Received: 6 June 2022 | Accepted: 31 August 2022

DOI: 10.1111/2041-210X.14003

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Domain-specific neural networks improve automated bird sound recognition already with small amount of local data

Patrik Lauha[1] | Panu Somervuo[1] | Petteri Lehikoinen[1] | Lisa Geres[2,3] | Tobias Richter[4,2] | Sebastian Seibold[4,2] | Otso Ovaskainen[1,5,6]

[1]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland; [2]Berchtesgaden National Park, Berchtesgaden, Germany; [3]Goethe University Frankfurt, Faculty of Biological Sciences, Institute for Ecology, Evolution and Diversity, Conservation Biology, Frankfurt am Main, Germany; [4]TUM School of Life Sciences, Ecosystem Dynamics and Forest Management, Technical University of Munich, Freising, Germany; [5]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland and [6]Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Patrik Lauha
Email: patrik.lauha@helsinki.fi

Funding information
H2020 European Research Council, Grant/Award Number: 856506; Jane ja Aatos Erkon Säätiö; Norges Forskningsråd, Grant/Award Number: 223257; Suomen Akatemia, Grant/Award Number: 309581

Handling Editor: Sarab Sethi

Abstract

1. An automatic bird sound recognition system is a useful tool for collecting data of different bird species for ecological analysis. Together with autonomous recording units (ARUs), such a system provides a possibility to collect bird observations on a scale that no human observer could ever match. During the last decades, progress has been made in the field of automatic bird sound recognition, but recognizing bird species from untargeted soundscape recordings remains a challenge.
2. In this article, we demonstrate the workflow for building a global identification model and adjusting it to perform well on the data of autonomous recorders from a specific region. We show how data augmentation and a combination of global and local data can be used to train a convolutional neural network to classify vocalizations of 101 bird species. We construct a model and train it with a global data set to obtain a base model. The base model is then fine-tuned with local data from Southern Finland in order to adapt it to the sound environment of a specific location and tested with two data sets: one originating from the same Southern Finnish region and another originating from a different region in German Alps.
3. Our results suggest that fine-tuning with local data significantly improves the network performance. Classification accuracy was improved for test recordings from the same area as the local training data (Southern Finland) but not for recordings from a different region (German Alps). Data augmentation enables training with a limited number of training data and even with few local data samples significant improvement over the base model can be achieved. Our model outperforms the current state-of-the-art tool for automatic bird sound classification.

4. Using local data to adjust the recognition model for the target domain leads to improvement over general non-tailored solutions. The process introduced in this article can be applied to build a fine-tuned bird sound classification model for a specific environment.

KEYWORDS
autonomous recording units, bioacoustics, bio-monitoring, bird sound recognition, convolutional neural networks, deep learning, model fine-tuning

1 | INTRODUCTION

Birds are important indicators of the state of the environment (Carignan & Villard, 2002; Fraixedas et al., 2020; O'Connell et al., 2000) and one of the most well-studied groups of taxa. As ecosystems are suffering from ongoing global biodiversity loss, bird populations all over the world are also decreasing at an alarming rate (Burns et al., 2021; Rosenberg et al., 2019; Sanderson et al., 2006; Xu et al., 2019). Collecting information about the changes in bird communities to efficiently target conservation measures is thus especially important at the moment.

Birds are easiest to detect by their vocalizations and passive acoustic monitoring has recently been applied in various studies (Aide et al., 2013; Farina et al., 2011; Frommolt, 2017; Furnas & Callas, 2015; Matsubayashi et al., 2017; Sethi et al., 2022; Shonfield & Bayne, 2017). Autonomous recording units (ARU) are a cost-effective tool for collecting audio data and the reliability of the species occurrence data produced by them is comparable to traditionally used human-made point counts (Darras et al., 2018; Shonfield & Bayne, 2017). ARUs may even outperform human observers in logistically challenging terrain such as mountains, because they can be set-up anytime while observers need to get on site each morning. Another advantage of ARUs is that they can stay on site for days or weeks and thus provide a better coverage of bird species than human observers. However, even with the ARUs producing species level identifications for audio data often requires a human expert and can be a laborious task especially when the data collection period is long. The poor scalability of human-based annotation production motivates the demand for reliable and efficient automatic identification systems.

Automated bird sound identification has been studied since the late 1990s (Anderson et al., 1996; Kogan & Margoliash, 1998). Various techniques, such as multilayer perceptrons (McIlraith & Card, 1997), decision trees (Lasseck, 2015), support vector machines (Fagerlund, 2007), classification of singular vectors (Hansson-Sandsten, 2015) or sinusoidal modelling (Härmä, 2003) have been applied and proven to work at least for a limited set of species. However, expanding to dozens or hundreds of species has been considered difficult (Priyadarshani et al., 2018).

Currently convolutional neural networks (CNN) are the best performing and most widely used method in automated bird sound recognition (Joly et al., 2021; Kahl et al., 2021). Typically CNNs are applied on spectrogram images, which are obtained from audio data through short-time Fourier transform. An overview of the development of bird sound identification methods during recent years can be obtained by observing the results of the annual BirdCLEF challenge organized as a part of the Cross Language Evaluation Forum (CLEF). The BirdCLEF challenge has been organized since 2014 to promote the development of machine learning algorithms identifying bird vocalizations. The main task in the competition was initially to build a model for identifying the foreground species in a recording that is primarily targeted at a single bird (Joly et al., 2014), but has afterwards been extended to identifying and localizing all bird species from soundscapes of multiple possibly overlapping vocalizations from varying distances. In the early years of the BirdCLEF competition, the best performing solutions were based on calculating some low-level statistics from short audio segments and using these as input for a classifier (Joly et al., 2014, 2015). CNNs were first applied in the challenge in 2016 and have been victorious ever since (Joly et al., 2016, 2017, 2021). Already in the year 2017, all submitted solutions applied CNNs (Joly et al., 2017).

The current state-of-the-art solution for general bird species identification tasks covering hundreds of species is BirdNET (Kahl et al., 2021) by the Cornell Lab of Ornithology. The number of species (2400 by May, 2022) included in BirdNET clearly exceeds all other existing methods and in terms of mean average precision and F0.5 score, the performance is still comparable with other applications (LeBien et al., 2020; Ruff et al., 2020). Another version of BirdNET, BirdNET-Lite (Kahl, 2020) achieves an even greater number of classes, covering impressive 6000 species, which is more than half of all bird species of the world.

Targeted single-species recordings can be accurately classified with current CNN-based solutions (Joly et al., 2021; Kahl et al., 2021). However, being able to generalize to noisy soundscapes with overlapping vocalizations from varying distances and adapting to different acoustic environments remains the key challenge in automatic bird sound identification. In this work, we propose a solution to the problem through parameter fine-tuning.

