Remote Sensing in Ecology and Conservation
Open Access
ZSL LET'S WORK FOR WILDLIFE

ORIGINAL RESEARCH

Automated identification of avian vocalizations with deep convolutional neural networks

Zachary J. Ruff[1], Damon B. Lesmeister[1,2], Leila S. Duchac[1,2,3], Bharath K. Padmaraju[4] & Christopher M. Sullivan[4]

[1]Pacific Northwest Research Station, USDA Forest Service, Corvallis, Oregon
[2]Department of Fisheries and Wildlife, Oregon State University, Corvallis, Oregon
[3]Oregon Cooperative Fish and Wildlife Research Unit, Corvallis, Oregon
[4]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon

Keywords
Acoustic monitoring, avian vocalization, Bioacoustics, machine learning, neural networks, spotted owls

Correspondence
Zachary J. Ruff, Pacific Northwest Research Station, USDA Forest Service, Corvallis, OR.
Tel: 541 750 7286; Fax: 541 750 7329;
E-mail: zjruff@gmail.com

Editor: Nathalie Pettorelli
Associate Editor: Vincent Lecours

Received: 21 May 2019; Revised: 16 July 2019; Accepted: 2 August 2019

doi: 10.1002/rse2.125

Abstract

Passive acoustic monitoring is an emerging approach to wildlife monitoring that leverages recent improvements in automated recording units and other technologies. A central challenge of this approach is the task of locating and identifying target species vocalizations in large volumes of audio data. To address this issue, we developed an efficient data processing pipeline using a deep convolutional neural network (CNN) to automate the detection of owl vocalizations in spectrograms generated from unprocessed field recordings. While the project was initially focused on spotted and barred owls, we also trained the network to recognize northern saw-whet owl, great horned owl, northern pygmy-owl, and western screech-owl. Although classification performance varies across species, initial results are promising. Recall, or the proportion of calls in the dataset that are detected and correctly identified, ranged from 63.1% for barred owl to 91.5% for spotted owl based on raw network output. Precision, the rate of true positives among apparent detections, ranged from 0.4% for spotted owl to 77.1% for northern saw-whet owl based on raw output. In limited tests, the CNN performed as well as or better than human technicians at detecting owl calls. Our model output is suitable for developing species encounter histories for occupancy models and other analyses. We believe our approach is sufficiently general to support long-term, large-scale monitoring of a broad range of species beyond our target species list, including birds, mammals, and others.

Introduction

Passive acoustic monitoring is an emerging alternative to traditional surveys for wildlife monitoring. Modern autonomous recording units (ARUs) can record continuously for days or weeks at a time, generating large amounts of audio data. Any species that makes characteristic sounds may be a good candidate for acoustic monitoring, and this approach has been successfully applied in studies of insects (Ganchev and Potamitis 2007), amphibians (Alonso et al. 2017), bats (Russo and Jones 2003), cetaceans (Luo et al. 2019), elephants (Wrege et al. 2017), primates (Heinicke et al. 2015), and various avian species (Figueira et al. 2015; Campos-Cerqueira and Aide 2016; Shonfield et al. 2018; Wood et al. 2019).

Researchers are typically interested in isolating a particular signal within the data, such as the vocalizations of some target species. Locating and identifying these signals within a large body of field recordings is a necessary first step in any analysis. Previous work has explored various methods for automating the detection of target signals, including hidden Markov models (Trifa et al. 2008), template matching with dynamic time warping (Brown and Miller 2007; Somervuo 2018), and artificial neural networks (Wood et al. 2019). Here we demonstrate the use of a deep convolutional neural network (CNN) for automating the detection of owl vocalizations in spectrograms generated from field recordings.

Our work follows from a recent effort to evaluate the effectiveness of passive bioacoustics for monitoring

northern spotted owls *Strix occidentalis caurina* (hereafter 'spotted owl') and for studying their competitive interactions with barred owls *S. varia*. The spotted owl was listed in 1990 as threatened under the US Endangered Species Act (US Fish and Wildlife Service 1990), and monitoring of populations as directed by the Northwest Forest Plan (US Department of Agriculture and US Department of Interior 1994) has revealed ongoing population declines due to a range of environmental stressors (Dugger et al. 2016, Lesmeister et al. 2018). The barred owl is native to eastern North America but has become established throughout the Pacific Northwest since the 1970s (Mazur and James 2000); this range expansion has brought barred owls into competition with spotted owls for territory and food resources (Gutiérrez et al. 2007). Being larger, more aggressive, and more generalist in its prey and cover selection, the barred owl has become a major contributor to the decline of the spotted owl (Wiens et al. 2014, Dugger et al. 2016, Lesmeister et al. 2018).

Following a successful 2017 study employing ARUs to monitor spotted and barred owls at 30 field sites (L. Duchac, unpublished data), the Northwest Forest Plan monitoring program approved the expansion of ARU deployments to 208 field sites in 2018. While in the previous study we searched recordings semi-manually for target species vocalizations, we felt that an automated approach would scale up better to accommodate the increasing pace of the data collection and would eventually require less human supervision.

CNNs have undergone rapid development in recent years (e.g., Kahl et al. 2017), initially spurred by the performance of 'AlexNet' (Krizhevsky et al. 2012) in the ImageNet Large Scale Visual Recognition Challenge competition (http://www.image-net.org/challenges/LSVRC/). Similar networks continue to define the state of the art in computer vision and image classification, with commercial applications in areas such as facial recognition (Taigman et al. 2014) and autonomous vehicles (Nvidia 2019). The suitability of CNNs for image classification stems from their structure, conceptualized as a stack of layers in which the output (or activation) of each layer is passed as input to the following layer. Activations in higher layers can represent increasingly complex features of the original input, enabling the network to parse an image as a composition of meaningful elements rather than a collection of unrelated pixels (LeCun 2015). Another appealing aspect of CNNs is that the visual features used to discriminate between image classes need not be explicitly programmed. Rather, the network learns these features automatically from labeled examples through a supervised training process. Thus researchers can bypass a great deal of tedious and error-prone coding, provided sufficient training data are available. The availability of large pre-labeled training datasets has helped drive the refinement of such models, as have improved methods for training deep neural networks on graphics processing units.

CNNs fulfill all the basic requirements for automated detection software: they process inputs efficiently, can generate class scores for an arbitrary number of target classes, and can incorporate new target classes through minor structural changes and the addition of new training data. Accuracy tends to improve with the addition of new training data for existing target classes, allowing for continual improvements in classification performance. Additionally, CNNs can be readily implemented using free and open-source software, affording substantial flexibility to fine-tune network behavior and performance to suit project objectives.

Materials and Methods

Target species

For the present analysis we had six focal species: spotted owl, barred owl, northern saw-whet owl *Aegolius acadicus*, great horned owl *Bubo virginianus*, northern pygmy-owl *Glaucidium gnoma*, and western screech-owl *Megascops kennicottii*. We included the non-*Strix* owls in the hope of producing new insights into the behavior of the forest owl assemblage as a whole. Furthermore, as all our target species are nocturnally active and vocalize at low frequencies, we believed that including these additional species would improve the CNN's discriminative ability for the *Strix* owls as well.

Audio data collection

We collected audio from three historic spotted owl study areas in Oregon (Coast Range and Klamath) and Washington (Olympic Peninsula). Field sites were selected from a uniform grid of 5 km^2 hexagons covering all three study areas. In 2017 we selected 10 non-adjacent hexagons in each study area, preferring hexagons where nesting spotted owls were reported the previous year. In 2018 we collected audio from only the Olympic Peninsula and Coast Range study areas. We generated a pool of hexagons that were >50% federally owned and >50% forested and randomly selected 120 non-adjacent hexagons in the Coast Range and 88 non-adjacent hexagons in the Olympic Peninsula for sampling. Within each hexagon we deployed five ARUs at random locations which were constrained to be on middle and upper slopes, ≥200 m from the hexagon edge, ≥50 m from roads and trails, and with ≥500 m between locations (e.g., Fig. 1). These rules were designed to randomly sample the hexagons, maximize detectability for each species, avoid double-counting

birds that might move between adjacent hexagons, and minimize noise from roads and streams.

