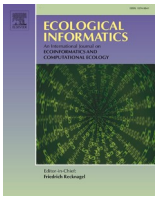
Ecological Informatics 61 (2021) 101236

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



BirdNET: A deep learning solution for avian diversity monitoring

Stefan Kahl [a,*], Connor M. Wood [a], Maximilian Eibl [b], Holger Klinck [a]

[a] Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, NY 14850, USA
[b] Technische Universität Chemnitz, D-09111 Chemnitz, Germany

ARTICLE INFO

Keywords:
Bioacoustics
Deep learning
Convolutional neural networks
Bird sound recognition
Avian diversity
Passive acoustic monitoring
Conservation

ABSTRACT

Variation in avian diversity in space and time is commonly used as a metric to assess environmental changes. Conventionally, such data were collected by expert observers, but passively collected acoustic data is rapidly emerging as an alternative survey technique. However, efficiently extracting accurate species richness data from large audio datasets has proven challenging. Recent advances in deep artificial neural networks (DNNs) have transformed the field of machine learning, frequently outperforming traditional signal processing techniques in the domain of acoustic event detection and classification. We developed a DNN, called BirdNET, capable of identifying 984 North American and European bird species by sound. Our task-specific model architecture was derived from the family of residual networks (ResNets), consisted of 157 layers with more than 27 million parameters, and was trained using extensive data pre-processing, augmentation, and mixup. We tested the model against three independent datasets: (a) 22,960 single-species recordings; (b) 286 h of fully annotated soundscape data collected by an array of autonomous recording units in a design analogous to what researchers might use to measure avian diversity in a field setting; and (c) 33,670 h of soundscape data from a single high-quality omnidirectional microphone deployed near four eBird hotspots frequented by expert birders. We found that domain-specific data augmentation is key to build models that are robust against high ambient noise levels and can cope with overlapping vocalizations. Task-specific model designs and training regimes for audio event recognition perform on-par with very complex architectures used in other domains (e.g., object detection in images). We also found that high temporal resolution of input spectrograms (short FFT window length) improves the classification performance for bird sounds. In summary, BirdNET achieved a mean average precision of 0.791 for single-species recordings, a F0.5 score of 0.414 for annotated soundscapes, and an average correlation of 0.251 with hotspot observation across 121 species and 4 years of audio data. By enabling the efficient extraction of the vocalizations of many hundreds of bird species from potentially vast amounts of audio data, BirdNET and similar tools have the potential to add tremendous value to existing and future passively collected audio datasets and may transform the field of avian ecology and conservation.

1. Introduction

Monitoring the status and trends of animal diversity and population levels of indicator species is critical to assess ecosystem health, to identify conservation priorities, and to guide decision making in conservation (Fitzpatrick and Rodewald, 2016; McComb et al., 2010). Birds are widely used as monitoring targets because they live in most environments and occupy almost every niche within those environments. Importantly, they are also conspicuous relative to other taxa that could be sensitive to similar ecological factors. As a result, birds are comparatively well studied, making many bird species' population trends the de facto baseline of ecosystem health. Intensive avian research efforts have also turned some avian species into model organisms, as the wealth of available data enables the development of novel quantitative methods that can then be applied beyond ornithology (Gutierrez, 2008). Many bird species also have broad cultural significance and popular appeal, making them excellent flagship species to communicate environmental issues (Cox and Gaston, 2016). In short, birds can serve as sentinel species, umbrella species, model organisms, and flagship species. Yet their utility in these roles is frequently predicated on adequate monitoring.

Traditionally, avian diversity has been monitored using point counts.

* Corresponding author.
E-mail addresses: sk2487@cornell.edu (S. Kahl), cmw289@cornell.edu (C.M. Wood), eibl@informatik.tu-chemnitz.de (M. Eibl), holger.klinck@cornell.edu (H. Klinck).

https://doi.org/10.1016/j.ecoinf.2021.101236
Received 13 November 2020; Received in revised form 18 December 2020; Accepted 18 December 2020
Available online 27 January 2021

This method requires a domain expert to visually and aurally identify and count birds in the field during 5–10-min intervals at sampling locations along a transect. Point count accuracy can be negatively affected by variability in detectability among species and sites, and when weather conditions suppress the birds' movement and calling activity (e.g., during rain, heavy fog, or windy conditions) (Ralph et al., 1995). Another disadvantage of this method is that monitoring large areas with high temporal resolution is typically not possible because of logistic and financial constraints. The results of these surveys might also be biased by the experience level of the human observers.

In contrast, passive acoustic monitoring (PAM) uses autonomous recording units (ARUs) to monitor the acoustic environment, often continuously, in the vicinity of the deployment location over extended periods (weeks to months). Data collection using PAM is cost-effective and has become a widely used tool in ecological research and monitoring over the past decade (Shonfield and Bayne, 2017; Wood et al., 2019b). Recent studies have shown that well- designed and executed PAM surveys produce diversity estimates mathcing those derived from traditional point count surveys (Darras et al., 2018). Moreover, the acoustic datasets are permanent records with much higher temporal resolution than point counts. This enables researchers to revisit the data to conduct additional analyses or manually verify automatically detected signals, thus reducing biases and uncertainties associated with the extracted observations. However, the analysis of extensive sound archives is still challenging. The application of automated analysis approaches has yielded limited success in the past. Thus, researchers have generally been forced to choose between extracting avian diversity data manually from a reduced number of recordings (e.g., (Furnas, 2020; Swiston and Mennill, 2009)) or to extract just a few species from vast quantities of data (e.g., (Potamitis et al., 2014)). New analytical approaches (e.g., machine learning algorithms) are needed to efficiently realize the full potential of passive acoustic data (Wood et al., 2019a).

Since their recent inception, deep artificial neural networks (DNNs) have transformed the field of machine learning. They frequently outperform traditional techniques in the domain of bioacoustic event detection and classification (Shiu et al., 2020; Salamon and Bello, 2017; Kahl et al., 2017a). This has also been exemplified by the outcomes of the annual LifeCLEF Bird Detection Challenge (BirdCLEF). BirdCLEF launched in 2014 as part of the Cross-Language Evaluation Forum (CLEF) and has since become one of the largest bird sound identification challenges in terms of dataset size and species diversity (up to 1500 bird species; (Kahl et al., 2019)). In early editions, mel-frequency cepstral coefficients (MFCC) were commonly used by the participating groups to reduce the dimensionality of feature vectors. These features were calculated for short time segments and fed into a classifier to identify avian calls in the data sets. Classifiers were based on support vector machines (Martinez et al., 2014; Leng et al., 2014; Joly et al., 2015), decision trees and random forests (Stowell and Plumbley, 2014; Lasseck, 2015; Stowell, 2015), or nearest neighbor clustering methods (Joly et al., 2014; Northcott, 2014). However, the large number of training and test recordings often forced participants to reduce the amount of data used to extract features due to computational constraints. The deficiencies of low-level feature classification restricted significant progress, and the need for more advanced classifiers became apparent.

