*Article*

# A Randomized bag-of-birds Approach to Study Robustness of Automated Audio Based Bird Species Classification

**Burooj Ghani [1,2,*] and Sarah Hallerberg [2]**

[1] Bernstein Center for Computational Neuroscience, Third Institute of Physics, University of Göttingen, Germany
[2] Faculty for Engineering and Computer Science, Hamburg University of Applied Sciences, Germany
* Correspondence: burooj.ghani@haw-hamburg.de

**Abstract:** The automatic classification of bird sounds is an ongoing research topic and several results have been reported for the classification of selected bird species. In this contribution we use an artificial neural network fed with pre-computed sound features to study the robustness of bird sound classification. We investigate in detail if and how classification results are dependent on the number of species and the selection of species in the subsets presented to the classifier. In more detail, a bag-of-birds approach is employed to randomly create balanced subsets of sounds from different species for repeated classification runs. The number of species present in each subset is varied between 10 and 300, randomly drawing sounds of species from a dataset of 659 bird species taken from Xeno-Canto database. We observe that the shallow artificial neural network trained on pre-computed sound features is able to classify the bird sounds relatively well. The classification performance is evaluated using several common measures such as precision, recall, accuracy, mean average precision and area under the receiver operator characteristics curve. All of these measures indicate a decrease in classification success as the number of species present in the subsets is increased. We analyze this dependence in detail and compare the computed results to an analytic explanation assuming dependencies for an idealized perfect classifier. Moreover, we observe that the classification performance depends on the individual composition of the subset and varies across 20 randomly drawn subsets.

**Keywords:** Bioacoustics, Machine Hearing, Bird sound recognition, Artificial Neural Networks, Audio Signal Processing

## 1. Introduction

The audio based automatic recognition of bird species has become an increasingly common and effective method in the context of bird species monitoring, studying the behavior of birds, and understanding their communication patterns [1,2]. Notwithstanding the advantages of using bird vocalizations to infer ecologically relevant information there are certain challenges associated with processing field recordings to produce robust results. Unattended field recordings can be quite noisy, depending on the distance from the recording device sound clips can be faded or distorted, recordings can include overlapping sounds from the same or different bird species. Several authors [3,4] have addressed the influences of noise, by adding artificial noise to recordings. On surveying the literature available on automatic audio based recognition of bird species [5–9], it has been found that most analysis have been performed using less noisy recordings and relatively small datasets, as has already also been pointed out in the review paper [2]. Also, the measures for classification success which different authors use to report their results vary, which makes it difficult to compare the performance of different classifiers [8,10–14]. However, the way in which a set of species is selected for the classification experiment can influence the classification success. It is obviously easier to distinguish bird vocalizations which are qualitatively very different types of sounds, such as sounds produced by crows or song-birds.

One of the main focus of this contribution is to look into the robustness of classification results and the reliability of accuracy measures, when the bird species are selected in a



**Citation:** Ghani, B.; Hallerberg, S. A Randomized bag-of-birds Approach to Study Robustness of Automated Audio Based Bird Species Classification. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

 doi:10.20944/preprints202108.0277.v1



randomized way. Additionally the influence of the number of selected species (classes) on different measures for multi-class classification success is investigated (see Sec. 3.4). For this work, we use the dataset curated by the organizers of the BirdClef 2019 challenge [36]. This dataset contains recordings of 659 bird species from South and North America and has originally been drawn from the Xeno-Canto repository for bird sounds [15].

In this contribution, out of 659 available species, $n$ species are randomly drawn with varying $n$ incrementally between 10 and 300. For each $n$ we generate 20 randomly composed lists of species and for each species we choose 200 recordings of comparable length to generate balanced subsets (as will be explained in more detail in Sec. 2.1). The robustness of classification results is then accessed by training a feed-forward neural network on each subset (see Sec. 2.3). Our motivations to choose a shallow feed-forward neural network over very deep networks lie mainly in the model simplicity, the lower computational costs, and the relatively small amount of data required to train such networks. We wanted to analyze the classification performance using a simple model that can be trained with handcrafted sound features. The performance of classification is accessed using several measures for classification success such as accuracy, precision, recall, area under the receiver-operator-characteristics curve, and mean average precision (see Sec. 2.4). As a consequence of these repeated randomized classification experiments, we can hence also provide box-and-whisker plots for each performance metric (see Sec. 3). Additionally it is possible to infer functional relations between the number of classes and the classification success (see Sec. 3.2). Furthermore, a discussion is included on how confidently the probabilistic classifier classifies different species and how the alignment of probability confidence with classification frequencies can be used as a measure to evaluate the performance of the classifier (see Sec. 3.3).

## 2. Methods

Our audio based bird classification framework comprises of three modules: data preparation, feature extraction and model construction.

### *2.1. Bags-of-Birds Approach: Performing Randomized Classification Experiments*

In this contribution we use a dataset containing birds sounds of 659 species provided within the BirdClef 2019 challenge [36], originally drawn from the Xeno-Canto [15] repository for bird sounds. The data is prepared, for our analysis, by splitting all sound recordings of varying lengths into 5-second chunks. The audio clips are then resampled to 22050 Hz with Librosa 0.6 audio processing package [16]. It has been reported that most birds vocalize in the frequency range of 0.5 kHz to 10 kHz [17]. The resampling is followed by peak normalisation. To get rid of sound samples that do not contain any bird sounds, a simple signal-to-noise ratio based estimate is employed. This estimate ensures with high probability that 5-second clips that do not contain bird sounds are discarded [18].

Since one of the aims of this work is to estimate robustness of the classification approach, for each classification run of $n$ species, the subsets of species are not carefully chosen. Rather a random number generator is employed to construct 20 bags of species. Analogously to *bag-of-words* approaches, one can maybe refer to this procedure as a *bag-of-birds* approach. Each *bags-of-birds* is basically a set of randomly drawn species without repetition from a complete list of 659 bird species. The idea is to repeat the computation 20 times to have a reliable estimate of classification performance given a certain number of species. Fig. 1 illustrates the idea of this numerical experiment. A balanced dataset was curated for the classification task in which 200 sound samples were randomly drawn for each bird species. Finally, the dataset was divided into training and testing sets such that training sets contain 75% and test sets 25% of data. Given that for each bird species we are using 200 sound samples, the test set for each analyzed species will consequently contain $m = 50$ sound samples. This is described in detail in Sec. 2.4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



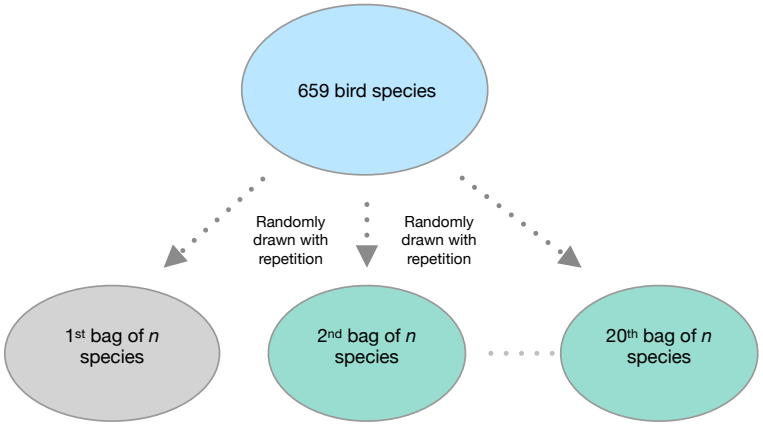


**Figure 1.** Schemetic for the *bags-of-birds* approach.

### 2.2. Feature Extraction

The next step entails extracting audio features from time series of audio signals. Extracting features allows us to obtain lower-dimensional compact statistical representations while preserving the distinguishing characteristics of signal in a non-redundant manner. In addition to reducing the computational costs feature extraction maximizes classification performance of the system. Studies have shown that aggregated features allow us to achieve a better classification performance compared to a single feature [19]. In this contribution we employ the spectral centroid, the spectral rolloff, the zero-crossing-rate, Spectral Bandwidth, the root-mean-square energy (RMSE), the Mel-frequency-cepstral-coefficients (MFCCs) and MFCCs as features.

