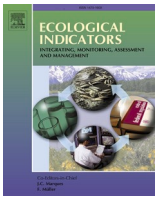
Ecological Indicators 124 (2021) 107419

Contents lists available at ScienceDirect

Ecological Indicators

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



Workflow and convolutional neural network for automated identification of animal sounds

Zachary J. Ruff [a,b,*], Damon B. Lesmeister [a,c], Cara L. Appel [a,c], Christopher M. Sullivan [d]

[a] Pacific Northwest Research Station, USDA Forest Service, Corvallis, OR, United States
[b] Oak Ridge Institute for Science and Education, Oak Ridge, TN, United States
[c] Department of Fisheries and Wildlife, Oregon State University, Corvallis, OR, United States
[d] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, United States

ARTICLE INFO

Keywords:
Bioacoustics
Machine learning
Wildlife
Ecology
Passive acoustic monitoring
Artificial intelligence

ABSTRACT

The use of passive acoustic monitoring in wildlife ecology has increased dramatically in recent years as researchers take advantage of improvements in autonomous recording units and analytical methods. These technologies have allowed researchers to collect large quantities of acoustic data which must then be processed to extract meaningful information, e.g. target species detections. A persistent issue in acoustic monitoring is the challenge of efficiently automating the detection of species of interest, and deep learning has emerged as a powerful approach to accomplish this task. Here we report on the development and application of a deep convolutional neural network for the automated detection of 14 forest-adapted birds and mammals by classifying spectrogram images generated from short audio clips. The neural network performed well for most species, with precision exceeding 90% and recall exceeding 50% at high score thresholds, indicating high power to detect these species when they were present and vocally active, combined with a low proportion of false positives. We describe a multi-step workflow that integrates this neural network to efficiently process large volumes of audio data with a combination of automated detection and human review. This workflow reduces the necessary human effort by > 99% compared to full manual review of the data. As an optional component of this workflow, we developed a graphical interface for the neural network that can be run through RStudio using the Shiny package, creating a portable and user-friendly way for field biologists and managers to efficiently process audio data and detect these target species close to the point of collection and with minimal delays using consumer-grade computers.

1. Introduction

Artificial intelligence technologies are increasingly being applied to issues in ecological research and conservation. In the field of wildlife ecology, the use of artificial intelligence in combination with recent advances in survey techniques has enabled researchers to collect data on species occurrences at much broader spatial and temporal scales than were previously possible. Passive monitoring methods such as camera traps and bioacoustics have greatly improved the capacity of researchers to survey for wildlife, but the resulting large datasets require substantial processing to extract useful information. The task of quickly and accurately locating signals of interest (e.g., target species detections) within large audio or photograph datasets remains a persistent issue. Automated or semi-automated approaches can greatly reduce the amount of time and effort required to extract and classify detections of target animals (Norouzzadeh et al., 2017; Willi et al., 2018). In turn, this allows researchers to proceed more quickly from data collection to analysis, generating insights on the underlying ecological processes in closer to real time and enabling more timely management responses to changes in the system.

Recent efforts have successfully employed image recognition algorithms to describe, count, and identify animals of interest in passively collected data (Weinstein, 2018). Deep convolutional neural networks (CNN), a family of algorithms commonly used for image recognition, have proven particularly useful in analysis of ecological data (Brodrick et al., 2019; Ruff et al., 2020). Several researchers have used CNNs to classify images from remote camera traps with high accuracy, facilitating the detection of various animal species captured therein (Gomez

* Corresponding author at: 3200 SW Jefferson Way, Corvallis, OR 97331, United States.
E-mail address: zachary.ruff@usda.gov (Z.J. Ruff).

https://doi.org/10.1016/j.ecolind.2021.107419
Received 6 July 2020; Received in revised form 7 January 2021; Accepted 13 January 2021
Available online 28 January 2021

Villa et al., 2017; Norouzzadeh et al., 2017; Tabak et al., 2019).

Computer vision techniques can also be used to extract detections of vocalizing or otherwise audible species from acoustic recordings. If sound files are converted into spectrograms, a visual representation of the energy present over a range of acoustic frequencies as a function of time, then algorithms developed for image classification can be used to identify distinctive sounds, which appear as visual patterns in the resulting images. Recordings of avian vocalizations have been analyzed using a variety of automated recognizer algorithms (e.g., Sebastián-González et al., 2015; Knight et al., 2017; Venier et al., 2017; Ovaskainen et al., 2018; Stowell et al., 2019), including CNNs specifically (Salamon et al., 2016; Salamon and Bello, 2017; Ruff et al., 2020; LeBien et al., 2020).

The problem of efficient automated detection of target species is especially important for the monitoring of rare and elusive species. The task of monitoring such species, which may be widely dispersed on the landscape, is more tractable when researchers can leverage technology to collect data at large temporal and spatial scales and quickly sift through the resulting datasets to locate the species of interest. In the Pacific Northwest, USA, the northern spotted owl (*Strix occidentalis caurina*) is an important conservation and management species that is facing ongoing population declines and various threats to persistence (Jenkins et al., 2019; Wiens et al., 2019; Lesmeister et al., 2018). Monitoring of spotted owls has historically consisted of species-specific callback surveys and mark-recapture methods to collect vital rate information and assess population trends (e.g., Forsman et al., 2011; Dugger et al., 2016). However, as spotted owl populations continue to decline, managers and researchers are seeking alternative survey methods to track spotted owl populations without capturing owls or eliciting territorial responses. Given these population declines, the resulting decrease in population densities, and multiple threats to persistence, spotted owl monitoring efforts require an approach that can yield greater spatial and temporal coverage, as well as insights on interspecific interactions between spotted owls and their competitors, prey, and potential predators in the forest community.

Passive acoustic monitoring is an appealing method for accomplishing these objectives, as researchers can, at minimum, collect detection/non-detection data to use in occupancy models with minimal disturbance to spotted owls. Additionally, the increased efficiency of deploying autonomous recording units compared with conducting callback surveys facilitates an increase in spatial and temporal sampling effort, making broadscale multispecies monitoring possible. Passive acoustic monitoring has proven to be effective at detecting northern spotted owls in this system, with detection probabilities exceeding 95% after as little as 3 weeks (Duchac et al., 2020), prompting the deployment of large-scale acoustic monitoring for this and other species (Lesmeister et al., 2019). To process the large sets of audio recordings generated by autonomous recording units, Ruff et al. (2020) developed a CNN that identified vocalizations of spotted owls along with five other owl species. Human verification of apparent detections was an integral part of their process, and will likely continue to be for high priority species. However, model classification performance and efficiency of verification methods have improved, reducing the amount of time and labor needed to manually review recordings.

A major advantage of passive monitoring over species-specific surveys is the ability to collect data on multiple target and incidental species without many of the associated biases of individual species surveys. In the Pacific Northwest, it has become increasingly important to understand the occurrence patterns of a suite of forest species in addition to northern spotted owls, including barred owls (*Strix varia*), which compete with spotted owls for resources, and potential nest predators like common ravens (*Corvus corax*) (Dugger et al., 2016; Gutiérrez et al., 2020). Other species of interest include the complete forest owl guild as well as cavity-nesting birds and small mammals that are prey for raptors and mammalian predators. Accordingly, researchers must be able to conduct multispecies surveys over a large geographical area, rather than just focusing on historic spotted owl territories, and in a randomized design that allows inference for many species with varied life history traits.

Long-term and broadscale passive acoustic monitoring programs need workflows for quick turnaround times between data collection and processing to allow sufficiently short timelines of analyses. Further, given their conservation status, it is important to assess northern spotted owl population trends on a frequent basis and rapidly detect significant changes, such as displacement of breeding pairs by competing species. Agencies must have current information to meet monitoring obligations for northern spotted owls, including regular formal status assessments (e.g. Forsman et al., 2011; Dugger et al., 2016). Moreover, population declines and threats to persistence are not evenly distributed across the subspecies' range (Lesmeister et al., 2018), and having current information allows agencies to target intensive monitoring or interventions where they are likely to be most effective.

Automated detection of target species in large acoustic datasets often requires extensive time to process data and so is typically conducted after the end of the field season to take advantage of high-performance computers. A primary disadvantage of this approach is the resulting delays between data collection and analysis. To address this issue, we sought to develop a tool that would allow biologists and managers in the field to process data in a distributed fashion, close to the point of collection and with minimal delays in the output of detections. While packaging our code as a simple command-line tool would have been equally efficient in terms of processing speed, we felt that users with less experience using specialized software might find it less useful, so an objective for our project was to develop a graphical interface for the sake of greater accessibility. In addition to potentially reducing training time, providing a graphical interface could expand the pool of potential users, allowing this tool to have more widespread benefits in wildlife monitoring using passive data collection methods.

In this paper we report a streamlined workflow to collect and process passive acoustic monitoring data for multiple forest birds and mammals in the Pacific Northwest. Specifically, we describe advancements to a previous CNN (Ruff et al., 2020), including expansion of the target species list to include 12 bird species and 2 small mammal species, and development of a desktop application with a Shiny graphical interface which can be run through RStudio (RStudio Team, 2020), creating a highly portable solution that allows efficient processing of audio data on consumer-grade desktop computers. We used this workflow to process audio data recorded in forested landscapes of Oregon and Washington, USA, during the spring and summer of 2018 and 2019. We assessed model performance by calculating the precision, recall, and F1 score of our CNN using an independent test dataset. We demonstrate the use of this workflow for processing large amounts of audio recordings and generating weekly encounter histories for target species. Finally, although the tools that we developed are technically limited to the 14 species we targeted, we discuss potential uses of our broader workflow for large-scale passive monitoring projects, the advantages of our desktop application for distributed data processing, and specific applications of our CNN.