We used Song Meter SM4 ARUs (Wildlife Acoustics, Maynard, MA, USA), each equipped with two omnidirectional microphones with sensitivity of −33.5 dB ± 3 dB and a signal-to-noise ratio of 80 dB at 1 kHz. Audio data were stored as hour-long WAV files with sampling rate of 32 kHz. In 2017 ARUs recorded from one hour before sunset to two hours after sunrise each night, producing 11–15 h of recordings per 24-h period. ARUs were deployed for 2 to 4 months between mid-March and late July and collected ca. 150 000 h of recordings. In 2018 ARUs recorded from 1 h before sunset to 3 h after sunset and from 2 h before sunrise to 2 h after sunrise, producing 8 h of recordings per 24-h period. ARUs were deployed at each site for approximately 6 weeks between March and August and collected ca. 350 000 h of recordings.

Training data compilation

Technicians located target species vocalizations in the 2017 data using the Simple Clustering feature of Kaleidoscope Pro software (version 5.0, Wildlife Acoustics). This function detects sounds that meet user-defined criteria and clusters sounds by similarity using a hidden Markov model. We selected for sounds 0.5–7.5 s in duration, 0 – 1.2 kHz in frequency, with a maximum inter-syllable gap of 2 s. We developed these parameters empirically to maximize the detection of spotted and barred owl calls but successfully used them to detect all target species. We reviewed the resulting clusters, tagging calls from our target species. Each tagged call corresponds to a record which includes the source file, timestamp within the file, duration, and a manually assigned species identification field.

We constructed our CNN training set using tagged records from the 2017 data, selecting a single call type for each target species. We chose call types that were highly stereotyped and diagnostic to each species, preferring calls that were produced frequently (Fig. 2). For barred owl we used 3920 unique examples of the two-phrased hoot call (Odom and Mennill 2010), which typically consists of eight notes and ends with a drawn-out, descending 'hooahhh'. For spotted owl we used 3,801 examples of the four-note location call (Forsman et al. 1984), which consists of an initial note, a pause, a closely-spaced pair of notes, another pause, and a terminal note. In practice the initial note is often omitted; our training set included the typical four-note version and the three-note variant. For northern saw-whet owl, we used 3338 examples of the advertising call (Rasmussen et al. 2008), an extended series of whistled notes given at a steady rate of 2–3 s^{-1}. For great horned owl, we used 3353 examples of the territorial hoot (Artuso et al. 2013), a low-pitched call consisting of three to six notes. For northern pygmy-owl, we used 3337 examples of the primary call (Holt and Petersen 2000), which is similar to the saw-whet owl's advertising call but slower, with intervals of 1–2 s between notes. For western screech-owl, we used 3346 examples of the 'bouncing ball' call and the closely related double trill call (Cannings et al. 2017), both consisting of a rapid series of very brief hoots.

The training set for the CNN consisted of spectrograms which we generated using SoX (version 14.4,

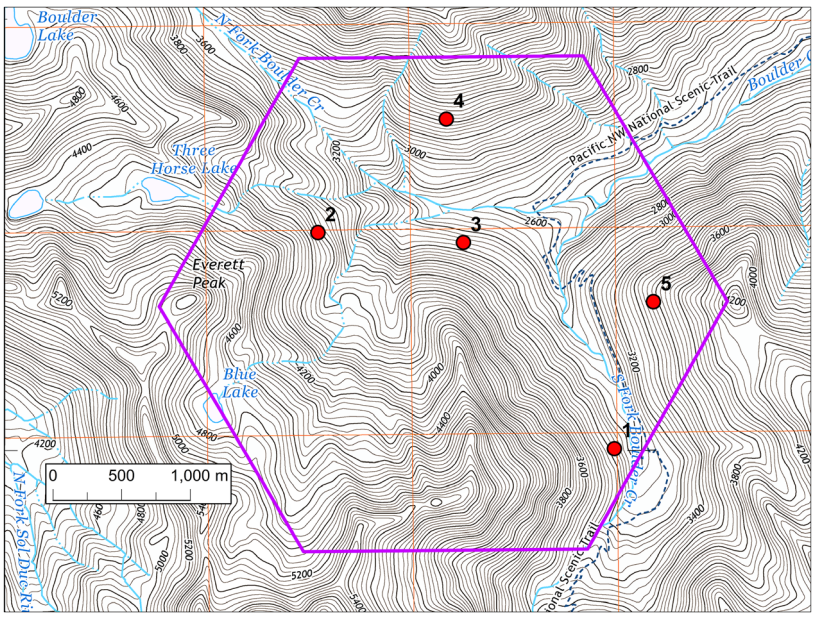

Figure 1. Example survey hexagon from the Olympic Peninsula study area, Washington, USA. Each 5 km^2 hexagon (purple polygon) contains five survey stations (red dots) randomly placed within the hexagon, avoiding areas with low topographic position (e.g. valley bottoms), with ≥200 m between each station and the hexagon edge, ≥50 m between the station and any road or trail, and ≥500 m between any two points. Elevation of topographic contours is given in feet above sea level. This hexagon is shown for illustrative purposes and was not analyzed for the present study.

http://sox.sourceforge.net/). To reflect the variation found in field recordings, we generated multiple spectrograms with different parameters for each unique clip. Each call to SoX included four distinct commands which were executed in sequence. The 'trim' command isolated a 12-s segment of the source audio beginning at time *t*, calculated as the call's original timestamp minus $x^*(12 - [\text{duration}])$ seconds, where *x* was a random number between zero and one. Hence, each spectrogram contained a complete call, but at a somewhat random position. The 'remix' command isolated a single channel of the audio. The 'rate' command resampled the audio at a rate of 6 kHz. Finally, the 'spectrogram' command generated an image from the processed audio. We used the –z option to randomize the dynamic range of the spectrogram to a value between −100 and −90 decibels below full scale as the lower end of the intensity scale. This affected the level of contrast in the resulting image, mimicking the effect of the call being louder or quieter relative to the background noise. Each spectrogram represented 12 s of audio in the frequency range 0–3 kHz; the upper frequency limit represents the Nyquist frequency of the 6 kHz sample rate. Spectrograms were generated using a Hann window with a window length of 2048, 50% window overlap, and a DFT size of 256. Spectrograms were saved as grayscale images at 500 × 129 resolution.

For the non-*Strix* owls we repeated the above process with three sets of randomized values, producing three images for each unique call, using only one channel of the audio. For spotted owl and barred owl, we repeated the process three times for each channel of the audio, producing six images for each unique call. For classification purposes we also created a Noise class, which served as a catch-all for clips containing no target species vocalizations. We produced training data for the Noise class by creating one spectrogram of a 12 s clip at a random offset within each audio file recorded at several sites. We reviewed all spectrograms visually to ensure that each image included visible call signatures of only the labeled class. Our final training data set included spectrograms for all seven target classes: northern saw-whet owl (n = 10 003), great horned owl (n = 9999), northern pygmy-owl (n = 10 003), western screech-owl (n = 10 004), spotted owl (n = 22 373), barred owl (n = 22 204), and Noise (n = 10 003).