Since signal statistics change rapidly, bird sounds, like audio signals in general, are non-stationary signals. For this reason, the feature extraction is carried out in a short-term processing manner where the signal is chunked into short *analysis frames* of 92.8 ms. The analysis frames are assumed to be in a quasi-stationary state [20]. Therefore, the spectral features described below are computed frame wise. In case an additional splitting into even shorter windows within each analysis frame was needed (e.g. for the MFCC), the temporal average of the feature is computed to generate a single value which is associated with the respective analysis frame. For each MFCC, also the variance of the coefficients within the analysis frame is computed and used as an additional feature.

In more detail, all features are computed using Librosa 0.6 audio processing package [16] and can be described as follows [20]:

- *Spectral Centroid:* The spectral centroid measures the frequency where energy of a spectrum is centered. In other words it localizes the center of mass of the spectrum and is calculated as a weighted mean of the frequencies the signal is composed of:

$$s_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)}, \tag{1}$$

  where S(k) is the spectral magnitude at frequency bin k and f(k) represents the center frequency of the bin [21].
- *Spectral Rolloff:* The spectral rolloff gives the frequency f(k) below which a pre-defined percentage (usually set to 0.85) of the total spectral energy is concentrated [22].
- *Zero-Crossing Rate:* The zero-crossing rate $r$ measures the smoothness of a signal. It is the rate at which signal changes its sign from negative to positive or vice versa [23].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



- *The root-mean-square energy (RMSE):* The root-mean-square energy of a signal gives the signal's total energy and is defined as:

$$e_{rms} = \sqrt{2\sum_{k}|S(k)|^2}, \tag{2}$$

  where $S(k)$ is the spectral magnitude at frequency bin $k$ [16].
- *Spectral Bandwith*: measures if the power spectrum is concentrated around the spectral centroid or spread across the spectrum. It is computed as:

$$s_b = \sqrt{\frac{\sum_k (k - s_c)^2 \cdot |S(k)|^2}{\sum_k |S(k)|^2}}, \tag{3}$$

  where $s_c$ is the spectral centroid, $S(k)$ is the spectral magnitude at frequency bin $k$ [24].
- *Mel-Frequency Cepstral Coefficients (MFCCs):* Mel-Frequency Cepstral Coefficients are inspired by human auditory perception. After computing the Fourier transform of a signal, the magnitude spectrum is projected to Mel scale that emphasizes relevant frequencies in a non-linear way – small bandwidth at low frequencies and large bandwidth at high frequencies. Mel scale approximates human auditory response better than linearly spaced frequency bands. The output is log transformed and MFCCs are obtained by taking a discrete cosine transform of the logarithmic outputs [25]. In this contribution the analysis frames are splitted into windows of lengths 512 and compute the first 20 MFCC values as features in our system.

In total the dimension of the feature space is 45, containing first 20 time averaged MFCCs, variances of first 20 MFCCs, and time averaged zero-crossing-rate, the spectral rolloff, the spectral centroid, root-mean-square energy and the spectral bandwidth.

### *2.3. Classification Model*

The features are then fed into a feed forward neural network, which is constructed using the sequential model within the Tensor Flow framework [26]. Feed forward neural

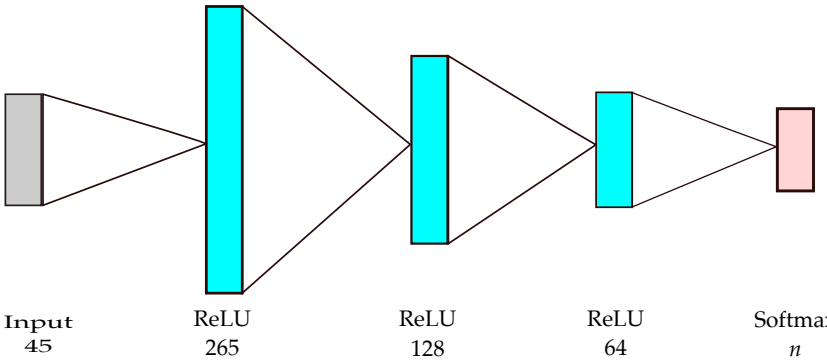


**Figure 2.** Architecture of the feed forward neural network used for bird classification. The input $x \in \mathbb{R}^{45}$ maps through three intermediate layers with $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$ and ReLU transfer functions. The output layer maps to $n$ independent classes with a softmax transfer function where $n$ is the number of bird species.

networks are archetypal models for machine learning [27,28] . In contrast to deep learning approaches used to classify bird sounds, the network consists of only four layers as illustrated in Fig. 2. These constitute a sequence of layers where each layer is an affine transformation followed by a non-linear transfer function $\sigma$:

$$f_i(\vec{x} \mid \theta) := \sigma_i\left(\vec{w}_i^{\top}\vec{x} + \vec{b}_i\right),$$

$\theta = (\vec{w}_i, \vec{b}_i, \sigma_i)$ constitutes the parameter space where $\vec{w}_i$ are weights, $\vec{b}_i$ are biases and $\vec{\sigma}_i$ are transfer functions for different layers $i$. The goal of the neural network is to learn the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



value of parameters $\vec{\theta}$ that generates the best function approximation [20,29]. The input to the neural network is the feature vector $\vec{x} \in \mathbb{R}^{45}$ which maps through three intermediate layers with $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$ hidden units respectively and is amplified using rectified linear units (ReLU) [29]. Finally the output layer maps to $n$ independent classes with a softmax transfer function [30,31] where $n$ is the number of bird species.

During training the model optimizes cross-entropy loss using Adam stochastic optimization algorithm [32]. We use a constant learning rate of 0.001. To identify the parameter setting that increases the likelihood of predictions the model is trained for 100 epochs. We observed, from our experiments, that the loss converged to a minimum toward the end of 100 epochs.

### *2.4. Measuring Classification Success*

This section introduces the indices used to measure classification performance of bird vocalizations [33][34]. In each numerical experiment we consider $n$ species and a dataset of a total of $4m$ sound samples. In more detail, $3m$ samples are used for training, whereas the remaining $m$ samples are employed to evaluate the quality of classifications. Within the dataset for $n$ randomly selected species, for each species $j = 1, 2, ..., n$ a dataset of $4m$ samples is processed. Also here $3m$ samples are used for training, whereas the remaining $m$ samples are employed to evaluate the quality of the classifications. These evaluations are then done by computing a confusion matrix for each species, containing:

$c_{tp}(j)$ the number of classifications which are true positives for species $j$,
$c_{tn}(j)$ the number of classifications which are true negatives for species $j$,
$c_{fp}(j)$ the number of classifications which are false positives for species $j$ and,
$c_{fn}(j)$ the number of classifications which are false negatives for species $j$.

