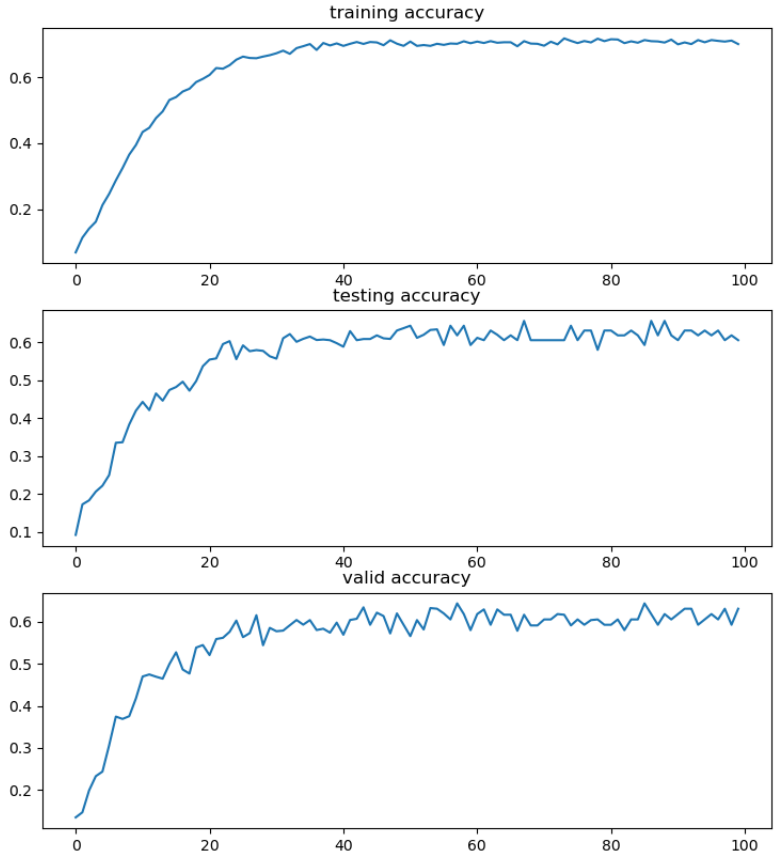


FIGURE 36 – Training, Testing, Validation accuracies for wdl=0.002, lr=0.001, bs = 64, 43 species

Different parameters were tested on the same training set to compare results (wdl = 0.005, 80 epochs).

15.1 Training for different locations

In order to verify the model validity and try to improve its validity, three similar models were trained for different locations. The first one, NH, is composed of every recording from wet environments above the 46.7° latitude, the second one, SH, is composed of wet environments below the 46.7° latitude. The last model, called 'other', was trained on every other sample. Nordic wetland model training set is to small to draw conclusions on it.

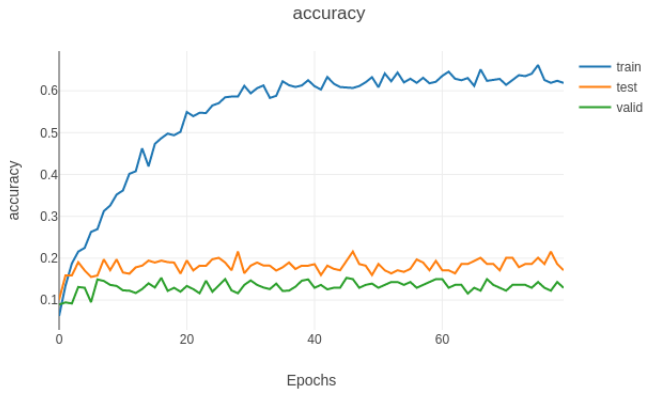


FIGURE 37 – Training, Testing, Validation accuracies for wdl=0.002, lr=0.001, bs = 64, 43 species on SH

