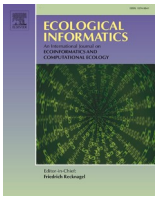
Ecological Informatics 87 (2025) 103122

Contents lists available at ScienceDirect

Ecological Informatics

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



HawkEars: A regional, high-performance avian acoustic classifier

Jan Huus [a,*], Kevin G. Kelly [b], Erin M. Bayne [b,c], Elly C. Knight [b,c]

[a] University of Alberta, Canada
[b] Alberta Biodiversity Monitoring Institute, University of Alberta, Edmonton, AB T6G 2E9, Canada
[c] Department of Biological Sciences, CW 405 Biological Sciences Building, University of Alberta, Edmonton, AB T6G 2E9, Canada

ARTICLE INFO

Keywords:
Autonomous recording unit
Bioacoustics
Bird sound recognition
Convolutional neural networks
Domain
Passive acoustic monitoring

ABSTRACT

Passive acoustic monitoring is rapidly emerging as a dominant approach for studying acoustic wildlife, with neural networks used as an increasingly common and promising approach for extracting detections of particular species from acoustic recordings. Existing options for avian classifiers include small custom models for focal species or large models that attempt to classify the entire global avian community, which suggests a possible tradeoff between classifier performance and species coverage. We argue that building domain-specific classifiers for particular geographic regions provides improved performance in exchange for reduced species coverage and present HawkEars, a regional avian classifier for Canada that includes 314 bird and 13 amphibian species. A major challenge in classifier development is the weak labeling of open access datasets. We developed a novel solution, using embedding-based search to efficiently generate strong labels. We evaluated HawkEars performance for bird species relative to two prominent avian community classifiers: BirdNET, and Perch for two datasets representing two applications: bird community surveys and studies of vocal activity rate. We found HawkEars had substantially higher performance across all metrics, detected on average two more species per recording minute in our community evaluation dataset, and had a recall of nearly twice Perch and four times BirdNET, given a precision of 0.9, for our vocal activity evaluation dataset. We suggest HawkEars provides better classification performance because a smaller species pool allows for more resources allocated per species to training and tuning and reduces the risk of class overlap, and our strong labeling method ensures high-quality training data. While our classifier, HawkEars, is a substantial improvement for practitioners studying acoustic wildlife in Canada and the northern United States, practitioners in other regions can use the HawkEars open-source code to build classifiers for other geographic regions. By continuing to improve deep-learning classification performance, HawkEars has the potential to substantially improve the efficiency and utility of passive acoustic monitoring studies.

1. Introduction

Sound is one of the primary modes of communication in the animal kingdom (Catchpole and Slater, 2008), and recent advances in acoustic recording technology have transformed the way we study acoustic animals through passive acoustic monitoring (PAM) (Gibb et al., 2018; Shonfield and Bayne, 2017; Sugai et al., 2018). PAM uses autonomous recording units (ARUs) in the field, rather than active observers, to survey for animals that communicate acoustically. Because ARUs can be left in the field for extended periods of time while recording on a pre-programmed schedule, they can allow for greater survey coverage of an area temporally, and with multiple ARUs large spatial extents can also be surveyed (Shonfield and Bayne, 2017). The recordings produced by ARUs can also be stored and re-processed in the future with new tools to identify the species on the recordings (Gibb et al., 2018; Shonfield and Bayne, 2017).

The challenge of PAM is that the recordings can require a lot of digital storage space and produce many more hours of surveys than active surveyors may produce in the same time period, leading to the challenge of how to process this large amount of data (Shonfield and Bayne, 2017; Sugai et al., 2018). Some research programs collect hundreds of thousands of hours of audio in a season (Kelly et al., 2023; Roe et al., 2021), which cannot feasibly be processed without the use of computer assisted recognizers to identify the species vocalizing. These automatic recognizers have evolved rapidly from traditional machine learning methods, which require optimal inputs and custom

* Corresponding author.
E-mail address: jhuus1@gmail.com (J. Huus).

https://doi.org/10.1016/j.ecoinf.2025.103122
Received 8 October 2024; Received in revised form 25 March 2025; Accepted 25 March 2025
Available online 29 March 2025

programming, to deep learning methods including convolutional neural networks (CNNs) (Borowiec et al., 2022), which produce much more accurate and usable results (Stowell, 2022; Xie et al., 2023). CNNs in particular have been proposed because they are able to better recognize distortions and shifts in the timing and frequency of an image pattern, such as a spectrogram of a birdsong, than previous machine-learning models (Cakir et al., 2017; Knight et al., 2017; Salamon and Bello, 2017).

Recently, deep learning classifiers have been scaled up to the entire *Aves* class, enabling, for the first time, automated processing of the entire avian community for large acoustic datasets. Two large avian audio classifiers currently exist: BirdNET (Kahl et al., 2021), which has more than 6000 species as of June 2023 and Perch, which has over 10,000 species (Google Inc, 2023). BirdNET has been shown to effectively identify the presence of bird species in large datasets (Brunk et al., 2023; Wood et al., 2024), particularly for cryptic species (Bota et al., 2023; Kelly et al., 2023). These large classifiers are typically optimized for high precision to minimize the amount of post-processing verification required to weed out false positives (Pérez-Granados, 2023). The output is therefore more likely to be sufficient for statistical applications like occupancy modeling that can account for imperfect detection (i.e., false negatives) (Balantic and Donovan, 2019; Chambert et al., 2018; Rhinehart et al., 2022); however, the recall is likely insufficient to achieve accurate parameters estimates for some applications (Pérez-Granados, 2023; Wood and Kahl, 2024). To achieve higher recall, researchers typically develop targeted classifiers, which specialize in the bird species present in a certain geographic area and have been shown to outperform the large-scale classifiers (Höchst et al., 2022; Lauha et al., 2022).

Instead of global *Aves* classifiers, classification of avian communities within a specific geographic region (i.e., domain-specific classifiers) may provide classifiers with higher recall. As with any classification exercise, performance typically declines as the number of classes in the model increases due to two mechanisms (Ho and Basu, 2002). First, the random probability of correct classification decreases with increasing number of classes. Second, the probability that two classes are highly similar also increases with an increasing number of classes. Perhaps more importantly, regional classifiers may result in higher classification performance because the training data for each species can be more carefully reviewed. Recordings available for training from sources such as Macaulay Library, Xeno-Canto and iNaturalist are weakly labeled (Michaud et al., 2023). That is, a primary species is identified per recording, but no information is given at the segment level, and secondary species may occur without being identified. Deep learning birdsong classifiers are trained on fixed-length spectrograms, for example 3 s in BirdNET (Kahl et al., 2021). Without complete segment-level labels, the classifier training data may be contaminated, causing training segments to be classified incorrectly and lowering overall classifier performance (Hanjalic et al., 2016; Schlüter, 2021; Shugaev et al., 2021).

Many methods have been developed to address the problem of unlabeled or weakly labeled source data, such as selecting random segments (Martynov and Uematsu, 2022; Schlüter, 2021; Shugaev et al., 2021), semi-supervised learning (Martynov and Uematsu, 2022; Schlüter, 2021; Shugaev et al., 2021), combining semi-supervised learning with manual labeling (Martynov and Uematsu, 2022; Schlüter, 2021; Shugaev et al., 2021), active learning (Osta et al., 2023), self-supervised learning (Wei et al., 2024) and unsupervised classification (Michaud et al., 2023). These techniques span a gradient of effort, which is generally a tradeoff with label strength and accuracy, from completely manual annotation to self-labeling (summarized by Michaud et al., 2023). For a regional scope, we suggest the ideal annotation approach is likely a semi-supervised method that optimizes this tradeoff by requiring limited manual review while quickly identifying relevant segments, even when sparsely distributed in large datasets, and excluding non-target species.