In general, the task of constructing a well-performing classifier with limited amount of training data is challenging. Transfer learning and fine-tuning are commonly used methods to address the problem. Transfer learning often refers to re-using a model trained for one task for constructing another model for another task, while fine-tuning means that a model (or part of it) is simply re-trained with new data to adjust the model output for better performance with the new target data (Chollet, 2017). However, transfer learning and

fine-tuning as such are general concepts and there are many ways to implement the details in practice. One of the early papers to investigate the issue is Yosinski et al. (2014). Different fine-tuning strategies for CNNs have been compared by Chu et al. (2016) and Pittaras et al. (2017). An example of applying transfer learning for a network pre-trained on ImageNet (Deng et al., 2009) and fine-tuned with mel spectrogram images computed from the bird audio data is given by Lasseck (2018) and Sevilla and Glotin (2017).

In this article, we introduce a workflow for combining global data from existing sound libraries and local data from a specific target domain, to build bird sound identification models for recognizing bird species in a particular region with a specific set of species and a unique acoustic environment. We start by building an identification model based on global training data and show how this so-called global model can be fine-tuned with local data on Finnish birds to create a local model. Local fine-tuning is applied by freezing the convolutional layers of the network and retraining the parameters of the remaining two layers. We demonstrate the performance of our model and the advantage given by local fine-tuning through a case study of 101 bird species from Southern Finland. We also show that the improvement in classification accuracy only applies for test data from the same Southern Finnish domain and not for another test data set from the German Alps. Our solution can be applied for automatic collection of bird occurrence data in different locations around the world, by constructing and fine-tuning several site-specific recognition models. One example of this kind of scheme is the international LIFEPLAN Project (2022), which collects bird and bat vocalization data with 1000 globally distributed ARUs and has motivated the methodological development reported in this paper.

2 | MATERIALS AND METHODS

The process of training the classification model consisted of extracting bird vocalizations from raw audio data, converting the sound data into spectrograms, and using these as inputs for a convolutional neural network. First we extracted 1000 example vocalizations for 101 bird species from a global bird sound library and trained a so-called global base model with these data. This global model was then fine-tuned with a local data set, which contains data originating from the same source as the final test data. The local data set contained samples of 72 bird species, on average 146 samples and up to 1089 samples per species. For those 29 species, for which no local data were available, we reused the global data in the fine-tuning phase. To enable neural network training with a limited number of training samples, we applied data augmentation during both training and fine-tuning.

2.1 | Audio data

We used two main sources of audio data: Macaulay Library (2021) as "global data" and Kerttu data (Lehikoinen et al., 2022), that were collected specifically for the purpose of this project, as "local data". Macaulay Library (2021) is the world's largest archive of animal sounds and contains over 30 million media files of more than 10,000 bird species and other animals. From the Macaulay Library, we obtained high quality recordings for 101 Finnish bird species (12–1186 recordings per species). The length of the recordings ranged from a few seconds to several minutes and all recordings were labelled for one main species but could also contain other background species. The 101 species were selected based on their potentiality to appear in the Finnish test data set from Kerttu.

The Kerttu data set (Lehikoinen et al., 2022) was collected in Southern Finland during the summer of 2018 with ARUs and contains 1.8 million minutes of audio. The data were labelled in the citizen science platform Kerttu producing two types of data: species-specific short time- and frequency-specified templates with binary labels (target species present or not) and longer 10 s clips labelled with a list of occurring species. We used the as additional local training data for the models and the latter as a test data set. The labelled data for local fine-tuning contained 13,898 samples (0–1089 per species). The test data contained 2039 samples excluding clips that were flagged by data annotators to be poor quality and were thus rejected.

As a third data set, we used ARU recordings collected in Berchtesgaden National Park, Southern Germany during spring and summer 2021. Recordings were part of a biodiversity monitoring program conducted by national park staff (L. Geres, T. Richter, S. Seibold). To avoid unintended recording of people, notification signs were installed at each site. No further permissions for field work were required. The data contained 2318 two-minute recordings from 215 sites in forests and open habitats with species labels generated by expert ornithologists. The data set contained 47 species of the 101 species that were included in our model and was used as an additional test data set to evaluate the model performance.

The geographical distribution of different data sets is visualized in Supplement I. The species included in our model and in the different data sets are listed in Supplement II.

2.2 | Data preprocessing

To extract the training data from Macaulay data, we applied the Matlab implementation of the Animal Sound Identifier (ASI) (Ovaskainen et al., 2018). Thus, we scanned through the spectrograms of all recordings searching for patterns that based on cross-correlation repeat within the recordings of one species but not in the recordings of other species. A more detailed description of the process is given in Supplement III.

We used the ASI algorithm to find 1000 vocalization candidates for each species. These candidates were used in the training data set as such. In addition, we selected 10 best candidates for each species to be verified by human experts. To do so, we used the cross-correlation as a ranking score to the best match within the same file, penalized by correlation to already selected candidates. The 10

automatically extracted candidates were refined and verified manually to ensure that they belong to the correct species and form a set that represents the repertoire of the specific species as well as possible. These 10 manually verified candidates were used as a reserve set to ensure that there were at least some manually curated samples available even for those species that were not found from the fine-tuning data set.

We formed the training data set by extracting 4-s clips around the selected vocalizations from Macaulay and Kerttu data sets. The training data was randomly split into a training set and a validation set to monitor the training process of the model. Each clip was further split into three 3-s subclips with an overlap of 2.5 s between consecutive frames, in order to increase the size of the training data set and introduce temporal variation to the location of vocalizations within the final spectrogram images.

We transformed the audio data to spectrogram images using Python library librosa (McFee et al., 2015). All audio files were resampled to 22,050 Hz and converted to logarithmic mel-scale spectrograms with default parameters of librosa functions *librosa.load* (sr = 22,050), *librosa.feature.melspectrogram* (n_fft = 2048, hop_length = 512) and *librosa.power_to_db* (ref = max). Pixel values of each individual sample were standardized to have zero mean and unit variance. By standardizing each sample individually, information about the intensity differences between the samples were lost. However, this was not considered to be harmful, because the intensity of the signal depends mainly on the distance between the microphone and the vocalizing bird and not on the species to be classified.

2.3 | Classification model

We used a simple convolutional neural network as the classification model. For model input, we used 128×129 matrices, which correspond to audio clips of 3 s. The inputs can be interpreted as images, where each pixel value corresponds to the intensity of the certain frequency bin at a certain time point. The model structure follows the basic CNN architecture (Chollet, 2017) with four convolutional layers and a classification head of two fully connected layers. The structure of the model is described in Supplement IV.