Data processing

To process the data we used Python (version 2.7, Python Foundation) to segment the raw audio files into non-overlapping 12-s clips, then used SoX to generate a spectrogram from each clip. We chose a 12-s interval as it cleanly divides an hour-long field recording, creates a tractable number of images given the volume of data we have to work with, and is long enough to fully contain any of the owl calls. Spectrograms were generated from a single channel of the audio using the same parameters as the training data, except that the lower limit of the intensity scale was fixed at −90 decibels below full scale. The audio channel used corresponds to the left microphone of each ARU; the choice of which channel to analyze was

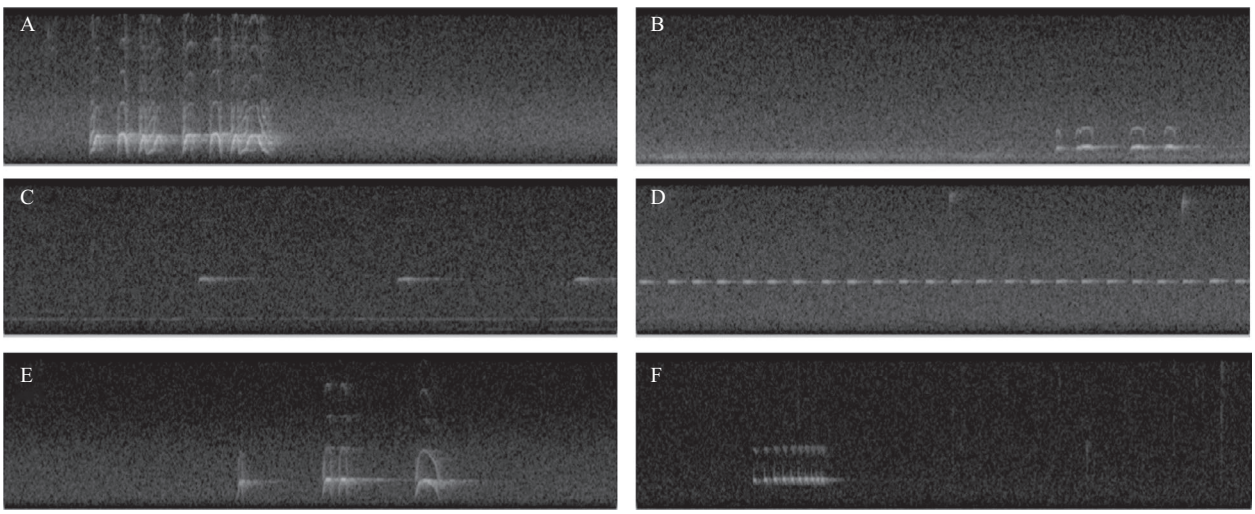

Figure 2. Example spectrograms of each target species call. A = Barred owl, B = Great horned owl, C = Northern pygmy-owl, D = Northern saw-whet owl, E = Spotted owl, F = Western screech-owl. Spectrograms plot the energy present across a range of combinations of time (on the *x*-axis) and frequency (on the *y*-axis), with lighter colors representing higher levels of energy. The lowest level of each call is the base frequency, which carries the most energy. Images A, B, E, and F include visible overtones, indicating that these calls have a high signal-to-noise ratio. Each spectrogram is 500 × 129 resolution and represents 12 s of audio in the frequency range 0–3 kHz.

arbitrary. The spectrograms were then fed into the trained CNN, which generated class scores for each image.

Convolutional neural network model

We implemented the CNN in Python using Keras (Chollet 2015), an application programming interface to the open-source TensorFlow software library developed by Google (Abadi et al. 2015). Our CNN contained four convolutional layers followed by two fully connected layers. The first and second convolutional layer of our network each contained 32 3 × 3 pixel filters and the third and fourth convolutional layer each contained 64 3 × 3 pixel filters. Each convolutional layer used rectified linear unit (ReLU) activation and was followed by 2 × 2 max pooling and 20% dropout. After pooling and dropout, output from the fourth convolutional layer was flattened and passed to a 64-unit fully connected layer with L2 regularization, ReLU activation, and 50% dropout. Our final layer was a seven-unit fully connected layer with softmax activation, whose activation tensor comprised the predicted class scores for our target classes.

Convolutional layers are so called because they perform convolution, which transforms an input (i.e., image) using a small matrix of weights, or filter, to produce an activation map. Each element of the activation map is valued as the dot product of the filter and an equal-sized region of the input, termed the receptive field. This value is highest when the values in the receptive field are similar to those of the filter, hence the activation map for each filter encodes the location of matching features within the input. A convolutional layer systematically applies a number of such filters (each with a different set of learnable weights) over its input and concatenates the resulting activation maps into an activation volume. Fully connected layers do not perform convolution. Each unit in a fully connected layer processes the entire activation volume of the previous layer through a single set of learnable weights; these layers translate the features detected by the convolutional layers to a set of predictions, that is, class scores.

Output from each layer passes through an activation function, which provides nonlinearity to contribute to the CNN's learning ability. The ReLU activation function outputs zero for negative inputs and leaves nonnegative inputs unchanged. Softmax activation normalizes activations from a layer so that they lie between zero and one and sum to one across all units in the layer.

Max pooling downsamples the activation volume by dividing each activation map into non-overlapping regions (e.g., squares two units wide and two units high) and retaining only the highest value from each; this substantially reduces the number of weights required for the following layer (and the overall number of trainable model parameters) while preserving coarse information on feature locations. Dropout randomly omits some proportion of units during training; this forces the model to develop redundancy and helps to prevent overfitting. L2 regularization prevents overfitting during training by adjusting the loss function by the squared Euclidean norm of the weights of the preceding layers.

We trained the CNN for 100 epochs on a set of ca. 95 000 labeled images with a 4:1 training-validation split. Training data were weakly labeled, that is, we provided the correct label for each image but not the location of any relevant features within the image. Loss was calculated as categorical cross-entropy, and we saved the model only after epochs in which validation loss decreased in order to prevent overfitting. We used the Adam optimization algorithm (Kingma and Ba 2015) with a learning rate of 0.0001.

Model performance and verification

We report preliminary results from the first set of data from the 2018 field season to be processed by the CNN, covering approximately 4976 h of recordings from 14 ARUs in three hexagons in the Coast Range study area. We first performed a 'naïve' classification by labeling each clip as the class with the highest class score (p_{Max}). We reviewed all clips that were labeled as target species under this classification scheme (some with p_{Max} as low as 0.190) as well as 1% of clips that were labeled as Noise; the subset of Noise clips to be reviewed were randomly selected, allowing us to estimate the number of false negatives. Reviewing all apparent detections is onerous and unnecessary for most applications; Chambert et al. (2018) concluded that review of as little as 1% of apparent detections could yield unbiased and reasonably precise estimates of site occupancy, provided that researchers employ a modeling approach that explicitly accounts for false detections.

Clips selected for review were extracted from the original recordings as 12-s clips and searched for target species vocalizations using the Kaleidoscope viewer. Although we assigned exactly one label to each clip based on class scores, technicians could assign multiple labels if the clip contained calls from multiple species. We considered a 'hit' to be a real detection if the labels assigned by a human technician included the class to which the CNN assigned the highest class score. Although the CNN assigned class scores based on the spectrogram alone, technicians could also listen to the recording in order to identify species.

We compared the labels assigned by human technicians to the CNN's class scores to calculate precision and recall.

Precision is defined as the proportion of true positives among apparent detections for each species, calculated as [True Positives]/[True Positives + False Positives]. Recall is defined as the proportion of real target vocalizations in the dataset that are detected and correctly labeled, calculated as [True Positives]/[True Positives + False Negatives]. We multiplied the number of clips of each target species that the CNN incorrectly labeled as Noise by 100 and added the result to the denominator when calculating recall. We first calculated precision and recall for the naïve classification, then repeated the calculations, considering only clips for which p_{Max} exceeded an increasingly selective threshold. This enabled us to explore the tradeoff involved in reviewing only a subset of apparent detections, which reduces the need for human labor but may result in lower recall. Following recommendations by Knight et al. (2017), we also report F1 score across the range of thresholds and present receiver operating characteristic and precision-recall curves, for comparison with other research. F1 score combines precision and recall to measure overall model performance; we used the unweighted version, calculated as 2*[Precision*Recall]/[Precision + Recall].

To confirm that the CNN's detection power for our target species was at least comparable to our previous analytical approach, we processed recordings from these hexagons using the same methods as the 2017 pilot study and compared the number of real detections that the two methods produced for each species.

We also wanted to examine the effect that increasing selectivity (i.e., threshold) might have on data used in an occupancy-based framework (e.g., MacKenzie et al. 2018). To this end, we generated weekly encounter histories for each target species at the hexagon level, first based on naïve classification, and then successively filtering the detections by maximum class score against an increasing threshold. This admittedly basic example illustrates how automated detection data may inform useful ecological analyses with a range of manual review effort. For a more in-depth treatment we direct readers to Chambert et al. (2018).