We present HawkEars, a regional avian community classifier for Canada. We developed a method for building strongly labeled training datasets using the embedding space of a preliminary model. Using this tool, we screened potential training clips for 314 bird species and 13 amphibian species from multiple sources and trained a CNN in PyTorch to classify those species. We subsequently tuned the model and incorporated several inference options to improve classification performance and efficiency. Finally, we evaluated HawkEars classification performance for birds relative to two prominent avian community classifiers: BirdNET, and Perch. We discuss the improvements of HawkEars over those classifiers and present several ways that practitioners can implement HawkEars for their own acoustic datasets in other regions. Our classifier framework provides several new and novel open-source tools for building and implementing classifiers, and provides evidence that regionally-scoped classifiers may lead to higher classification performance.

2. Methods

2.1. Geographic scope

Canada is a good candidate region for a high quality community classifier because the bird community is small enough (~450 species) to facilitate careful training and tuning attention to each species with domain-specific expertise, and because many of those species have breeding ranges that span the country, reducing the risk of class overlap (Fig. 1). Furthermore, there are large amounts of weakly labeled training data available for Canada because much of the country is remote and best monitored by passive acoustic monitoring (Barker et al., 2015; Wilgenburg et al., 2015, 2020). Many of these recordings have been annotated by expert human listeners and can be used to supplement existing sound libraries for rare species.

2.2. Training data

2.2.1. Acquisition

We sourced approximately 425,000 focal training samples for HawkEars from six sources: the Macaulay Library at the Cornell Lab of Ornithology (50.4 %; www.macaulaylibrary.org/), Xeno-Canto (44.0 %; www.xeno-canto.org), iNaturalist (3.8 %; www.inaturalist.org), WildTrax (1.0 %; www.wildtrax.ca), the Google Audioset (Gemmeke et al., 2017) and the Hamilton Bioacoustics Field Recordings (0.2 %; archive.org/details/hamiltonbioacousticsfieldrecordings), We used the Macaulay Library and Xeno-Canto as sources of bird sounds and the Google Audioset exclusively for non-avian sounds, including mammals (domestic dogs, coyotes, wolves and squirrels), amphibians, noise (wind, rain and microphone noise), and other sounds (human voices, vehicles, sirens, insects). We used iNaturalist, Wildtrax and the Hamilton Bioacoustics Field Recordings for both avian and non-avian sounds. We filtered duplicates between data sources by comparing embeddings of recordings of the same primary species and duration from a preliminary model (see "2.2.2. Selection" below).

2.2.2. Selection

From our acquired focal recordings, we selected training data using a two-stage approach: manual, then embeddings (Fig. 2). In this context, an embedding is a relatively small 1-dimensional representation of a 2-dimensional image, generated by a deep learning model (McGinn et al., 2023). Embeddings encode relevant features but not irrelevant features, such as small horizontal shifts in the case of spectrograms. As a result, the distance between embeddings is a useful measure of image similarity.

First, we manually screened a selection of focal recordings for 50 of the most common species in our candidate poll, plus the "noise" and "other sound" classes. For this purpose we randomly selected about 200 recordings per species. We plotted the spectrogram of the first 10 s of

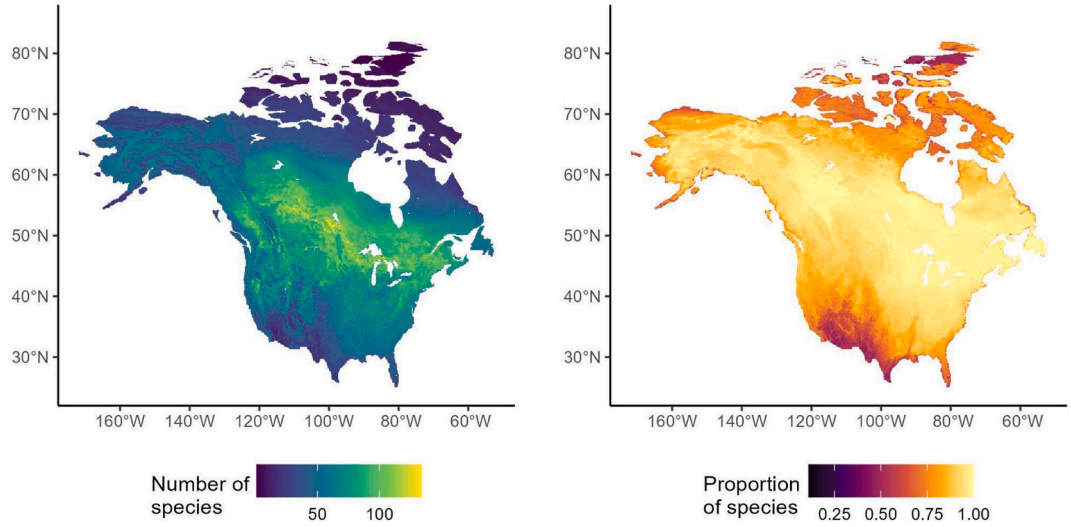


Fig. 1. Spatial distribution of number of breeding bird species (left) and the proportion of total North American breeding bird species available in HawkEars (right). Range maps sourced from eBird (Fink et al., 2023) exclude 152 species on the American Birding Association breeding bird list (Pyle et al., 2023) that are rare or difficult to model, of which 9 are available in HawkEars.

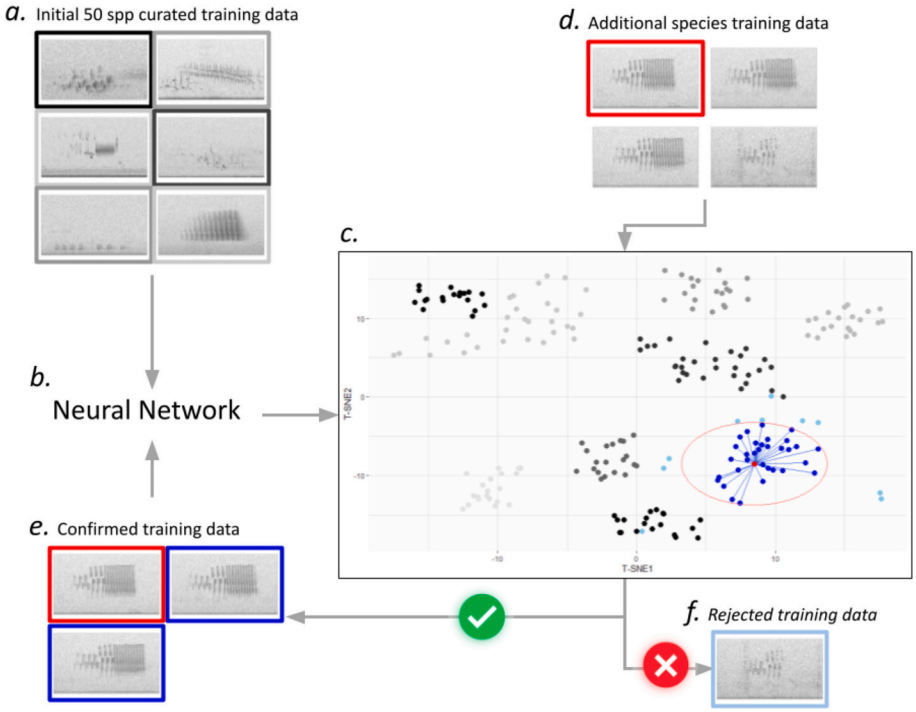


Fig. 2. Embedding-based search tool for training data selection. Manually curated training data for 50 species (a.) were used to train the initial version of the classifier model (b.) The embeddings for this initial model were plotted (c. gray-scale points) alongside the embeddings of potential training data for new species (c. dark and light blue points) and a representative spectrogram (c. red point) for each species. Potential training data within cosine distance 0.5 of the representative sample (c. dark blue points within the red circle) were manually confirmed (e. dark blue), and added to the training dataset. Those outside of 0.50 were discarded (f., light blue points). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each selected recording, chose recordings with only the target sound with minimal other noise, plotted the entire recording in 3 s spectrograms, and then selected approximately 600 of those 3 s spectrograms that met our criteria for selection (minimal noise, clear primary species and no secondary species). We then trained a preliminary model (see "2.3.1 Architecture" below for details) with the approximately 30,000 spectrograms from manual review.