2. Materials and methods

2.1. Target species and training data

Our CNN included 17 target classes comprising calls and sounds produced by 14 target species: northern spotted owl, barred owl, great horned owl *Bubo virginianus*, northern pygmy-owl *Glaucidium gnoma*, northern saw-whet owl *Aegolius acadicus*, western screech-owl *Megascops kennicottii*, band-tailed pigeon *Patagioenas fasciata*, common raven *Corvus corax*, Douglas' squirrel *Tamiasciurus douglasii*, mountain quail *Oreortyx pictus*, pileated woodpecker *Dryocopus pileatus*, red-breasted sapsucker *Sphyrapicus ruber*, Steller's jay *Cyanocitta stelleri*, and Townsend's chipmunk *Neotamias townsendii*.

Some of the target classes were included because the species in question were of ecological or management interest due to potential competitive or predatory interactions with other species. For example, Townsend's chipmunks are important prey for many raptors and mammalian predators. Other classes were added because previous verification efforts indicated that they were likely to produce false positives for existing target classes. Some target species fulfilled more than one of these criteria; band-tailed pigeon is a managed game species (Sanders, 2015) that was extremely common at our study sites, and was a major source of false positive detections for great horned owl in the Ruff et al. (2020) study, and Townsend's chipmunk calls are easily confused with northern saw-whet owl calls.

We compiled our training dataset from 53,292 unique clips of vocalizations from the 14 target species (Table 1). For eight of these species, including band-tailed pigeon, great horned owl, mountain quail, northern pygmy-owl, northern saw-whet owl, northern spotted owl, red-breasted sapsucker, and Townsend's chipmunk, the examples in the training set included only one highly-stereotyped call type or sound. For another four species, namely common raven, pileated woodpecker, Steller's jay, and western screech-owl, the training set included multiple call types, but the call types for each species were lumped into one class because we felt the component syllables were sufficiently similar that including closely related call types would not hinder identification. For the remaining two species, barred owl and Douglas' squirrel, we included two call types but incorporated each call type as a separate class. We also included a catch-all "Noise" class for images that contained none of our target species.

Training images were generated from annotated records of calls recorded during a survey effort for northern spotted owls and barred owls (see Duchac et al., 2020). Those calls were detected semi-manually using the simple clustering feature of Kaleidoscope Pro software (version 5.0, Wildlife Acoustics). The dataset from which our training data were drawn was collected during Mar – July 2017 at 30 field sites on three historic northern spotted owl demographic study areas (Coast Range, Klamath, Olympic Peninsula) in the Pacific Northwest, USA. See Forsman et al. (2011) for study area descriptions. For each class we included training examples from at least two and in most cases all three study areas, to ensure that the CNN's internal representation of each class would be less influenced by systematic regional differences in vocalizations or background noise.

To augment the unique sound clips included in the training set we generated multiple variant spectrograms with randomized offset and dynamic range, producing three to six distinct images for each unique call, using the same procedure detailed in Ruff et al. (2020). Spectrograms (Fig. 1) consisted of grayscale images in portable network graphic (PNG) format with resolution of 500 × 129 pixels and a bit depth of eight.

We generated these images using the spectrogram command in SoX (version 14.4, http://sox.sourceforge.net). After creating the images, we reviewed the training set to ensure that each image contained a visible signature of sounds corresponding to exactly one of our target classes. We reserved a randomly selected 20% of these images to be used for validation and used the other 80% of images to train the CNN. The training set included spectrograms used to train the Ruff et al. (2020) CNN, as well as many additional images for those original target species and the ten additional classes. Because many sounds that were previously considered "noise" corresponded to one of the new target classes, we reviewed images included in the previous training set again to remove any spectrograms that contained calls of multiple species. Following the generation of the variant spectrograms, the final training dataset comprised 173,964 images representing our 17 classes (Table 1).

2.2. Convolutional neural network training

We compiled and trained the CNN model in Python (version 2.7, Python Foundation) using Keras (Chollet, 2015), an open-source, machine learning-focused application programming interface to Google's TensorFlow software library (Abadi et al., 2015). The CNN contained six trainable layers, including four convolutional layers and two fully connected layers. The first convolutional layer contained 32 5x5 filters, the second layer contained 32 3x3 filters, and the third and fourth layers each contained 64 3x3 filters. Each convolutional layer had a stride length of one and used Rectified Linear Unit (ReLU) activation. Each convolutional layer was followed by 2x2 max pooling and 20% dropout. The first fully connected layer contained 256 units using ReLU activation and L2 regularization and was followed by 50% dropout. The L2

Table 1
Target species and the characteristic sounds used to train the convolutional neural network and construct the test set. Each row denotes a separate class and corresponds to one node in the network's output layer. We had a number of unique audio clips representing each class, from which we generated spectrogram images that we used to train the network. Each class was a specific call type or group of vocalizations with similar syllables. We generated three to six variant spectrograms with slightly different parameters for each clip to increase the volume of training data. During processing, long audio clips are segmented into spectrograms, each representing 12 s of audio. For each image the trained network outputs a vector of 17 class scores, each between zero and one, representing the strength of the match between the image and each of the target classes. The test set ($n = 131,767$) was comprised of images generated from examples of the same call types that were not used to generate spectrograms for the training or validation set. Some images in the test set contained calls from > 1 target class.

Common name	Sound[a]	Uniquesound clips in training set	Training images	Images in test set
Band-tailed pigeon	Song	3,044	12,000	3,745
Barred owl	Inspection call	1,760	10,329	4,000
Barred owl	Two-phrase hoot	3,249	18,240	5,355
Common raven	Demonstrative calls	3,355	13,162	3,276
Douglas' squirrel	Rattle call	2,243	8,529	316
Douglas' squirrel	Chirp call	953	3,727	413
Great horned owl	Territorial hoot	3,242	9,643	2,600
Mountain quail	Advertisement call	1,267	4,679	1,709
Northern pygmy-owl	Primary call	3,435	10,195	11,471
Northern saw-whet owl	Advertising call	3,315	9,931	10,270
Northern spotted owl	Four-note location call	3,068	17,070	5,001
Pileated woodpecker	"Wek" call	2,157	8,190	1,366
Red-breasted sapsucker	Drum	2,722	10,350	1,512
Steller's jay	"Wek" series call, "wah" alarm call	2,642	10,073	3,652
Townsend's chipmunk	Chip call	1,619	6,256	3,224
Western screech-owl	Bouncing ball call, double-trill call	3,221	9,590	7,792
Noise	Many	12,000	12,000	70,183

[a] Sources for sound descriptions, by species: Keppie and Braun, 2000; Odom and Mennill, 2010; Boarman and Heinrich, 1999; Smith, 1978; Artuso et al., 2013; Gutiérrez and Delehanty, 1999; Holt and Petersen, 2000; Rasmussen et al., 2008; Forsman et al., 1984; Bull and Jackson, 2011; Walters et al., 2014; Walker et al., 2016; Brand, 1976; Cannings et al., 2017.