The resulting elements of the confusion matrix enter into computation of more advanced metrics for measuring classification success, such as precision, recall, accuracy, mean-average-precision and receiver-operator characteristics. To understand how the entries of each species' confusion matrix influence the outcomes of these summarizing measures, more detailed descriptions of their computations are introduced in the following:

- *Precision:* The metric gives the measure of reliability of our predictions. The formula to compute precision for a bird species $j$ is

$$p(j) := \frac{c_{tp}(j)}{c_{tp}(j) + c_{fp}(j)}. \tag{4}$$

  Therefore, precision for a species $j$ indicates how many true positives the model predicted out of all positives. Therefore, higher the precision the more confident a model is about its predictions. In order to compute precision for entire test dataset we average over all species

$$P = \frac{1}{n} \sum_{j=1}^{n} p_j. \tag{5}$$

- *Recall:* This metric gives the measure of predictive power of a model. The formula to compute recall for each bird species $j$ is

$$r(j) := \frac{c_{tp}(j)}{c_{tp}(j) + c_{fn}(j)}. \tag{6}$$

  so recall for a class species $j$ would mean, of all actual positives in the test dataset how many did the model predict as positive. Therefore, the higher the recall the more positive samples model correctly classified as positive. In order to compute recall for entire test dataset we average over all species

$$R = \frac{1}{n} \sum_{j=1}^{n} r_j. \tag{7}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



- *Accuracy:* While precision and recall are computed for each class separately in a multi-class classification problem, the accuracy $A$ is computed for the entire test dataset using

$$A := \frac{\sum_{j=1}^{n} c_{tp}(j)}{m \cdot n}. \tag{8}$$

so out of all test samples how many were correctly classified.

- *Area Under ROC Curve (AUC):* An ROC curve shows performance of a classification model at different classification thresholds. The curve is computed by plotting *true positive rate* ($r_{tp}$) against *false positive rate* ($r_{fp}$) at these thresholds. The *true positive rate* for a bird species $j$ is defined as:

$$r_{tp}(j,\rho) := \frac{c_{tp}(j,\rho)}{c_{tp}(j,\rho) + c_{fn}(j,\rho)}, \tag{9}$$

and the *false positive rate* for a bird species $j$ is defined as

$$r_{fp}(j,\rho) := \frac{c_{fp}(j,\rho)}{c_{fp}(j,\rho) + c_{tn}(j,\rho)}, \tag{10}$$

with $\rho$ denoting a probability threshold that is varied from 0 to 1 in order to obtain the ROC curve. The area under the ROC curve (AUC) gives an aggregate measure of classification performance. The ROC was originally developed for a binary classifier and has later been generalized for multi-class classification system [35]. The test set labels are binarized by employing either the one-vs-one or the one-vs-rest configuration. We have employed the one-vs-one configuration for our task. In more detail, different sound samples are ranked by their probabilities and then false positive and true positive rates are computed by choosing different probability cut-offs $\rho$ to generate the ROC curve. AUC is computed as the area under the ROC curve. In the end an average across species is computed to get one AUC value for the entire data set, i.e.

$$AUC = \frac{1}{n} \sum_{\rho} \sum_{j=1}^{n} r_{tp}(j,\rho). \tag{11}$$

- *Mean Average Precision (mAP):* The evaluation metric gives us a way of characterizing the performance of a classifier by monitoring how precision changes on varying the classification probability threshold, one the model uses to make a decision if a bird sound sample belongs to a class $j$. A good classifier will maintain a high precision as recall increases while a poor classifier will take a hit on precision as recall increases with changes in threshold.
In more detail, to compute Average Precision for a species $j$, a list of probabilities is generated in which the discrimination probabilities our model has assigned to all test samples for class $j$ are stored. The list is then sorted by decreasing probabilities and each element is assigned a rank $k$. By varying the rank $k$ (by gradually lowering the probability threshold) a list of true positives and false positives is generated. Note that as the classification threshold is lowered, the model labels increasingly more samples as positive. This will lead to an increase in false positives. The list is consequently employed to come up with a list of precision values at different ranks $p(k)$. Considering all the $K$ cases in the list where the sound sample belongs to class $j$, the average precision is computed as:

$$P_A(j) := \frac{\sum_{k=1}^{K} p(k)\mathbf{1}(k)}{c_{tp}(j)}, \tag{12}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



where $\mathbf{1}(k)$ is an indicator function that equals unity if the sample at threshold $k$ is a true positive. The mean average precision $P_{mA}$ is then computed by averaging over all classes (species) [36].

$$P_{mA} := \frac{\sum_{j=1}^{n} P_A(j)}{n}. \tag{13}$$

## 3. Results and Discussion

We evaluated the performance of our classification algorithm by testing it for 20 trials on $n$ randomly selected species, (out of 659 species in the selected dataset) with $n$ varying between $n = 10$ and $n = 300$. The results for precision, AUC, mAP, recall and accuracy are summarized in Fig. 3. Box plots were estimated from 20 different randomized data sets for each $n$. Box plots also known as box-and-whisker plots provide robust statistical summaries for the data, if the sample size is relatively small, i.e. here 20. The box plot divides data into quartiles or fourths – 2 box panels and 2 whiskers. The middle 50% of data is spanned by the box with 25th percentile or 25% of data falling below lower edge of the box (first quartile) and 75% of data falling below the upper edge of box (third quartile). The edges of the box are often referred to as *hinges* and the length of the box is called the *interquartile range* (IQR). The median is indicated by the middle line of box. The whiskers mark the extremes for the remaining 50% of data [37]. Surprisingly, we find no increase in the size of the interquartile range for the performance measures with increasing $n$.

There are several aspects of these results which deserve to be addressed in more detail.

### 3.1. Variations Due to Randomized Sub-sets

As observed in Fig. 3 we can see that for each choice of a subset of $n$ species, we get a range of performance values, depending on the particular random selection of species. Our results show that the interquartile-ranges vary as much as 12% in some cases. For instance, in case of $n = 30$ in Fig. 3(c) we see that the mean average precision (mAP) varies between 0.72 for one subset of 30 species to 0.84 for another subset of randomly drawn 30 species. Similarly for $n$ = 70, the mAP varies between 0.6 and 0.71. We can see a similar trend in the figures for other metrics considered in this work. As mentioned earlier the experiment has been repeated 20 times for different randomized selections of $n$ species. The variation in results within different $n$ species' trials show that classification results can vary significantly depending on the choice of species chosen for the analysis. Consequently, it can be inferred that generic claims about the performance of a certain algorithm for a certain number of non-randomly selected species, must be interpreted with caution. The results might not generalize for another set of $n$ species, even when the species are drawn from the same dataset.

One possible reason, among others, that could explain the variability in performance between different subsets or ensembles of randomly drawn sound samples from $n$ species (*bag-of-birds*) is the possible degrees of similarity of sounds, or the lack of it, between species of different subsets. Ambiguity can be a consequence of similarity of the sounds of species within an ensemble. Therefore one possible explanation for these results is that sounds of species in the bags leading to lower performance measures, have a higher degree of similarity compared to bags that generate higher classification performance.

### 3.2. The Dependence on the Number of Species

All performance measures decrease with increased number of species as is visible in Figs. 3 and 4. An intuitive explanation for this could be that species are more difficult to distinguish when more species are added to the classification task.