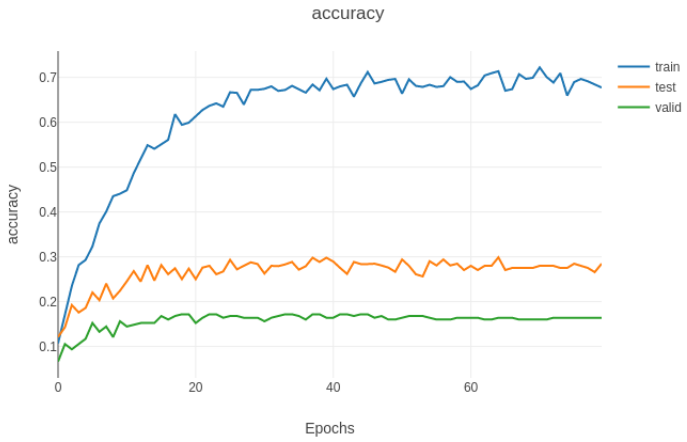


FIGURE 38 – Training, Testing, Validation accuracies for wdl=0.002, lr=0.001, bs = 64, 43 species on other

Figure 37 and 42 show lower prediction accuracy than in Figure 36. This is due to an identification of the background noise in the added layer. Training and testing sets were not properly separated. However, we plan to separate them by taking different days on the same sites.

15.2 Training for different days

Another way to split the data up to increase the difference between the trainset and the testset and thus the model's ability to generalize is to create different sets for different days. In this subsection the training set is made up of the even days of the month, and the test set and validation sets of the odd days of the month. Multiple experiments were made but down below is the one with the best scores.

In order to obtain these scores, every sample was filtered with a band pass filter (400Hz, 9kHz), and the trainset was complexified by overlapping different 10s samples one over another to create new background noises and choruses.

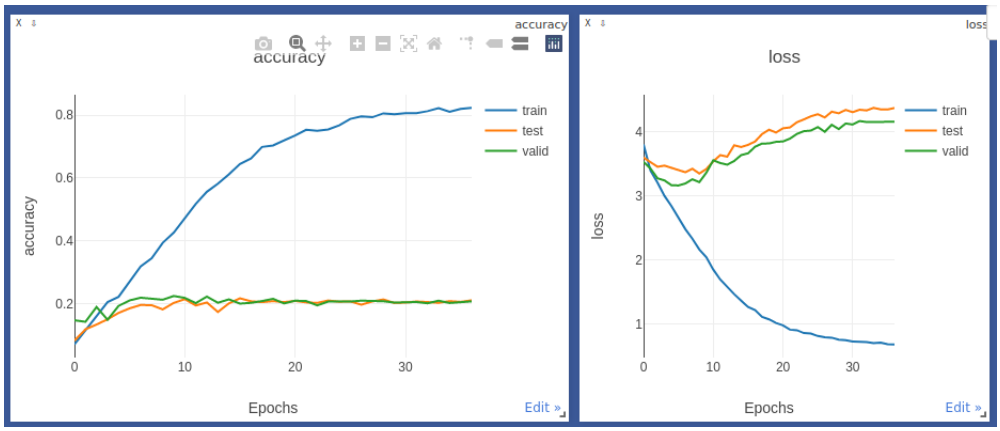


FIGURE 39 – Training, Testing, Validation accuracies for filtered samples, trained on even days and tested on odd days of the month

Down below are the three confusion matrices associated with the above experiment, but split into each type of environment.