Since any number of bird species can be present at the same time in a recording, we have a multilabel, multiclass classification problem. Consequently, we used sigmoid activations in the output layer. In hidden layers, we used rectified linear unit (ReLU) activations. The model was trained for 10 epochs with binary cross-entropy loss function and RMSprop optimizer with learning rate 0.0001 using batch size 64. To train the local model, the two last fully connected layers of the global model were fine-tuned with the local data for 20 epochs, while other training parameters and the rest of the model weights remained unchanged. The number of epochs was selected by observing the development of validation accuracy over the epochs. The training was stopped when the validation accuracy plateaued.

2.4 | Data augmentation

Data augmentation was applied in the training phase mainly for three reasons: to increase the size and diversity of the training data set, to avoid overfitting and to modify the training data to resemble the test data in terms of background noise and signal-to-noise ratio. Augmentation was applied for spectrogram images before log-transformation and standardizing. The set of selected augmentations consisted of methods that are meaningful with audio data. For example, rotation and flipping, which are commonly used for natural images, were not applied since they would result in a spectrogram that no longer represents the same vocalization as the original one. The augmentation methods applied were horizontal and vertical stretch, horizontal and vertical shift, raising to random power, adding noise, mixup and horizontal and vertical masking. Descriptions and visualizations of included augmentation methods are provided in Supplement V. Similar methods have been used by various authors such as Lasseck (2018), Kahl et al. (2021) and several participants in the Cornell Birdcall Identification Contest (2020).

2.5 | Model evaluation

The models were tested with the Kerttu test data set. Model predictions were produced by splitting the 10-s test files into slightly overlapping 3-s chunks, predicting for all of them and selecting the maximum prediction for each species. From these predictions, we calculated the area under the ROC-curve (AUC) for each species and compared the performance of our models with the current state-of-the-art model BirdNET (Kahl et al., 2021). Since the prevalence of different species varied significantly in the test data, we selected for each species all samples where the species was present and an equal number of randomly selected samples, where the species was not present. Species-specific AUCs were calculated for these subsets. The final results were averaged over 20 runs to minimize the random effect regarding the selection of the evaluation set.

2.6 | Quantifying the domain shift

To understand the difference between global and local data sets, we used the activations of the last convolutional layer of the global model to represent the data, that is each data point was represented by a 512-dimensional feature vector. Domain shift was quantified by calculating the average Euclidean distance between the local data and the species-specific mean of the global data.

3 | RESULTS

Our results show that local data are very useful for training the identification model and already 50 samples per class can significantly improve the model performance and help outperform non-localized

models. Fine-tuning only improves model performance with test data originating from the same domain as fine-tuning data and not for test data from another source.

3.1 | Performance of global and local models

We compared the performance of our models with the current state-of-the-art bird sound recognition model BirdNET. We used two versions of BirdNET, the recently published BirdNET-Analyser (Kahl, 2022) and the older BirdNET-Lite (Kahl, 2020). BirdNET-Lite was included in the comparisons, since BirdNET-Analyser does not cover all of the species that were studied. The models were tested with those 43 species that occurred in the test data at least 15 times, to ensure the reliability of test results. When comparing species-specific AUCs of different models, the global model matches and exceeds the performance of BirdNET and fine-tuning with local data improves the performance even further (Figure 1d). The mean of the AUCs was 0.835 for BirdNET-Lite, 0.830 for BirdNET-Analyser, 0.862 for the global model and 0.903 for the local model. Local model produced the best results for almost all species (Figure 1a).

We also studied the connection between the number of fine-tuning samples and the improvement obtained compared to the global model. The amount of local data needed to improve the AUC depends on the species. Even 50–100 fine-tuning samples can lead to significant improvement in model performance and help overcome non-fine-tuned methods (Figure 1c). The same effect can be seen in Figure 2, which shows the results from fitting the models with restricted amounts of local data both with and without global pre-training. Even small amounts of local data increase the average of the species-specific AUCs. When the amount of local data is large enough, the performance of the model trained with local data only matches the performance of the globally pre-trained model. However, when there is not enough local data (for example less than a few hundred samples per species), pre-training with global data is the best solution.

To examine the reasons for improved performance of the localized model, we tested the models with the data collected from Berchtesgaden National Park, Southern Germany. Overall, Berchtesgaden data contained vocalizations of 81 species. We evaluated the model performance for those 37 bird species that were included in the Finnish models and occurred in the Berchtesgaden data at least 15 times.

Our results show that with German data the performance of the Finnish local model does not exceed the performance of the global model (Figure 1e). This suggests that the improvement achieved with the localized model is not only caused by covering the domain shift between targeted recordings and ARU recordings, but also due to adapting to a specific type of sound environment and location-dependent vocalization types.

Species-specific AUCs are not directly comparable between Finnish Kerttu data and German data from Berchtesgaden, since the data sets differ from each other in terms of recording duration and quality of labels. However, different models can still be compared in relation to each other. Whereas with Kerttu data set our global model performed better than BirdNET and locally fine-tuned model still better than the global model, with Berchtesgaden data there are no differences between our models and BirdNET-Lite. BirdNET-Analyser performs better for most species than other models, but for some species the performance is very poor.