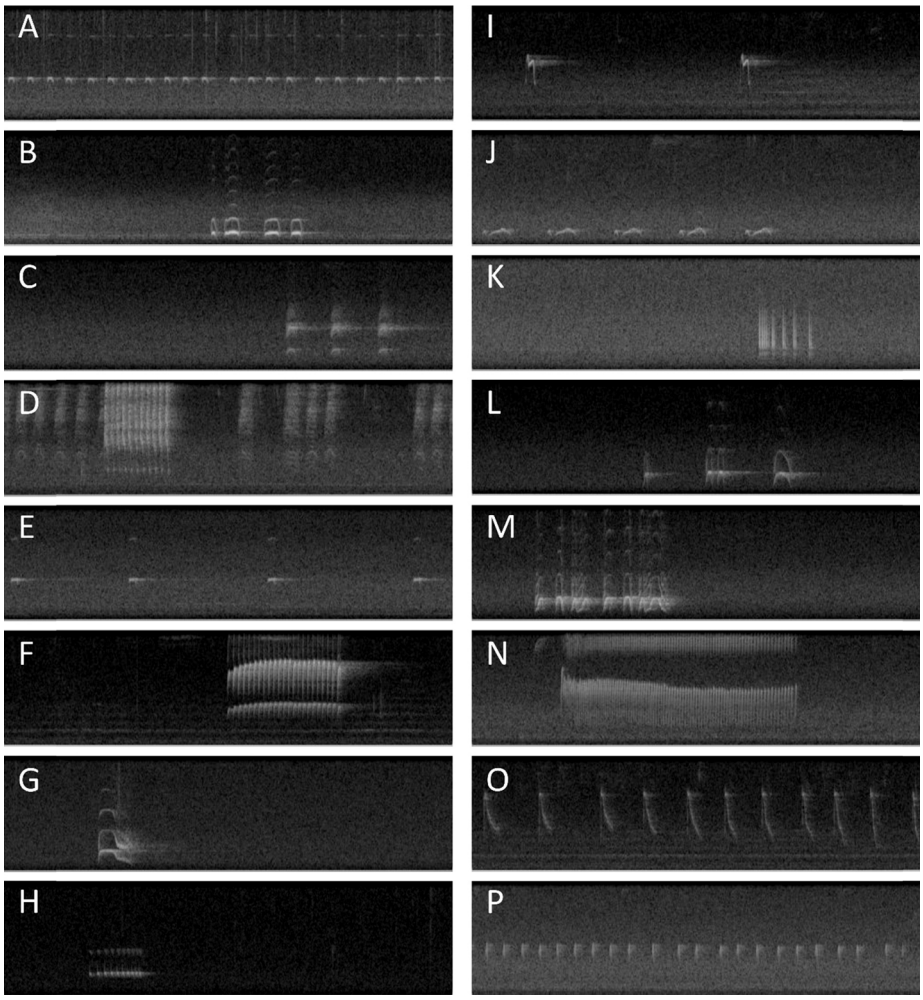


Fig. 1. Examples of spectrograms representing each of 16 target classes. Spectrograms plot sound energy as a function of time (horizontal axis) and frequency (vertical axis); lighter colors represent higher energy intensity. These spectrograms represent 12 s of sound in the frequency range 0–3000 Hz. A. Northern saw-whet owl advertising call. B. Great horned owl territorial call. C. Common raven demonstrative calls. D. Steller's jay "wek" series call and "wah" alarm call. E. Northern pygmy-owl primary call. F. Pileated woodpecker "wek" call. G. Barred owl inspection call. H. Western screech-owl "bouncing ball" call. I. Mountain quail advertisement call. J. Band-tailed pigeon song. K. Red-breasted sapsucker drum. L. Northern spotted owl four-note location call. M. Barred owl two-phrase hoot. N. Douglas' squirrel rattle call. O. Douglas' squirrel chirp call. P. Townsend's chipmunk chip call. We also used a Noise class (not pictured) for images containing only background noise or sounds that did not correspond to any target class.

regularization was the squared Euclidean norm of the weight matrix of the hidden layer, or the sum of all such squared norms, in the case of multiple hidden layers, and including the output layer. The use of regularization techniques such as L2 may have been redundant to other optimization techniques with adaptive learning rates (Loschilov and Hutter, 2019). The second fully connected layer, which was the output layer of the model, contained 17 units using sigmoid activation. Hence, the output from the model was a 17-element vector of class scores, each of which was between zero and one. Sigmoid activation is not normalized and thus the class scores that comprise the output of this CNN do not sum to one for a given input, theoretically allowing for multi-label classification.

We trained the CNN for 100 epochs using a batch size of 128 images. We measured loss using the binary cross-entropy function and used the Adam optimization algorithm (Kingma and Ba, 2015) with an initial learning rate of 0.001. To prevent overfitting, we saved the model after epochs in which validation loss decreased. We also included a stepdown function to adjust the learning rate during training: if the validation loss did not decrease by at least 0.025 for five epochs, the learning rate was reduced by half. This was followed by a cooldown period of six epochs; hence, the learning rate could diminish at a maximum rate of once every ten epochs. We implemented the cooldown period based on the observation that improvements in model performance during training are stochastic and therefore it might take several epochs to realize the potential benefit of a given learning rate. We trained the CNN using an IBM POWER8 high-performance computer running the IBM OpenPOWER Linux-based OS with two Nvidia Tesla P100-SXM2-16GB general-purpose graphics processing units.

2.3. Model testing

To evaluate the performance of the CNN, we compiled an independent test set of 131,767 images for which the correct labels were already known and which had not been part of the training or validation set (Table 1). This test set contained examples of all 17 target classes, of which approximately half ($n = 70,183$) were in the Noise class. The test set included 4082 images containing multiple target classes; of these, 4047 contained two target classes, 34 contained three target classes, and one contained four target classes. Most of the calls included in the test set were located opportunistically during human verification of output from the Ruff et al. (2020) CNN; these data had been collected in Mar – Sept 2018 in the Coast Range study area in Oregon and the Olympic Peninsula study area in Washington, USA. During the verification process we tagged images by identifying species by ear and by visually examining the spectrogram. After assembling the test set, we used the CNN to classify the images, and we calculated performance metrics based on the class scores that the CNN assigned to each image.

We calculated the same performance metrics that were reported by Ruff et al. (2020), namely precision, recall, and F1 score, and using the same definition of detection threshold, i.e. an image was considered a potential detection for a class if the class score for that image exceeded the threshold. Precision is defined as True Positives/(True Positives + False Positives) and represents the proportion of apparent 'hits' that are real detections. Recall is defined as True Positives/(True Positives + False Negatives) and represents the proportion of real examples in the dataset that are detected and correctly labeled. The F1 score is intended to measure overall performance by a balanced combination of precision

and recall, and can be weighted to emphasize either precision or recall. We calculated the unweighted F1 scores, calculated as (2 * Precision * Recall)/(Precision + Recall), and present them here for comparison with other work.

For easier comparison to other models and in keeping with recommendations by Knight et al. (2017) we present receiver operating characteristic (ROC) curves for each species, which are plots of the true positive rate (recall), calculated as [True Positives]/[True Positives + False Negatives] against the false positive rate, calculated as [False Positives]/[False Positives + True Negatives]. We calculated the Area Under the Curve (AUC) values for each species for the ROC curves, which was the probability that the model assigned a higher score to a randomly chosen positive example than to a randomly chosen negative example. We also plotted Precision-Recall (PR) curves for each species, which plot precision (True Positives/[True Positives + False Positives]) against recall (True Positives/[True Positives + False Negatives]) to illustrate a given model's performance and the magnitude of the trade-off between these metrics across a range of detection thresholds. We calculated the AUC for the PR curves; these values had no intuitive interpretation but were useful in comparing different models or comparing a model's performance on different classes. We plotted the ROC and PR curves using the PRROC package in R to interpolate values of precision, recall, true positive rate and false positive rate across a comprehensive range of threshold values.

2.4. Data processing and verification workflow

The basic data processing workflow (Fig. 2) began with a set of hour-long audio recordings in WAV files (waveform audio file format) which were typically brought in from the field on SD (Secure Digital) memory cards and transferred to external USB (universal serial bus) hard drives for processing. We generated spectrograms for each non-overlapping 12-second segment of audio in the dataset, so each hour of audio data yielded 300 images. We generated spectrograms programmatically using SoX as described for the CNN training set. After generating the full set of spectrograms for an audio dataset, we used the trained CNN to process the images and output an array of class scores for each. This array, along with the filename of the image, was written to a text file in comma-separated value format. We then used a score thresholding procedure to determine which portions of the dataset were likely to contain target species vocalizations and therefore warranted review by human technicians. We first extracted all rows with a class score ≥ 0.25 for northern spotted owl; these were considered potential spotted owl detections, regardless of the scores assigned for other classes. We then extracted rows with a class score ≥ 0.95 for any other non-Noise class, which we considered potential detections for the class with the highest score. We extracted these potential detections from the original recordings in the form of 12-second audio clips, which we sorted into directories by target class and by recording station. We further divided potential detections by time, creating a folder for each week that an autonomous recording unit was deployed. We reviewed these short audio clips and corresponding spectrograms using Kaleidoscope Pro software and assigned tags for each species detected in each clip.

We reviewed all clips flagged as potential spotted owl detections at the score threshold of 0.25 in order to maximize detections for this species, at the expense of generating more false positives. For all other classes, we verified only enough clips (3–5) to confidently establish the presence of the species at a recording station in each week. The set of clips that were reviewed and the species tags applied to each were stored as a simple comma-separated text files, allowing us to quickly generate weekly encounter histories for each species in a format suitable for use in occupancy modeling programs.

We found that technicians could comfortably review several thousand clips per day on average, depending on the number of vocal species present and other environmental factors. The level of review effort can be tailored to suit specific research questions as well as the available time and personnel. Shiu et al. (2020) recommend that detection threshold (and consequently recall) be tailored to a realistic level of review effort by examining the relationship between recall and the number of false positives produced per hour of audio. While reviewing

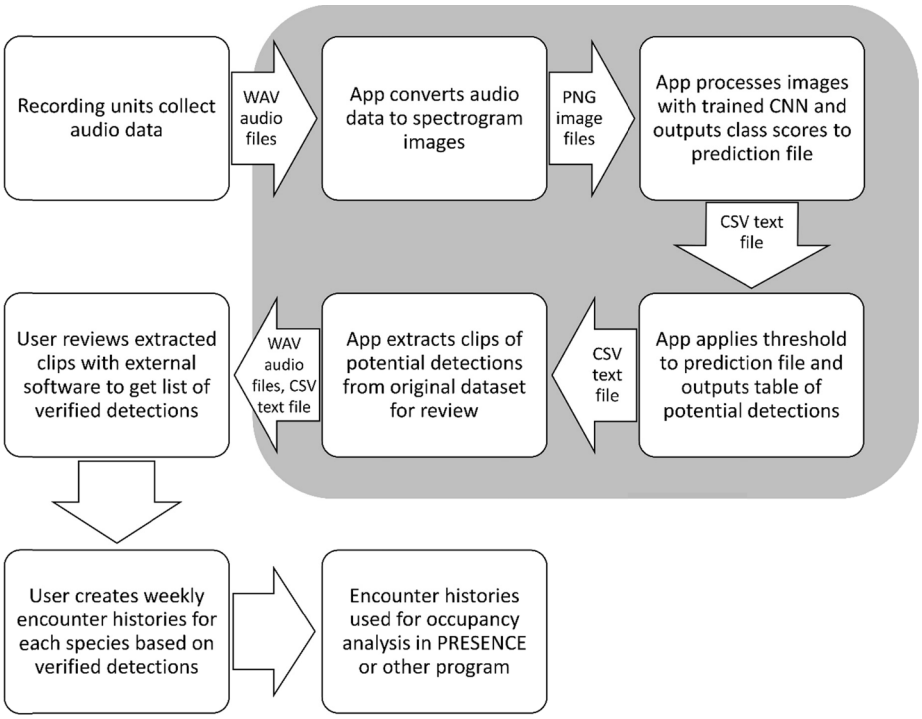


Fig. 2. Diagram of steps involved in audio processing and target species detection workflow. Processing steps in the shaded region can be performed using the desktop application provided, while those outside must be performed using external software or equipment. Arrows between boxes indicate the type and format of the data that are produced during each step and passed to the next step.