In the BirdCLEF 2016 challenge, (Sprengel et al., 2016) introduced a novel convolutional neural network (CNN) classifier trained on spectrograms that instantly outperformed all competing systems by more than 10%. The approach included splitting up data into 5-s chunks, extracting mel-scale spectrograms for each chunk, and pre-filtering the audio data by an elaborate signal-to-noise estimation based on morphological operations optimized for spectrograms of bird sounds to select salient chunks for training. Further, data augmentation was applied to all training samples consisting of pitch and time shifts, and adding additional noise samples from rejected segments (i.e., training data that did not contain a bird sound). Since 2017, every team participating in the challenge presented a solution based on a CNN trained on spectrograms (Goëau et al., 2017, 2018). Interestingly, well-designed, shallow networks with only a few layers (Kahl et al., 2017b; Schlüter, 2018) performed comparably to very deep architectures (Sevilla and Glotin, 2017; Lasseck, 2018) that are commonly used in other domains like object recognition in images. It became apparent that complex classifiers and task-specific training regimes that focus on the overall characteristics of bird sounds are equally important to successfully classify many hundreds of incredibly diverse acoustic classes such as avian vocalizations.

Here we present a new algorithm, called BirdNET, that builds on these previous successes using CNNs and spectrogram data to classify 984 bird species. We discuss the necessary steps required to develop such a system and provide insights into our training and testing workflow of a deep artificial neural network. Finally, we give an outlook on how this technology can support ornithologists and conservation biologists in their work to identify and address the environmental challenges of our time.

2. Methods

Our overall workflow for training a deep neural network consisted of the acquisition of large amounts of audio data, the pre-processing of this data to generate visual representations of sound, the augmentation of these visualizations, and, finally, the training of a complex model architecture with ~27 million trainable parameters. We ran inference on independent validation and test splits of the acquired data and applied our model to continuous soundscape recordings, which are part of an ongoing monitoring project in Ithaca, NY, USA.

2.1. Data acquisition

We used three primary data sources to assemble the training data set. First, a list of the most common North American ($n = 595$) and European ($n = 555$) species was assembled using eBird (eBird, 2020) bar chart data and additional input from expert ornithologists. As several species (e.g., the house sparrow, *Passer domesticus*) occur on both continents, the total number of species initially considered for the model was 1049. Secondly, we collected data from Xeno-canto, a community-curated collection of recordings from around the world for these species. This collection features more than 500,000 recordings of over 10,000 bird species totaling over 7000 h of audio data (Xeno-canto, 2020)). Finally, we extracted additional recordings from the Macaulay Library of Natural Sounds, which is part of the Cornell Laboratory of Ornithology in Ithaca, NY, USA, and contains over 750,000 audio recordings of bird vocalizations covering more than 10,000 species (Macaulay, 2020). Every recording featured extensive metadata, including a quality score. We used this information to select high-quality recordings for training. However, because of the massive amounts of data, manual verification of the species labels was not practical. Therefore, there might have been incorrectly labeled data included in the training process, which could have negatively impacted the model performance. We retrieved a maximum of 500 recordings for each species from the pool of available data, resulting in a total number of 226,078 audio files. We eliminated species for which fewer than 10 recordings were available, which resulted in a final species list comprising a total of 984 bird species (out of the initially selected 1049 species).

We also included non-event classes to train the neural network to ignore non-bird signals. The most common sources of false-positive detections were other vocalizing animals (e.g., insects, anurans, mammals), geophysical noise (e.g., wind, rain, thunder), human vocal and non-vocal sounds (e.g., whistling, footsteps, speech), anthropogenic sounds typically encountered in urban areas (e.g., cars, airplanes, sirens), and electronic recorder noise. The Google AudioSet is one of the largest collections of human-labeled sounds that span a wide range of classes that are organized in an ontology (Gemmeke et al., 2017). We used 16 classes from the AudioSet and enriched them with recordings

from the Freefield1010 (Stowell and Plumbley, 2013) and WarblR datasets, which are both part of the DCASE bird detection challenge and did not contain bird sounds (Mesaros et al., 2018; Stowell et al., 2016). Non-bird animal sounds were downloaded from the Macaulay Library archive and smaller private collections. We combined non-event sounds into three classes: 'other animals' (~400 recordings), 'human' (~7800 recordings), and 'environmental noise' (~10,500 recordings). The final dataset featured almost 1000 different bird and non-bird classes totaling over 3978 h of recordings. This collection was split into training (80%), validation (10%), and testing (10%) datasets to ensure evaluation with mostly uncorrelated samples (i.e., recordings that were collected with different instruments and at different locations and times).

The omnidirectional recordings obtained during ARU surveys are often characterized by overlapping vocalizations of multiple bird species and, depending on the recording situation, low signal-to-noise ratios (SNR) because of increased distance between source and receiver as well as varying ambient sound levels. Using these soundscape recordings as training data would require expert birders to annotate a significant portion of the data before a model can be trained. However, it would be challenging to annotate enough data collected in many places to ensure proper representation of all possible bird calls and their variations (e.g., regional dialects). A rich and diverse set of training data is essential to generate a model that generalizes well enough to allow the application to new acoustic environments without the need for extensive adaption. Fortunately, a variety of community projects and professional audio collections provide a vast amount of open data on bird vocalizations. These recordings are typically obtained using (semi-) professional gear (handheld directional microphones) and are characterized by a high SNR. These so-called 'focal' recordings often contain only a single, clearly audible bird species and no overlapping vocalizations. However, it should be noted that the domain shift from focal to soundscape data presents a significant challenge. A well-performing model trained on focal recordings might not perform well when applied to much more complex soundscape recordings.

While the datasets we used for developing the model were solely comprised of focal recordings, we also assembled two long-term continuous sets (soundscape recordings) for further evaluation of the model. Both datasets were collected in the Sapsucker Woods Sanctuary in Ithaca, NY, USA. The first set contained 286 h of fully annotated soundscapes (almost 12 days) recorded by an array of 30 ARUs (Swift recorders, Cornell Lab of Ornithology) between March and July of 2017. Expert analysts annotated more than 80,000 vocalizations that covered 84 bird species. Annotations were merged into species lists for each 5-s chunks of audio, which we used as ground truth to assess the real-world applicability of our recognition system. The same data was used during the 2019 BirdCLEF challenge (Kahl et al., 2019). Additionally, we used 134,683 sound files of 15-min duration recorded between 2016 and 2019 as part of the Sapsucker Woods live audio stream project. These recordings were complemented with eBird 'hotspot' observations of the same area to assess the correlation between automated acoustic and traditional observer point counts. All of these data were collected at a sampling rate of 48 kHz and 16 bits resolution.

2.2. Data pre-processing

Following previous approaches, the input of our neural network is a spectrogram that we treat as a monochrome image. The computation of spectrograms from audio files has various degrees of freedom for frequency and magnitude scaling and the number and length of time steps. Recent studies have demonstrated that mel-spectrograms—typically used for human speech processing—perform well for acoustic event recognition (Kiyokawa et al., 2019; Delphin-Poulat and Plapous, 2019). However, we propose that avian vocal and auditory capabilities should be considered when tuning the spectrogram computation parameters instead of simply re-scaling arbitrary values to fit the input size of the model. Due to their superior temporal integration, birds outperform humans in their ability to distinguish the gap between two consecutive tones that differ in frequency (Dooling et al., 2000). This fact should be reflected in spectrogram computations (e.g., increased temporal resolution) to some extent since these acoustic features play an essential role in avian communication, including caller identification.