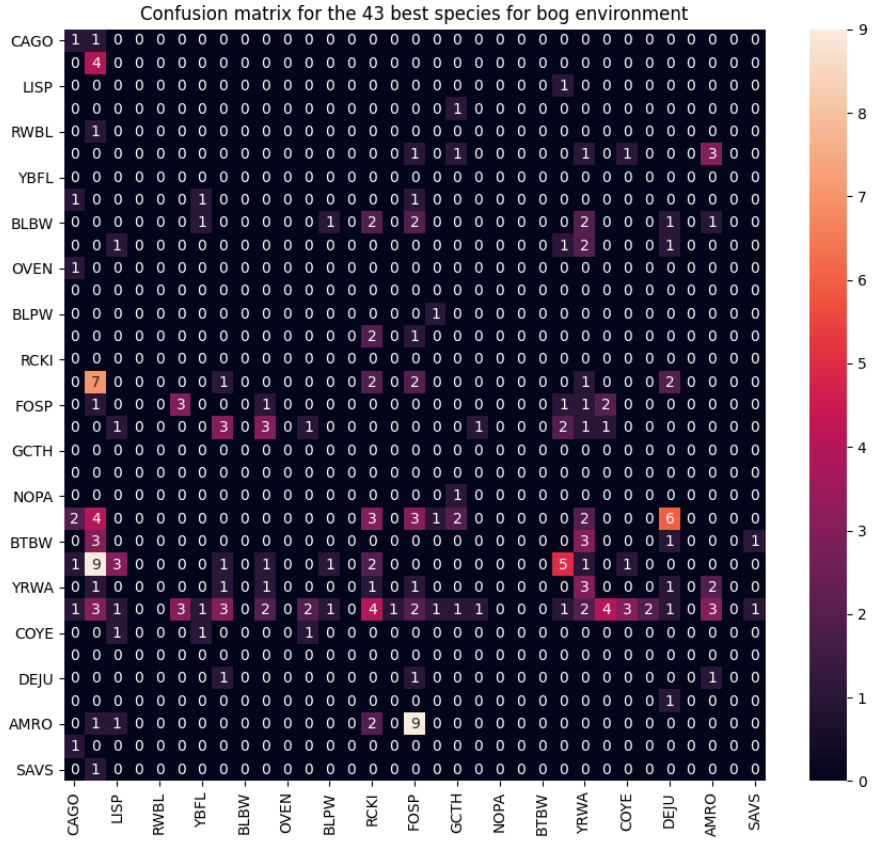


FIGURE 40 – Confusion matrix for test set and validation set samples from the bog environment on model that performed best on test set

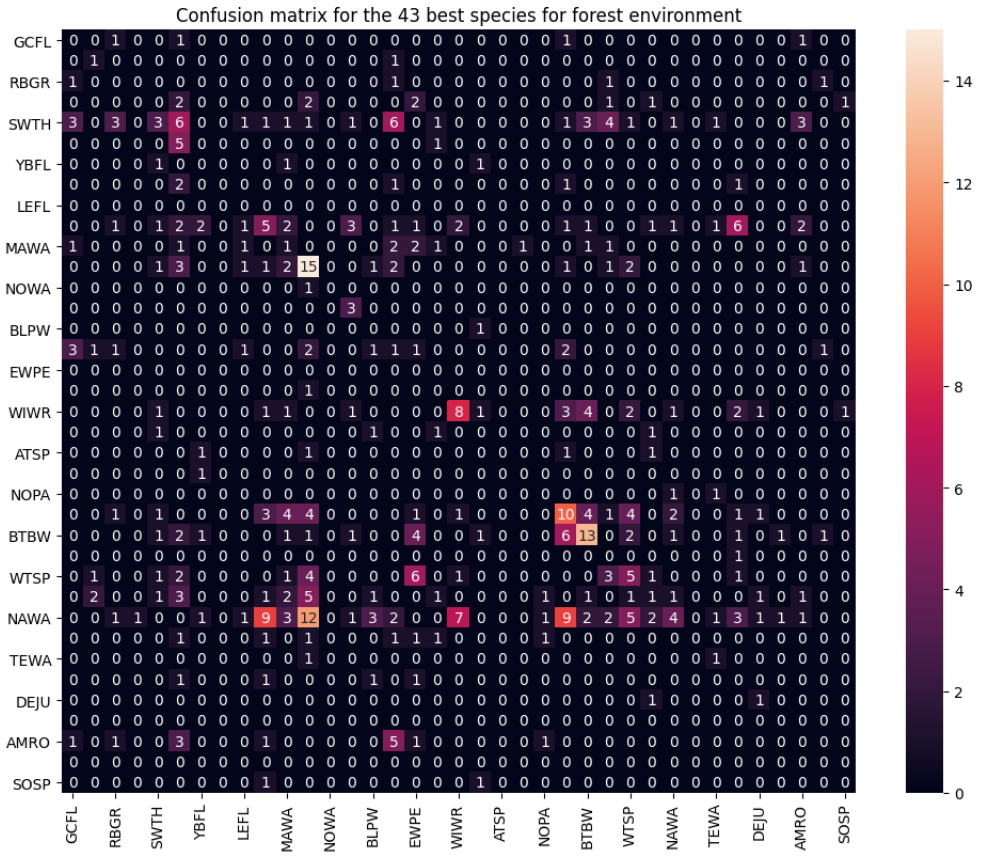


FIGURE 41 – Confusion matrix for test set and validation set samples from the forest environment on model that performed best on test set