Next, we selected several representative spectrograms for each bird species we wanted to include in our classifier, with multiple examples of each call/song type where available (Pieplow, 2017, 2019). We used our initial model to generate embeddings (Chen et al., 2018) for those representative spectrograms and every 3 s spectrogram of our acquired focal recordings. We measured the cosine distance of the embeddings (Singhal, 2001) between each representative spectrogram and each acquired focal recording spectrogram of that species. Finally, we plotted the closest recording spectrograms in ascending order of cosine distance up to a maximum distance (~0.50) and visually reviewed them, removing those that did not meet our selection criteria. Those remaining

were then added to the training set for the final model.

2.2.3. Preprocessing

We resampled all selected training recordings to a standard sampling rate. For stereo recordings, we selected the channel with the least noise by extracting a spectrogram from the beginning of each channel, summing the spectrogram values, and selecting the channel with a lower value unless the sum was zero, indicating a non-functioning channel. We then converted each training recording to 3-s mel spectrograms (Ghaffari and Devos, 2024) of width 384 and height 192 (Appendix A). We used an FFT window length of 2048, minimum frequency of 200 Hz and a maximum of 13 kHz. We selected 200 Hz as our minimum frequency to reduce low-frequency recorder noise and because all target species had higher frequency bands except Ruffed Grouse (*Bonasa umbellus*) drumming, for which we built a separate model (see "2.3.4. Ruffed Grouse submodel" below). We extracted separate linear-scale spectrograms for Ruffed Grouse from 0 to 200 Hz.

2.2.4. Data augmentation

Most of the training data was sourced from focal recordings in which overall noise levels were lower than typically real-world soundscapes, which can cause domain mismatch between training and test data (Merriënboer et al., 2024; Shugaev et al., 2021). We therefore applied a variety of data augmentations that are known to be effective (Kumar et al., 2024) to training data to minimize that domain mismatch (Table 1).

2.3. Model training

2.3.1. Architecture

We implemented HawkEars in Python, using PyTorch (Paszke et al., 2019) for its deep learning functionality, with models from the timm library (Wightman, 2019). We used an ensemble of primary models and a separate Ruffed Grouse sub-model, discussed below. The primary ensemble consisted of five HGNet models (Appendix A) (Wang et al., 2022). We also used an EfficientNet V2 model (Tan and Le, 2021) to generate embeddings for spectrogram searching. Although we did not use a custom model design, we did use a custom configuration of each model, to size it appropriately. We found that models with 5-6 M parameters worked best given our current number of classes, training records, and spectrogram resolution; larger models were slower in training and inference and did not improve our test results.

2.3.2. Model training

We trained the primary models on ~425,000 spectrograms, from ~200 to ~3800 per class, with a median count of ~1200 (Appendix B). We used model weights to address class imbalance by setting the weight for each class to $\sqrt{\frac{T}{N*C}}$ where T is the total number of training records, N is the number of classes and C is the record count for the class. We trained all models using an Adam optimizer (Kingma and Ba, 2014) for 16 epochs, with cosine learning rate decay and label smoothing set to 0.125.

Table 1
Data augmentations applied to training data. All selection of random variables used uniform distributions.

Augmentation	Proportion of training data augmented	Settings
Adding noise	30 %	Using training data labeled as noise
Speckling	10 %	Multiplication of spectrogram with dense Gaussian noise
Horizontal shifting	100 %	Random between −8 and 8 pixels
Merging	35 %	Combining two different classes (excluding noise classes)
Sound reduction when adding noise	20 %	Random between 20 % and 95 % (reduction of non-noise spectrogram when adding noise)
Final sound reduction	100 %	Random between 0 % and 90 % (reduction of maximum sound level after augmentation and normalization)

2.3.3. Model tuning

We performed two types of tests to tune HawkEars. First, we tested for misclassification within groups of species with similar sound structure (Appendix C). For example, a "trill" test included recordings for seven species whose songs can be described as trills, which can be hard to distinguish. We tested those species on short, focal recordings that were excluded from our training set and that contained only one or two of the target species. Second, we reviewed inference results on a diverse set of soundscape recordings from a range of locations and seasons. Here our assessments were based on a mix of factors and techniques, including our familiarity with common species, species occurrence data, comparing against confirmed recordings of the identified species, using embedding-based search to identify and review the most similar training segments, and comparing against BirdNET results on the same recording. When a test revealed low performance for a particular species, we tuned the model in two ways to improve performance. During initial model tuning, we tuned the model by adjusting parameters (e.g., spectrogram or data augmentation parameters) on small models of approximately 50 species. We also tuned the model by sourcing and adding training data to help differentiate the species with low precision or recall.

2.3.4. Ruffed grouse submodel

Ruffed Grouse (*Bonasa umbellus*) is a nonmigratory game bird, common from coast to coast in Canada (Rusch et al., 2020). Its drumming is often heard, and occurs at a very low frequency, with maximum intensity near 50 Hz. While testing our initial models, we found that the recall for this species was almost zero. To address this, we trained a binary classifier using linear-scale spectrograms from 0 to 200 Hz of 539 drumming samples and 1960 non-drumming samples, including low-frequency sounds similar to drumming collected from environmental recordings. We augmented the training data following the same protocol as for the primary models, except that horizontal shifting was the only data augmentation applied, with a probability of 20 %. We used a DLA model (Yu et al., 2018) configured to have ~1.5 million parameters.

2.4. Model inference

We ran inference using all models in the primary ensemble and estimated the score for each detection as the mean of the scores across the models. We then ran inference on low-band spectrograms with the Ruffed Grouse submodel and used the maximum of the primary ensemble score and the submodel score for Ruffed Grouse. We set the default inference to use the channel selection logic used during training data preprocessing (Section 2.1.3 above). As with other classifiers (Kahl et al., 2021), we set spectrogram overlap of 50 % (i.e. 1.5 s) as the default during inference to improve classifier performance.

Similar to BirdNET, we provided the option to ignore species based on eBird (Sullivan et al., 2009) occurrence records. That is, given a location and date, we retrieve a value representing the probability that a species is present. Species whose occurrence value falls below a configured threshold are then excluded from the output. Like BirdNET, we included the ability to set the date and location for an inference run. However, we also provided an option to include an input file during inference that specifies the latitude, longitude and date separately for each recording.

We also provided a species-pool adjustment option to use to improve recall, whereby the score threshold for a particular species is reduced when a species is detected in a user-defined number of seconds above a

particular raised threshold. With this option, the default inference settings lower the threshold for a species in a recording by 40 % when it is detected for at least 8 s of the recording with a score greater than or equal to a threshold of 0.85 (overall default threshold = 0.75).

Finally, we provided an option to use low-pass and high-pass filters during inference. Rather than applying filters to audio before creating spectrograms, we applied filters directly to the spectrograms for efficiency. Using the band-pass filters generates three sets of predictions: without filters, with only the low-pass filter and with only the high-pass filter. The inference algorithm then reports the maximum score for each class in each window. We set the default for the low-pass filter with a cutoff at ~4200 Hz and a high-pass filter with a cutoff at ~3000 Hz; however, the filters are configurable by the user.

We set the model inference default to use a GPU for processing, and provide output in the form of Audacity label files (Audacity Team, 2024).

2.5. Classifier evaluation

To evaluate the real-world performance of our classifier, we used HawkEars to process several datasets of long-duration soundscape recordings annotated by expert human listeners. We also processed those same datasets with BirdNET and Perch to compare the performance of HawkEars relative to other available community classifiers. We used version 2.4 of BirdNET from June 2023 and version 1.0 of Perch from March 2023. Since human annotation is also subject to error (Bart and Schoultz, 1984; Brauer et al., 2016; McClintock et al., 2010), we reviewed the false positives above a threshold of 0.8 for all evaluation datasets and corrected the evaluation dataset in any instances where the human observer was wrong (Knight et al., 2017). We sourced all evaluation data from publicly available datasets in WildTrax (www.wildTrax.ca), an online platform which allows users to store, manage, process and share acoustic datasets which includes manually tagged vocalizing birds within the recordings. We ran all inferences on a desktop PC with an AMD Ryzen 95,900× CPU and an Nvidia RTX 3080ti GPU (Appendix A). We ran BirdNET 2.4 and HawkEars 0.1.0 as per the instructions on their respective Github repositories and Perch 1.0 from the Open Soundscape model zoo (Lapp et al., 2023), using the default settings for all three programs. We performed threshold-invariant evaluation in Python version 3.12.3 using scikit-learn 1.5.0 (Pedregosa et al., 2012) with NumPy 1.26.4 (Harris et al., 2020) and pandas 2.2.2 (McKinney, 2010) threshold-variant evaluation in R version 4.2.3 (R Core Team, 2023). See Appendix A for additional details about test configuration and performance.