Preserving as much information as possible while maintaining a good overall compression in the final visual representa- tion of an acoustic signal is vital for further processing—especially when considering the computational costs associated with large input vectors. Acknowledging that a high temporal resolution of the spectrogram might be important, we selected a Fast Fourier Transform (FFT) window size of 10.7 ms (512 samples at 48 kHz sampling rate) and an overlap of 25%, each frame representing a time step of 8 ms. The frequency range of most bird vocalizations is limited between 250 Hz and 8.3 kHz (Hu and Cardoso, 2009, supplemental material). Therefore, we restricted the frequency range of the spectrogram to values between 150 Hz and 15 kHz covering the frequency range of the vast majority of bird vocalizations but also leaving room for pitch shifts during data augmentation. We performed frequency compression using a mel scale with 64 bands and a break frequency at 1750 Hz—considerably above the original proposal (Stevens et al., 1937) to achieve approximate linear scaling up to 500 Hz. According to the work of Schlüter (2018), using a nonlinear magnitude scale seems to be the most appropriate choice for bird call recognition in noisy environments even though this type of compression is neither adaptive nor widely used (Schlüter, 2018). The choice of duration of an audio signal visualized in a spectrogram has to reflect the expected duration of a bird vocalization to avoid cropping. We decided to use 3-s chunks of audio based on empirical evaluation of almost 80,000 human-annotated bird sounds, which revealed an average duration of 1.94 s across a large set of species (Kahl, 2019, p. 56). Additionally, temporal shifting during data augmentation is useful to diversify the training data, and the 3-s window length enables the implementation of this data augmentation method.

When training on weakly labeled samples of varying lengths, the biggest challenge is to extract segments of audio recordings that contain the target signal. Typically, the focal data (i.e., the recording) we used for training included a variable number of calls or songs of a single species. While the metadata provided information on which species was recorded, information on the exact location of the bird signals in the recording was not provided. Therefore, we developed a simple detector based on signal strength to extract segments containing vocalizations from the recordings. Since focal recordings are of high quality (high SNR), this method is very reliable. During the 2016 BirdCLEF edition, (Sprengel et al., 2016) established a method of signal-strength estimation to determine the presence of a bird sound in a spectrogram using morphological features to distinguish salient audio segments from chunks that only contain background sounds. Several approaches in the same domain have successfully adopted this process (Chou and To, 2018), including our BirdCLEF submissions (Kahl et al., 2017b, 2018).

2.3. Data augmentation

Domain-specific data augmentation is essential to account for unforeseen variations in real-world samples, especially when considering the shift in acoustic domains between focal and soundscape recordings. Our set of augmentations (Fig. 1) was derived from (Lasseck, 2019) and comprised current best practices from the acoustic domain. It included (a) random shifts in frequency and time (vertical and horizontal roll); (b) random partial stretching in time and frequency which are often used in human speech recognition (e.g., warping (Park et al., 2019)); and (c) the addition of noise from samples that were rejected during pre-processing (i.e., non-salient chunks of audio from the training data). Each method represents acoustic variations between training and test data, such as changes in the vocal output of birds depending on environmental factors, high levels of ambient noise in soundscape recordings, and lack of training sample diversity. All augmentation techniques were applied to

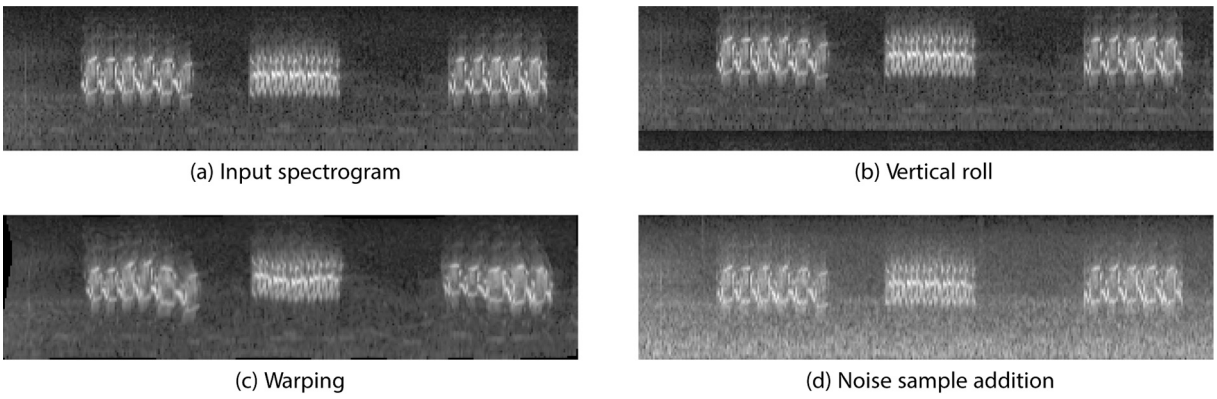


Fig. 1. Different domain-specific augmentation methods. Shifting bird song in frequency leads to improved scores across all evaluation scenarios (b). Warping has become popular in human speech recognition and can also be applied to bird sounds (c). The most powerful augmentation method was the randomly weighted addition of ambient noise extracted from audio chunks that do not contain a bird vocalization (d).

the original spectrogram during the training process with a probability of 0.5 and a maximum of three augmentations per sample.

2.4. Model architecture

Our model is based on a residual network or ResNet. ResNets are easy to implement and can be scaled in both width and depth to fit the needs of the training scenario best. (Zagoruyko and Komodakis, 2016) showed that wide residual networks provided similar performance compared to extremely deep architectures despite limited depth. This can be explained by improved regularization in the residual blocks and scaling of network in width (scaling factor K) and depth (scaling factor N) to identify the best possible layout. We followed this design and chose scaling factors of K = 4 and $N = 3$ to build our network, which resulted in a sequence of 157 total layers of which 36 are weighted. Three core components formed the succession of layers in our model. First, a pre-processing block transformed the original input spectrogram before it was passed through a series of residual stacks. Second, this sequence of residual stacks—consisting of downsampling and regular residual blocks—extracted features that were eventually passed through the third component, the classification block (see Table 1). The pre-processing block consisted of a single convolution with a slightly larger receptive field (with 5 × 5 kernels) than the original implementation, followed by 1 × 2 max pooling. Our design of residual blocks followed the original Wide ResNet design. Our downsampling blocks employed the changes suggested in (Xie et al., 2018). The third and final component, the classification block, was derived from (Schlüter, 2018) and resulted in probability predictions for all 987 classes per second (i.e., three predictions per 3-s input spectrogram) followed by global log-mean-exponential pooling and sigmoid activation. Most convolutions use 3 × 3 kernels with padding and are succeeded by batch normalization (Ioffe and Szegedy, 2015) and ReLU activation (if not noted differently). Because we treated spectrograms as monochrome input images, the input was a single channel mel spectrogram of size 64 × 384, visualizing three seconds of audio at a sampling rate of 48 kHz.