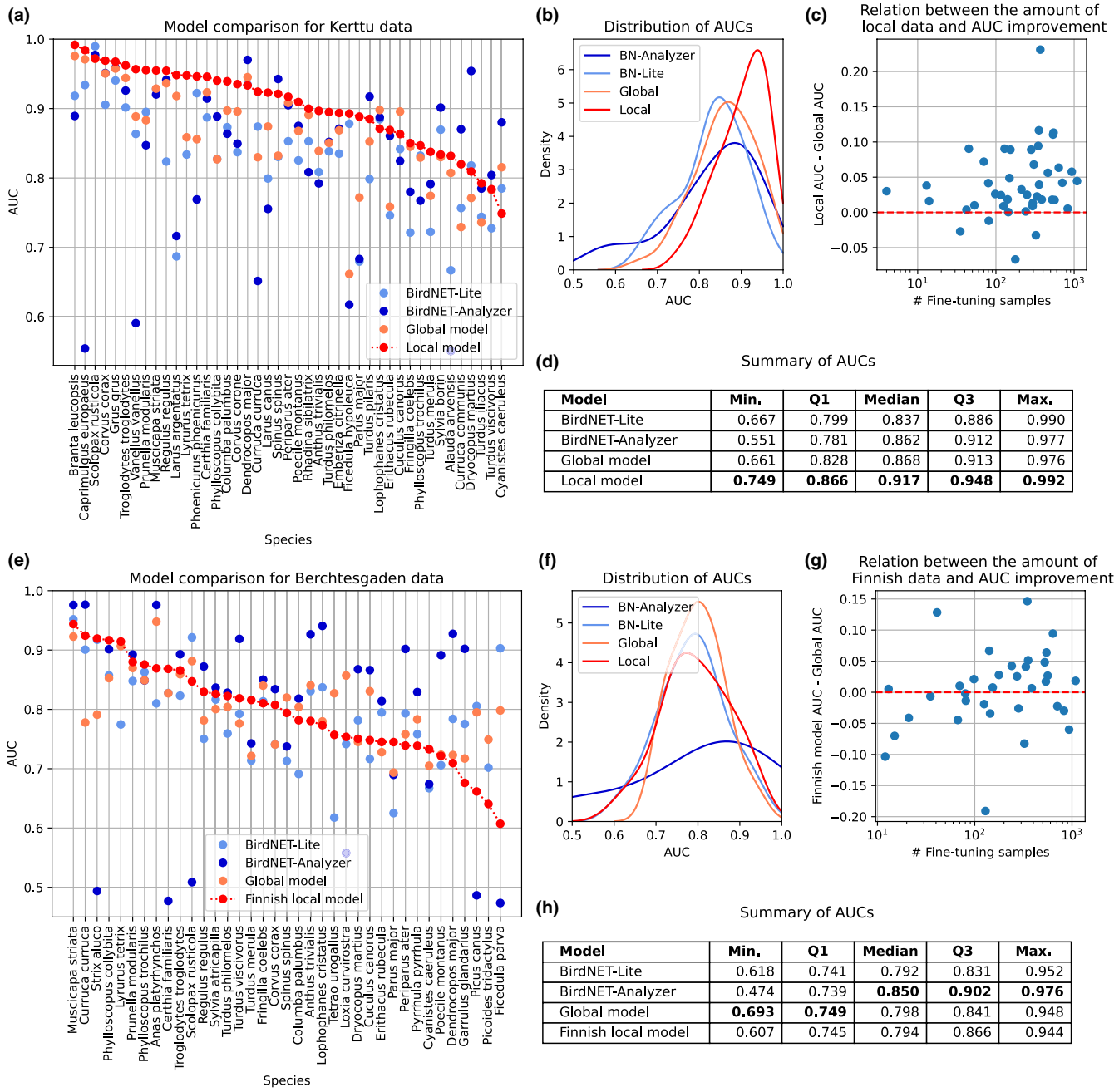
Compared to truly global BirdNET, which is trained for 2400 bird species worldwide (BirdNET-Lite for over 6000), our "global" model benefits from being trained for only those species that occur in the Finnish data and using local data for fine-tuning further improves the results. When applying the models on data collected in a different location, these advantages vanish.

3.2 | Inspection of model results

While selecting the set of augmentation methods, we tested different combinations of methods with a restricted data set of four species (*Fringilla coelebs*, *Phylloscopus trochilus*, *Turdus merula* and *Turdus philomelos*) to understand how individual augmentation methods affect the model results. Augmentation trials revealed that none of the augmentation techniques were very effective by themselves, but when random combinations of the techniques were applied together, data augmentation improved the species-wise AUCs by 8–22 percentage points. The most significant improvement was caused by adding noise and mixup. Mixup did not improve the validation accuracy in the validation data set originating from the same domain as training data, but yielded significant improvement for soundscape data and was therefore included. Heavy use of data augmentation provided two important advantages. Firstly, models could be fit and eligible results obtained with a very limited number of training data. Secondly, it helps to avoid overfitting.

We also investigated the difference between local and global data. Local data can be considered as a subset of the global data. We used t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten & Hinton, 2008) to project the data into 2-dimensional space. As an input to the t-SNE, we used 512-dimensional activations of the last convolutional layer of the network. In all data projections, local data form dense and visually distinct clusters among the broader set of global data, typically located at the edge of global data. We quantified the magnitude of the domain shift as explained in section 2.6 and studied how the domain shift affects the improvement achieved by local fine-tuning. The correlation between domain shift and AUC improvement was weak (0.314), but statistically significant ($p = 0.026$). Visualizations of t-SNE embeddings and the relation between domain shift and AUC improvement are provided in Supplement VI.

Further inspection of model predictions shows that even though no object detection techniques were explicitly applied, our models are capable of focusing on the important parts of the spectrogram. We studied how different parts of the test set spectrogram affect the prediction produced by the models with perturbation-based class



Model	Min.	Q1	Median	Q3	Max.
BirdNET-Lite	0.667	0.799	0.837	0.886	0.990
BirdNET-Analyzer	0.551	0.781	0.862	0.912	0.977
Global model	0.661	0.828	0.868	0.913	0.976
Local model	**0.749**	**0.866**	**0.917**	**0.948**	**0.992**

Model	Min.	Q1	Median	Q3	Max.
BirdNET-Lite	0.618	0.741	0.792	0.831	0.952
BirdNET-Analyzer	0.474	0.739	**0.850**	**0.902**	**0.976**
Global model	**0.693**	**0.749**	0.798	0.841	0.948
Finnish local model	0.607	0.745	0.794	0.866	0.944

FIGURE 1 Species-specific model comparison for test data from Southern Finland (a–d) and German Alps (e–h). Panels A and E show species-specific AUCs for all models ordered according to the AUCs of the locally fine-tuned model. With Finnish test data our global model exceeds the performance of BirdNET and fine-tuning with local data further improves the performance for almost all species. Fine-tuning with Finnish data does not improve the performance with German test data. Panels (b and f) show the distributions of AUCs across different species for all models and panels (d and h) the summary statistics for the same distributions. Panels (c and g) show how the number of local fine-tuning samples affects the improvement of AUCs compared to the global model. There is no obvious connection between the number of fine-tuning samples and improvement in local model performance. In all analysis the model performance was only evaluated for those species that occurred at least 15 times in the test data.

activation maps (Zeiler & Fergus, 2014). The class activation maps were created by covering different parts of the spectrogram by setting all pixels of a certain area to zero and comparing the model outputs for the original spectrogram and modified spectrograms. The class activation map reveals, if masking a certain area considerably affects the model output, which means that this area is especially important for producing the original classification. Figure 3 shows how the local model performs on an example clip, where vocalizations of several birds temporally overlap with each other.