2.5.1. Community dataset

We used recordings from the Bioacoustic Unit and Alberta Biodiversity Monitoring Institute in the WildTrax database to build a dataset to evaluate classifier performance for community composition per minute of recording. We selected recordings annotated by the top three rated WildTrax observers (based on regular evaluation of human listening performance; *A. MacPhail, personal communication*). We also limited our evaluation dataset to recordings where each minute of each recording was annotated so that we could calculate our performance metrics per minute of recording effort as opposed to the entire recording. For the bird species available within those recordings, we selected at least 10 recording minutes for each species that HawkEars can classify for inclusion in our evaluation dataset. For species that did not have 10 recording minutes available, we included all recording minutes for that species (Appendix B). We excluded amphibian classes from our evaluation because most of the 13 species available in HawkEars were not available in Perch or BirdNET at the time of evaluation.

The resultant dataset contained 799 recording minutes from 623 stereo recordings in mp3 format. All recordings were collected with Song Meters (Wildlife Acoustics SM2, SM2+, SM3, SM4 units) with a minimum sampling rate of 32 kHz, and between the years 2013 and 2019. The recordings were collected both during the day (82 %), primarily at dawn, and at night (18 %) to ensure representation by both diurnal and nocturnal bird species. Similarly, most recordings were collected during the breeding songbird season in Canada from mid-May to mid-July (89 %) with the remaining recordings collected from mid-February to mid-May (11 %) to capture early breeding resident species like owls. In total, our community dataset contained 198 of the 314 species that HawkEars classifies, with anywhere from one to 229 recording minutes per species (Appendix B).

We processed the recordings from the subsequent dataset with a threshold of zero with each of the three classifiers of interest and using the default inference settings for each configuration option except the species-pool adjustment option, which we disabled to enable calculation of threshold-invariant performance metrics. We also ran HawkEars with a number of different inference settings to compare the contribution of each setting to overall performance (Appendix D). We then matched the detection windows of each of the classifiers to the minute bins of the evaluation dataset. We calculated the area under the curve (AUC) of the receiver operating characteristic (ROC), mean average precision (MAP), and macro-averaged MAP (cMAP; MAP averaged across classes to overcome class imbalance). We also calculated precision, recall, F1-score, and species richness across thresholds from 0.01 to 0.99 for each recording (Knight et al., 2017). We also calculated precision, recall, and F1-score across the same thresholds for each species. We compared the maximum F1-score per species between classifiers to evaluate the higher-performing classifier for each species in the community dataset (Appendix B).

2.5.2. Vocal activity dataset

We also compiled a dataset from WildTrax of recordings from the Bioacoustic Unit where every single detection for a specific species was annotated in each recording so that we could evaluate the performance of the classifiers for higher temporal resolution applications like vocal activity. There were nine different datasets available, each for a single species (Appendix B).

The resultant dataset contained 4106 1-min recordings. There was a mean duration of 456 recording minutes per species, with a minimum of 43 for Common Yellowthroat and a maximum of 1250 for Ruffed Grouse. The recordings were sampled across the circadian period for most target species to capture variation in calling behavior, with the exception of Common Yellowthroat, Black-throated Green Warbler, and Ruffed Grouse, which were sampled primarily at dawn. As before, most recordings were collected during the breeding songbird season in Canada from mid-May to mid-July (81 %) with the remaining recordings collected from mid-February to mid-May (19 %) to capture Barred Owl calling behavior.

We matched the 3 s windows of each of the classifiers to each detection in the vocal activity dataset. We calculated precision, recall, and F1-score across thresholds from 0.10 to 0.99 for each species.

3. Results

HawkEars outperformed both of the other classifiers across all score-

Table 2
Area under the curve (AUC) for three threshold invariant performance metrics (micro-averaged mean average precision (MAP), macro-averaged mean average precision (MAPc), and micro-averaged receiver operating characteristic (ROC)) for three multispecies avian classifiers using default settings. Metrics were derived from a test dataset of 623 three-minute recordings processed by expert human listeners for the presence of all bird species in one minute intervals.

Classifier	MAP	MAPc	ROC AUC
HawkEars	0.604	0.533	0.894
BirdNET	0.346	0.390	0.803
Perch	0.220	0.218	0.730

invariant metrics for the community dataset (Table 2). The various inference settings each added approximately 0.01 to the ROC AUC (Appendix D). BirdNET and Perch had higher mean precision across recordings at most score thresholds (< 0.67, < 0.90, respectively); however, HawkEars had higher precision and F1-score at higher thresholds and higher mean recall across all thresholds under 0.96 (Fig. 3). At the default universal threshold for HawkEars (0.75), mean precision was 0.79 and mean recall was 0.48. HawkEars achieved a maximum average recall near 0.98 at low score thresholds (0.01) and a maximum average precision of 0.99 at high score thresholds (0.95). Of the 188 species included in the community dataset, 28 had higher maximum F1-score values for BirdNET and 7 had higher maximum F1-score values for Perch. The remaining 160 species had higher values for HawkEars (Appendix B). Species richness reported by HawkEars was also consistently higher than the other two classifiers. HawkEars detected a mean of 3.29 (SD = 1.92) species per recording minute at the default threshold of 0.75. In contrast, BirdNET and Perch detected a mean of 0.66 (SD = 0.76) and 1.44 (SD = 1.14) species per recording minute at thresholds with similar precision (thresholds of 0.84 and 0.23 for precision = 0.79).

HawkEars also outperformed both of the other classifiers for the vocal activity dataset (Fig. 4). At a precision of 0.9, the mean recall across all nine species was 0.31 (SD = 0.29), 0.64 (SD = 0.20), and 0.15 (SD = 0.13) for BirdNET, HawkEars, and Perch, respectively. HawkEars performed substantially better for Ruffed Grouse in particular, for which BirdNET detected none of the 1250 drums, Perch had a maximum recall of 0.15, and HawkEars had a maximum recall 0.72.

4. Discussion

We trained a regional avian community deep-learning classifier for bird species that breed in Canada. When compared to existing global avian community classifiers, our classifier HawkEars provided a substantial improvement in performance. HawkEars had higher performance than BirdNET and Perch across all metrics for two evaluation datasets. Notably, HawkEars detected on average two more species per