Table 1
Our model design follows the Wide ResNet approach by Zagoruyko and Komodakis and consists of 157 layers, of which 36 are weighted.

Group	Name	Input shape	Output shape
Pre-processing	5×5 Conv+BN+ReLU	(1×64×384)	(32×64×384)
	Ma× pooling	(32×64×384)	(32×64×192)
ResStack 1	Downsampling block	(32×64×192)	(64×32×96)
	2 × ResBlock	(64×32×96)	(64×32×96)
ResStack 2	Downsampling block	(64×32×96)	(128×16×48)
	2 × ResBlock	(128×16×48)	(128×16×48)
ResStack 3	Downsampling block	(128×16×48)	(256×8×24)
	2 × ResBlock	(256×8×24)	(256×8×24)
ResStack 4	Downsampling block	(256×8×24)	(512×4×12)
	2 × ResBlock	(512×4×12)	(512×4×12)
Classification	4×10 Conv+BN + ReLU + DO	(512×4×12)	(512×1×3)
	1×1 Conv+BN + ReLU + DO	(512×1×3)	(1024×1×3)
	1×1 Conv+BN + DO	(1024×1×3)	(987×1×3)
	Global LME pooling	(987×1×3)	(987×1)
	Sigmoid activation	(987×1)	(987×1)

With its ~27 million trainable parameters, the model has sufficient capacity to classify 984 bird species and three non-event classes. Like Schlüter (2018) proposed in his work, the classification block employs time-step predictions and log-mean-exponential pooling to preserve scores across multiple intervals.
BN = Batch normalization, DO = Dropout, LME = Log-mean-exponential.

2.5. Training

We trained our model using ~1.5 million spectrograms extracted from the training dataset with a maximum of 3500 samples (i.e., 3-s spectrograms) per class—using more samples did not improve the performance. Oversampling of underrepresented classes to at least 10% of the maximum amount of training samples per species was applied to reduce class imbalance. We tried other methods of cost-sensitive learning (e.g., class weights or focal loss mentioned in (Lin et al., 2017)), but none of them improved the overall model performance. To simulate simultaneously vocalizing bird species, we employed mixup training (Zhang et al., 2017) by randomly combining up to three spectrograms into one sample. We used the ADAM optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001 and a batch size of 32. We reduced the learning rate by a factor of 0.5 according to a step-wise schedule whenever the validation loss stalled. This also included a step-wise reduction of dropout probabilities by 0.1, starting at an initial probability of 0.5. Early stopping with a cooldown of 3 epochs prevented overfitting. We started with a maximum amount of 500 sample spectrograms per class and increased that amount by 1000 whenever training was finished—pre-trained snapshots were used to initialize the weights. Finally, an approach of knowledge distillation (Hinton et al., 2015) was used to train a born-again network (Furlanello et al., 2018) using the best previous snapshot as a teacher. Doing so helped to improve the overall performance even though the capacity of the model remained unchanged. We did not experiment with ensemble techniques due to the high computational demand and the respective lack of real-world applicability despite their popularity for this kind of task. Our overall process mostly followed current best practices in deep learning which are well summarized in (Chollet, 2017).

2.6. Inference

Our model accepts 3-s spectrograms as input, which requires converting input audio data to spectrograms during inference. Each audio

file has to be loaded, split, and re-sampled to 48 kHz if applicable. Each audio chunk is then converted to a mel scale spectrogram and passed through our model, which yields prediction probabilities for all 987 classes. Processing audio from different domains requires different techniques during inference. We employed two main methods at test time:

2.6.1. Focal recording analysis

Audio files from our Xeno-cano and Macaulay Library testing split were processed with a sliding-window approach with a 1-s overlap between consecutive spectrograms. We averaged all predictions for every species across the entire recording as part of a global pooling step. By doing so, we derived a ranked list of species for the entire recording (weak labels), reflecting the fact that some of the recordings contain labels for foreground and background species.

2.6.2. Soundscape analysis

Test recordings were split into 3-s chunks, and species probabilities were retrieved and stored individually for each segment. During inference, mainly three settings can be adjusted: the prediction sensitivity, overlap of segments, and the minimum required confidence. Sensitivity in our model can be altered by adjusting the sigmoid scaling to flatten the activation function—which requires less activation to cross a prediction threshold. We also experimented with different sharpness values in the log-mean-exponential pooling function (see (Schlüter, 2018) eq. 2). A higher sharpness acted more like max pooling and worked better in densely crowded acoustic scenes (e.g., dawn chorus). At the same time, smaller values acted more like average pooling, which resulted in better predictions for faint vocalizations. Overlapping segments increased the detection resolution, which also helped to cope with dense acoustic scenes. But a small step size for the sliding prediction widow significantly increased the processing time. Pooling scores across multiple segments can help to reduce the number of occasional false positives that were triggered by non-bird signals by averaging scores. Since we employed a ranking metric, no lower confidence threshold was set.

3. Results

For comparability reasons, we evaluated our system using sample-wise (mAP) and class-wise mean average precision (cmAP) metrics. Both metrics were used in past editions of BirdCLEF and worked well for multi-label classification scenarios (Kahl et al., 2019). As for the training dataset, test samples were unevenly distributed across classes. While sample-wise evaluation benefited from this imbalance by assigning common species a higher rank, it also reasonably represented real-world use cases as common species are much more likely to occur in a recording. A class-wise evaluation provided a balanced measure in which the true positive rate for each species contributed equally to the score. This better reflected the need for precise primary predictions. Additionally, we used several complementary metrics like top-1 accuracy, F0.5 score, and area under the ROC curve (AUC) to assess the suitability of our approach for a variety of application scenarios where precision might be more desirable than recall.

3.1. Focal recording evaluation

Our test data contained 22,960 focal recordings from both Xeno-canto and Macaluay Library with labels for 984 different foreground and background species. These recordings were independent of the training data, and our model scored a mAP of 0.791 and cmAP of 0.694 in a multi-label setting with pooled scores across each recording. Considering only primary predictions, our model achieved a top-1 accuracy of 0.777 with an AUC of 0.974 across all test files. Pooling single predictions over time increased the confidence score despite moments of silence between vocalizations. Our system showed a mean confidence value of 0.627 for all primary detections. However, soundscapes posed a significant challenge due to the discussed domain shift. We observed that our model performed well on focal recordings that closely resembled the training data, despite being independent in location and time.