CNN output requires some specialized knowledge and skill, we found that technicians could become proficient in the most common sounds in our study areas within 1–2 weeks of supervised work.

2.5. Desktop processing

We created a desktop application to run the CNN on a personal computer through RStudio using the Shiny interface. This application is written in R but makes use of other software, primarily SoX, through system calls. Spectrograms can be generated from audio data and then classified using the pretrained CNN through a straightforward graphical user interface. Other required software includes Anaconda or Miniconda (Anaconda, Inc.). In our limited testing we obtained speeds of approximately 100 h of audio data processed per hour of processing time on inexpensive laptop computers with 2- and 4-core central processing units. See Appendix A for detailed instructions on how to install and use the desktop application to process and classify audio data and to extract clips for manual review following processing.

3. Results

3.1. Model training

Model performance as reported by the Keras training procedure continued to improve throughout training; we saved the final model configuration after epoch 100 with training loss of 0.0182, validation loss of 0.0139, training accuracy of 0.9954, and validation accuracy of 0.9969. The learning rate stepdown function was invoked a total of 10 times (i.e., as often as was possible given the patience and cooldown periods that we specified), the last being at epoch 97, and the final learning rate was 9.77×10^{-7}, a 1000-fold reduction from the initial learning rate of 0.001. The full training run took approximately 12.5 h.

3.2. Test set performance

Precision (Fig. 3), recall (Fig. 4) and F1 scores (Fig. 5) for the CNN's performance on the test set varied for the 16 non-Noise classes across a range of thresholds. Performance was generally stronger for owls than for the other species. Among the owls, precision was highest for northern saw-whet owl and northern pygmy-owl, although precision at higher thresholds exceeded 90% for all owl species except great horned owl, for which precision was noticeably lower at all thresholds (Fig. 3). Among other species, precision was highest for Townsend's chipmunk, band-tailed pigeon, Steller's jay, and common raven (Fig. 3). Precision was relatively low for Douglas' squirrel chirp call, mountain quail, and red-breasted sapsucker (Fig. 3). Precision was lowest for the Douglas' squirrel rattle call and did not exceed 50% for this class even at the highest threshold of 0.99 (Fig. 3).

Among the owl species, recall was best for spotted owl, barred owl (both call types), and northern saw-whet owl, somewhat lower for northern pygmy-owl and great horned owl, and lowest for western screech-owl across most of the range of thresholds, although recall was well above 50% for most species even at thresholds of 0.9 or more (Fig. 4). Recall for other species was less consistent and was highest for pileated woodpecker and Townsend's chipmunk, moderate for band-tailed pigeon and common raven, lower for mountain quail and Steller's jay, and lowest for red-breasted sapsucker and Douglas' squirrel (Fig. 4). Most classes showed recall above 50% at thresholds > 0.9.

The plots of F1 score versus threshold indicated that the CNN had a balanced mix of precision and recall for most owl classes across a broad range of thresholds, as demonstrated by the flatness of the curves at moderate threshold values (Fig. 5). Great horned owls had markedly better F1 scores at higher thresholds, which may be attributable to the low precision for this class across most of the range of thresholds. F1 scores peaked at low threshold values for several owls, including northern pygmy-owl, northern saw-whet owl, and western screech-owl

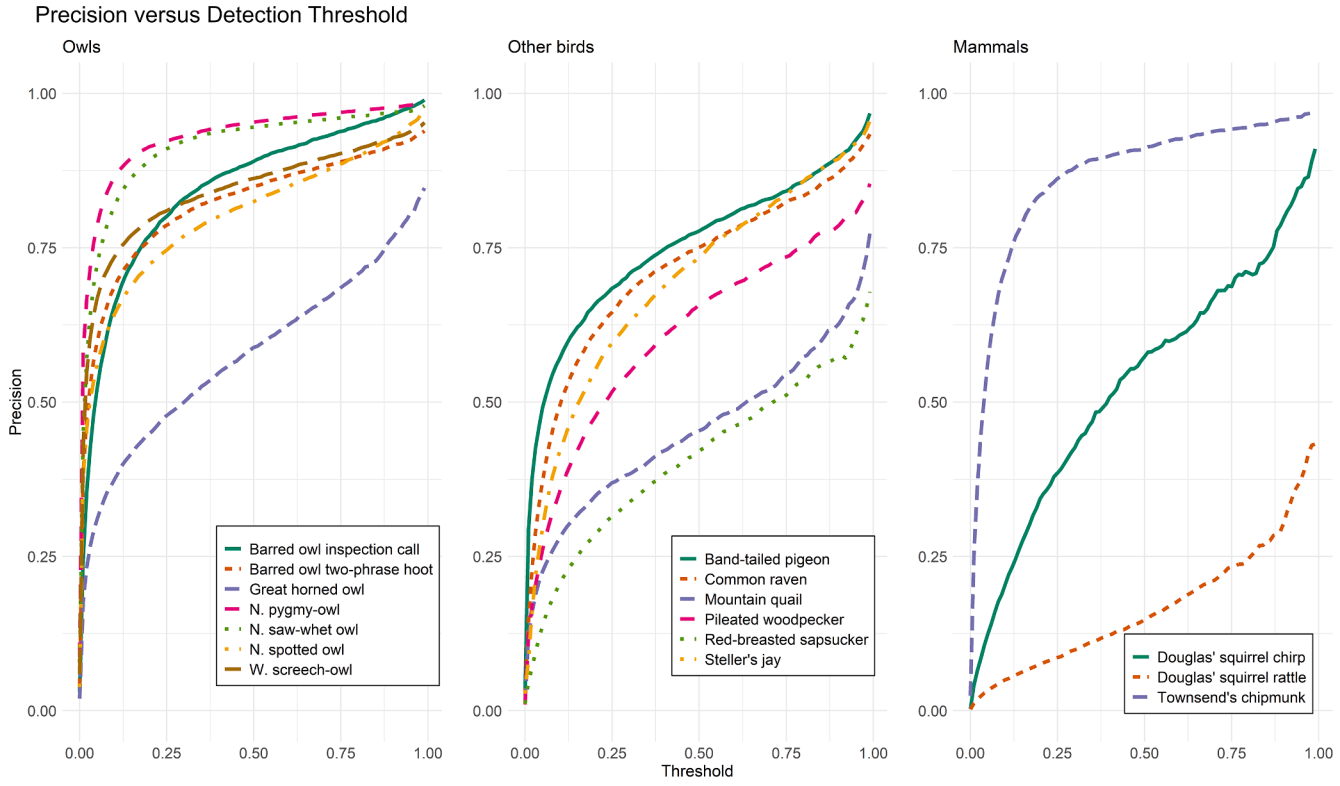


Fig. 3. Precision versus detection threshold for 16 sounds produced by 14 avian and mammalian species. Precision or specificity is calculated as [True Positives]/[True Positives + False Positives], considering only clips with class score exceeding the detection threshold for each target class. Precision represents the proportion of apparent "hits" that correspond to real instances of the class in question.