However, looking at the definitions of the performance measures (Eqs. 4 - 13) we tried whether it is possible to understand the numerical results by some analytic reasoning. Consider e.g. the precision $P(n)$ which is defined in Eq. (5). If each precision per species $p(j)$ contributing to the average was constant and not depending on $n$ (i.e. $p(j) \sim c$), one should expect $P(n) \sim c$. This is obviously not what is observed in Fig. 4. Therefore, one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



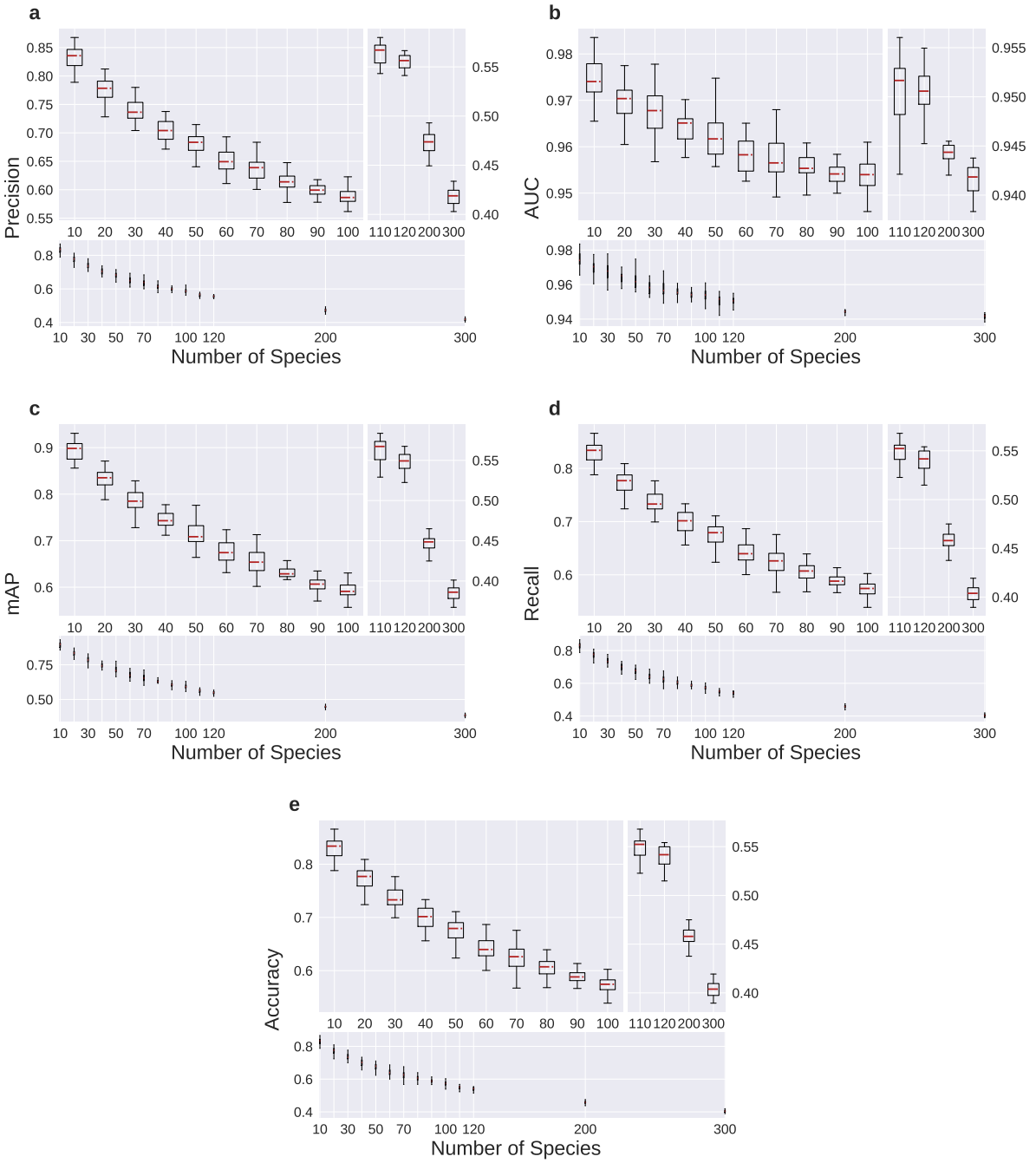


**Figure 3.** The performance measures a) Precision b) AUC c) mAP d) Recall and e) Accuracy decrease as the number of species in each subset is varied between $n = 10$ and $n = 300$. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for each value of $n$. The red marking in each box represents the median and the boxes indicate the middle 50% of the results.

must assume that $p(j)$ is dependent on $n$, although this is not explicitly visible in Eq. (4). To investigate this implicit dependence on $n$ we have visualized the average numbers of true positives

$$a_{tp} = \frac{1}{n} \sum_{j=1}^{n} c_{tp}(j), \tag{14}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