Overall, the models seem to perform well with most foreground vocalizations. Local model produces more confident predictions and seems to perform better with the background species than the other

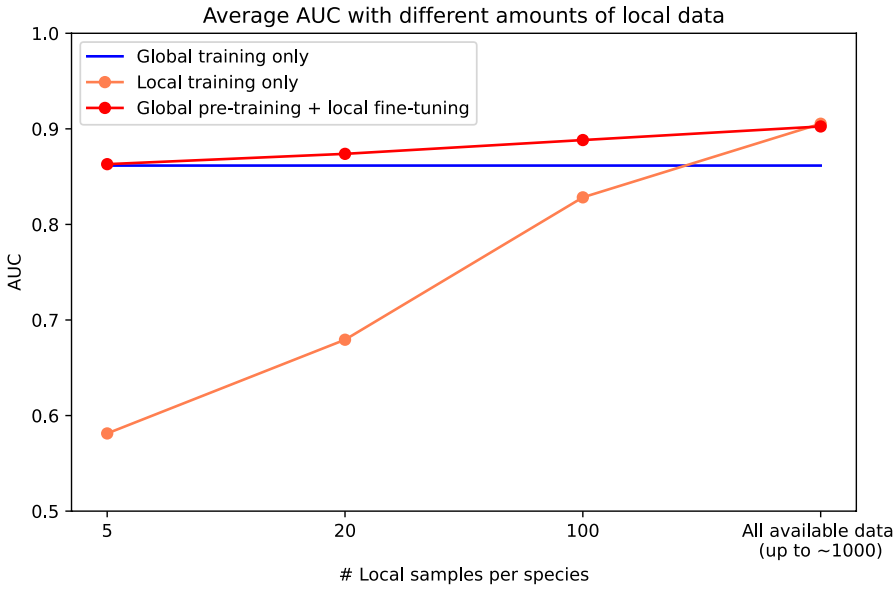


FIGURE 2 The importance of local and global data. Red and orange curves show the average AUCs of the model, when the size of the training set was restricted to 5/20/100/ all available samples per species. The more local data are available for fine-tuning, the better are the results, but already 20 samples per species brings some improvement over the global model. The model performance was evaluated for those species that occurred at least 15 times in the test data, and the AUC was calculated as an average of the species-specific AUCs.

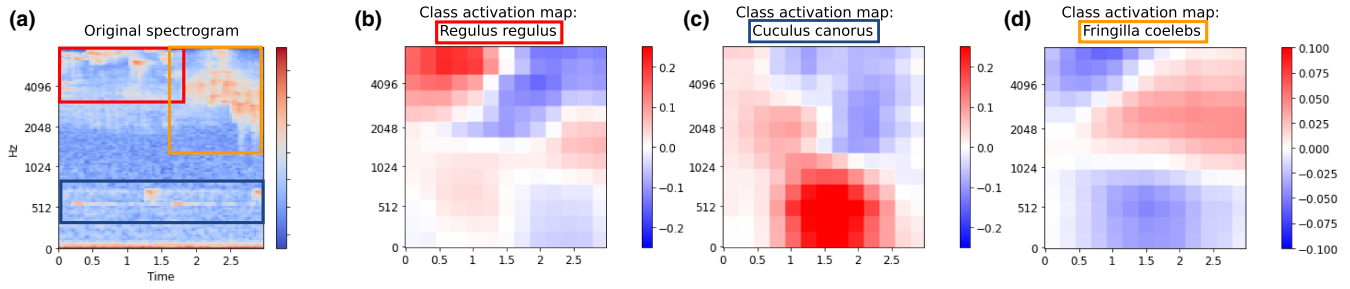


FIGURE 3 Class activation maps for the locally fine-tuned model for an example sample extracted from a 10 s test recording. The sample shown on panel (a) contains vocalizations of three species highlighted with manually drawn boxes. *Regulus regulus* (goldcrest) sings on the high frequencies, ending at 1.5 s. *Cuculus canorus* (common cuckoo) vocalizes through the whole sample on the lower frequencies and *Fringilla coelebs* (common chaffinch) starts to sing at 1.5 s on the higher frequencies just below the goldcrest. Panels (b–d) show class activation heatmaps for each species given by the local model. Red areas indicate which parts of the spectrogram have yielded higher predictions for the given class and should roughly correspond to the boxes on panel (a). Locally fine-tuned model gives predictions 0.324 for goldcrest, 0.485 for common cuckoo and 0.070 for common chaffinch (0.672, 0.732 and 0.256 for the whole 10 s clip). All predictions are based on correct parts of the spectrogram.

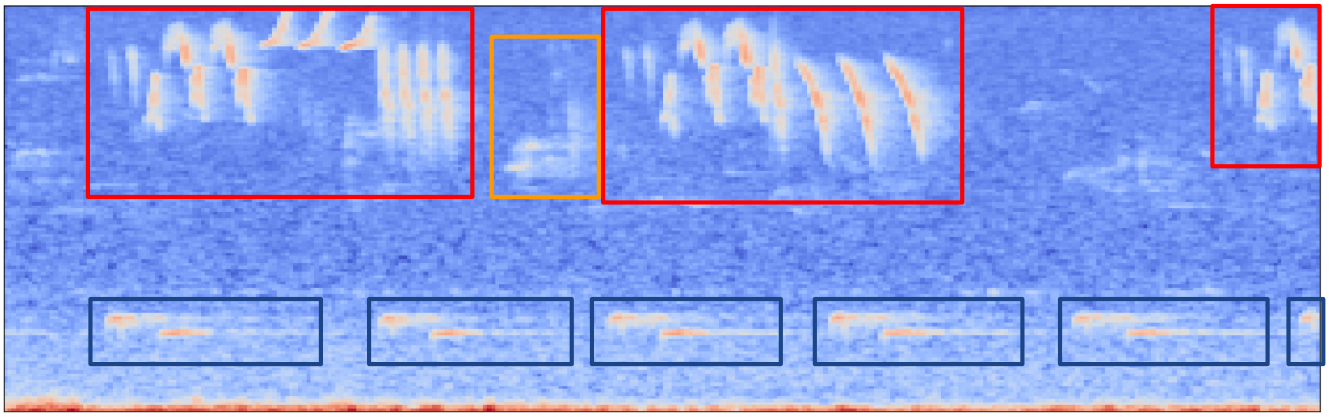
models. However, predicted probabilities for background species are still typically quite low. Moreover, the predictions of all models seem to be rather conservative and the predicted probabilities are often too low. Comparison of model predictions with two 10 s example recordings from the test set is shown in Figure 4.