3.2. Soundscape recording evaluation

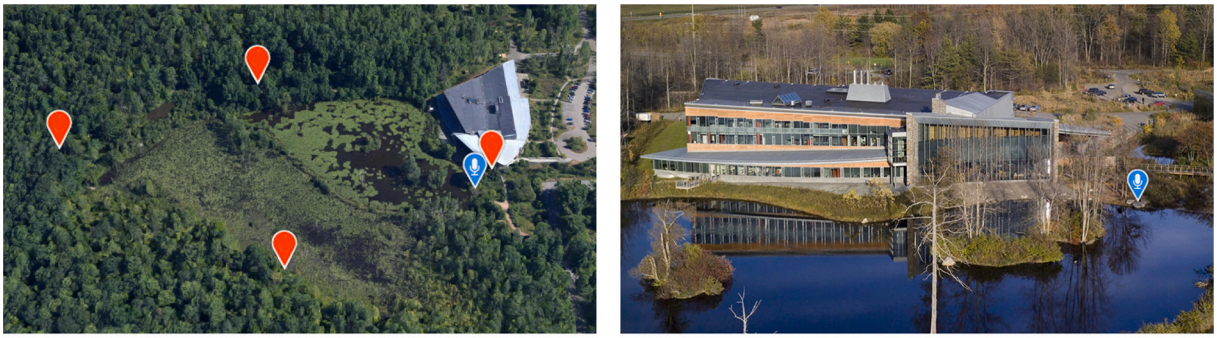
The observed performance drastically decreased in the soundscape domain. Our model was optimized to yield the best possible F0.5 score for which we assumed better representation of real-world applicability than mAP and cmAP due to the increased focus on precision over recall. Our detection system achieved a F0.5 score of 0.414 on the 2019 BirdCLEF test dataset spanning 12 fully annotated days of soundscape recordings. Considering only dawn chorus soundscapes—chosen as a subset which only contained recordings that were made one hour before and one hour after sunrise, our model maintained a F0.5 measure of 0.395 despite the vast increase in vocal activity. This also indicated that the detection system is agnostic to non-bird events, which usually occur during the quiet times of the day (e.g., during the night). However, an overall AUC of 0.596 also illustrated that post-processing confidence values could not easily reduce false positives. Other techniques like post-filtering species based on eBird bar chart data or other location-based metadata might yield better results.

The soundscape data used for evaluating the model performance equaled the 2019 BirdCLEF hidden test dataset. Lasseck (2019) submitted the best performing single model which achieved a F0.5 score of 0.260 (Lasseck, 2019). The performance of this model improved significantly when trainined on soundscape validation data provided by the organizers—a 3-day subset of the hidden test set recorded at the same location. Single-model performance improved to a F0.5 score of 0.416 partly due to overfitting to the test data that was well-represented by the validation recordings. Our model was not trained on soundscape data and significantly outperformed the best single model that was not trained on the provided validation samples. It also matched the best single-model performance (trained on validation data) without the need for extensive on-site training data. However, training data differed in both cases. The BirdCLEF 2019 training set contained significantly less species (659) and fewer focal recordings per species (max. 100 per species, 50,153 total). Yet, we concluded that the performance of our model provides competitive scores for focal and soundscape audio data.

3.3. Evaluation on continuous stream data

Our soundscape test dataset is one of the largest fully annotated datasets for bird sound recognition to date—containing significantly more labeled audio data than other recently published datasets (e.g., (Lostanlen et al., 2018; Morfi et al., 2019)). However, we aimed to investigate whether the performance of our model would suffice to detect seasonal changes in avian diversity and match human observer performance even though the achieved soundscape scores (see Paragraph 3.2) implied shortcomings of our approach. Our fully annotated soundscape test dataset only covered 12 days randomly selected from a 5-month period of a single year. Therefore, we analyzed four continuous years of audio data recorded with a single, omnidirectional outdoor microphone located just outside the Cornell Lab of Ornithology (Fig. 2). For validation, we compared our results with observations collected by eBird users (often experts from the lab) at certain hotspots throughout the area. We pooled the data from four point count locations in the proximity of the microphone (~300 m radius) and compared our detections with the occurrence frequency on eBird checklist data.

Recordings for this experiment were stored as flac-encoded audio files with a duration of 15 min each. In total, we analyzed 134,683 files (~33,670 h), which covered 96.1% of the observed period (01/2016–12/2019). We achieved 60× real-time analysis speed on a consumer-like workstation by using non-overlapping analysis windows of 3-s duration. The model sensitivity was empirically set to 0.85 (instead of 1.0), which flattened the sigmoid activation curve in the last



(a) Hotspots and mic location (Google Maps) (b) Lab building and mic location (Robert Barker)

Fig. 2. eBird observational data is usually collected at so-called hotspots (red) where expert birders and citizen scientists regularly conduct point counts. The Cornell Lab of Ornithology is surrounded by dense vegetation and a large pond. We considered four hotspots to be in the acoustic range of our live stream microphone (blue) for our evaluation.

layer and required less activation to achieve a score above the detection threshold. Additionally, differences in scores became more meaningful due to the less 'binary' characteristics of the sigmoid slope. For the analysis in this paper, we only considered primary detections (rank one of all probabilities) with a confidence score of at least 0.5 as valid and accumulated all of them for each week for every species. Even this relatively restrictive approach led to results that match human observer performance without accounting for overlapping vocalizations during a busy dawn chorus.

The eBird checklist frequency represents the relative weekly occurrence of a particular bird species across all submitted checklists at a site. By combining four observation locations, the minimum number of checklists for a given week was 7, the maximum 83. For this analysis, we focused on species with at least four checklist entries with a frequency above 0.05 between 2016 and 2019. The total number of species was 121 and included migratory and non-migratory birds across many genera. We determined the correlation coefficient between the weekly checklist frequency and the total number of detections for that period. Across all 121 species, we achieved an average correlation (Pearson's r) of 0.251 with a correlation of 0.5 and more for many migratory species (see Fig. 3). We were able to reproduce the seasonal occurrence pattern by using acoustics only. A low correlation coefficient mostly resulted from an offset between observed occurrences and measured vocal activity peaks or as a result of a higher sampling resolution when human observations were sparse (e.g., brown thrasher, *Toxostoma rufum*). Hotspot observations were based on visual and aural cues and—most importantly—were not continuous. The obtained results indicated that the inferred occurrence patterns correlate with those observed by humans for many migratory species. The correlation of occurrence patterns for non-migratory species—especially for those species known to be regularly feeder visitors—is often considerably lower. The Cornell Lab of Ornithology's feeder garden is located only a few yards away from our recording location, which benefits the detection process as birds are known to interact and vocalize around feeders. However, it also biases the results towards these local feeder species. For example, we observed higher levels of vocal activity during the winter months for some of these species (e.g., black-capped chickadees, *Poecile atricapillus*), which does not necessarily represent the expected vocal output during this time of the year. (See Fig. 4.)

4. Discussion

Our results showed that the trained model was capable of replicating patterns generated by human observer data despite the high number of classes and the acoustic domain mismatch between training and test recordings (focal vs. soundscape data). As expected, the quality of the soundscape recordings affected overall detection performance, with lower SNRs of received calls causing a reduction in performance. Using high-end recording devices (e.g., the outdoor microphone used for our study described in Section 3.3) appeared to yield more reliable results than lower cost devices typically used for acoustic surveys (e.g., ARUs like the SWIFT recorder). This implies that the gap in acoustic domains still poses a significant challenge despite sophisticated data augmentation. We observed that our model performed well for most species. We investigated the deficiencies in more detail and assessed the detection performance and its correlation with repertoire size, training sample quality, and quantity. It became apparent that species repertoire size alone is not a conclusive predictor of recognition performance. For extremely versatile species with vast repertoires, confusion with other species could potentially affect the scores. However, when good quality training data is available, repertoire size does not affect the classification scores—our model achieves high scores for common nightingale (*Luscinia megarhynchos*) and brown thrasher (*Toxostoma rufum*). The poor results for species known to incorporate hetero-specific material into their vocalizations like the European starling (*Sturnus vulgaris*) and northern mockingbird (*Mimus polyglottos*) implied that imitation is a significant challenge for automated recognition systems and scores remain low despite good quality training data. Additionally, noisy training data (i.e., low SNR) significantly affected overall performance and mostly led to decreased performance for any species. Due to weakly labeled samples, non-bird events might find their way into the training data, which has a considerable impact on species with little training data. It may have also affected the evaluation process due to the lack of a true gold standard for individual time segments.