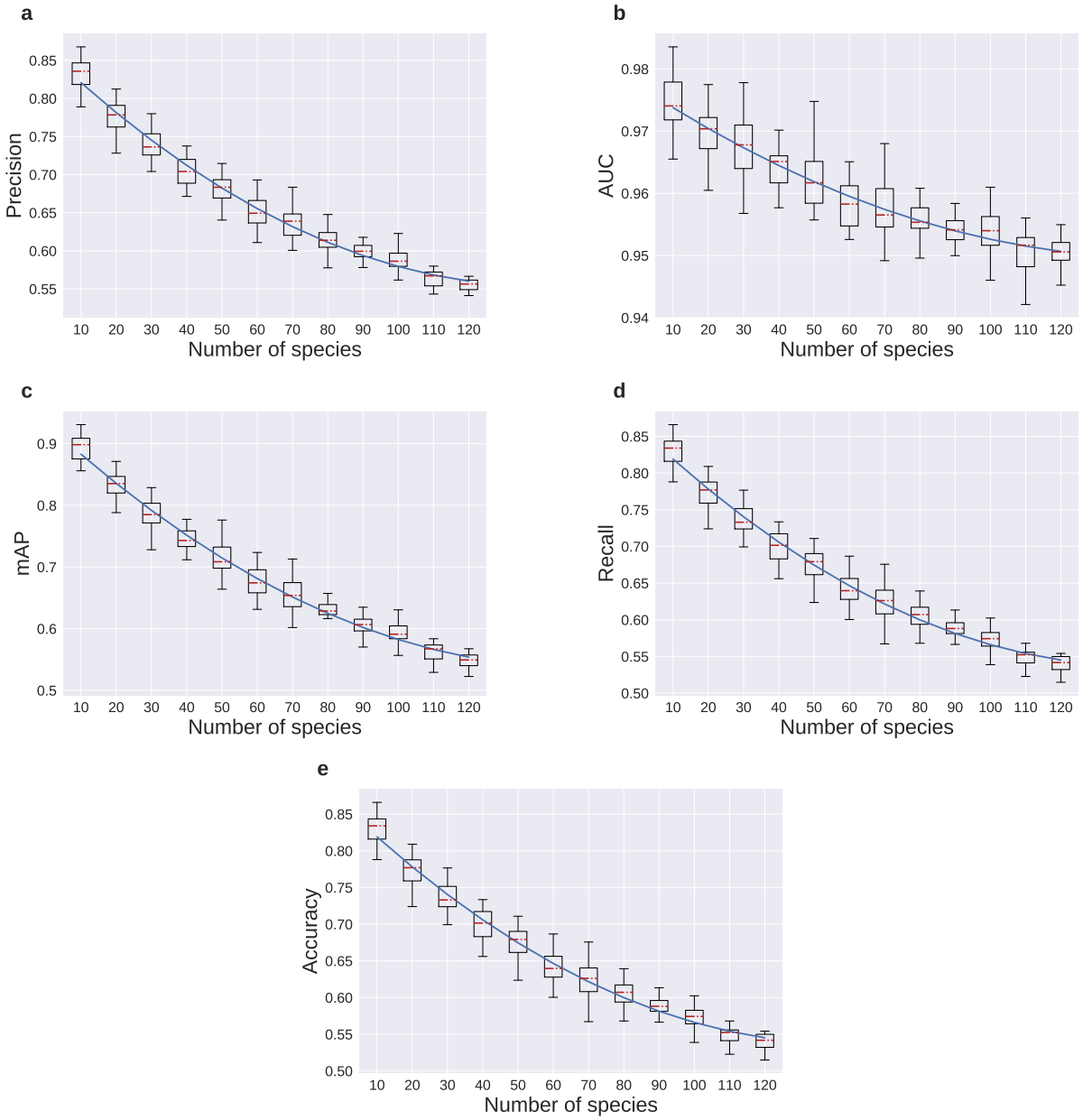


**Figure 4.** The $n$-dependence of a) Precision, b) AUC, c) mAP, d) Recall and e) Accuracy can be described by fitting quadratic functions (line). The error bars (whiskers) represent the ranges from best result to worst result obtained for 20 different randomly drawn subsets for each value of $n$. The red marking in each box indicates the median and the boxes show the range of the middle 50% of the results.

for each $n$ and in a similar way the averaged numbers of false positives, true negatives and false negatives in Fig. 5(a-c). These elements of a confusion matrix enter (in an non-averaged form) into the computed performance measures and thus their dependence on $n$ influences the performance measures. The averaging in Fig. 5 was done, since the amount of sound samples in each trial obviously depends linearly on $n$. Therefore this trivial dependence was removed and we can monitor a non-trivial implicit dependence on $n$. As one can see the dependence of the averaged numbers of true positive, false positives and false negatives can be described relatively well by a quadratic function, whereas the averaged number of true negatives increases linearly with increasing $n$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



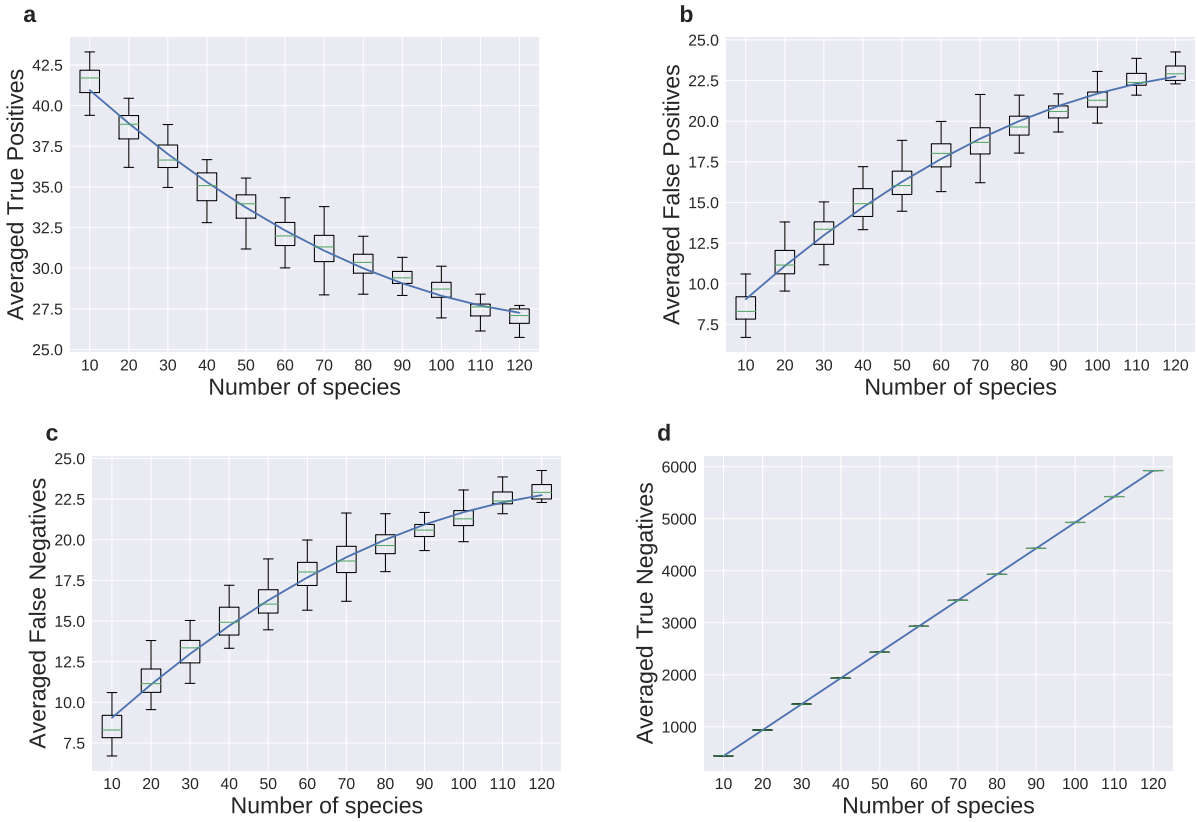


**Figure 5.** The elements of a confusion matrix ( a) averaged numbers of true positives, b) averaged numbers of false positives, c) averaged numbers of true negatives and d) averaged numbers of false negatives) display a linear (true negatives) and quadratic (all other elements) dependence on $n$. The error bars (whiskers) represent the ranges from best result to worst result obtained for 20 different randomly drawn subsets for each value of $n$. The green marking in each box indicates the median and the boxes show the range of the middle 50% of the results.

The fact that true negatives, as can be seen in Fig. 5(d), behave differently than the other elements of the confusion matrix can be understood by considering the way true negatives are computed in a multi-class classification problem, using a one-vs-all configuration. Each time a sound sample was correctly not classified as the particular species $j$ under consideration, the count of true negatives is increased by one. Therefore, e.g. in a subset of $m \cdot n = 500$ sound samples recorded from $n = 10$ different species and each species being represented by $m = 50$ sound samples, a perfect algorithm would classify 50 samples correctly as belonging to species $j$. Consequently the count of true positives would be $c_{tp}(j) = 50$ and the count of true negatives $c_{tn}(j) = 450$ for a perfect classifier. In other words we can expect $c_{tn}(j) = nm - m$ with $m$ being the sample size, as specified before, in case of a perfect classifier

$$a_{tn} = \frac{1}{n}\sum_{i=1}^{n} nm - m = \frac{n(nm-m)}{n} = m(n-1). \tag{15}$$

The results of the prediction experiments in this contribution with an obviously not perfect classifier, reveal that $a_{tn}$ can be fitted by a linear function $a_{tn}(n) = 49.88n - 59.27$. Note that the two coefficients are relatively close to the true sample size $m = 50$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



The dependence of the other elements of the confusion matrix on $n$ are more subtle with respect to the range in which this numbers vary and the dependence can be described by quadratic functions

$$a_{tp}(n) = d_2\,n^2 - d_1\,n + c_0, \tag{16}$$
$$a_{fp}(n) = -d_2 n^2 + d_1 n + d_0 \quad \text{and}, \tag{17}$$
$$a_{fn}(n) = -d_2 n^2 + d_1 n + d_0, \tag{18}$$

with $d_2 = 8.02$, $d_1 = 22.87$, $d_0 = 6.84$ and $c_0 = 43.16$. Note that the first two coefficients $d_1$ and $d_2$ of $a_{tp}$, $a_{fp}$ and $a_{fn}$ have either the same values (up to the first 8 digits which are not shown here), or just differ in sign, but not in value. These coefficients are shown in detail here, since we will in the following demonstrate a connection between Eqs. (16)-(18) and the functions describing the dependence of the overall performance measures.