4 | DISCUSSION

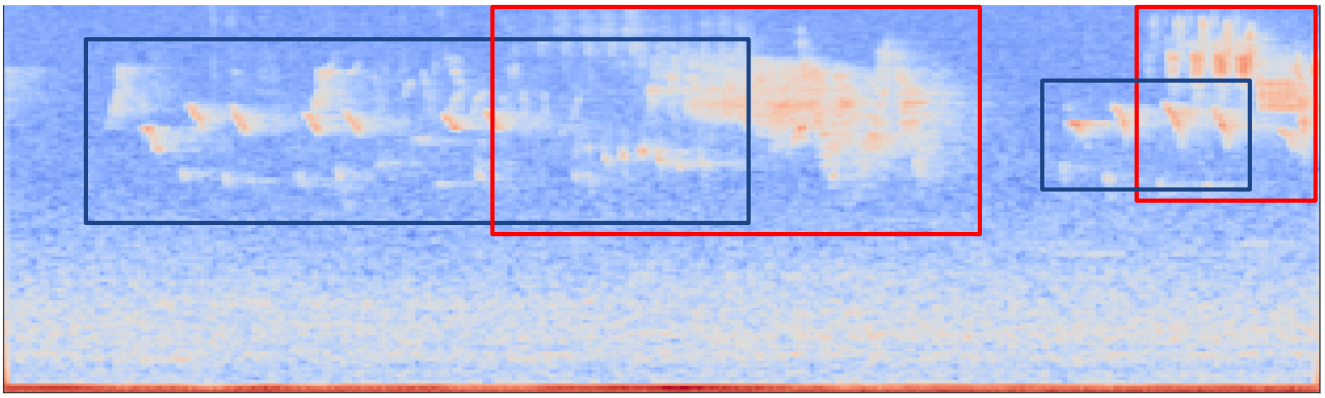
The results of our case study demonstrate that our approach can be used to construct bird sound identification models with improved performance compared to existing non-localized solutions. The same steps could in principle be repeated with different data to build a localized model for specific type of data, such as data from another location or from specific type of microphones.

The number of layers and parameters in our model is small compared to very deep networks used for image classification, such as Inception (Szegedy et al., 2015) or Resnet (He et al., 2016) architectures. Also BirdNET (Kahl et al., 2021) has a more complex model architecture than our model. However, spectrograms have a simpler structure than natural images, and our results indicate that even a shallow model has enough capacity to learn a high number of different vocalizations.

Another notable difference between our approach and widely used models such as BirdNET (Kahl et al., 2021) is the processing and use of training data. We have used a relatively small number of training samples selected by either human expert (10 hand curated samples for each species from Macaulay data & 0–1089 samples for each species from Kerttu data) or machine (1000 training data samples per species from Macaulay data), whereas BirdNET uses enormous amounts of uncurated data for training (Kahl et al., 2021). Generally neural networks benefit from large amounts of data, but with weakly labelled bird sound data, inclusive data selection might cause problems. For example, common species often occur in the background of the recordings labelled to other species. When background calls of common species are



True labels:	BirdNET-Lite predictions:	BirdNET-Analyzer predictions:	Global model predictions:	Local model predictions:
Anthus trivialis	Anthus trivialis : 0.7046	Anthus trivialis : 0.9321	Anthus trivialis : 0.7019	Anthus trivialis : 0.8865
Cuculus canorus	Cuculus canorus : 0.3871	Cuculus canorus : 0.326	Cuculus canorus : 0.6533	Cuculus canorus : 0.6931
Turdus philomelos	Motacilla alba : 0.2649		Periparus ater : 0.3391	Turdus philomelos : 0.2351
	Saxicola rubetra : 0.1712		Prunella modularis : 0.1659	Periparus ater : 0.2232
	Carpodacus erythrinus : 0.1512		Phylloscopus collybita : 0.1536	Columba palumbus : 0.1958
			Columba palumbus : 0.1518	



True labels:	BirdNET-Lite predictions:	BirdNET-Analyzer predictions:	Global model predictions:	Local model predictions:
Fringilla coelebs	Ficedula parva : 0.3734	Turdus viscivorus : 0.2903	Fringilla coelebs : 0.4494	Fringilla coelebs : 0.8662
Turdus philomelos	Tetrao urogallus : 0.2905	Rhadina sibilatrix : 0.2169	Turdus philomelos : 0.2059	Turdus philomelos : 0.6634
	Turdus philomelos : 0.2263	Ficedula hypoleuca : 0.2008	Chloris chloris : 0.2046	Turdus merula : 0.305
	Rhadina sibilatrix : 0.2209	Turdus philomelos : 0.1581	Accipiter gentilis : 0.1713	
	Accipiter gentilis : 0.1691			
	Luscinia luscinia : 0.1682			
	Dryocopus martius : 0.1609			

FIGURE 4 Two examples of model predictions for 10-s field recordings from the test data. The species occurring in the recordings are listed on the left panel and highlighted in the spectrogram with manually drawn coloured boxes. All species that were predicted with a confidence higher than 0.15 are listed for all models. In the first example, all models give fairly high predictions for both foreground species *Cuculus canorus* (common cuckoo) and *Anthus trivialis* (tree pipit), but the localized model is generally more confident than other models. The localized model is the only model to recognize also the background species *Turdus philomelos* (song thrush), but with a significantly lower confidence. In the second example, both global and local models recognize both species, but the confidence of the predictions is significantly higher with the local model. BirdNET fails to recognize *Fringilla coelebs* (common chaffinch), probably because it is a very common and widespread species and the model has learned to ignore it in the background of other vocalizations.

included in training samples of different classes, the model learns to ignore them. Our results show that even a relatively small amount of data can be useful if the quality of the data and the labels is good enough.

The models presented in this paper are based on the idea of division between *global* and *local* data. Ideally, the *global* data should contain a comprehensive set of vocalizations for all species with a diverse variety of song types from different locations possibly recorded with different types of microphones, while *local* data would contain typical vocalizations from the target domain. The global data would be used to fit a reasonably good model, that would work as a good baseline and produce satisfactory results for all species, including rare species for which there are few or none fine-tuning samples available. However, the ideal conditions of global data were not fully met in our model. The geographical distribution of the global training data was very skewed, which means that the training set can not be considered as a comprehensive global set, but more as a restricted non-local data set instead. A more extensive global data set would most likely lead to improvement in both global and local models. Also, it is worth noting that the "global" model presented here is not completely global, since the noise data that is applied in the augmentation originates from the target domain and thus helps the model to adapt to local conditions even before actual local vocalizations are included.

While the global data are used to guarantee tolerable performance with rarely encountered species, the local data play a key role with the most common species. As our results from Section 3.1 suggest, if local data contain several hundreds of samples for each species, global pre-training is no longer needed. However, collecting this much data is often not feasible for rare species, and global pre-training is thus essential for these species. For common species, hundreds of local vocalizations can be obtained more easily and the role of global pre-training becomes less important. In this case the local data enable adjusting the model to perform as well as possible with vocalization types and species that it most often encounters.

There are several possible explanations for why the local data seem to be so useful. One very plausible reason could be that the data augmentation can not fully cover the domain shift between targeted training data recordings and omnidirectional test data soundscapes, and local data are thus needed to teach the model how distant vocalizations sound in reality. However, the fact that the localized fine-tuning is not beneficial with test data from a different country suggests that the main advantage of local fine-tuning is not due to adapting to untargeted recordings. Therefore, a more plausible explanation could be that local data contain a lot of information about the local sound environment including e.g. the background noise. In addition, the vocalizations of some species might vary across different areas and local data would thus enable the model to learn the local "dialect" of the birds.