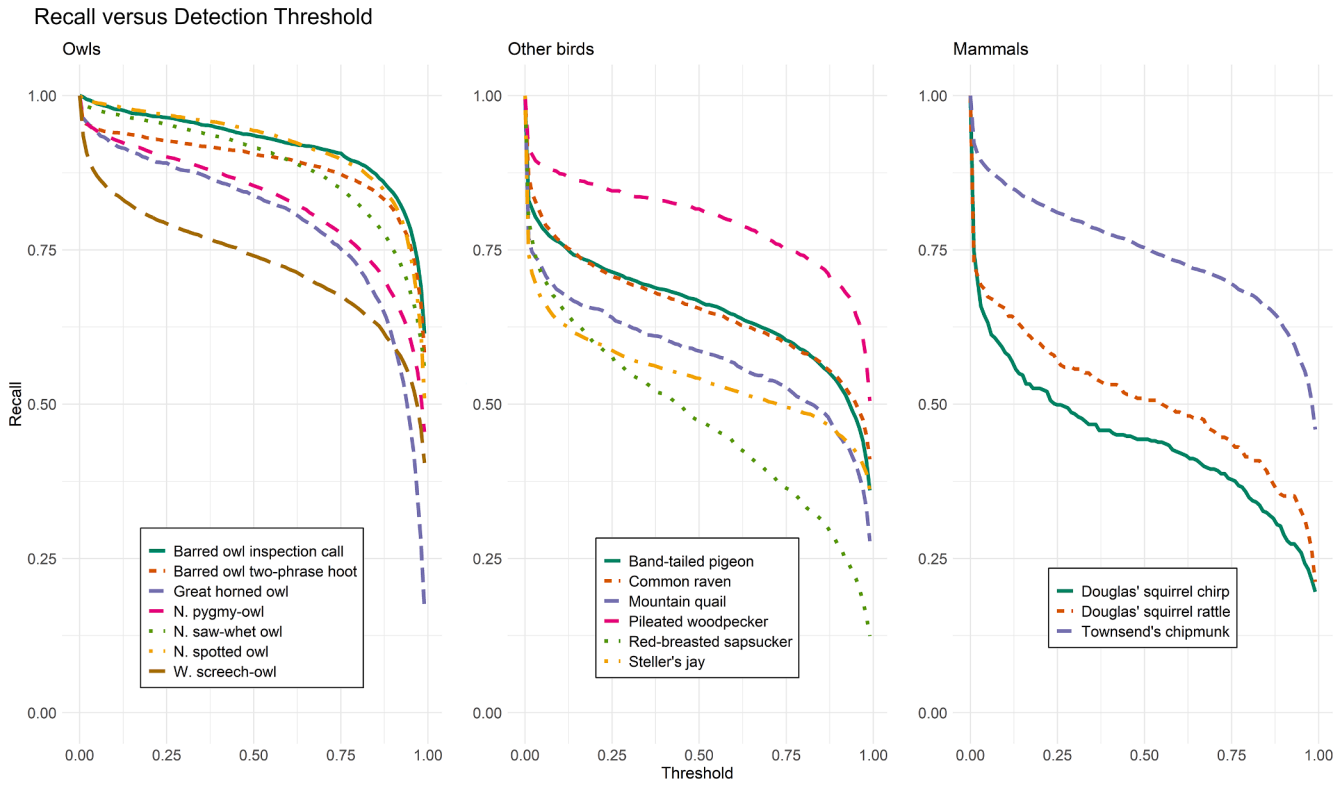


Fig. 4. Recall versus detection threshold for 16 sounds produced by 14 avian and mammalian species. Recall or sensitivity is calculated as [True Positives]/[True Positives + False Negatives], considering only clips with class score exceeding the detection threshold for each target class. Recall represents the proportion of real examples present in the dataset that are detected and correctly identified at a given detection threshold.

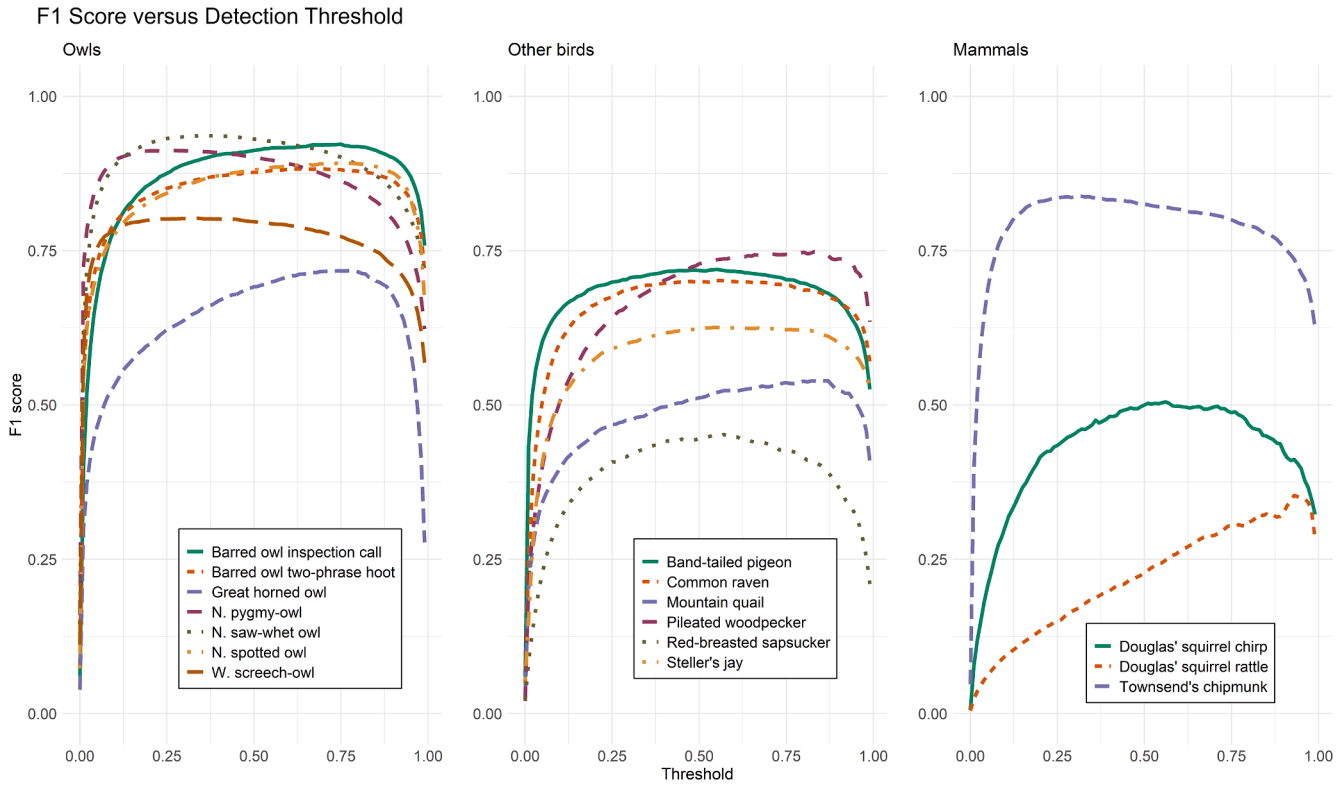


Fig. 5. F1 score versus detection threshold for 16 sounds produced by 14 avian and mammalian species. F1 score is calculated as [2*Precision*Recall]/[Precision + Recall], with both precision and recall calculated at a specific detection threshold. F1 score is intended as a balance of precision and recall and is used to gauge overall model performance.

(Fig. 5). These species had high precision at low threshold values which appeared to then be offset by diminishing recall at higher thresholds. Similar patterns were visible for non-owl avian species, although these covered a broader range of values. We observed the best F1 scores for pileated woodpecker, common raven, and band-tailed pigeon, depending on threshold (Fig. 5). Mammals showed a wide range of F1 scores; Townsend's chipmunk was comparable to the owls, while Douglas' squirrel showed low F1 for the chirp call and lower F1 for the rattle call (Fig. 5).

The ROC curves showed close to ideal performance for most owl species, good performance for several of the other bird species including band-tailed pigeon and pileated woodpecker, as well as Townsend's chipmunk, and somewhat weaker performance for common raven, Steller's jay, and both Douglas' squirrel classes (Fig. 6). The AUC values for ROC curves indicated generally good discriminative ability of the CNN across all classes. CNN performance as measured by ROC-AUC was especially high for all owls: barred owl inspection call = 0.998, barred owl two-phrase hoot = 0.973, great horned owl = 0.979, northern pygmy-owl = 0.978, northern saw-whet owl = 0.991, northern spotted owl = 0.995, and western screech-owl = 0.971. ROC-AUC values for most other species also indicated good predictive power; band-tailed pigeon = 0.892, common raven = 0.925, Douglas' squirrel chirp call = 0.849, Douglas' squirrel rattle call = 0.838, mountain quail = 0.870, pileated woodpecker = 0.946, red-breasted sapsucker = 0.870, Steller's jay = 0.818, Townsend's chipmunk = 0.962.

The PR curves were more variable than the ROC curves, but showed strong CNN performance for most owls, which was also supported by PR-AUC values: barred owl inspection call = 0.967, barred owl two-phrase hoot = 0.875, great horned owl = 0.715, northern pygmy-owl = 0.945, northern saw-whet owl = 0.961, northern spotted owl = 0.948, and western screech-owl = 0.852. The PR-AUC values for other species showed varied performance: band-tailed pigeon = 0.736, common raven = 0.720, Douglas' squirrel chirp call = 0.444, Douglas' squirrel rattle call = 0.267, mountain quail = 0.483, pileated woodpecker = 0.735, red-breasted sapsucker = 0.361, Steller's jay = 0.602, Townsend's chipmunk = 0.864 (Fig. 7).