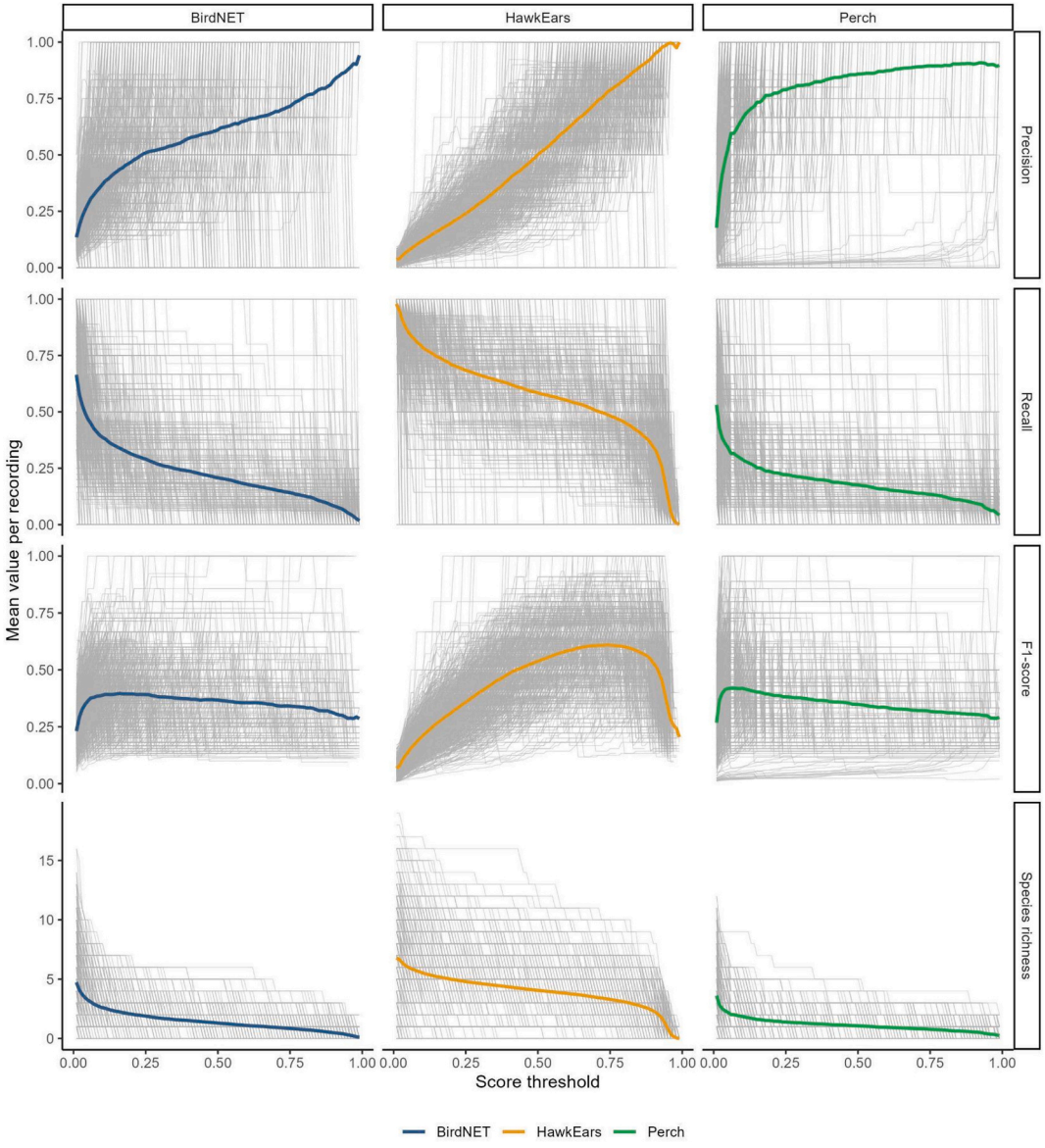


Fig. 3. Mean precision, recall, F1-score and species richness per recording minute reported by three multispecies avian classifiers. Metrics were derived from a test dataset of 623 three-minute recordings processed by expert human listeners for the presence of all bird species in one minute intervals; therefore, precision, recall and F1-score metrics should be interpreted at the resolution of one minute of recording. Colored lines represent the mean across recordings and individual gray lines represent individual recordings.

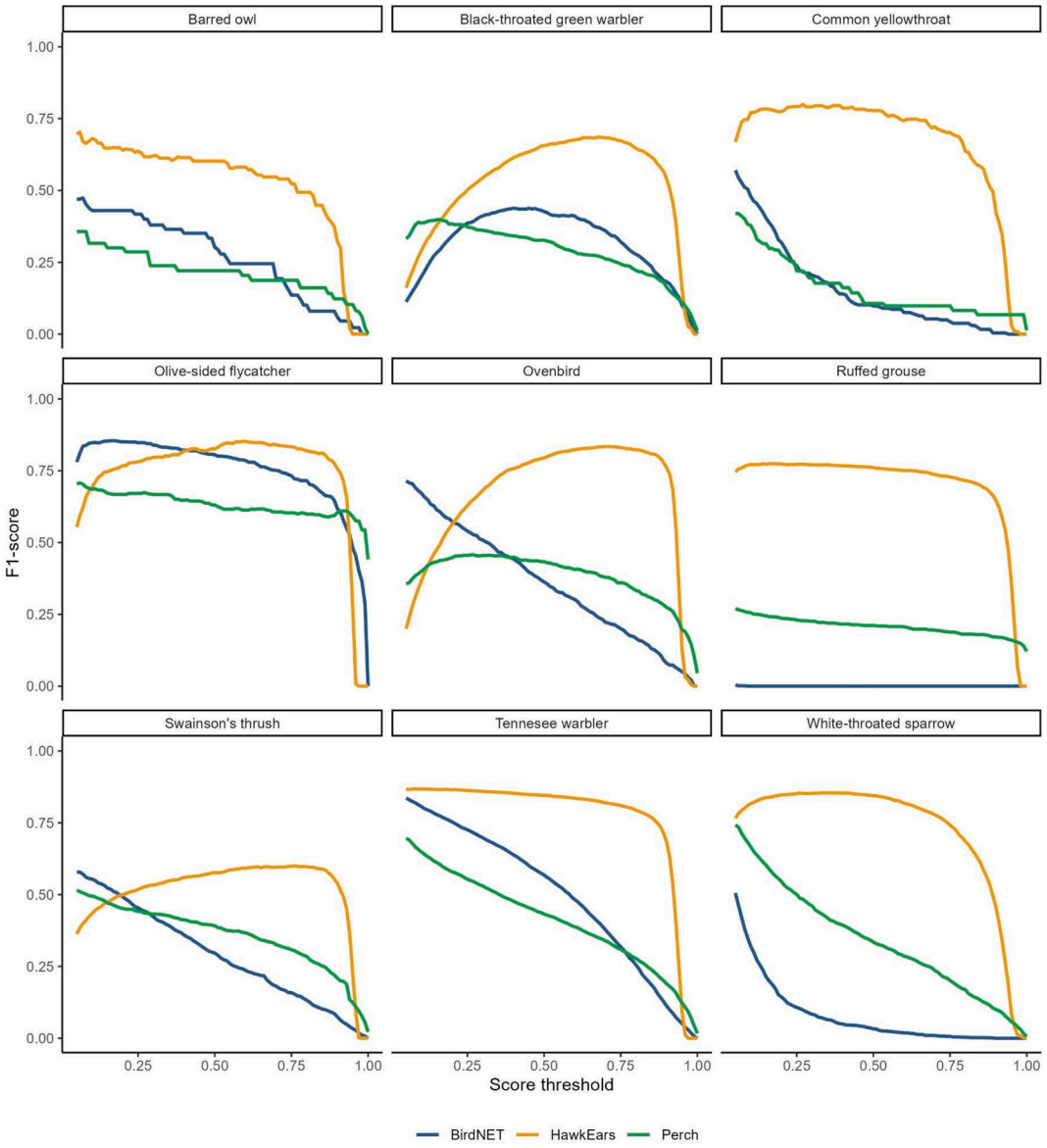


Fig. 4. F1-score reported by three multispecies avian classifiers for the detection of individual sounds of nine bird species.

recording minute than BirdNET and Perch in our community evaluation dataset, and had a recall of nearly twice Perch and four times BirdNET, given a precision of 0.9, for the nine bird species in our vocal activity evaluation dataset. We suggest these improvements in classification performance are primarily due to the regional scope of our classifier, which enabled careful attention to the acquisition of strongly labeled training data and model tuning for the 314 bird species in HawkEars. We discuss these advantages and other potential contributions to the performance of our classifier below.