In our view, the most interesting questions regarding future development concern expanding the model to an even broader set of species and utilizing recently invented machine learning techniques (Chen et al., 2020; Locatello et al., 2020; Wisdom et al., 2020) to acquire a more developed structure for the global (and local) model. In terms of applicability of the model to practical wildlife monitoring scenarios, an important question is, to how many classes could the model be extended. 101 species is enough to cover practically all species that can be expected to appear in a particular habitat in Southern Finland during the summer months, but for more biodiverse locations, such as rainforests, there might be a need for covering hundreds of species. In this paper our focus was mostly in the data and not on the variety of possible modelling approaches. However, the field of machine learning and neural networks is evolving extremely fast and there are several techniques that could be applied to improve the classification model itself. For example, self-supervised pre-training with contrastive learning (Al-Tahan & Mohsenzadeh, 2021; Chen et al., 2020) might enable training better base models with unlabeled data, which is much easier to acquire than labelled data. Sound source separation through mixture invariant training (Denton et al., 2021; Wisdom et al., 2020) or object detection (Locatello et al., 2020) also holds great potential, since current methods perform very well with targeted one-class samples. If the raw audio data could be split to several channels according to the source of vocalization, the classification task would become substantially easier.

5 | CONCLUSIONS

We fitted a neural network for classifying bird vocalizations with global data and fine-tuned it with local data to improve the model performance in specific conditions. The model was trained for 101 species with 1000 automatically selected training samples per class and fine-tuned with 0–1089 hand-selected local samples per class.

Our results suggest that using appropriate data augmentation techniques while training a CNN for bird sound recognition improves the results of the model and enables training with a small number of training samples.

Using additional fine-tuning data from the same domain where the model will eventually be used, such as recordings from the same location and same type of microphones, is very useful for improving the model performance. Fine-tuning with just 50–100 samples per class can significantly improve the results.

AUTHOR CONTRIBUTIONS

Patrik Lauha, Otso Ovaskainen and Panu Somervuo conceived the original idea; Patrik Lauha designed, trained and tested the models with guidance of Panu Somervuo; Otso Ovaskainen supervised the project and extracted the automatically selected global training samples from Macaulay data; Petteri Lehikoinen verified the hand selected global samples from Macaulay data and substantially advanced the labeling of Kerttu data; Sebastian Seibold, Lisa Geres and Tobias Richter collected and prepared the Berchtesgaden test data; Patrik Lauha wrote the first version of the manuscript. All authors contributed to the manuscript and gave final approval for the publication.

ACKNOWLEDGEMENTS

We thank Meeri Rannisto for her essential contribution in enabling the collection of annotated data, Macaulay Library and Matthew Medler for providing access to the recordings of the target species, Stefan Kahl for assistance on the use of BirdNET and fruitful conversations, as well as Tommi Mononen, Shayan Gharib and Graham Taylor for stimulating discussions that helped us to improve on this work. Otso Ovaskainen was funded by the Academy of Finland (grant no. 309581), Jane and Aatos Erkko Foundation, Research Council of Norway through its Centres of Excellence Funding Scheme (223257), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506; ERC-synergy project LIFEPLAN).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14003.

DATA AVAILABILITY STATEMENT

The recognition models described in this article as well as the Berchtesgaden test data are available in the Dryad Repository (Lauha et al. (2022), https://doi.org/10.5061/dryad.2bvq83btd). Kerttu data are available in the Zenodo Repository (Lehikoinen et al. (2022), https://doi.org/10.5281/zenodo.7030863).

ORCID

Patrik Lauha https://orcid.org/0000-0001-7204-8255
Panu Somervuo https://orcid.org/0000-0003-3121-4047
Petteri Lehikoinen https://orcid.org/0000-0002-2272-024X
Sebastian Seibold https://orcid.org/0000-0002-7968-4489
Otso Ovaskainen https://orcid.org/0000-0001-9750-4421

REFERENCES

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, *1*, e103. https://doi.org/10.7717/peerj.103

Al-Tahan, H., & Mohsenzadeh, Y. (2021). CLAR: contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics* (pp. 2530–2538). Proceedings of Machine Learning Research.

Anderson, S. E., Dave, A. S., & Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, *100*(2), 1209–1219. https://doi.org/10.1121/1.415968

Burns, F., Eaton, M. A., Burfield, I. J., Klvaňová, A., Šilarová, E., Staneva, A., & Gregory, R. D. (2021). Abundance decline in the avifauna of the European Union reveals cross-continental similarities in biodiversity change. *Ecology and Evolution*, *11*(23), 16647–16660. https://doi.org/10.1002/ece3.8282

Carignan, V., & Villard, M. A. (2002). Selecting indicator species to monitor ecological integrity: A review. *Environmental Monitoring and Assessment*, *78*(1), 45–61. https://doi.org/10.1023/A:1016136723584

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). Proceedings of Machine Learning Research.

Chollet, F. (2017). Deep learning with Python.

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., & Darrell, T. (2016). Best practices for fine-tuning visual classifiers to new domains. In *European conference on computer vision* (pp. 435–442). Springer. https://doi.org/10.1007/978-3-319-49409-8_34

Cornell Birdcall Identification Contest. (2020). https://www.kaggle.com/c/birdsong-recognition/

Darras, K., Batáry, P., Furnas, B., Celis-Murillo, A., Van Wilgenburg, S. L., Mulyani, Y. A., & Tscharntke, T. (2018). Comparing the sampling performance of sound recorders versus point counts in bird surveys: A meta-analysis. *Journal of Applied Ecology*, *55*(6), 2575–2586. https://doi.org/10.1111/1365-2664.13229

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Denton, T., Wisdom, S., & Hershey, J. R. (2021). Improving bird classification with unsupervised sound separation. arXiv preprint arXiv:2110.03209.