3.3. Workflow effectiveness and usability

We used the workflow described above to process ca. 350,000 h of audio data collected in 2018 and a comparable volume of data collected in 2019. Technicians tagged 866,628 clips collected across 290 field sites in 2019, for an average of approximately 3,000 clips per site. As each clip comprised 12 s of audio, this means that technicians typically reviewed approximately 10 h of audio per site, or 0.7% of the ca. 1,500 h of audio collected at each field site over a six-week deployment. In other words, this workflow enabled us to extract meaningful information on our target species with > 99% efficiency compared to full manual review of the data collected. Assuming each technician could review approximately 5,000 clips per day, this equates to roughly two months of full-time work for a team of four technicians. Under the verification scheme described in Section 2.4, the case requiring a minimum of verification effort would be if all 16 target classes were present at all four recording stations at a field site during each week of a six-week deployment, and their vocalizations were detected with near-perfect precision. Verifying these detections would still require the examination of 16*4*6*3 = 1152 audio clips. Although the verification process clearly yielded less than ideal efficiency when applied to real data,

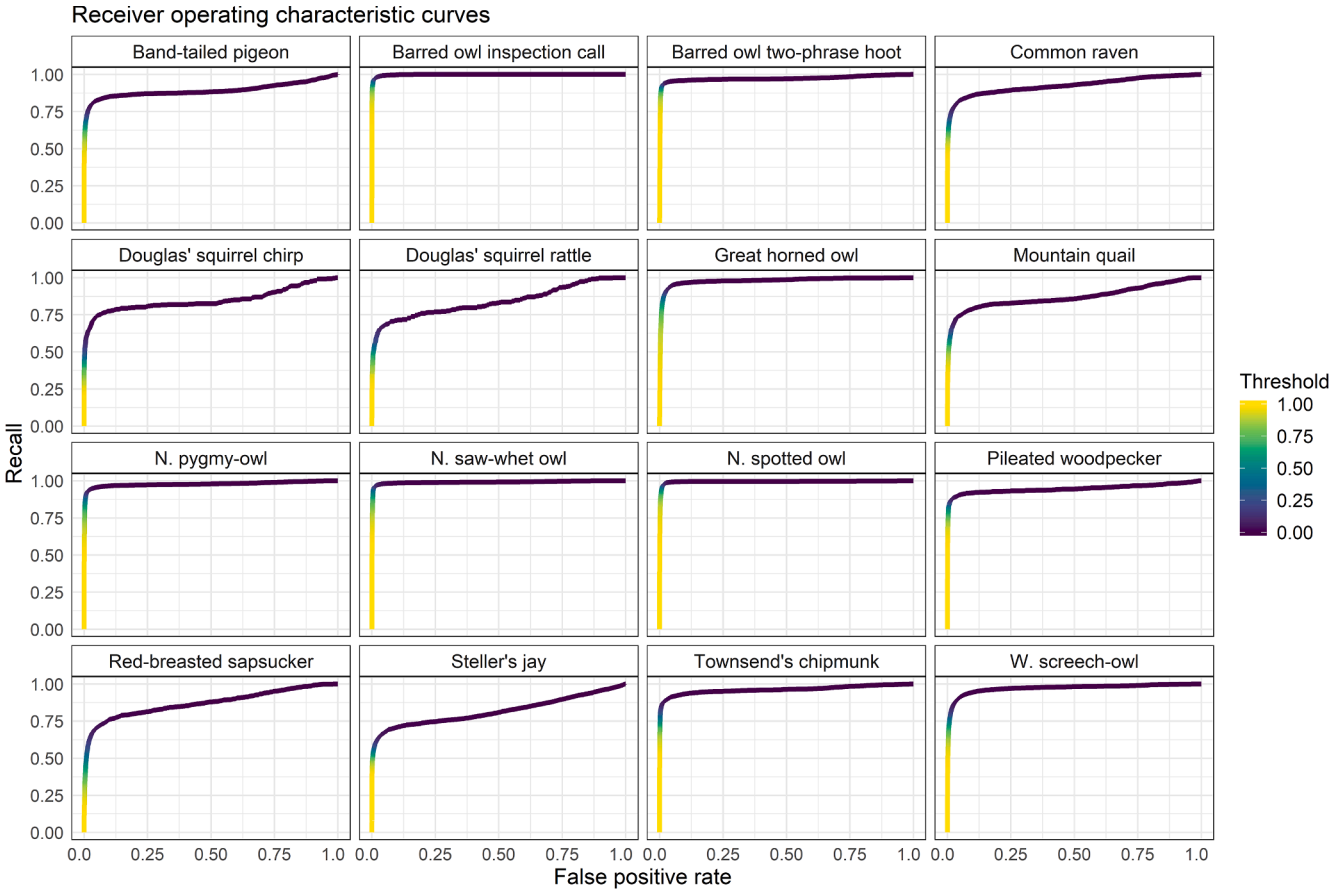


Fig. 6. Receiver operating characteristic curves for 16 sounds produced by 14 avian and mammalian species. The receiver operating characteristic curve plots the true positive rate (recall), calculated as [True Positives]/[True Positives + False Negatives] against the false positive rate, calculated as [False Positives]/[False Positives + True Negatives].

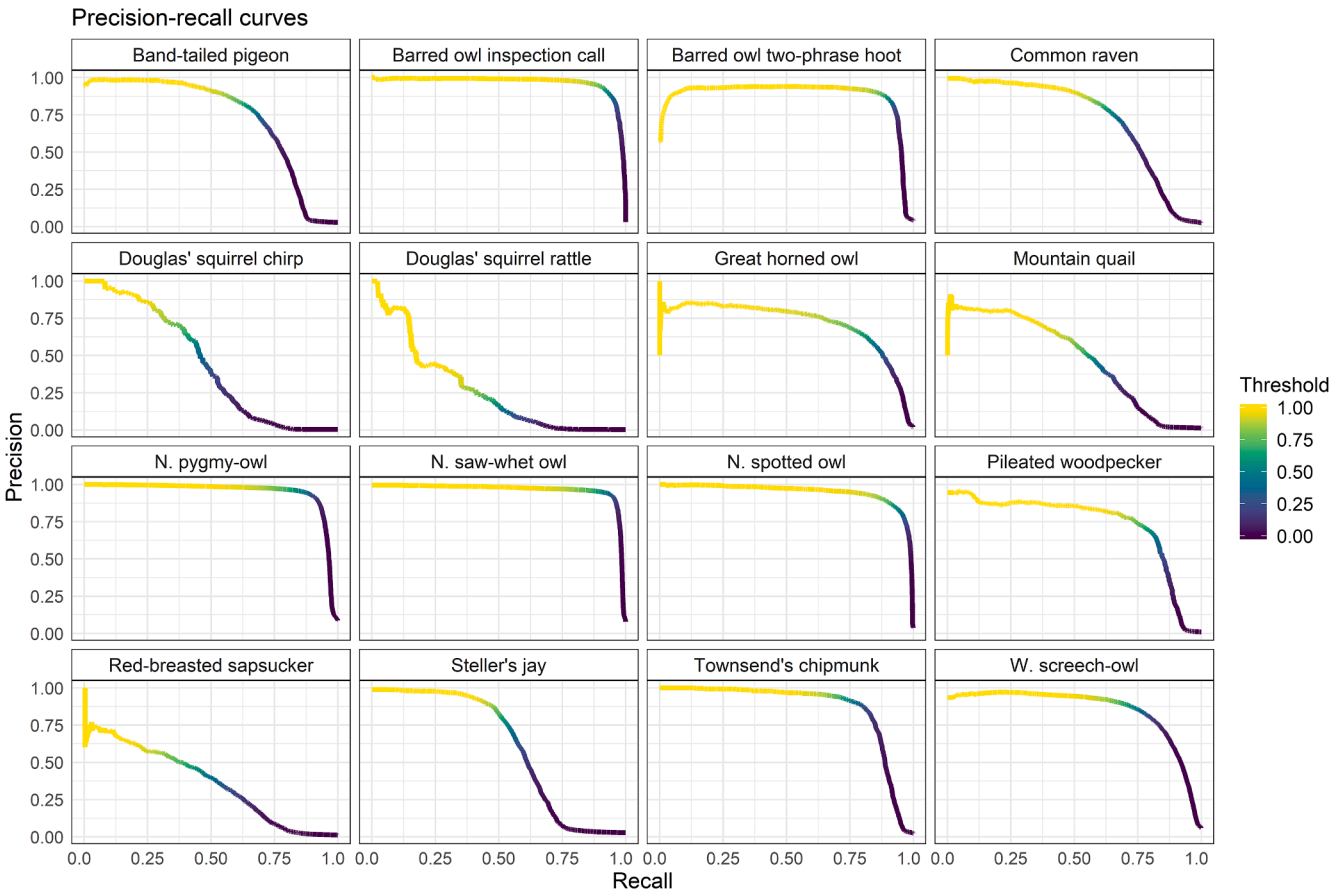


Fig. 7. Precision-recall curves for 16 sounds produced by 14 avian and mammalian species. The precision-recall curve plots precision (True positives/[True Positives + False Positives]) against recall (True Positives/[True Positives + False Negatives]) to illustrate a given model's performance and the magnitude of the trade-off between these metrics across a range of detection thresholds.

partly due to the intentionally low threshold used for northern spotted owl, we felt it still represented a reasonable investment of time. Most of these data were processed using high-performance computers, which yielded much reduced processing time relative to desktop computers. However, assuming even modest desktop processing speeds of ca. 100 h of audio per hour of processing time, the data from the 2019 season could have been processed in a matter of several weeks if spread across several inexpensive laptop computers.