Being able to describe $a_{tp}(n)$, $a_{fp}(n)$ and $a_{fn}(n)$ one can now try to understand the dependencies of the performance measures. Assuming that each species is classified equally well by a perfect classifier, one would expect $a_{tp}(n) = c_{tp}(j,n)$ for all $j$ and similar for $a_{fp} = c_{fp}(j,n)$ and $a_{fn} = c_{fn}$. Inserting Eq. (16) and Eq. (17) in Eq. (4) holds

$$p(j,n) \approx \frac{a_{tp}(n)}{a_{tp}(n) + a_{fp}(n)} \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0}, \tag{19}$$

since the non-constant terms in the denominator cancel each other. Inserting this in the equation for the overall precision (Eq. 5) holds

$$P(n) \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0} \sim a_{tp}(n), \tag{20}$$

since all terms $p(j)$ are identical for the perfect classifier. Consequently one should be able to predict the scaling of $P(n)$, knowing the coefficients $d_2, d_1, d_0$ and $c_0$.

Fitting the coefficients for the quadratic function describing $P(n)$ as in Fig. 4, one obtains

$$P(n) \approx g_2 n^2 + g_1 n + g_0, \tag{21}$$

with $g_2 = 0.16$, $g_1 = -0.44$, $g_0 = 0.86$. Note that these coefficients are very close to the coefficients of $a_{tp}$ multiplied with a factor $\frac{1}{d_0+c_0} = \frac{1}{50}$ as indicated by Eq. (20). Hence, we could confirm numerically that the dependence of the precision on the number of classes follows the dependence of $a_{tp}$ up to a scaling factor of $\frac{1}{d_0+c_0} = \frac{1}{50}$.

Following the same assumptions and reasoning one obtains

$$R(n) \approx r(j) \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0} \sim a_{tp}(n), \tag{22}$$

for the recall. Also here, the relation between the fitting coefficients of $a_{tp}$ and $R$ is confirmed by the quadratic function fitted to $R$ in Fig. 4. Note that for the prediction experiments in this study the same quadratic function is able to describe the $n$-dependence of precision and recall.

Extending the above reasoning (i.e. $c_{tp}(j,n) \approx atp(n)$) to explain the $n$-dependence of the accuracy as given by Eq. (8) yields

$$A(n) \approx \frac{1}{m}\left(d_2 n^2 - d_1 n + c_0\right) \sim a_{tp}(n). \tag{23}$$

Also this relation was numerically confirmed by comparing the coefficients for the polynomials describing $A$ and $a_{tp}$.

Discussing the $n$-dependence of the multi-class AUC and the mAP analytically is not as straightforward as the previous considerations, therefore only numerical results are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



presented in this contribution. As one can see in Fig. 4, the $n$-dependence of AUC and mAP can be also described by quadratic functions. Additionally, we observe that the coefficients for the linear and the quadratic term of the function describing the mAP resemble the coefficients describing $P(n)$ in value. Consequently one can argue that the above discussion for $P(n)$ could possibly also explain the $n$-dependence of mAP. Nevertheless, the constant term added to the function describing $P_{mA}(n)$ is higher than the constant offset of the precision.

Summarizing we can relate the n-dependence of several measures for the classification success to the n-dependences of the confusion matrix, assuming the behaviour of a perfect classifier and we fit functions describing these dependencies. Note that this does not imply that we claim our classifier to be a perfect classifier, neither do we claim that scaling with $n$ which we obtain here, is universal in the sense that it will be observed for any other classifier. The latter aspect is a question which needs to be tested in future contributions, but it is out of the scope of this work.

### *3.3. Metric of Confidence*

The decisions made by the classifier are based on probabilities which are estimated (through the ANN) for each species. The predicted label is then assigned to the species

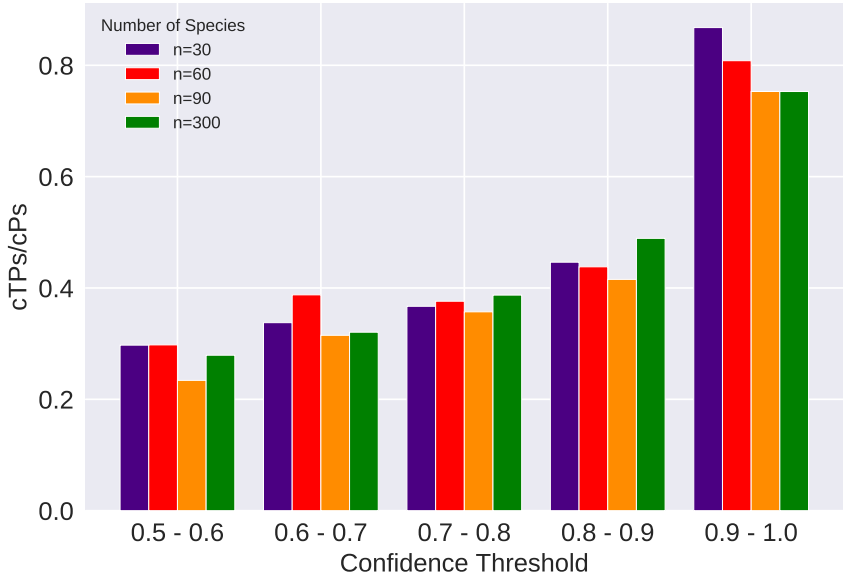


**Figure 6.** The precision (cTPs/cPs) increases as the confidence threshold (probability a sample needs to be assigned by the model in order for it to be classified as a positive prediction) is varied between 0.5 and 1.0. Shown are precision values for $n = 30, 60, 90$ and 300. As we can see for different $n$ as the confidence threshold tends toward 1.0 precision also increases significantly.

with the highest probability. Here, we analyze the effect of introducing a *confidence threshold* requiring the assigned probability to be above the threshold in order to accept the classification. In Fig. 6 one can see that the precision (cTPS/cPs), i.e. the ratio of true positives to all classified positives changes as the confidence threshold is varied between 0.5 and 1.0. We see that precision increases as the confidence threshold is increased. And for instance, for $n = 30$ species, the precision for confidence threshold in range 0.9-1.0 is more than 0.8. Similar results can be seen for other $n$. This basically shows when the model is assigning high confidence to its predictions, the predictions are mostly correct which should be expected from a good classifier.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



### 3.4. Comparing Different Measures for Classification Success

In this contribution, we use several common measures for evaluating classification performance and compare their results. The primary reason for this is that different indices

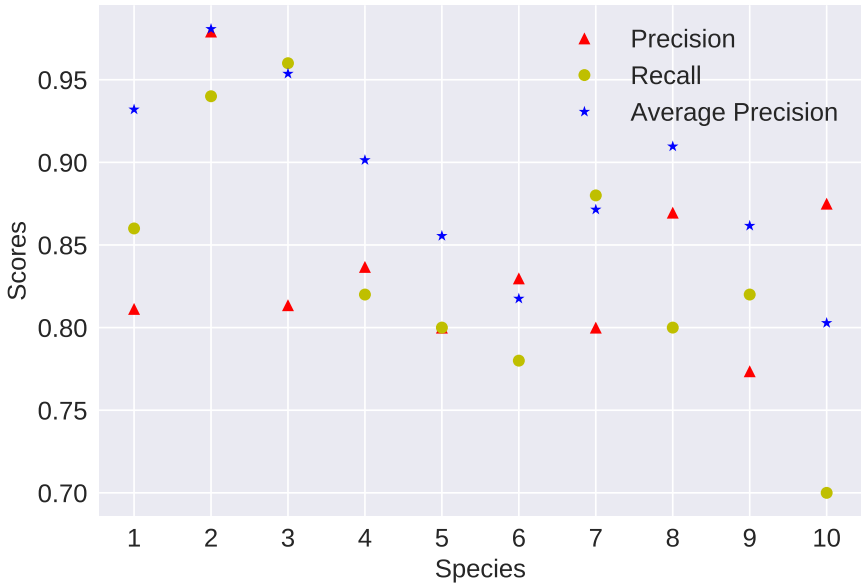


**Figure 7.** The precision, the recall, and the average precision for individual species classified in a subset composed of 10 randomly selected species are compared. The largest and the smallest value of each performance measure indicate a relatively large variation and strong dependence on the particular species under study.

encapsulate different aspect of the classification performance. Secondly, as mentioned earlier, there seems to be no consensus in the literature available on the choice of evaluation metric for the audio based bird species classification task. This compelled us to study a set of indices and not rely on a specific metric.