Results

The reviewed dataset included 164 210 clips. Technicians confirmed 71 963 clips as containing calls from target species. These included clips containing western screech-owl (n = 29 252), northern pygmy-owl (n = 27 458), northern saw-whet owl (n = 12 342), barred owl (n = 5387), great horned owl (n = 1643), and spotted owl (n = 94) calls. Great horned owls were detected at two hexagons, while the other target species were detected at all three hexagons. A total of 4033 clips contained two target species, and 90 clips contained three target species; of the 4123 clips containing multiple target species, 3876 contained western screech-owl and 2932 contained northern saw-whet owl. A total of 89 clips originally labeled Noise contained target species, including western screech-owl (n = 62), northern pygmy-owl (n = 20), barred owl (n = 3), great horned owl (n = 3), and northern saw-whet owl (n = 2).

Recall was consistently highest for northern saw-whet owl, northern pygmy-owl, and spotted owl (Table 1, Fig. 3), whereas precision was highest for northern saw-whet owl and western screech-owl (Table 1, Fig. 4). F1 score was highest for the three smallest owls and lowest for spotted owls; the threshold at which the CNN performed best was different for each species (Table 1, Fig. 5). False positives were thinned more than real detections at higher thresholds, indicating that the CNN generally assigned higher scores to real calls than to similar, non-target sounds (Fig. 6). Receiver operating characteristic curves indicated generally good performance at distinguishing true positives (target species vocalizations) from true negatives (Noise; Fig. 7). Precision-recall curves indicated that the CNN performed best for the three small species, while performance was mixed for barred owl and poorer for great horned owl and spotted owl (Fig. 8).

Compared to our previous approach, with human technicians tagging output from the Simple Clustering feature of Kaleidoscope, the CNN detected and correctly labeled as many or more vocalizations from all target species, although both methods detected the same set of species at each hexagon. At the level of the hexagons, naïve occupancy (≥1 detection at a hexagon) was consistently unchanged after increasing the detection threshold from 0 to 0.99; weekly encounter histories changed somewhat with increasing threshold, with some initial detections occurring later in the season, but the overall patterns were highly consistent (Table 2).

Discussion

Our results suggest that CNNs can be highly successful at classifying owl calls and may detect more vocalizations than other analytical methods. This finding is encouraging given the desire of many researchers and management agencies to use bioacoustic methods for broad-scale, long-term monitoring of avian populations. In a review of existing literature on automated birdsong recognition, Priyadarshani et al. (2018) found that while many researchers have reported strong results for single species in short recordings, efficient automatic recognition of multiple species in noisy, long-form field recordings remains elusive. Work on automatic recognizers with owls has been limited; Wood et al. (2019) obtained precision of 40% and recall of 87% for California spotted owls

Table 1. Precision, recall, and F1 score for six owl species at select threshold levels

Metric	Threshold	Barred owl	Great horned owl	N. Pygmy-owl	N. Saw-whet owl	Spotted owl	W. Screech-owl
Precision	None	0.257	0.065	0.571	0.771	0.004	0.726
Precision	0.50	0.317	0.076	0.605	0.815	0.005	0.797
Precision	0.75	0.484	0.112	0.683	0.895	0.009	0.852
Precision	0.90	0.596	0.155	0.743	0.935	0.012	0.872
Precision	0.95	0.649	0.186	0.776	0.949	0.015	0.879
Precision	0.99	0.734	0.229	0.842	0.965	0.023	0.892
Recall	None	0.667	0.917	0.980	0.918	0.915	0.814
Recall	0.50	0.638	0.895	0.962	0.901	0.894	0.789
Recall	0.75	0.551	0.766	0.900	0.855	0.840	0.700
Recall	0.90	0.475	0.623	0.823	0.801	0.745	0.614
Recall	0.95	0.416	0.528	0.766	0.760	0.691	0.559
Recall	0.99	0.304	0.321	0.624	0.678	0.617	0.446
F1 score	None	0.371	0.121	0.721	0.838	0.009	0.768
F1 score	0.50	0.423	0.140	0.743	0.856	0.011	0.793
F1 score	0.75	0.515	0.195	0.776	0.875	0.018	0.769
F1 score	0.90	0.529	0.249	0.781	0.863	0.024	0.721
F1 score	0.95	0.507	0.276	0.771	0.844	0.029	0.683
F1 score	0.99	0.429	0.267	0.717	0.796	0.044	0.595

Performance metrics for the convolutional neural network are given for a naïve classification (no threshold), in which each clip was assigned the label corresponding to the highest predicted class score (p_{Max}), and for increasingly selective classification, in which we consider only clips for which p_{Max} equals or exceeds some threshold. Precision is the proportion of apparent 'hits' that represent real detections for a given species and is calculated as [True Positives]/[True Positives + False Positives]. Recall is the proportion of real calls present in the dataset that are detected and correctly identified by a recognizer and is calculated as [True Positives]/[True Positives + False Negatives]. F1 score is a measure of overall model performance, calculated as 2 * [Precision * Recall]/[Precision + Recall].

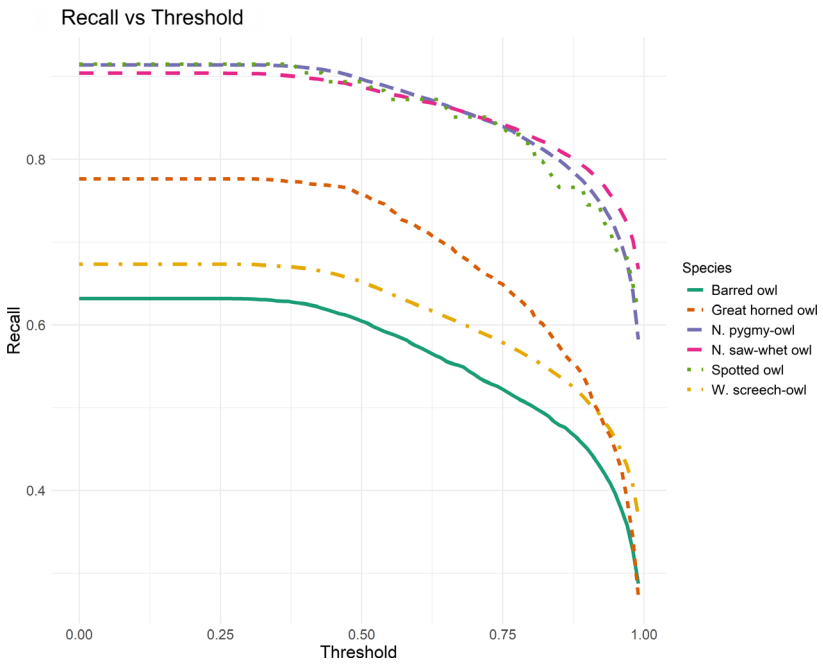

Figure 3. Recall vs threshold for six owl species. Recall is calculated as [True Positives]/[True Positives + False Negatives], considering only clips with $p_{Max} \geq$ Threshold, where p_{Max} is the maximum class score predicted by the convolutional neural network. Recall represents the proportion of target species calls present in the dataset that were detected and correctly labeled by the convolutional neural network.

considering calls exceeding a template matching score of 0.75 using Raven Pro. Shonfield et al. (2018) report precision of 1.7%, 72%, and 99% for barred owl, great horned owl, and boreal owl respectively using template matching in Song Scope, although that study did not assess recall at the level of individual detections.

Our results, with recall ranging from 63.1 to 91.5% from only a single training of the CNN, demonstrate this

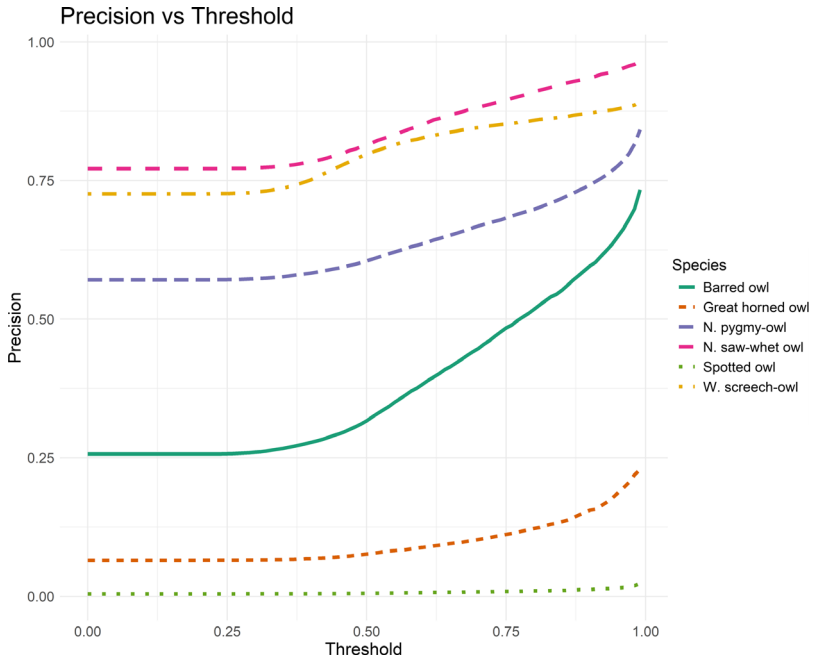

Figure 4. Precision vs threshold for six owl species. Precision or True Positive Rate is calculated as [True Positives]/[True Positives + False Positives], considering only clips with $p_{Max} \geq$ Threshold, where p_{Max} is the maximum class score predicted by the convolutional neural network. Precision represents the proportion of apparent detections that correspond to real target species calls.