4. Discussion

Here we present an original workflow designed for processing large amounts of audio data from passive acoustic monitoring for northern spotted owls and other forest-adapted bird and mammal species in the Pacific Northwest. We expanded on previous proof-of-concept work to train a CNN on 14 target species, resulting in higher performance compared to the network previously reported in Ruff et al. (2020). Additionally, we developed a way for data to be easily processed in a distributed fashion by packaging the CNN as a desktop application run through Rstudio, making the benefits of this tool available to field biologists and practitioners in a portable and user-friendly interface that requires only free and widely available software. Our workflow allowed us to generate weekly encounter histories for 14 species from approximately 1000 recording stations in a few months. These successes demonstrate the utility of our passive acoustic monitoring design and processing workflows for broad-scale, multispecies monitoring, with a quick turnaround time between data collection and analysis necessary for monitoring sensitive species like northern spotted owls. The encounter histories generated from this workflow will allow us to achieve our monitoring goals by estimating spotted owl population changes using an occupancy framework (Lesmeister et al., in press).

Our CNN showed strong performance for six owl species according to area under curve values for both the receiver operating characteristic (ROC) and precision-recall (PR) curves for each species, demonstrating substantial improvements to previous models (Ruff et al., 2020). We also incorporated eight additional species not targeted in previous models, for which performance was mixed but still strong enough to use CNN output to generate encounter histories for occupancy analyses when detections were verified by human technicians. Performance for the non-owl classes was comparable to that previously observed for six owl classes as reported by Ruff et al. (2020), suggesting that further improvements are likely with future efforts. Manual review of apparent detections by humans has the side benefit of producing training data which can be fed back into the network to improve performance in an iterative fashion by periodically retraining the model. In spite of some noted issues, the performance of our CNN was broadly comparable to that of another recent CNN which achieved precision ranging from 0.13 to 1.00 and recall ranging from 0.25 to 1.00 at a detection threshold of 0.99 for 24 classes of avian and anuran vocalizations from Puerto Rico (LeBien et al., 2020).

The effectiveness of our workflow showed several limitations which may have stemmed from the structure and behavior of the trained model, characteristics of our training set, or a combination of the two. We used sigmoid activation in the output layer of the presented CNN in the hope that it would allow for accurate classification of images containing multiple target classes. As our list of target species expands in future versions, clips containing multiple species will comprise a growing proportion of the data being processed. However, in practice,

we did not find that the CNN reliably assigned high scores to each appropriate class when multiple species were present in an image. This may be because our training set contained only singly labeled images, suggesting that effort should be made to include images with multiple correct labels—perhaps even generating them artificially by combining multiple single-class images—to train the CNN to more reliably recognize multi-class images. Alternatively, recent work has obtained strong multi-label performance using "pseudo-labeled" training data, in which each training example had one class labeled as present or absent and all other classes labeled as unknown, and a custom loss function which penalized incorrect prediction for only the labeled class (Zhong et al., 2020).

We found that recall for non-owl birds and mammals was well below 100% even at very low thresholds (e.g. 0.05), suggesting that the CNN commonly assigned a very low score to the correct class in a number of cases. This was also true for western screech-owl, though less dramatically so. A larger proportion of test images of the non-owl classes featured calls of multiple species and therefore had multiple correct labels. Because the CNN did not reliably assign high scores to every class that was present in multi-class images, the lower recall observed for non-owls may be an artifact of our test set rather than a feature of the CNN itself.

The time resolution of our spectrograms may also have limited the CNN's performance on some target classes. Precision was noticeably weak for the Douglas' squirrel rattle call even at high thresholds. This call consists of an extended sequence of rapidly repeated ($\sim$15 s^{-1}) notes. The speed of the call combined with the time resolution of our spectrogram images (500 pixels representing 12 s of recording time) means that even in cases with a high signal-to-noise ratio, individual chirps may be separated by as little as one pixel, and this separation may effectively vanish when combined with echoes, scattering, and ambient sounds. A potential solution to improve performance for this call would be to generate spectrograms at higher resolution and retrain the CNN with the larger images, but incorporating this would also decrease processing speed, as larger spectrograms would take more time to generate.

The workflow presented here allowed us to achieve our goal of using passive acoustic monitoring to monitor northern spotted owl populations, while also conducting broadscale multispecies surveys. Further, with our development of the Shiny App this model is quite portable and can be run by biologists with little computational background with relatively short turnaround time for data processing, allowing users to quickly produce encounter histories to use in occupancy analyses. While training deep CNNs is computationally intensive and benefits from high-performance computer hardware, the actual processing of audio data can be done at a reasonable speed on consumer-grade computers. Because the task of generating spectrograms can be parallelized to a substantial degree, this task makes efficient use of multi-core processors, and the availability of inexpensive 8- and 12-core central processing units makes desktop processing increasingly attractive. Classifying the resulting images is not computationally demanding and can be run at reasonable speed without relying on powerful graphics processing units. Depending on the specific hardware configuration, processing speed may be limited either by the data connection or the read-write speeds of the storage media. We obtained the best processing speeds with data stored on internal solid-state drives with high-speed data connections; however, external hard drives connected by universal serial bus still offered satisfactory performance. The desktop application also included functions for extracting rows representing potential detections from the raw results file and for extracting short clips corresponding to these rows for subsequent verification, which remained an important part of our workflow and depended on third-party software. Encounter histories could be generated directly from the CNN output using scripts or modifications to the desktop application, but some false detections may occur without verification.

We demonstrated the use of our current workflow for processing our data from broad-scale passive acoustic monitoring and suggest that this workflow may be a useful template for others designing similar projects and tools to bridge the gap between data collection and analysis. Our model was trained on 14 target species found in forests of the Pacific Northwest and may be used by other researchers interested in classifying audio recordings containing any of these target species, with expected performance comparable to what we report here, or the model structure and training procedure may serve as a starting point for other model training. Our code may serve as a starting point for those who wish to train a CNN on their own audio data, with their own list of target species. Our CNN can also be directly adapted by researchers interested in automating the detection of other species, by replacing the output layer of the CNN and retraining on a different set of training data, which allows the reuse of learned features from lower layers.

The combination of accurate automated detection and rapid human review enables highly efficient extraction of ecological information from vast quantities of acoustic data; this is a prerequisite to the successful application of passive acoustic monitoring. Moving beyond the ability to process audio on consumer-grade computers, the advantages of large-scale passive acoustic monitoring may only be fully realized when the raw data can be processed in close to real time (i.e., as they are collected) and salient results communicated quickly to biologists and managers. This may entail processing data in a distributed fashion using small, inexpensive system-on-chip processing devices coupled to the recording device, or the use of purpose-made recording devices with software for onboard processing, e.g. AudioMoth (Hill et al., 2017; Prince et al., 2019). Such distributed processing nodes could communicate potential detections to biologists remotely over mobile data networks, streamlining the process of retrieving data from the field and allowing for very rapid responses to emergent issues at field sites (e.g. the Rainforest Connection project; https://rfcx.org). However, engineering such an all-in-one solution to remain active for weeks at a time and to withstand the environmental conditions typical of many field sites is a non-trivial challenge, and these developments will require multi-disciplinary collaborations between ecologists, computer scientists, and engineers. These advancements will enhance our ability to monitor target species in close to real time. This is especially important for species such as northern spotted owls which are rare, elusive, and have habitats that are subject to land management actions with economic and ecological implications.

5. Data statement

The code used to compile and train the CNN as well as the source code for the desktop CNN app are available from the Zenodo repository at https://doi.org/10.5281/zenodo.3923030.

CRediT authorship contribution statement

Zachary J. Ruff: Formal analysis, Investigation, Methodology, Software, Writing - original draft. **Damon B. Lesmeister:** Conceptualization, Funding acquisition, Writing - review & editing. **Cara L. Appel:** Data curation, Methodology, Writing - original draft, Writing - review & editing. **Christopher M. Sullivan:** Resources, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding for this research was provided by the USDA Forest Service and USDI Bureau of Land Management, and we thank R. Davis, B. Hollen, and G. McFadden for facilitating that funding. We thank the

many field biologists and lab technicians that collected and validated much of the data presented here, including C. Cardillo, D. Culp, L. Duchac, Z. Farrand, T. Garrido, E. Guzman, A. Ingrassia, D. Jackobsma, E. Johnston, R. Justice, K. McLaughlin, A. Munes, P. Papajcik, J. Runjaic, and S. Sabin. The primary computer system used for the development of the convolutional neural network was owned and administered by the Center for Genome Research and Biocomputing at Oregon State University. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official U.S. Department of Agriculture 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. Government of any product or service.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2021.107419.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudler, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. chuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. Tensorflow: large-scale machine learning on heterogeneous systems. [online].

Artuso, C., Houston, C.S., Smith, D.G., Rohner, C., 2013. In: Great Horned Owl (*Bubo virginianus*), Version 2.0. In The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bna.372.

Boarman, W.I., and B. Heinrich. 1999. Common Raven (Corvus corax), version 2.0. in A. F. Poole and F.B. Gill, eds. The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bna.476.

Brand, L.R., 1976. The vocal repertoire of chipmunks (genus *Eutamias*) in California. Anim. Behav. 24, 319–335. https://doi.org/10.1016/S0003-3472(76)80040-1.

Brodrick, P.G., Davies, A.B., Asner, G.P., 2019. Uncovering ecological patterns with convolutional neural networks. Trends Ecol. Evol. 2523, 1–12. https://doi.org/10.1016/j.tree.2019.03.006.