As one can see in Fig. 3 the precision and recall for $n < 100$ do not show much disparity, rather look quite similar. Although, by definition, these two indices encapsulate different aspects of model performance. This can be clearly seen in Fig. 7. Here we see that for different species in one classification run of $n = 10$ the precision and recall values differ. There are species where precision is higher than recall (e.g. species 10) while others where recall is higher than precision (e.g. species 3). But it seems for $n < 100$, the precision and recall values more or less equalize when an average is taken over species. But as the number of species increases ($n > 100$) we see that the precision increases slightly compared to recall, at least for some computation runs. For precision to be higher than recall, one can infer that the model has done a better job at classifying the samples correctly than in labeling the samples as positives.

From Figs. 3 and 4 one can additionally see that the accuracy is exactly the same as recall, since the equations of recall and accuracy become the same when an averages are taken over all classes.

Additionally, the mean average precision (mAP) was used to evaluate classification success. Increasingly number of works in the recent years have been using this metric to state the classification performance of their models. Note that average precision is one way of measuring the area under the precision-recall curve. Compared to precision and recall that are computed for one probability threshold, average precision is computed cumulatively by varying the threshold. We see in Fig. 3 that although it follows a similar downward quadratic trend as recall and precision, the mAP values are slightly higher

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



than the precision and recall values for different $n$. For instance, the range for $n = 10$ species for the 20 runs is between 0.86 and 0.94, whereas the precision and recall ranges are between 0.79 and 0.87. Therefore, as per this metric our model performs better compared to evaluating using the other two metrics. This observation also reflected in the offset of the functions describing the $n$-dependence as mentioned above.

Another commonly used metric for classification success is the area under the Receiver Operating Characteristics curve (AUC). Our model achieves a high score on the AUC metric as can be seen in Figs. 3 and 4. Although the AUC score decreases with the increase in number of species $n$, the score is nevertheless unexpectedly high. For instance, the AUC score for $n = 300$ for one run is 0.94 which is unexpected for such a large number of species. (Note that as per the definition of the AUC a random classifier making randomized decisions should give a score of 0.5).

In our understanding, the multi-class nature of our problem explains this result. As mentioned earlier, the AUC metric is essentially designed for a binary classifier and has later been generalized for multi-class classification problems [35]. Therefore in case of multi-class problems one needs to binarize the class labels to compute the AUC score, such that the problem is transformed into a binary classification problem with $\frac{n(n-1)}{2}$ binary classifiers (where $n$ is number of classes). Using a one-vs-one configuration [35] [38], as recommended by tutorials of many software packages, an AUC score is then computed for each of these binary classifiers and finally an average is computed to get a final AUC score for the entire set of $n$ classes. For an actual binary classifier that classifies poorly, the miss-classifications will reflect in significant enough values of false positives and false negatives to give us a low true positive rate and high false positive rate as per Eq. 9 and 10. This will result in a low AUC score. But in the multi-class scenario with one-vs-one configuration we observethat a classifier distributing miss-classifications sparsely across several classes leads to small number of false positives and small number of false negatives for these artificially assumed binary classifiers. One should note that this will happen even if the classifier fairs poorly i.e. miss-classifies with a high rate. An example for this can be seen in Fig. 8 which shows a confusion matrix for a classification run with 20 species. It can be seen that the miss-classifications are spread throughout the rows and columns of the confusion matrix. Consequently, less numbers of false positives and false negatives will amount to high true positive rate and low false positive rate for individual binary comparisons. And therefore a high AUC score (refer to Eqs. (9) and (10)). This is exactly what is reflected on averaging the individual AUC scores to compute the total AUC score for $n$ classes. The classifier is distributing the false predictions sparsely across several classes and the one-vs-one generalization is unable to capture the actual performance of the model. This leads us to the conclusion that ROC is not a suitable performance measure for multi-class classification tasks. Especially in cases where the miss-classifications are distributed rather evenly among several classes, it is very likely to obtain overestimated AUC scores.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2021 doi:10.20944/preprints202108.0277.v1



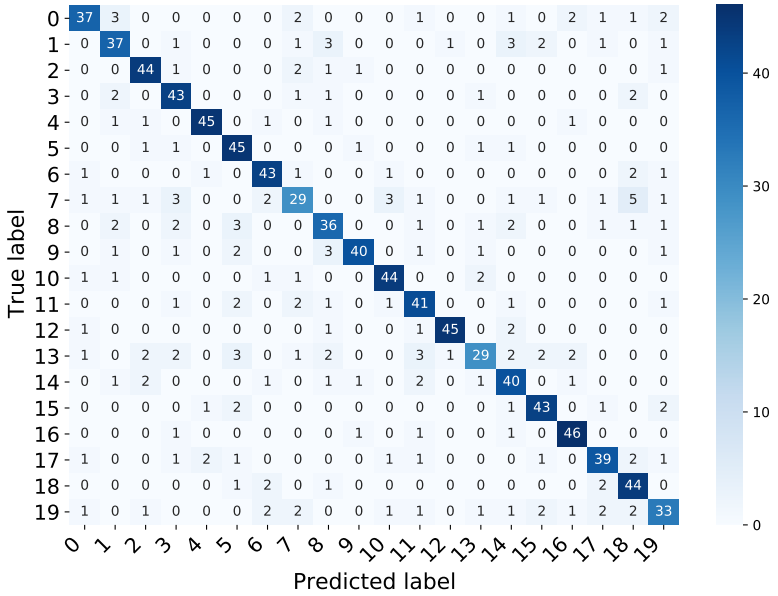


**Figure 8.** The confusion matrix for a classification run with $n = 20$ species displays that the misclassifications are spread among several classes. This can result in an overestimation of the averaged AUC for multi-class classifications.

## 4. Conclusions

The novelty of this work lies in studying the dependence of classification success on the number of species for bird sound classification. Furthermore, the idea is to illustrate how these classification results are heavily contingent on the composition of bird species subsets. Therefore we employ balanced subsets of bird sounds for $n$ species, drawing the species randomly from a larger dataset containing 659 species, where $n$ is varied between 10 and 300. For each $n$ we repeat the whole procedure (composition of the subset, training of the classifier and testing) 20 times to come up with a reliable estimate of the performance given a certain number of bird species.

The classification is performed using a shallow feed forward neural network trained on 45 pre-computed sound features. We have used a shallow neural network to conduct our analysis primarily due to its model simplicity, less computational costs and relatively less amount of data that is required to train such networks vis-a-vis the deep neural networks. We wanted to benchmark the classification performance and perform our analysis using a simple model that can be trained using hand crafted sound features.