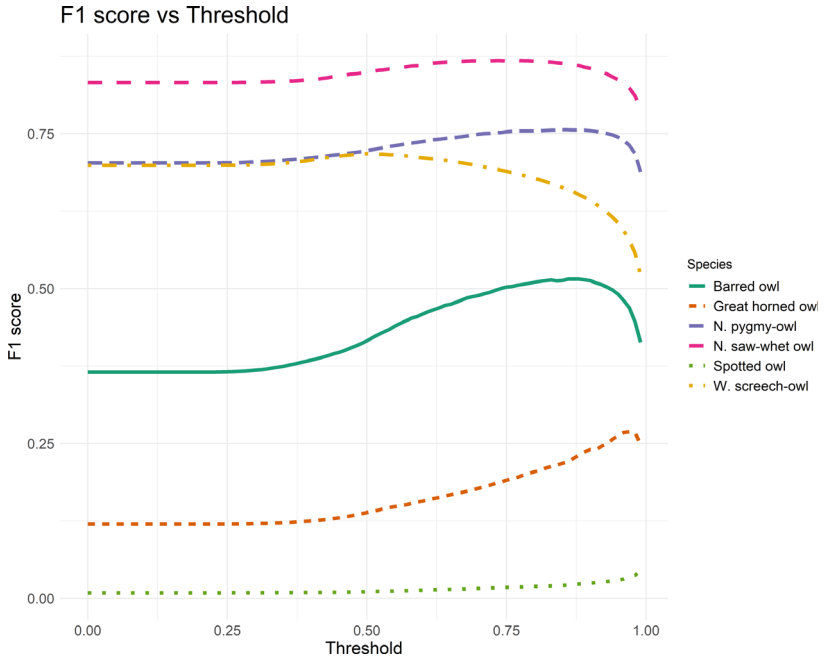

Figure 5. F1 score vs threshold for six owl species. F1 score is interpreted as a balanced measure of overall classifier performance combining precision and recall. The F1 score can be weighted to emphasize either precision or recall; we have plotted the unweighted version, calculated as 2*[Precision * Recall]/[Precision + Recall], considering only clips with $p_{Max} \geq$ Threshold, where p_{Max} is the maximum class score predicted by the convolutional neural network. The highest point on each curve represents the threshold at which the model gives the best overall performance for each species if we consider precision and recall to be equally important.

technology's potential to generate useful ecological information, but given the large percentage of false positives (range 22.9–99.6%), further refinement of the CNN will be required before we can be confident that the output accurately reflects the number of calls of a given species. The misclassification of irrelevant sounds as target vocalizations is a major issue for several target species, especially spotted owl and great horned owl, and subsequent trainings will be necessary to improve precision. Human review will likely remain a necessary step between data processing and analysis, but the effort required can be reduced by reviewing only detections with high classification scores, which disproportionately thins false positives.

Our reported precision was low for several species, particularly spotted owl. It should be noted that the number of real detections for this species was very small, only 94

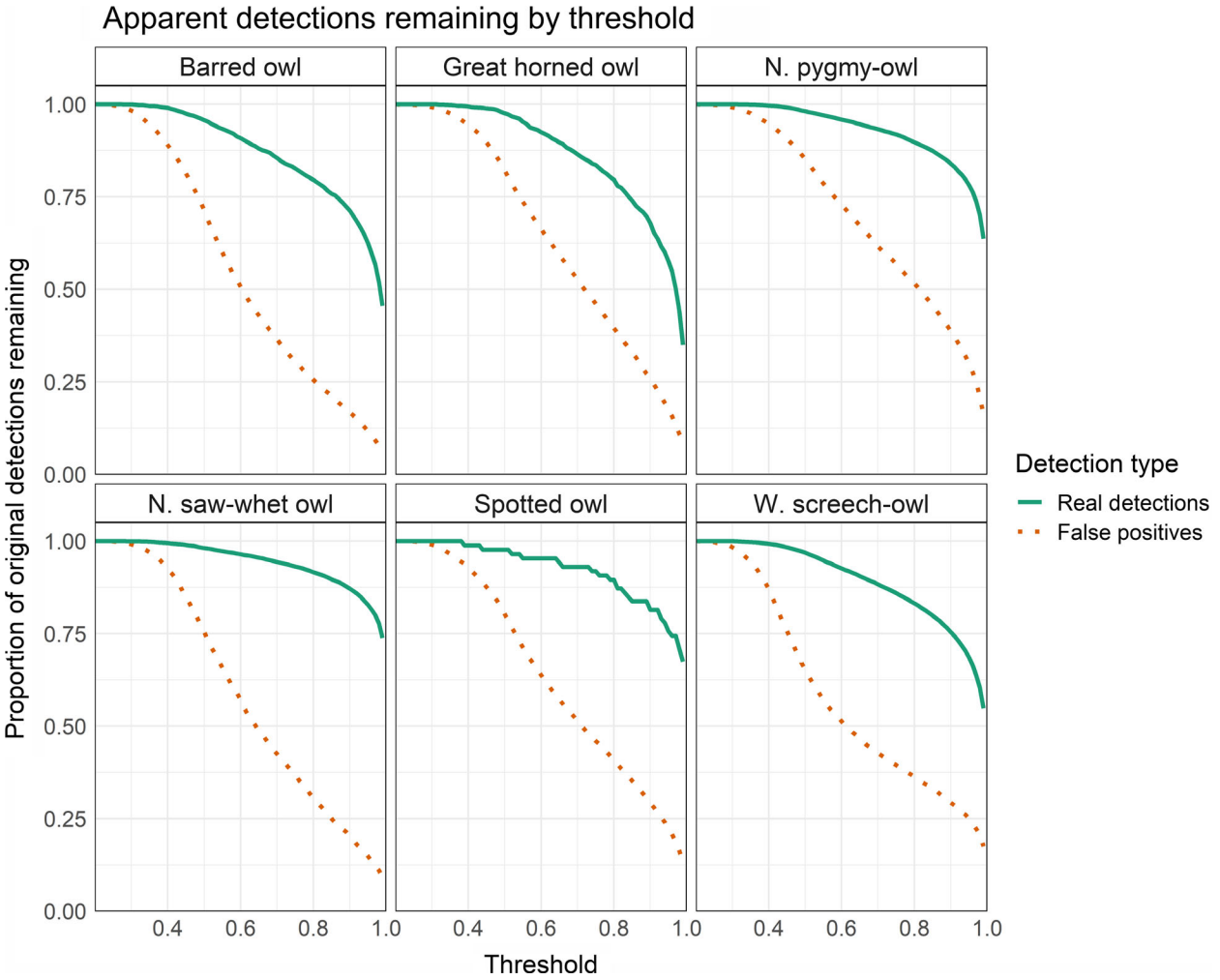

Figure 6. Proportion of apparent detections remaining at various thresholds. This figure illustrates the effect of thresholding on the number of apparent detections for each species, including both real detections and false positives, considering only clips with $p_{Max} \geq$ Threshold, where p_{Max} is the maximum class score predicted by the convolutional neural network. A value of 1.00 on the *y*-axis is the original number of apparent detections of each type generated for each species by assigning each clip to the class with the highest class score. False positives are thinned quickly by thresholding, while the majority of real detections remain even at thresholds of 90% or more.

across the three field sites. Given the level of calling activity that we have observed at known spotted owl territories and nest sites, and considering spotted owl calls may be audible at ranges of >1 km (Forsman et al. 1984), it is unlikely that these sites were occupied by territorial spotted owls. Because precision is calculated as the proportion of true positives among apparent detections, even a low rate of false positives would overwhelm real detections of an uncommon species when processing large amounts of data. Hence, our reported precision for spotted owl should be interpreted cautiously. Indeed, given the motivations for this study, it is encouraging that the CNN successfully detected such a tenuous spotted owl presence at multiple sites. For species of conservation concern, which may be present at low densities on the landscape, recall is more important than precision; the need for high detection power justifies additional human effort in reviewing apparent detections.