Bull, E.L., and J.A. Jackson. 2011. Pileated Woodpecker (Dryocopus pileatus), version 2.0. In A.F. Poole, ed. The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bna.148.

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

Chollet, F. 2015. Keras. [Online].

Duchac, L.S., Lesmeister, D.B., Dugger, K.M., Ruff, Z.J., Davis, R.J., 2020. Passive acoustic monitoring effectively detects northern spotted owls and barred owls over a range of forest conditions. Condor 122, 1–22. https://doi.org/10.1093/condor/duaa017.

Dugger, K.M., Forsman, E.D., Franklin, A.B., Davis, R.J., White, G.C., Schwarz, C.J., Burnham, K.P., Nichols, J.D., Hines, J.E., Yackulic, C.B., Doherty Jr., P.F., Bailey, L. L., Clark, D.A., Ackers, S.H., Andrews, L.S., Augustine, B., Biswell, B.L., Blakesley, J. A., Carlson, P.C., Clement, M.J., Diller, L.V., Glenn, E.M., Green, A., Gremel, S.A., Herter, D.R., Higley, J.M., Hobson, J., Horn, R.B., Huyvaert, K.P., McCafferty, C., McDonald, T.L., McDonnell, K., Olson, G.S., Reid, J.A., Rockweit, J., Ruiz, V., Saenz, J., Sovern, S.G., 2016. The effects of habitat, climate and Barred Owls on the long-term population demographics of Northern Spotted Owls. Condor 118, 57–116. https://doi.org/10.1650/CONDOR-15-24.1.

Forsman, E.D., Meslow, E.C., Wight, H.M., 1984. Distribution and biology of the spotted owl in Oregon. Wildlife Monographs 87, 3–64.

Forsman, E.D., Anthony, R.G., Dugger, K.M., Glenn, E.M., Franklin, A.B., White, G.C., Schwartz, C.J., Burnham, K.P., Anderson, D.R., Nichols, J.D., Hines, J.E., Lint, J.B., Davis, R.J., Ackers, S.H., Andrews, L.S., Biswell, B.L., Carlson, P.C., Diller, L.V., Gremel, S.A., Herter, D.R., Higley, J.M., Horn, R.B., Reid, J.A., Rockweit, J., Schaberl, J.P., Snetsinger, T.J., Sovern, S.G., 2011. Population demography of northern spotted owls. Stud. Avian Biol. 40, 1–106. https://doi.org/10.1525/9780520950597.

Gomez Villa, A., Salazar, A., Vargas, F., 2017. Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Ecol. Inform. 41, 24–32. https://doi.org/10.1016/j.ecoinf.2017.07.004.

Gutiérrez, R.J., and D.J. Delehanty. 1999. Mountain Quail (Oreortyx pictus), version 1.0. In A.F. Poole and F.B. Gill, eds. The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bna.457.

Gutiérrez, R.J., A.B. Franklin, and W.S. Lahaye. 2020. Spotted owl (Strix occidentalis), version 1.0. In A.F. Poole and F.B. Gill, eds. Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bow.spoowl.01.

Hill, A.P., Prince, P., Piña Covarrubias, E., Doncaster, C.P., Snaddon, J.L., Rogers, A., 2017. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods Ecol. Evol. 9, 1199–1211. https://doi.org/10.1111/2041-210X.12955.

Holt, D.W., Petersen, J.L., 2000. In: Northern Pygmy-Owl (*Glaucidium gnoma*), version 2.0. In The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bna.494.

Jenkins, J.M.A., Lesmeister, D.B., Wiens, J.D., Kane, J.T., Kane, V.R., Verschuyl, J., 2019. Three-dimensional partitioning of resources by congeneric forest predators with recent sympatry. Sci. Rep. 9, 6036. https://doi.org/10.1038/s41598-019-42426-0.

Keppie, D.M., and C.E. Braun. 2000. Band-tailed pigeon (Patagioenas fasciata), version 2.0. In A.F. Poole and F.B. Gill, eds. The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bna.530.

Knight, E.C., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R.M., Bayne, E., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conserv. Ecol. 12 (2), 14. https://doi.org/10.5751/ACE-01114-120214.

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

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

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

Lesmeister, D.B., R.J. Davis, L.S. Duchac, and Z.J. Ruff. 2019. Research update on using passive acoustics to monitor northern spotted owl populations in Washington and Oregon. 2018 annual research report. USDA Forest Service, Pacific Northwest Research Station. Corvallis, OR. 21 p.

Lesmeister, D.B., C.L. Appel, R.J. Davis, C.B. Yackulic, and Z.J. Ruff. In Press. Simulating effort necessary to detect changes in northern spotted owl (Strix occidnetalis caurina) populations using passive acoustic monitoring. Research Paper PNW-RP-XXX. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Loschilov, I., and F. Hutter. 2019. Decoupled weight decay regularization. International Conference on Learning Representation 2019, New Orleans, Louisiana.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., Clune, J., 2017. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. U.S.A. 115, E5716–E5725. https://doi.org/10.1073/pnas.1719367115.

Odom, K.J., Mennill, D.J., 2010. A quantitative description of the vocalizations and vocal activity of the barred owl. Condor 112, 549–560. https://doi.org/10.1525/cond.2010.090163.

Prince, P., Hill, A., Piña Covarrubias, E., Doncaster, P., Snaddon, J.L., Rogers, A., 2019. Deploying acoustic detection algorithms on low-cost, open-source acoustic sensors for environmental modeling. Sensors 19, 553. https://doi.org/10.3390/s19030553.

Rasmussen, J.L., Sealy, S.G., Cannings, R.J., 2008. In: Northern Saw-whet Owl (*Aegolius acadicus*), version 2.0. In The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bna.42.

Rstudio Team, 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.

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, 79–92. https://doi.org/10.1002/rse2.125.

Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. 24, 279–283. https://doi.org/10.1109/LSP.2017.2657381.

Salamon, J., Bello, J.P., Farnsworth, A., Robbins, M., Keen, S., Klinck, H., Kelling, S., 2016. Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS ONE 11, e0166866. https://doi.org/10.1371/journal.pone.0166866.

Sanders, T.A. 2015. Band-tailed pigeon population status, 2015. U.S. Department of the Interior, Fish and Wildlife Service, Division of Migratory Bird Management, Washington, D.C.

Sebastián-González, E., Pang-Ching, J., Barbosa, J.M., Hart, P., 2015. Bioacoustics for species management: two case studies with a Hawaiian forest bird. Ecol. Evol. 5, 4696–4705. https://doi.org/10.1002/ece3.1743.

Shiu, Y., Palmer, K.J., 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, 607. https://doi.org/10.1038/s41598-020-57549-y.

Smith, C.C., 1978. Structure and function of the vocalizations of tree squirrels (*Tamiasciurus*). J. Mammal. 59, 793–808. https://doi.org/10.2307/1380144.

Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., Glotin, H., 2019. Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. Methods Ecol. Evol. 10, 368–380. https://doi.org/10.1111/2041-210X.13103.

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Halseth, J.M., Sweeney, S.J., Vercauteren, K.C., Snow, N.P., Di, P.A., Jesse, S., Michael, S.L., Ben, D.W., Beasley, J. C., Schlichting, P.E., Boughton, R.K., Wight, B., Newkirk, E.S., Ivan, J.S., Odell, E.A., Brook, R.K., Lukacs, P.M., Moeller, A.K., Mandeville, E.G., Clune, J., Miller, R.S., 2019. Machine learning to classify animal species in camera trap images: applications in ecology. Methods Ecol. Evol. 10, 585–590. https://doi.org/10.1111/2041-210X.13120.

Venier, L.A., Mazerolle, M.J., Rodgers, A., McIlwrick, K.A., Holmes, S., Thompson, D., 2017. Comparison of semiautomated bird song recognition with manual detection of

recorded bird song samples. Avian Conserv. Ecol. 12, 2. https://doi.org/10.5751/ACE-01029-120202.

Walker, L.E., P. Pyle, M.A. Patten, E. Greene, W. Davison, and V.R. Muehter. 2016. Steller's Jay (Cyanocitta stelleri), version 2.0. In P.G. Rodewald, ed. The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA. [Online] 10.2173/bna.343.

Walters, E.L., Miller, E.H., Lowther, P.E., 2014. Red-breasted sapsucker (*Sphyrapicus ruber*), version 2.0. In: Poole, In A.F. (Ed.), The Birds of North America. Cornell Lab of Ornithology, Ithaca, NY, USA [Online].

Weinstein, B.G., 2018. A computer vision for animal ecology. J. Anim. Ecol. 87, 533–545. https://doi.org/10.1111/1365-2656.12780.

Wiens, J.D., Dilione, K.E., Eagles-Smith, C.A., Herring, G., Lesmeister, D.B., Gabriel, M. W., Wengert, G.M., Simon, D.C., 2019. Anticoagulant rodenticides in *Strix* owls indicate widespread exposure in west coast forests. Biol. Conserv. 238, 108238 https://doi.org/10.1016/j.biocon.2019.108238.

Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., Fortson, L., 2018. Identifying animal species in camera trap images using deep learning and citizen science. Methods Ecol. Evol. 10, 80–91. https://doi.org/10.1111/2041-210X.13099.

Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J.P., Aide, T.M., 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Appl. Acoust. 166, 107375 https://doi.org/10.1016/j.apacoust.2020.107375.