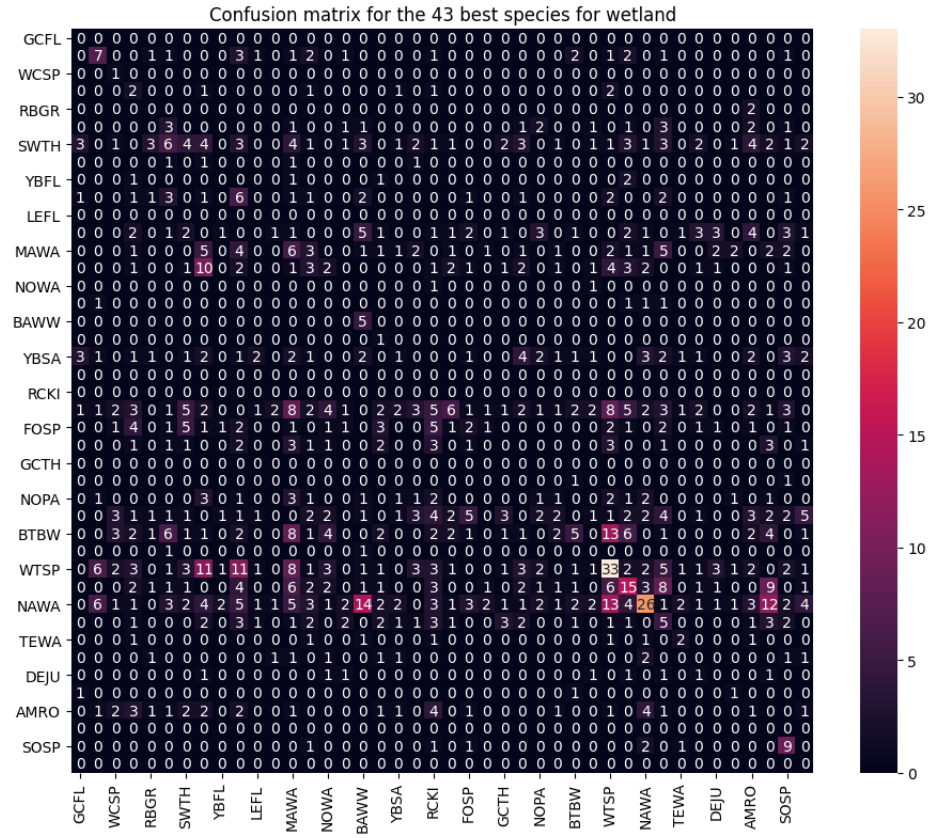


FIGURE 42 – Confusion matrix for test set and validation set samples from the humid environment on model that best performed on the test set

The model better classifies unseen data from the forest environment with an accuracy of 17%. For samples from the wet environment, the model has an accuracy of 13%, this is maybe due to the presence of frogs or choruses on the recordings. The model struggles with samples from the bog environment, with an accuracy of 8% that could be explain by the heavy influence of the wind and climate conditions on the quality of the recordings.

16 Conclusion

These experiments show that data augmentation fails to make the model generalize. This failure to generalize can have many roots - amount of samples, quality of samples, uniformity of samples, mislabeling of samples, incorrect time-frequency representation of the signal, etc. However, the model and python script were previously tested and corrected, furthermore, the five classes used contained more than 100 samples each which should be enough for some, if not good, generalization. It can then be concluded that the samples themselves, the 10 second signals, are the main root of the problem. Some samples contain multiple species at once, some species have many different calls, some resembling other species' vocalizations, which makes it difficult for both humans and

neural networks to differentiate the classes. The possible influence of the weather conditions was analyzed but led to inconclusive results.

In this report, many different ideas and methods were tested. Future perspectives lie in the detection and suppression of chorus, the detection of good quality recordings that can accurately be predicted by the CNN, and in the creation of distinct testing sets (grouped by sites, month) to avoid background noise identification.

Identification of real-world recordings of birds is not an easy task but our various attempts at it let us better understand where the difficulties lie. Every experiment brings us closer to an efficient classification tool.