Even for species with a stronger presence, precision can vary dramatically due to the presence of sounds similar to those made by a target species. These sounds, which the CNN may classify as target species with high confidence, greatly increase the need for human review to avoid biasing model results. Hence, efforts to refine the model will be most productive if we identify consistent sources of misclassification and work to counter these errors in future trainings. For example, the presented iteration of our CNN frequently misclassified ubiquitous band-tailed

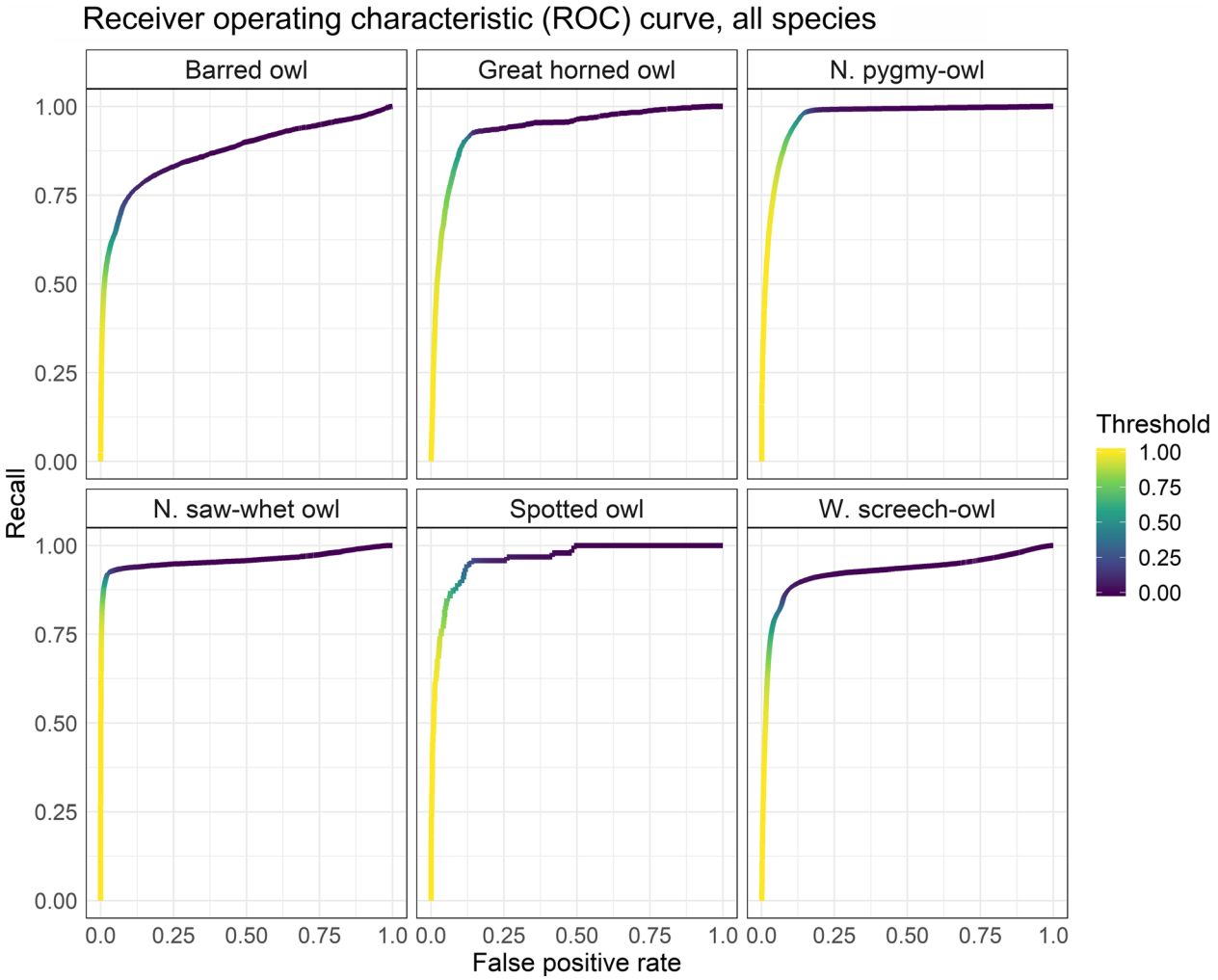

Figure 7. Receiver operating characteristic (ROC) curves for convolutional neural network classification of six owl species. The ROC curve plots the recall (AKA true positive rate) against the false positive rate. Recall is calculated as [True Positives]/[True Positives + False Negatives]. False positive rate is calculated as [False Positives]/[False Positives + True Negatives]. The area under the ROC curve (AUC) corresponds to the probability that the classifier will assign a higher score to a randomly chosen true positive than to a randomly chosen true negative. AUC values by species were: Barred owl, 0.872; Great horned owl, 0.934; Northern pygmy-owl, 0.967; Northern saw-whet owl, 0.959; Spotted owl, 0.961; Western screech-owl = 0.922. ROC curves were generated using the PRROC package in R, which interpolates values across the full range of threshold values.

pigeon *Patagioenas fasciata* calls as great horned owl, negatively affecting precision for the latter species. Hence, we will likely include band-tailed pigeon as a target class in future. CNNs are by nature modular; adding a new target class is as simple as increasing the number of units in the output layer and retraining the network. It is difficult to anticipate every sound that might be confused for a target species, but we can address common sources of error; human review of CNN output has the convenient side effect of producing training data for these non-target sounds in rough proportion to their prevalence.

Recall was generally good but was noticeably poorer for barred owl and western screech-owl. We believe several factors contributed to this result. First, the design of our CNN implicitly treats each 12-s clip as containing, at most, one target species. In reality, multiple target species may call simultaneously, allowing one species to mask the presence of another. This appears to have disproportionately affected recall for western screech-owl, which was present in the great majority (94%) of clips containing multiple species but received the highest class score in less than one-third of those cases, producing ca. 1900 missed detections for this species. In future we may use sigmoid activation in the