We evaluate the classification performance using several common measures for classification success and also analyze their dependence on $n$ in detail. We observed that the classification performance is relatively high, even when many different species are present in the datasets under study and using relatively less data. This is an interesting result, since many recent approaches are based on deep neural networks trained on much larger datasets of images of spectrograms without any feature selection. This suggests that shallow neural networks trained on pre-computed sound features can also provide a robust approach to bird classification which at the same time is inexpensive in terms of computational costs and the amount of data used.

Concerning the robustness of the approach we find that all measures of classification success show a decline in value if the number of species present in the subset is increased. For some of these measures this decline can be explained analytically knowing the n-dependence of the confusion matrix and assuming the behavior of an idealized perfect classifier.

Additionally, we observe that the classification success depends on the individual composition of the bird subsets and classification results can vary significantly depending on the choice of species chosen for the analysis. For this reason, it seems the generic claims about the performance of a certain algorithm for say $n$ species of non-randomly drawn species, must not be interpreted as a generalized measure of performance for any $n$ species. The classification results might not generalize for another set of $n$ species, even when the species are drawn from the same dataset.

**Author Contributions:** Conceptualization, methodology, coding, validation, writing, B.G.; Conceptualization, methodology, review, S.H. Both authors have read and agreed to the published version ofthe manuscript.

**Funding:** B.G. received financial support from the project titled *AuTag BeoFisch* (LFF-FV91) funded by the *Landesforschungsförderung Hamburg*.

**Data Availability Statement:** A publicly available dataset was analyzed in this study. The dataset has been taken from the Xeno-Canto repository for bird sounds: https://www.xeno-canto.org/.

**Acknowledgments:** We are grateful to the creators of Xeno-Canto repository for providing the excellent dataset of bird recordings which was the basis for this study. We thank Timo Gerkmann and Florentin Wörgötter for fruitful discussions and *Landesforschungsförderung Hamburg* for their financial support.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| RMSE | root-mean-square energy |
| MFCC | Mel-frequency-cepstral-coefficients |
| ReLU | Rectified linear units |
| AUC | Area Under ROC Curve |
| ROC | Reciever Operating Characterics |
| mAP | Mean Average Precision |
| IQR | Interquartile range |
| ANN | Artificial neural netwroks |

## References

1. Sutherland, W.J.; Newton, I.; Green, R. *Bird ecology and conservation: a handbook of techniques*; Vol. 1, OUP Oxford, 2004.
2. Priyadarshani, N.; Marsland, S.; Castro, I. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology* **2018**, *49*, jav–01447.
3. Zhang, X.; Li, Y. Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing* **2015**, *155*, 108–116.
4. Jančovič, P.; Köküer, M. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing* **2011**, *2011*, 982936.
5. Fox, E.J.; Roberts, J.D.; Bennamoun, M. Text-independent speaker identification in birds. Ninth International Conference on Spoken Language Processing, 2006.
6. Cai, J.; Ee, D.; Pham, B.; Roe, P.; Zhang, J. Sensor network for the monitoring of ecosystem: Bird species recognition. 2007 3rd international conference on intelligent sensors, sensor networks and information. IEEE, 2007, pp. 293–298.
7. Chen, Z.; Maher, R.C. Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America* **2006**, *120*, 2974–2984.
8. Jančovič, P.; Köküer, M. Acoustic recognition of multiple bird species based on penalized maximum likelihood. *IEEE Signal Processing Letters* **2015**, *22*, 1585–1589.
9. Wielgat, R.; Potempa, T.; Świętojański, P.; Król, D. On using prefiltration in HMM-based bird species recognition. 2012 International Conference on Signals and Electronic Systems (ICSES). IEEE, 2012, pp. 1–5.
10. Briggs, F.; Lakshminarayanan, B.; Neal, L.; Fern, X.Z.; Raich, R.; Hadley, S.J.; Hadley, A.S.; Betts, M.G. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America* **2012**, *131*, 4640–4650.
11. Juang, C.F.; Chen, T.M. Birdsong recognition using prediction-based recurrent neural fuzzy networks. *Neurocomputing* **2007**, *71*, 121–130.

 doi:10.20944/preprints202108.0277.v1




12. Sprengel, E.; Jaggi, M.; Kilcher, Y.; Hofmann, T. Audio based bird species identification using deep learning techniques. Technical report, 2016.
13. Bastas, S.; Majid, M.W.; Mirzaei, G.; Ross, J.; Jamali, M.M.; Gorsevski, P.V.; Frizado, J.; Bingman, V.P. A novel feature extraction algorithm for classification of bird flight calls. 2012 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2012, pp. 1676–1679.
14. Selin, A.; Turunen, J.; Tanttu, J.T. Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing* **2006**, *2007*, 1–9.
15. Xeno Canto, https://www.xeno-canto.org/.
16. McFee, B.; McVicar, M.; Balke, S.; Thomé, C.; Raffel, C.; Lee, D.; Nieto, O.; Battenberg, E.; Ellis, D.; Yamamoto, R.; others. librosa/librosa: 0.6. 3. *URL: https://doi. org/10.5281/zenodo* **2019**, *2564164*.
17. Marler, P.R.; Slabbekoorn, H. *Nature's music: the science of birdsong*; Elsevier, 2004.
18. Kahl, S.; Wilhelm-Stein, T.; Klinck, H.; Kowerko, D.; Eibl, M. Recognizing birds from sound-the 2018 BirdCLEF baseline system. *arXiv preprint arXiv:1804.07177* **2018**.
19. Xie, J.; Zhu, M. Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics* **2019**, *52*, 74–81.
20. Virtanen, T.; Plumbley, M.D.; Ellis, D. *Computational analysis of sound scenes and events*; Springer, 2018.
21. Klapuri, A.; Davy, M. Signal processing methods for music transcription **2007**.
22. Smith, J.O. *Spectral Audio Signal Processing*; http://ccrma.stanford.edu/ jos/sasp/, accessed <date>. online book, 2011 edition.
23. Giannakopoulos, T.; Pikrakis, A. *Introduction to Audio Analysis: a MATLAB® approach*; Academic Press, 2014.
24. Abreha, G.T. An environmental audio-based context recognition system using smartphones. Master's thesis, University of Twente, 2014.
25. Logan, B.; others. Mel frequency cepstral coefficients for music modeling. Ismir, 2000, Vol. 270, pp. 1–11.
26. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; others. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), 2016, pp. 265–283.
27. Bebis, G.; Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* **1994**, *13*, 27–31.
28. Fine, T.L. *Feedforward neural network methodology*; Springer Science & Business Media, 2006.
29. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep learning*; Vol. 1, MIT press Cambridge, 2016.
30. Goodfellow, I.; Bengio, Y.; Courville, A. Softmax Units for Multinoulli Output Distributions. Deep Learning, 2018.
31. Bishop, C.M. *Pattern recognition and machine learning*; springer, 2006.
32. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
33. Sunasra, M. Performance Metrics for Classification problems in Machine Learning, 2019.
34. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: an Overview, 2020, [arXiv:stat.ML/2008.05756].
35. Hand, D.J.; Till, R.J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* **2001**, *45*, 171–186.
36. Kahl, S.; Stöter, F.R.; Goëau, H.; Glotin, H.; Planque, R.; Vellinga, W.P.; Joly, A. Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes. Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum. CEUR, 2019, number 2380, pp. 1–9.
37. Nuzzo, R.L. The box plots alternative for visualizing quantitative data. *PM&R* **2016**, *8*, 268–272.
38. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.