The results of our study appeared to be competitive compared to other attempts in the same domain, especially considering the high number of covered species. We achieved a F0.5 score of 0.414 and a mAP of 0.791 across 984 species. In comparison,(Ruff et al., 2020) achieved an average F0.5 score of 0.500 across six owl species, (LeBien et al., 2020) reported a mAP of 0.893 across 24 species of birds and frogs, and (Cramer et al., 2020) could detect flight calls with a micro-average accuracy of 0.663 across 14 species. However, the lack of a standardized benchmark dataset, standardized metrics, and true gold standard for evaluation makes a direct comparison between approaches difficult. Deep artificial neural networks work particularly well for the task of acoustic event recognition and mostly generalize well. Yet, designing such networks is still often guided by intuition, and the training process is computationally costly for most phases. During our work in this domain, we also noticed that:

- High temporal resolution of input spectrograms (short FFT window length) improved the classification performance for bird sounds.
- Multi-label classification with mixup training increased the overall performance across all tasks.
- Deeper topologies (more layers) did not necessarily perform better than wider topologies (more filters).

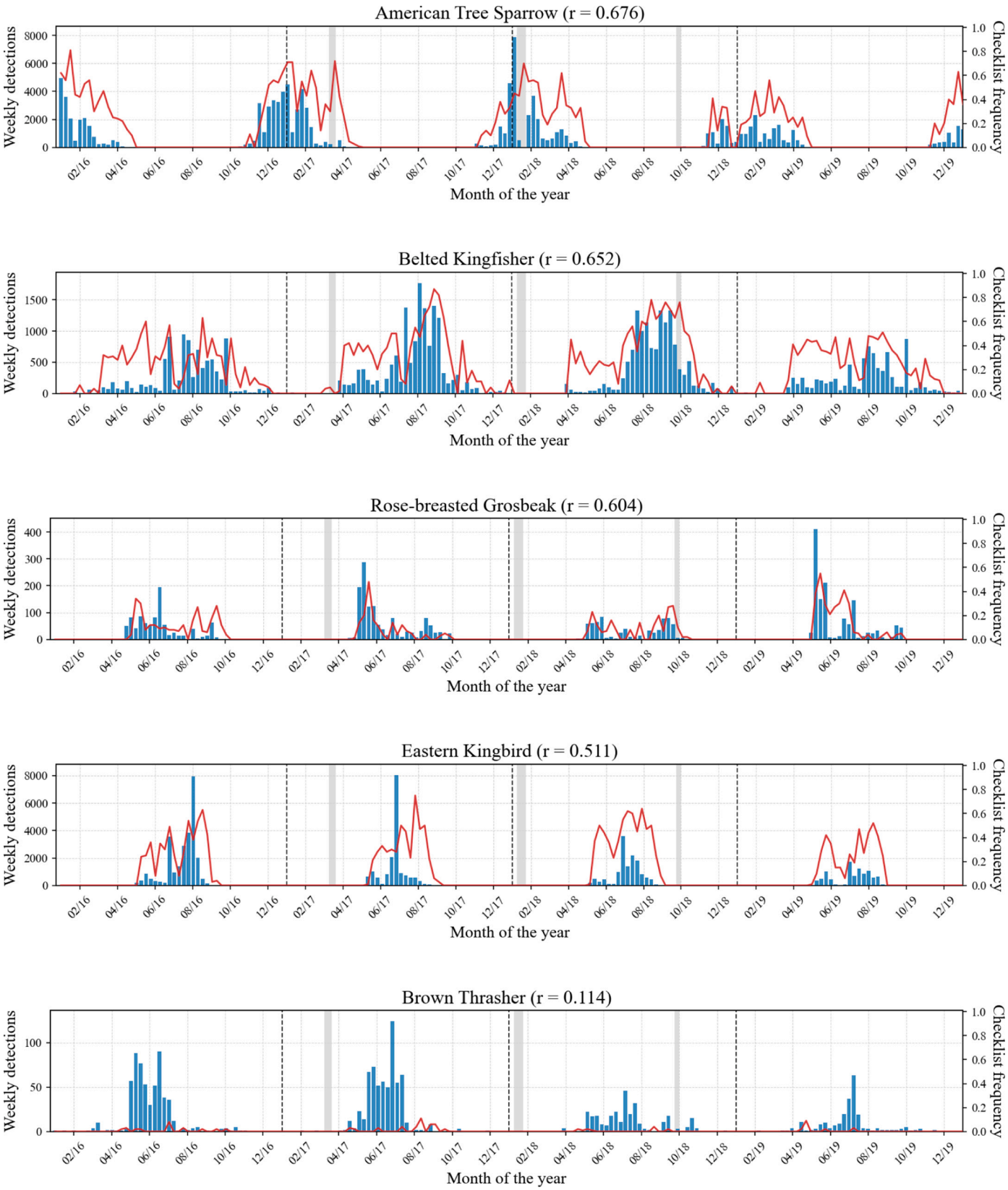


Fig. 3. Migratory species occurrence correlation (r) between weekly cumulative BirdNET detections (blue) and human point count observations (eBird checklist frequency, red). Gray areas indicate partially missing data. The detections closely resemble human observational performance. We achieved a high correlation for migratory species that vocalize frequently (i.e., multiple hundreds of detections per week). Even with low correlation, we showed that continuous detections may provide a better representation of seasonal changes in vocal activity when compared to sparse observer data (e.g., brown thrasher, *Toxostoma rufum*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