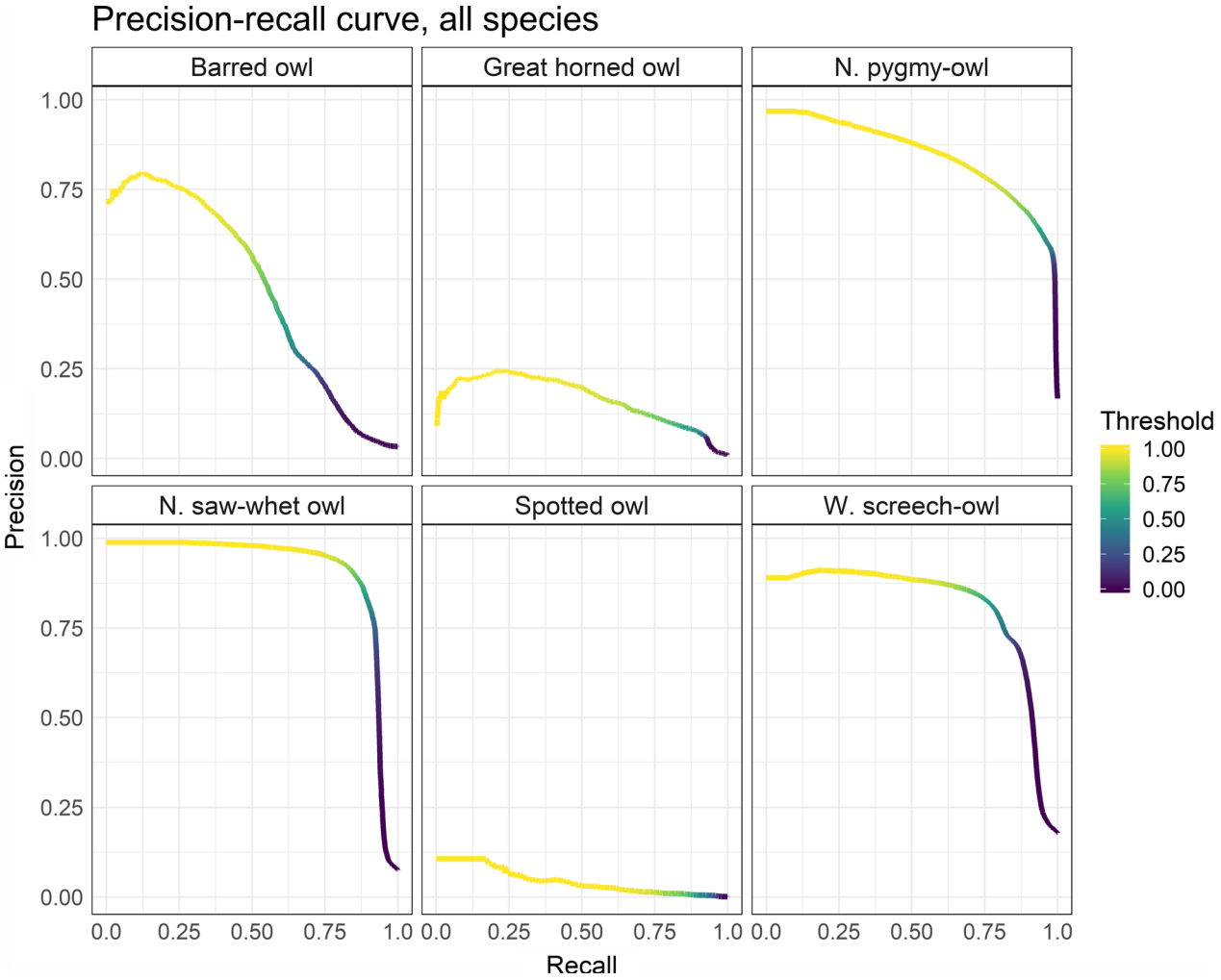

Figure 8. Precision-recall curves for convolutional neural network classification of six owl species. The precision-recall curve illustrates the tradeoff between sensitivity (recall) and specificity (precision). Area under the curve (AUC) serves as a general measure of model performance; a classifier with perfect precision and perfect recall would have AUC = 1. AUC values by species were: Barred owl, 0.473; Great horned owl, 0.166; Northern pygmy-owl, 0.847; Northern saw-whet owl, 0.908; Spotted owl, 0.043; Western screech-owl, 0.807. Precision-recall curves were generated using the PRROC package in R, which interpolates values across the full range of threshold values.

output layer of the CNN; sigmoid activation is not normalized across the layer and so is potentially conducive to multi-label classification. Western screech-owl was also the target species most often labeled as Noise, for reasons unknown. Additionally, while we trained the CNN on only one call type for each species, technicians labeled clips only to the species level, without noting call type. Barred owls have a diverse vocal repertoire, and barred owl calls other than the two-phrased hoot were frequently misclassified as other target species. In particular, the barred owl 'inspection call' (Odom and Mennill 2010) appears to have been a significant source of false positives for spotted owl, possibly because it resembles the terminal note of the spotted owl's four-note location call. We may include this call type as a separate class in future.

Because our approach depends on the recognition of visual patterns in spectrograms, it is useful to consider the nature of spectrograms and the variation inherent in this type of plot. To produce a spectrogram we take a sound recording, which represents energy as a periodic function of time, and apply a discrete-time Fourier transform, which decomposes the signal into its constituent frequencies to represent energy as a function of both frequency and time. The output is then plotted with time on the *x*-axis and frequency on the *y*-axis, with energy levels mapped to a color scale. Many signals of interest, such as bird calls, produce recognizable patterns, which trained human observers can recognize from cursory inspection of a spectrogram. Here we demonstrated significant strides in developing a neural network that can reliably accomplish the same task.

Table 2. Weekly encounter histories for six owl species with varying selectivity

	Threshold	Barred owl	Great horned owl	N. saw-whet owl	N. pygmy-owl	Spotted owl	W. screech-owl
Hexagon A	**None**	11111110	11111100	00001010	11111110	10011110	11111110
	0.50	11111110	11111100	00001010	11111110	**10001110**	11111110
	0.75	11111110	11111100	00001010	11111110	**00001110**	11111110
	0.90	11111110	11111100	00001010	11111110	00001110	11111110
	0.95	11111110	11111100	00001010	11111110	00001110	11111110
	0.99	11111110	11111100	00001010	11111110	00001110	11111110
Hexagon B	**None**	01111111	00000000	11111111	11110110	00010010	11111111
	0.50	01111111	00000000	11111111	11110110	00010010	11111111
	0.75	01111111	00000000	11111111	11110110	00010010	11111111
	0.90	01111111	00000000	11111111	11110110	00010010	11111111
	0.95	01111111	00000000	11111111	11110110	**00000010**	11111111
	0.99	01111111	00000000	11111111	11110110	00000010	11111111
Hexagon C	**None**	11111110	11111010	11101110	11111110	01011000	00111010
	0.50	11111110	11111010	11101110	11111110	01011000	00111010
	0.75	**11110110**	**11111000**	11101110	11111110	**00011000**	00111010
	0.90	**11110010**	11111000	11101110	11111110	00011000	00111010
	0.95	11110010	11111000	11101110	11111110	00011000	**00011010**
	0.99	**11110000**	11111000	11101110	11111110	00011000	**00010010**

We generated weekly encounter histories for six owl species at three study hexagons (each containing five survey stations), first based on naïve classification (considering all real detections that were correctly tagged for a given species), then considering only real detections that were correctly tagged with p_{Max} greater than or equal to some threshold, where p_{Max} is the maximum class score predicted by the CNN. Each encounter history indicates whether the species was detected (1) or not detected (0) at the site in each of eight consecutive weeks of recording. Bolded entries represent a change from the hexagon-level encounter history for a given species at the previous threshold level. Great horned owls were never detected at Hexagon B.

Compared to three-dimensional objects in a video or photograph, signals in a spectrogram have limited degrees of freedom; they occur only at a fixed size and orientation within a two-dimensional plane. However, there are several forms of variation which a CNN must disregard in order to make reliable predictions. Overall signal intensity varies greatly, as animals produce sound at inconsistent volume and at varying positions and orientations relative to the recorder; this effect is compounded by variation in background noise. Intensity also varies within each call as the vocalizing animal places more stress on some syllables or parts of syllables than others; less intense parts of the call may fade out entirely as the sound attenuates with distance, changing the apparent shape of the signal. The signal may be compressed or expanded slightly in time or frequency due to individual variation in sound production. The signal may also blur along the time axis due to echoes or scattering; this obscures the shape and separation between syllables, causing the call to appear cloudy or smeared. Affected calls might still be recognizable but will be more challenging to identify. When spectrograms represent non-overlapping segments of the audio, some portion of calls will be split between adjacent segments. If this consistently hinders identification, such calls could be under-counted; conversely, if the model can reliably identify partial calls, split calls might be double counted. These forms of variation will all be present in varying combinations in field recordings. To make the CNN more reliable, the training set should contain similar variation. This can be accomplished organically, by drawing training examples from recordings made under varying conditions, but it can also be simulated through data augmentation.

Despite the need for further development, the demonstrated effectiveness of this approach suggests that it may be a good choice to support long-term monitoring of a range of species at a large scale. We expect to achieve further improvements with increased volume of training data and number of target classes, experimentation with alternative model architectures, and fine-tuning the training procedure.