We argue that HawkEars had higher classification performance than the global avian classifiers tested here because it covers a regional scope. Lauha et al. (2022) developed a relatively simple classifier for 101 bird species found in southern Finland that outperformed BirdNET despite using a simple neural network with only four convolutional layers and relatively low-resolution spectrograms. Other authors have also suggested that the species pool in these global avian classifiers is too large for optimal classifier performance (Nolan et al., 2023). A smaller suite of species could result in higher classification performance both because there is a higher random probability of correct classification, and because there is a lower probability that two classes are highly similar, which has been recognised as the most harmful issue for pattern classification ("class overlap") (Santos et al., 2022). Indeed, we found during model training that adding a dissimilar species does not degrade performance; however, adding a species that is very similar to one or more species already supported often degrades detection of the existing species. In addition to a smaller species pool, regional models are less likely to have class overlap because species are adapted to avoid acoustic competition within an ecosystem, but not necessarily across ecosystems (Duellman and Pyles, 1983; Kleyn et al., 2021; Krause, 1993). Finally, a regional scope may have a lower probability of domain shift, or change in data distribution between training and application, than a global scope due to greater similarity of geophony and anthrophony between regions (Merriënboer et al., 2024; Moreno-Torres et al., 2012), as suggested as an explanation for differences in BirdNET precision between countries (Sethi et al., 2024). Future work should use experiments of incrementally increasing species pools and geographic extents combined with measures of classification complexity (Lorena et al., 2019) to confirm our hypotheses about regional vs global classifiers, to optimize classifier size, and to explore the link between the acoustic niche hypothesis and classifier performance. Experiments examining the relationship between the number of training clips and model performance would also be useful to guide development of training datasets.

Performance has been shown to significantly improve with the iterative addition of only ~100 samples/species (Eichinski et al., 2022) and large-scale tests of CNN recognizers using few-shot learning techniques like BirdCLEF have found weak correlations between the number of training samples and recognizer performance for individual species (Kahl et al., 2022).

Likely the most important driver of improved performance that is facilitated by the regional scope of HawkEars was the time allocated to training data acquisition and model tuning. In other words, supporting fewer species allows more time per species for data selection and tuning with a given set of resources, which in turn yields better classification performance. Classification complexity is inversely related to the quality of training data, often measured in terms of sparsity and dimensionality (Ho and Basu, 2002; Lorena et al., 2019). While deep-learning classifiers can be trained with incomplete, sparse, or weakly labeled data (e.g., few-shot learning) (Fu et al., 2023; Wang et al., 2020), the classification performance may be insufficient for many ecological applications. A regional scope may also facilitate better model tuning through more resources per species, for example species-specific frequency ranges and spectrogram resolution, or adding additional training data to address common pairwise misclassifications. Improved tuning may also occur for a regional scope because the modeler can provide more insight into misidentifications if they have domain-specific knowledge of the avian community they are attempting to classify, and it is difficult to be familiar with the entire global avian community. In our case, we included a separate submodel for Ruffed Grouse because the recall for this species was near zero in our preliminary models due to the extreme low frequency of this species' acoustic signal (~50 Hz). In Perch and BirdNET, which do not have species-specific tuning, recall for Ruffed Grouse was very low or zero, respectively. The decision about whether to include a species-specific model will depend upon priorities; however, we suggest that submodels should be incorporated into regional models for species that have recall near zero in preliminary models and whose acoustic signature is outside the range of the rest of the species pool.

Training data selection is a key step in the model development process, and we find our embedding-based search method to be effective. This approach essentially creates strongly labeled data from weakly labeled datasets by removing training clips that lack relevant vocalizations or contain vocalizations from non-target species. In machine learning, this is a form of "active learning", wherein labeled data is used to assign labels to unlabeled or weakly-labeled data (Qian et al., 2017; Settles, 2011). Embedding searches can also be performed on full length recordings to build training datasets for data-sparse or rare species (Hamer et al., 2023). In fact, embeddings from large classifiers are showing promise for a variety of applications including dialect or age classification and improving classification performance (Ghani et al., 2023; McGinn et al., 2023; Tolkova et al., 2021). Other authors have used a similar iterative approach by using model scores to find new training data (Eichinski et al., 2022; Lüers et al., 2024; Michaud et al., 2023). An embeddings search can potentially also be used to improve training datasets, including labeling ambiguous vocalizations that human annotators have difficulty classifying, removing annotator bias (Osta et al., 2023), and identifying underrepresented dialects within the training dataset that may be contributing to low classification performance for particular species (Ghani et al., 2023). We caution that using this embeddings approach to build training datasets may be difficult to scale to many thousands of species for a global classifier because it is a supervised approach that requires time to find representative sounds and review the embeddings, and suggest an embeddings search could be automated as an active or self-supervised learning loop, perhaps incorporating clustering algorithms (Kath et al., 2024; Martinsson et al., 2024; Wei et al., 2024). We further suggest future application of our tool for building training datasets should start with existing avian classifiers such as HawkEars (Appendix E) instead of a custom preliminary model to remove the requirement for an initial model trained with completely supervised training data annotation.

Our evaluation is likely a fair comparison of regional vs global classifiers, given that BirdNET performance in our tests was similar to studies that used BirdNET for identification of large groups of species, but the variability in methods between studies make direct comparisons difficult. Variation in acoustic classifier evaluation is common (Knight et al., 2017), e.g. studies often differ in the time frames of the recordings used to calculate performance (e.g., recall at the recording level vs. the individual call level) and in how the minimum confidence score is chosen. The score threshold has a direct impact on the precision, recall and subsequent F1 scores and adjusting it across the entire suite of species in the study (Funosas et al., 2024), or individually for each species (Schuster et al., 2024; Sethi et al., 2024; Singer et al., 2024) will alter any possible comparisons. In addition, recall was not always measured to compare the tradeoff of maximizing precision (Sethi et al., 2024). North America based BirdNET evaluations have included fewer species but showed similarly wide ranges in precision and recall, all influenced by the confidence score chosen (Cole et al., 2022; Schuster et al., 2024). Evaluating the trade-off between recall and precision requires an understanding of the goals of the research (Knight et al., 2017; Wood and Kahl, 2024). Taking the time to evaluate the precision and recall values for each species in each study is important in order to come to relevant conclusions (Pérez-Granados, 2023) and this paper serves as one of the first independent evaluations of the recall and precision for BirdNET for an entire regional bird community and appears to be the first published evaluation of Perch for a regional bird community. One of the surprising insights of our evaluation was the peak of the F1-score curve at low score thresholds for both BirdNET and Perch, suggesting threshold-variant metrics may not be directly comparable at the same threshold across the three classifiers presented here.

For the ecologist, the improved performance of HawkEars over global avian classifiers should greatly improve the ability to use PAM to monitor and research acoustic species. Although HawkEars was developed for Canadian birds, the classifier includes the majority of the bird community for the northern United States and was trained with dialects from across the range of all species. Low coverage of the bird community in the far north is primarily driven by the abundance of breeding seabird species in that region, which do not vocalize frequently outside of breeding colonies and so were excluded from the classifier. We note that model tuning and incorporation of noise classes was done explicitly with test data from Canada, and so as with any deep-learning classifier, users in the United States should be wary of potential domain shifts (Merriënboer et al., 2024; Moreno-Torres et al., 2012). Future development should explore transfer learning from the embeddings of BirdNET and Perch with the HawkEars training dataset to mitigate potential domain shifts (Ghani et al., 2023, 2024; Kath et al., 2024). HawkEars can also be used to extract detections of 13 amphibian species and canines from ARU recordings; however, we caution that the performance of those classes has not been evaluated.

The higher recall of HawkEars over BirdNET and Perch is likely to improve ecological studies. At bare minimum, it will allow practitioners to estimate occupancy model parameters with better precision (Chambert et al., 2018). The higher recall may also be adequate for some applications that are more sensitive to false negatives like behavioral research and density estimation (Buckland et al., 2001; Pérez-Granados, 2023; Wood and Kahl, 2024). For example, density estimation approaches that rely on the predictable relationship between call rate and abundance (Pérez-Granados and Traba, 2021) may be hindered if low recall disrupts that relationship (Navine et al., 2024). Several studies have suggested that if a classifier is trained with high-amplitude clips, as HawkEars has, then the difference in recall between automated signal recognition and human listening is simply the by-product of a smaller effective detection radius (Jahn et al., 2017; Knight and Bayne, 2018; Leseberg et al., 2022). Further work is needed to test whether the assumptions of various statistical analyses are violated when using data processed by deep-learning classifiers. Finally, the higher species richness recall of HawkEars suggests that it will be more efficient for general

avian diversity surveys, requiring either less recording effort to reach the asymptote of species rarefaction curves (Wood et al., 2021) or less supplemental human annotation (Ware et al., 2023).