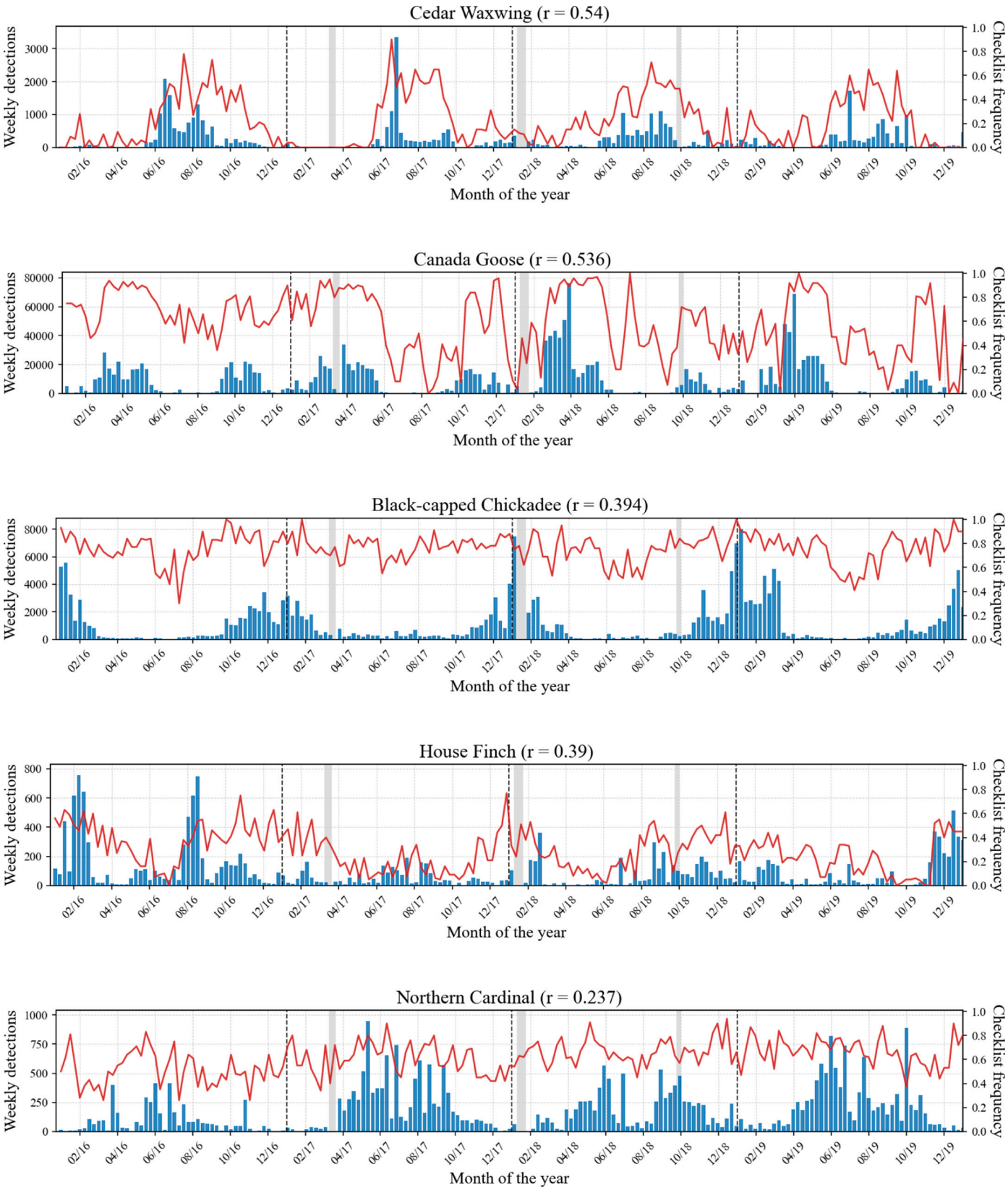


Fig. 4. Non-migratory species occurrence correlation (r) between weekly cumulative detections (blue) and human point count observations (eBird checklist frequency, red). Gray areas indicate partially missing data. For some species, our detections closely resembled human observer data. In other cases, our detections indicated varying vocal activity patterns despite consistently high observation counts. This indicated that certain survey design choices (and seasonal vocal behaviors), such as the placement of the microphone next to the Cornell Lab of Ornithology feeder garden, could be a source of bias (e.g., inflated estimates of regular feeder visitors like black-capped chickadees, *Poecile atricapillus*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- Deeper topologies outperformed shallow layouts when computational resources are limited (e.g., embedded systems).
- Except for oversampling, cost-sensitive learning did not improve the overall scores despite significant class imbalance.
- Task-specific model designs and training regimes for audio event recognition perform on-par with very complex architectures used in other domains (e.g., object detection in images).

Adjusting the detection system according to the task at hand usually leads to better performance—even with a fully trained and fixed model. Analyses of soundscape data will likely be improved by manually annotating a subset of the audio data to verify automated detections and to find the best possible parameter settings for a given application. Such an approach would still require considerably less manual labor than other techniques, and future developments in the domain of deep learning for bioacoustics might even render this task obsolete (e.g., by self-supervised or few-shot learning).

5. Conclusion

Decreasing hardware costs is likely to make PAM an increasingly popular approach in ecology. Many current applications focus on one or a few species (e.g., (Wood et al., 2020)), but the raw audio typically contains vocal signatures of many other species (Wood et al., 2019a), thus enabling retrospective community-level studies. The comparable performance of ARUs relative to point counts (Darras et al., 2018) suggests that PAM can indeed be an effective method of rapidly and efficiently surveying avian diversity. We have shown that BirdNET can identify vastly more bird species than any other similar programs, and it can do so without sacrificing processing efficiency or performance. The continued refinement of BirdNET and other programs capable of rapidly identifying entire avian communities will eventually enable avian ecology and conservation research at scales—in space, time, and diversity —previously unattainable.

Availability

The final detection system is available online at *https://github.com/kahst/BirdNET* and can be used to run model inference on any number of recordings. A demo of the live stream analysis can be accessed at *https://birdnet.cornell.edu/live/* and visualizes the results of the real-time detection process. More demos and in- formation on the project can be found on our project page at *https://birdnet.cornell.edu/* which also lists all supported species.

Declaration of Competing Interest

None.

Acknowledgements

This project is supported by Jake Holshuh (Cornell class of ‘69). The Arthur Vining Davis Foundations also kindly support our efforts. The European Union and the European Social Fund for Germany partially funded this research. This work was also partially funded by the German Federal Ministry of Education and Research in the program of Entrepreneurial Regions InnoProfileTransfer in the project group localizeIT (funding code 03IPT608X). We want to thank all expert birders who helped to annotate soundscapes with incredible effort: Cullen Hanks, Jay McGowan, Matt Young, Randy Little, and Sarah Dzielski.