Acknowledgments

The authors thank C. Cardillo, M. Corr, D. Culp, T. Garrido, E. Guzmán, A. Ingrassia, D. Jacobsma, E. Johnston, R. Justice, K. McLaughlin, P. Papajcik, and W. Swank for field assistance in collecting data and C. Cardillo, D. Culp, Z. Farrand, R. Justice, A. Munes, and S. Pruett for validating CNN output and locating additional training data. Three anonymous reviewers provided valuable insights and feedback which have greatly improved the manuscript. Funding and logistical support were provided by USDA Forest Service and USDI Bureau of Land

Management. The Center for Genome Research and Biocomputing, Oregon State University provided C. Sullivan salary and biocomputing infrastructure support. This work was partially supported through a Research Participation Program administered by Oak Ridge Institute for Science and Education (ORISE) and hosted by US Forest Service, Pacific Northwest Research Station. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. The use of trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Department of Agriculture of any product or service.

Data Availability

Our trained neural network, test data, and original code are archived on Zenodo, https://doi.org/10.5281/zenodo.3338550.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. 2015. Tensorflow: large-scale machine learning on heterogeneous systems. [Online] URL: https://www.tensorflow.org/

Alonso, J. B., J. Cabrera, R. Shyamnani, C. M. Travieso, F. Bolaños, A. García, et al. 2017. Automatic anuran identification using noise removal and audio activity detection. *Expert Syst. Appl.* **72**, 83–92.

Artuso, C., C. S. Houston, D. G. Smith, and C. Rohner. 2013. Great Horned Owl (*Bubo virginianus*), version 2.0. *in* A. F. Poole, ed. *The Birds of North America*. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] https://doi.org/10.2173/bna.372

Brown, J. C., and P. J. O. Miller. 2007. Automatic classification of killer whale vocalizations using dynamic time warping. *J. Acoust. Soc. Am.* **122**, 1201–1207.

Campos-Cerqueira, M., and T. M. Aide. 2016. Improving distribution data of threatened species by combining acoustic monitoring and occupancy modeling. *Methods Ecol. Evol.* **7**, 1340–1348.

Cannings, R. J., T. Angell, P. Pyle, and M. A. Patten. 2017. Western Screech-Owl (*Megascops kennicottii*), version 3.0. *in* P. G. Rodewald, ed. *The Birds of North America*. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] https://doi.org/10.2173/bna.wesowl1.03

Chambert, T., J. H. Waddle, D. A. W. Miller, S. C. Walls, and J. D. Nichols. 2018. A new framework for analysing automated acoustic species detection data: occupancy estimation and optimization of recordings post-processing. *Methods Ecol. Evol.* **9**, 560–570.

Chollet, F. 2015. Keras. [Online] https://keras.io

Dugger, K. M., E. D. Forsman, A. B. Franklin, R. J. Davis, G. C. White, C. J. Schwarz, et al. 2016. The effects of habitat, climate, and Barred Owls on long-term demography of Northern Spotted Owls. *Condor* **118**, 57–117.

Figueira, L., J. L. Tella, U. M. Camargo, and G. Ferraz. 2015. Autonomous sound monitoring shows higher use of Amazon old growth than secondary forest by parrots. *Biol. Cons.* **184**, 27–35.

Forsman, E. D., E. C. Meslow, and H. M. Wight. 1984. Distribution and Biology of the Spotted Owl in Oregon. *Wildlife Monographs* **87**, 3–64.

Ganchev, T., and I. Potamitis. 2007. Automatic acoustic identification of singing insects. *Bioacoustics* **16**, 281–328.

Gutiérrez, R. J., M. Cody, and S. Courtney. 2007. The invasion of barred owls and its potential effect on the spotted owl: a conservation conundrum. *Biol. Invasions* **9**, 181–196.

Heinicke, S., A. K. Kalan, O. J. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kuhl. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods Ecol. Evol.* **6**, 753–763.

Holt, D. W., and J. L. Petersen. 2000. Northern Pygmy-Owl (*Glaucidium gnoma*), version 2.0. *in* A. F. Poole, F. B. Gill, eds. *The Birds of North America*. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] https://doi.org/10.2173/bna.494

Kahl, S., T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, et al. 2017. Large-scale bird sound classification using convolutional neural networks. BirdCLEF 2017.

Kingma, D. P., and J. L. Ba. 2015. Adam: A method for stochastic optimization. International Conference on Learning Representation 2015, San Diego, California.

Knight, E. C., K. C. Hannah, G. J. Foley, C. D. Scott, R. M. Brigham, and E. Bayne. 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology* **12**, 14.

Krizhevsky, A., I. Sutskever, and G. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2010: Neural Information Processing Systems. Lake Tahoe, Nevada.

LeCun, Y. 2015. Deep Learning. *Nature* **521**, 436–444.

Lesmeister, D. B., R. J. Davis, P. H. Singleton, and J. D. Wiens. 2018. Northern spotted owl habitat and populations: status and threats. Pp. 245–298 *in* T. A. Spies, P. A. Stine, R. Gravenmier, J. W. Long, and M. J. Reilly, eds. *Synthesis of Science to Inform Land Management within the Northwest Forest Plan Area. PNW-GTR-966*. USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Luo, W., W. Yang, and Y. Zhang. 2019. Convolutional neural network for detecting odontocete echolocation clicks. *J. Acoust. Soc. Am.* **145**, 7–12.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2018. *Occupancy Estimation and*

Modeling: inferring Patterns and Dynamics of Species Occurrence, 2nd ed. Academic Press, Cambridge, MA.

Mazur, K. M., andP. C. James. 2000. Barred Owl (*Strix varia*), version 2.0. *in* A. F. Poole, F. B. Gill, eds. *The Birds of North America*. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] https://doi.org/10.2173/bna.508

Nvidia. 2019. NVIDIA DRIVE – Autonomous Vehicle Development Platforms. [Online] https://developer.nvidia.com/drive

Odom, K. J., and D. J. Mennill. 2010. A quantitative description of the vocalizations and vocal activity of the barred owl. *Condor* **112**, 549–560.

Priyadarshani, N., S. Marsland, and I. Castro. 2018. Automated birdsong recognition in complex acoustic environments: a review. *J. Avian Biol.* https://doi.org/10.1111/jav.01447

Rasmussen, J. L., S. G. Sealy, and R. J. Cannings. 2008. Northern Saw-whet Owl (*Aegolius acadicus*), version 2.0. *in* A. F. Poole, F. B. Gill, eds. *The Birds of North America*. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] https://doi.org/10.2173/bna.42

Russo, D., and G. Jones. 2003. Use of foraging habitats by bats in a Mediterranean area determined by acoustic surveys: conservation implications. *Ecography* **26**, 197–209.

Shonfield, J., S. Heemskerk, and E. M. Bayne. 2018. Utility of automated species recognition for acoustic monitoring of owls. *Journal of Raptor Research* **52**, 42–55.

Somervuo, P. 2018. Time-frequency warping of spectrograms applied to bird sound analysis. *Bioacoustics*. https://doi.org/10.1080/09524622.2018.1431958.

Taigman, Y., M. Yang, M. A. Ranzato, and L. Wolf. 2014. DeepFace: closing the gap to human-level performance in face verification. [Online] https://research.fb.com/wp-content/uploads/2016/11/deepface-closing-the-gap-to-human-level-performance-in-face-verification.pdf?

Trifa, V. M., A. N. G. Kirschel, C. E. Taylor, and E. E. Vallejo. 2008. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *J. Acoust. Soc. Am.* **123**, 2424–2431.

US Department of Agriculture and US Department of Interior. 1994. Final Supplemental Environmental Impact Statement on Management of Habitat for Late-Successional and Old-Growth Forest Related Species Within the Range of the Northern Spotted Owl. US Forest Service.

US Fish and Wildlife Service. 1990. Endangered and threatened wildlife and plants: determination of threatened status for the northern spotted owl. *Fed. Reg.* **55**, 26114–26194.

Wiens, J. D., R. G. Anthony, and E. D. Forsman. 2014. Competitive Interactions and Resource Partitioning Between Northern Spotted Owls and Barred Owls in Western Oregon. *Wildlife Monographs* **185**, 1–50.

Wood, C. M., V. D. Popescu, H. Klinck, J. J. Keane, R. J. Gutiérrez, S. C. Sawyer, et al. 2019. Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework. *Ecol. Ind.* **98**, 492–507.

Wrege, P. H., E. D. Rowland, S. Keen, and Y. Shiu. 2017. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods Ecol. Evol.* **8**, 1292–1301.