5. Conclusion

We conclude that building regional classifiers is likely to result in improved classifier performance over global classifiers, which in turn will likely improve statistical estimates of ecological parameters and extend the potential statistical analyses that can be conducted with PAM data processed with deep-learning classification. While our classifier, HawkEars, is a substantial improvement for practitioners studying acoustic wildlife in Canada and the northern United States, practitioners in other regions can use the HawkEars open-source code and tools like our embedding-based search tool for training data selection to build classifiers for other geographic regions (Appendix E). Existing global avian classifiers like BirdNet and Perch can also be leveraged for training regional classifiers via use of the embeddings to build training datasets and transfer learning from the embedding space to build classifiers that are robust to domain shift. The field of PAM is rapidly developing, and the training and subsequent application of custom deep learning classifiers is increasingly accessible for practitioners studying acoustic wildlife (Borowiec et al., 2022; Lapp et al., 2023; Stowell, 2022).

CRediT authorship contribution statement

Jan Huus: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Kevin G. Kelly:** Writing – original draft, Visualization. **Erin M. Bayne:** Writing – review & editing, Resources. **Elly C. Knight:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Conceptualization.

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.

Acknowledgements

We are deeply grateful to the many contributors of acoustic recordings to sound libraries across the world, without whom large bird community deep learning classifiers would be possible. We would also like to acknowledge the invaluable assistance and motivation provided by Doug Welch in the early stages of this project. Thanks to members of the Kitzes lab for their testing and feedback, and for development and support of opensoundscape.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103122.

Data availability

HawkEars is open-source and available through multiple providers. For users who want full customization, the model and instructions for inference are available via Github at https://github.com/jhuus/HawkEars. Users interested in the embedding search process, transfer learning from HawkEars, or any of the training code should also access the Github repository. Users who would like a streamlined wrapper for inference can find HawkEars within the Open Soundscape model zoo at https://opensoundscape.org/. Users who are interested in having the classifier run automatically over human-annotated recordings will be able to access it on the WildTrax platform at https://wildtrax.ca/ beginning in 2025.

References

Balantic, C., Donovan, T., 2019. Dynamic wildlife occupancy models using automated acoustic monitoring data. Ecol. Appl. 29, e01854. https://doi.org/10.1002/eap.1854.

Barker, N.K.S., Fontaine, P.C., Cumming, S.G., Stralberg, D., Westwood, A., Bayne, E.M., Sólymos, P., Schmiegelow, F.K.A., Song, S.J., Rugg, D.J., 2015. Ecological monitoring through harmonizing existing data: lessons from the boreal avian modelling project. Wildlife Soc. B 39, 480–487. https://doi.org/10.1002/wsb.567.

Bart, J., Schoultz, J.D., 1984. Reliability of singing bird surveys: changes in observer efficiency with avian density. Auk 101, 307–318.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. 13, 1640–1660. https://doi.org/10.1111/2041-210x.13901.

Bota, G., Manzano-Rubio, R., Catalán, L., Gómez-Catasús, J., Pérez-Granados, C., 2023. Hearing to the unseen: AudioMoth and BirdNET as a cheap and easy method for monitoring cryptic bird species. Sensors 23, 7176. https://doi.org/10.3390/s23167176.

Brauer, C.L., Donovan, T.M., Mickey, R.M., Katz, J., Mitchell, B.R., 2016. A comparison of acoustic monitoring methods for common anurans of the northeastern United States. Wildl. Soc. Bull. 40, 140–149. https://doi.org/10.1002/wsb.619.

Brunk, K.M., Gutiérrez, R.J., Peery, M.Z., Cansler, C.A., Kahl, S., Wood, C.M., 2023. Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests. Fire Ecol. 19, 19. https://doi.org/10.1186/s42408-023-00180-9.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., 2001. Distance Sampling: Estimating Abundance of Biological Populations. Springer. https://doi.org/10.2307/3802478.

Cakir, E., Adavanne, S., Parascandolo, G., Drossos, K., Virtanen, T., 2017. Convolutional Recurrent Neural Networks for Bird Audio Detection 1–3.

Catchpole, C.K., Slater, P., 2008. Bird song: biological themes and variations. Cambridge University Press.

Chambert, T., Waddle, J.H., Miller, D.A.W., Walls, S.C., Nichols, J.D., 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. https://doi.org/10.1111/2041-210x.12910.

Chen, H., Perozzi, B., Al-Rfou, R., Skiena, S., 2018. A Tutorial on Network Embeddings. arXiv. https://doi.org/10.48550/arxiv.1808.02590.

Cole, J.S., Michel, N.L., Emerson, S.A., Siegel, R.B., 2022. Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. Ornithol. Appl. 124, 1–15. https://doi.org/10.1093/ornithapp/duac003.

Duellman, W.E., Pyles, R.A., 1983. Acoustic resource partitioning in anuran communities. Copeia 1983, 639. https://doi.org/10.2307/1444328.

Eichinski, P., Alexander, C., Roe, P., Parsons, S., Fuller, S., 2022. A convolutional neural network bird species recognizer built from Little data by iteratively training, detecting, and labeling. Front. Ecol. Evol. 10, 810330. https://doi.org/10.3389/fevo.2022.810330.

Fink, Auer T., Johnston, A., Strimas-Mackey, M., Ligocki, S., Robinson, O., Hochachka, W., Jaromczyk, L., Crowley, C., Dunham, K., Stillman, A., Davies, I., Rodewald, A., Ruiz-Gutierrez, V., Wood, C., 2023. eBird Status and Trends, Data Version: 2022. Cornell Lab of Ornithology, Ithaca, New York.

Fu, Y., Yu, C., Zhang, Y., Lv, Danjv, Yin, Y., Lu, J., Lv, Dan, 2023. Classification of birdsong spectrograms based on DR-ACGAN and dynamic convolution. Ecol. Inform. 77, 102250. https://doi.org/10.1016/j.ecoinf.2023.102250.

Funosas, D., Barbaro, L., Schillé, L., Elger, A., Castagneyrol, B., Cauchoix, M., 2024. Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data. Ecol. Indic. 164, 112146. https://doi.org/10.1016/j.ecolind.2024.112146.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 776–780. https://doi.org/10.1109/icassp.2017.7952261.

Ghaffari, H., Devos, P., 2024. On the role of audio frontends in bird species recognition. Ecol. Inform. 81, 102573. https://doi.org/10.1016/j.ecoinf.2024.102573.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13, 22876. https://doi.org/10.1038/s41598-023-49989-z.

Ghani, B., Kalkman, V.J., Planqué, B., Vellinga, W.-P., Gill, L., Stowell, D., 2024. Generalization in birdsong classification: impact of transfer learning methods and dataset characteristics. arXiv. https://doi.org/10.48550/arxiv.2409.15383.

Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E., 2018. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 3. https://doi.org/10.1111/2041-210x.13101, 992–17.

Google Inc, 2023. Perch: Google Bird Vocalization Classifier: A Global Bird Embedding and Classification Model.

Hamer, J., Laber, R., Denton, T., 2023. CCAI Tutorial: Agile Modeling for Bioacoustic Monitoring. Google Inc.

Hanjalic, A., Snoek, C., Worring, M., Bulterman, D., Huet, B., Kelliher, A., Kompatsiaris, Y., Li, J., Kumar, A., Raj, B., 2016. Audio event detection using weakly labeled data. In: Proc. 24th ACM Int. Conf. Multimedia, pp. 1038–1047. https://doi.org/10.1145/2964284.2964310.

Harris, C.R., Millman, K.J., Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S.,

Kerkwijk, M.H., Brett, M., Haldane, A., Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Ho, T.K., Basu, M., 2002. Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. 24, 289–300. https://doi.org/10.1109/34.990132.