References

Chollet, F., 2017. Deep Learning with Python. Manning Publications Company.
Chou, J., To, C.-H., 2018. Cocktail party problem for bird sounds. In: CS230. Stanford University.
Cox, D.T., Gaston, K.J., 2016. Urban bird feeding: connecting people with nature. PLoS One 11 (7), e0158717.
Cramer, J., Lostanlen, V., Farnsworth, A., Salamon, J., Bello, J.P., 2020. Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In: ICASSP 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 901–905 (IEEE).
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. J. Appl. Ecol. 55 (6), 2575–2586.
Delphin-Poulat, L., Plapous, C., 2019. Mean Teacher with Data Augmentation for DCASE 2019 Task 4. Orange Labs Lannion, France, Tech. Rep.
Dooling, R.J., Lohr, B., Dent, M.L., 2000. Hearing in Birds and Reptiles. Springer.
eBird, 2020. Engaging Birders in Science and Conservation (https://ebird.org/about. Accessed: 2020-09-20).
Fitzpatrick, J.W., Rodewald, A.D., 2016. *Handbook of Bird Biology*, Chapter Bird Conservation. John Wiley & Sons.
Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A., 2018. Born again neural networks. arXiv Preprint. arXiv:1805.04770.
Furnas, B.J., 2020. Rapid and varied responses of songbirds to climate change in California coniferous forests. Biol. Conserv. 241, 108347.
Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. AudioSet: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780 (IEEE).
Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R., Joly, A., 2017. LifeCLEF bird identification task 2017. In: CLEF (Working Notes), 1866.
Goëau, H., Kahl, S., Glotin, H., Vellinga, W.-P., Planqué, R., Joly, A., 2018. Overview of BirdCLEF 2018: Monospecies vs. soundscape bird identification. In: CLEF (Working Notes), 2125.
Gutierrez, R.J., 2008. Spotted owl research: a quarter century of contributions to education, ornithology, ecology, and wildlife management. Condor 110 (4), 792–798.
Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv Preprint. *arXiv:1503.02531*.
Hu, Y., Cardoso, G.C., 2009. Are bird species that vocalize at higher frequencies preadapted to inhabit noisy urban areas? Behav. Ecol. 20 (6), 1268–1273.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv Preprint. arXiv:1502.03167.
Joly, A., Champ, J., Buisson, O., 2014. Instance-based bird species identification with undiscriminant features pruning. In: CLEF (Working Notes), 1180.
Joly, A., Leveau, V., Champ, J., Buisson, O., 2015. Shared nearest neighbors match kernel for bird songs identification-LifeCLEF 2015 challenge. In: CLEF (Working Notes), 1391.
Kahl, S., 2019. Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring. PhD thesis. Chemnitz University of Technology.
Kahl, S., Hussein, H., Fabian, E., Schloßhauer, J., Thangaraju, E., Kowerko, D., Eibl, M., 2017a. Acoustic event classification using convolutional neural networks. INFORMATIK 2017.
Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M., 2017b. Large-scale bird sound classification using convolutional neural networks. In: CLEF (Working Notes), 1866.
Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M., 2018. A baseline for large-scale bird species identification in field recordings. In: CLEF (Working Notes), 2125.
Kahl, S., Stöter, F.-R., Goëau, H., Glotin, H., Planque, R., Vellinga, W.-P., Joly, A., 2019. Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes. In: CLEF (Working Notes), 2380.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv Preprint (arXiv:1412.6980).
Kiyokawa, Y., Mishima, S., Toizumi, T., Sagi, K., Kondo, R., Nomura, T., 2019. Sound event detection with ResNet and self-mask module for DCASE 2019 task 4. Tech. Rep.
Lasseck, M., 2015. Improved automatic bird identification through decision tree based feature selection and bagging. In: CLEF (Working Notes), 1391.
Lasseck, M., 2018. Audio-based bird species identification with deep convolutional neural networks. In: CLEF (Working Notes), 2125.
Lasseck, M., 2019. Bird species identification in soundscapes. In: CLEF (Working Notes), 2380.
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. Ecol. Inform. 101113.
Leng, Y.R., Dennis, J.W., Dat, T.H., 2014. Bird classification using ensemble classifiers. In: CLEF (Working Notes), 1180.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.
Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, J.P., 2018. Birdvox-full-night: A dataset and benchmark for avian flight call detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 266–270 (IEEE).
Macaulay, 2020. The World’s Premier Scientific Archive of Natural History Audio, Video, and Photographs. https://www.macaulaylibrary.org/about/history/. (Accessed 20 September 2020).
Martinez, R., Silva, L., Olvera, T.E.V., Fuentes, G., Ruíz, I.V.M., 2014. SVM candidates and sparse representation for bird identification. In: CLEF (Working Notes), 1180.

McComb, B., Zuckerberg, B., Vesely, D., Jordan, C., 2010. Monitoring Animal Populations and their Habitats: A practitioner's Guide. CRC Press.
Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., Plumbley, M. D., 2018. Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (2), 379–393.
Morfi, V., Bas, Y., Pamuła, H., Glotin, H., Stowell, D., 2019. Nips4bplus: a richly annotated birdsong audio dataset. PeerJ Comp. Sci. 5, e223.
Northcott, J., 2014. Participation of group SCS to LifeCLEF bird identification challenge 2014. In: CLEF (Working Notes), 1180.
Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv Preprint. arXiv:1904.08779.
Potamitis, I., Ntalampiras, S., Jahn, O., Riede, K., 2014. Automatic bird sound detection in long real-field recordings: applications and tools. Appl. Acoust. 80, 1–9.
Ralph, C.J., Droege, S., Sauer, J.R., 1995. *Monitoring Bird Populations by Point Counts*, Volume 149, Chapter Managing and Monitoring Birds Using Point Counts: Standards and Applications, Pages 161–168. US Department of Agriculture, Forest Service, Pacific Southwest Research Station.
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 Sens. Ecol. Conserv. 6 (1), 79–92.
Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. 24 (3), 279–283.
Schlüter, J., 2018. Bird identification from timestamped, geotagged audio recordings. In: CLEF (Working Notes), 2125.
Sevilla, A., Glotin, H., 2017. Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: CLEF (Working Notes), 1866.
Shiu, Y., Palmer, K., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10 (1), 1–12.
Shonfield, J., Bayne, E., 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv. Ecol. 12 (1).
Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T., 2016. Audio based bird species identification using deep learning techniques. In: CLEF (Working Notes), 1609.
Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8 (3), 185–190.
Stowell, D., 2015. BirdCLEF 2015 submission: unsupervised feature learning from audio. In: CLEF (Working Notes), 1391.
Stowell, D., Plumbley, M.D., 2013. An open dataset for research on audio field recording archives: freefield1010. arXiv Preprint. arXiv:1309.5275.
Stowell, D., Plumbley, M.D., 2014. Audio-only bird classification using unsupervised feature learning. In: CLEF (Working Notes), 1180.
Stowell, D., Wood, M., Stylianou, Y., Glotin, H., 2016. Bird detection in audio: A survey and a challenge. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6 (IEEE).
Swiston, K.A., Mennill, D.J., 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. J. Field Ornithol. 80 (1), 42–50.
Wood, C.M., Gutiérrez, R.J., Peery, M.Z., 2019a. Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. Ecology 100 (9), e02764.
Wood, C.M., Popescu, V.D., Klinck, H., Keane, J.J., Gutiérrez, R., Sawyer, S.C., Peery, M. Z., 2019b. Detecting small changes in populations at landscape scales: A bioacoustic site-occupancy framework. Ecol. Indic. 98, 492–507.
Wood, C.M., Klinck, H., Gustafson, M., Keane, J.J., Sawyer, S.C., Gutiérrez, R., Peery, M. Z., 2020. Using the ecological significance of animal vocalizations to improve inference in acoustic monitoring programs. Conserv. Biol.
Xeno-canto, 2020. Sharing Bird Sounds from Around the World. https://www.xeno-canto.org/about/xeno-canto (Accessed: 2020-09-20).
Xie, J., He, T., Zhang, Z., Zhang, H., Zhang, Z., Li, M., 2018. Bag of tricks for image classification with convolutional neural networks. arXiv Preprint. *arXiv:1812.01187*.
Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv Preprint. *arXiv:1605.07146*.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv Preprint. arXiv:1710.09412.