Fagerlund, S. (2007). Bird species recognition using support vector machines (2007). In *EURASIP Journal on Advances in Signal Processing* (pp. 1–8). Springer. https://doi.org/10.1155/2007/38637

Farina, A., Pieretti, N., & Piccioli, L. (2011). The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study. *Ecological Informatics*, *6*(1), 354–363. https://doi.org/10.1016/j.ecoinf.2011.07.004

Fraixedas, S., Lindén, A., Piha, M., Cabeza, M., Gregory, R., & Lehikoinen, A. (2020). A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions. *Ecological Indicators*, *118*, 106728. https://doi.org/10.1016/j.ecolind.2020.106728

Frommolt, K. H. (2017). Information obtained from long-term acoustic recordings: applying bioacoustic techniques for monitoring wetland birds during breeding season. *Journal of Ornithology*, *158*, 659–668. https://doi.org/10.1007/s10336-016-1426-3

Furnas, B. J., & Callas, R. L. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *The Journal of Wildlife Management*, *79*(2), 325–337. https://doi.org/10.1002/jwmg.821

Hansson-Sandsten, M. (2015). Classification of bird song syllables using singular vectors of the multitaper spectrogram. In *2015 23rd European Signal Processing Conference (EUSIPCO)* (pp. 554–558). IEEE. https://doi.org/10.1109/EUSIPCO.2015.7362444

Härmä, A. (2003). Automatic identification of bird species based on sinusoidal modeling of syllables. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP-88)* (Vol. *5*). IEEE. https://doi.org/10.1109/ICASSP.2003.1200027

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Champ, J., Planqué, R., Palazzo, S., & Müller, H. (2016). LifeCLEF 2016: Multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 286–310). Springer. https://doi.org/10.1007/978-3-319-44564-9_26

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Lombardo, J. C., Planqué, R., Palazzo, S. & Müller, H. (2017). LifeCLEF 2017 lab overview: Multimedia species identification challenges. In *Experimental IR Meets Multilinguality, Multimodality,*

and Interaction. (CLEF 2017, pp. 255–274). Springer. https://doi.org/10.1007/978-3-319-65813-1_24

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Planqué, R., Rauber, A., Fisher, R., & Müller, H. (2014). Lifeclef 2014: multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 229–249). Springer. https://doi.org/10.1007/978-3-319-11382-1_20

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Planqué, R., Rauber, A., Palazzo, S., Fisher, R. & Müller, H. (2015). LifeCLEF 2015: Multimedia life species identification challenges. In *Experimental IR Meets Multilinguality, Multi-modality, and Interaction*. (CLEF 2015, pp. 462–483). Springer. https://doi.org/10.1007/978-3-319-24027-5_46

Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., Deneu, B., Servajean, M., Durso, A., Bolon, I., Glotin, H., Planqué, R., de Castañeda, R. R., Vellinga, W. P., Klinck, H., Denton, T., Eggel, I., Bonnet, P., & Müller, H. (2021). Overview of LifeCLEF 2021: An evaluation of machine-learning based species identification and species distribution prediction. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 371–393). Springer. https://doi.org/10.1007/978-3-030-85251-1_24

Kahl, S. (2020). BirdNET-Lite. https://github.com/kahst/BirdNET-Lite

Kahl, S. (2022). BirdNET-Analyzer. https://github.com/kahst/BirdNET-Analyzer

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kogan, J. A., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: a comparative study. *The Journal of the Acoustical Society of America*, *103*(4), 2185–2196. https://doi.org/10.1121/1.421364

Lasseck, M. (2015). Improved Automatic Bird Identification through Decision Tree based Feature Selection and Bagging. In *CLEF (Working Notes)* (p. 1391). CLEF.

Lasseck, M. (2018). Acoustic bird detection with deep convolutional neural networks. In *DCASE* (pp. 143–147). DCASE.

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., & Ovaskainen, O. (2022). Data from: Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.2bvq83btd

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*, 101113. https://doi.org/10.1016/j.ecoinf.2020.101113

Lehikoinen, P., Rannisto, M., Camargo, U., Aintila, A., Lauha, P., Piirainen, E., Somervuo, P., & Ovaskainen, O. (2022). Data from: Crowdsourcing training material for automated bird sound classification - a pilot study. *Zenodo Repository*. https://doi.org/10.5281/zenodo.7030863

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, *33*, 11525–11538.

Macaulay Library. (2021). A scientific archive for research, education, and conservation. https://www.macaulaylibrary.org/

Matsubayashi, S., Suzuki, R., Saito, F., Murate, T., Masuda, T., Yamamoto, K., Kojima, R., Nakadai, K., & Okuno, H. G. (2017). Acoustic monitoring of the great reed warbler using multiple microphone arrays and robot audition. *Journal of Robotics and Mechatronics*, *29*(1), 224–235. https://doi.org/10.20965/jrm.2017.p0224

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. *8*). SciPy 2015.

McIlraith, A. L., & Card, H. C. (1997). Birdsong recognition using backpropagation and multivariate statistics. *IEEE Transactions on Signal Processing*, *45*(11), 2740–2748. https://doi.org/10.1109/78.650100

O'Connell, T. J., Jackson, L. E., & Brooks, R. P. (2000). Bird guilds as indicators of ecological condition in the central Appalachians. *Ecological Applications*, *10*(6), 1706–1721. https://doi.org/10.1890/1051-0761(2000)010[1706:BGAIOE]2.0.CO

Ovaskainen, O., Moliterno de Camargo, U., & Somervuo, P. (2018). Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecology Letters*, *21*(8), 1244–1254. https://doi.org/10.1111/ele.13092

Pittaras, N., Markatopoulou, F., Mezaris, V., & Patras, I. (2017). Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In *International conference on multimedia modeling* (pp. 102–114). Springer. https://doi.org/10.1007/978-3-319-51811-49

Priyadarshani, N., Marsland, S., & Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, *49*(5), jav.01447. https://doi.org/10.1111/jav.01447

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., & Marra, P. P. (2019). Decline of the North American avifauna. *Science*, *366*(6461), 120–124. https://doi.org/10.1126/science.aaw1313

Ruff, Z. J., Lesmeister, D. B., Duchac, L. S., Padmaraju, B. K., & Sullivan, C. M. (2020). Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation*, *6*(1), 79–92. https://doi.org/10.1002/rse2.125

Sanderson, F. J., Donald, P. F., Pain, D. J., Burfield, I. J., & Van Bommel, F. P. (2006). Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation*, *131*(1), 93–105. https://doi.org/10.1016/j.biocon.2006.02.008

Sethi, S. S., Ewers, R. M., Jones, N. S., Sleutel, J., Shabrani, A., Zulkifli, N., & Picinali, L. (2022). Soundscapes predict species occurrence in tropical forests. *Oikos*, *2022*(3), e08525. https://doi.org/10.1111/oik.08525

Sevilla, A., & Glotin, H. (2017). Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In *CLEF (Working Notes)* (p. 1866). CLEF.

Shonfield, J., & Bayne, E. M. (2017). Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, *12*(1), 14. https://doi.org/10.5751/ACE-00974-120114

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). IEEE.

The LIFEPLAN Project. (2022). https://www2.helsinki.fi/en/projects/lifeplan.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., & Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, *33*, 3846–3857.

Xu, Y., Si, Y., Wang, Y., Zhang, Y., Prins, H. H., Cao, L., & de Boer, W. F. (2019). Loss of functional connectivity in migration networks induces population decline in migratory birds. *Ecological Applications*, *29*(7), e01960. https://doi.org/10.1002/eap.1960

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, *27*, 3320–3328.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., & Ovaskainen, O. (2022). Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution*, *13*, 2799–2810. https://doi.org/10.1111/2041-210X.14003