Höchst, J., Bellafkir, H., Lampe, P., Vogelbacher, M., Mühling, M., Schneider, D., Lindner, K., Rösner, S., Schabo, D.G., Farwig, N., Freisleben, B., 2022. Networked systems. In: 10th International Conference, NETYS 2022, Virtual Event, May 17–19, 2022, Proceedings. Lect. Notes Comput. Sci, 69–86. https://doi.org/10.1007/978-3-031-17436-0_6.

Jahn, O., Ganchev, T.D., Marques, M.I., Schuchmann, K.-L., 2017. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. PLoS One 12. https://doi.org/10.1371/journal.pone.0169041 e0169041-29.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: a deep learning solution for avian diversity monitoring. Ecol. Inform. 61, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236.

Kahl, S., Navine, A., Denton, T., Klinck, H., Hart, P., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R., Joly, A., 2022. Overview of BirdCLEF 2022: endangered bird species recognition in soundscape recordings. CLEF (Working Notes) 1929–1939.

Kath, H., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D., 2024. Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife. Ecol. Inform. 82, 102710. https://doi.org/10.1016/j.ecoinf.2024.102710.

Kelly, K.G., Wood, C.M., McGinn, K., Kramer, H.A., Sawyer, S.C., Whitmore, S., Reid, D., Kahl, S., Reiss, A., Eiseman, J., Berigan, W., Keane, J.J., Shaklee, P., Gallagher, L., Munton, T.E., Klinck, H., Gutiérrez, R.J., Peery, M.Z., 2023. Estimating population size for California spotted owls and barred owls across the Sierra Nevada ecosystem with bioacoustics. Ecol. Indic. 154, 110851. https://doi.org/10.1016/j.ecolind.2023.110851.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv. https://doi.org/10.48550/arxiv.1412.6980.

Kleyn, T., Kaizer, M.C., Passos, L.F., 2021. Sharing sound: avian acoustic niches in the Brazilian Atlantic Forest. Biotropica 53, 658–670. https://doi.org/10.1111/btp.12907.

Knight, E.C., Bayne, E.M., 2018. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics 28, 539–554. https://doi.org/10.1080/09524622.2018.1503971.

Knight, E.C., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R.M., Bayne, E.M., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conserv. Ecol. 12. https://doi.org/10.5751/ace-01114-120214 art14.

Krause, B., 1993. The niche hypothesis. Soundscape Newsl. 1–5.

Kumar, A.S., Schlosser, T., Kahl, S., Kowerko, D., 2024. Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. Ecol. Inform. 82, 102699. https://doi.org/10.1016/j.ecoinf.2024.102699.

Lapp, S., Rhinehart, T., Freeland-Haynes, L., Khilnani, J., Syunkova, A., Kitzes, J., 2023. OpenSoundscape: an open-source bioacoustics analysis package for Python. Methods Ecol. Evol. 14, 2321–2328. https://doi.org/10.1111/2041-210x.14196.

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., Ovaskainen, O., 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods Ecol. Evol. 13, 2799–2810. https://doi.org/10.1111/2041-210x.14003.

Leseberg, N.P., Venables, W.N., Murphy, S.A., Jackett, N.A., Watson, J.E.M., 2022. Accounting for both automated recording unit detection space and signal recognition performance in acoustic surveys: a protocol applied to the cryptic and critically endangered night parrot (Pezoporus occidentalis). Aust. Ecol. 47, 440–455. https://doi.org/10.1111/aec.13128.

Lorena, A.C., Garcia, L.P.F., Lehmann, J., Souto, M.C.P., Ho, T.K., 2019. How complex is your classification problem? ACM Comput. Surv. (CSUR) 52, 1–34. https://doi.org/10.1145/3347711.

Lüers, B., Serafini, P.P., Campos, I.B., Gouvêa, T.S., Sonntag, D., 2024. BirdNET-annotator: AI-assisted strong labelling of bird sound datasets. In: 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE). Vancouver, Canada.

Martinsson, J., Mogren, O., Sandsten, M., Virtanen, T., 2024. From Weak to Strong Sound Event Labels using Adaptive Change-Point Detection and Active Learning. arXiv. https://doi.org/10.48550/arxiv.2403.08525.

Martynov, E., Uematsu, Y., 2022. Dealing with class imbalance in bird sound classification. CLEF (Working Notes) 2151–2158.

McClintock, B.T., Bailey, L.L., Pollock, K.H., Simons, T.R., 2010. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. Ecology 91, 2446–2454. https://doi.org/10.1890/09-1287.1.

McGinn, K., Kahl, S., Peery, M.Z., Klinck, H., Wood, C.M., 2023. Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Ecol. Inform. 74, 101995. https://doi.org/10.1016/j.ecoinf.2023.101995.

McKinney, W., 2010. Data structures for statistical computing in Python. In: Proc. 9th Python Sci. Conf, pp. 56–61. https://doi.org/10.25080/majora-92bf1922-00a.

Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E., Denton, T., 2024. Birds, bats and beyond: evaluating generalization in bioacoustics models. Front. Bird Sci. 3, 1369756. https://doi.org/10.3389/fbirs.2024.1369756.

Michaud, F., Sueur, J., Cesne, M.L., Haupert, S., 2023. Unsupervised classification to improve the quality of a bird song recording dataset. Ecol. Inform. 74, 101952. https://doi.org/10.1016/j.ecoinf.2022.101952.

Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F., 2012. A unifying view on dataset shift in classification. Pattern Recogn. 45, 521–530. https://doi.org/10.1016/j.patcog.2011.06.019.

Navine, A.K., Camp, R.J., Weldy, M.J., Denton, T., Hart, P.J., 2024. Counting the chorus: A bioacoustic indicator of population density. Ecol. Indic. 169, 112930. https://doi.org/10.1016/j.ecolind.2024.112930.

Nolan, V., Scott, C., Yeiser, J.M., Wilhite, N., Howell, P.E., Ingram, D., Martin, J.A., 2023. The development of a convolutional neural network for the automatic detection of northern bobwhite Colinus virginianus covey calls. Remote Sens. Ecol. Conserv. 9, 46–61. https://doi.org/10.1002/rse2.294.

Osta, J.M., Dreis, B., Meyer, E., Grogan, L.F., Castley, J.G., 2023. An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (Poephila cincta cincta). Ecol. Inform. 77, 102233. https://doi.org/10.1016/j.ecoinf.2023.102233.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. https://doi.org/10.48550/arxiv.1912.01703.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: Machine Learning in Python. arXiv. https://doi.org/10.48550/arxiv.1201.0490.

Pérez-Granados, C., 2023. BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165, 1068–1075. https://doi.org/10.1111/ibi.13193.

Pérez-Granados, C., Traba, J., 2021. Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research. Ibis 163, 765–783. https://doi.org/10.1111/ibi.12944.

Pieplow, N., 2017. Peterson Field Guides to Bird Sounds of Eastern North America. Peterson Field Guides, New York, NY.

Pieplow, N., 2019. Peterson Field Guides to Bird Sounds of Western North America. Peterson Field Guides, New York, NY.

Pyle, DeCicco L., Gochfeld, D., Jaramillo, A., Kratter, A.W., Lockwood, M.W., Mutchler, M., Sibley, D., 2023. 34th report of the ABA checklist committee 2023. North Am. Birds 74, 4–9.

Qian, K., Zhang, Z., Baird, A., Schuller, B., 2017. Active learning for bird sound classification via a kernel-based extreme learning machine. J. Acoust. Soc. Am. 142, 1796–1804. https://doi.org/10.1121/1.5004570.

Rhinehart, T.A., Turek, D., Kitzes, J., 2022. A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. Methods Ecol. Evol. 13, 1778–1789. https://doi.org/10.1111/2041-210x.13905.

Roe, P., Eichinski, P., Fuller, R.A., McDonald, P.G., Schwarzkopf, L., Towsey, M., Truskinger, A., Tucker, D., Watson, D.M., 2021. The Australian acoustic observatory. Methods Ecol. Evol. 12, 1802–1808. https://doi.org/10.1111/2041-210x.13660.

Rusch, Destefano S., Reynolds, M.C., Lauten, D., 2020. Ruffed grouse (Bonasa umbellus). In: Poole, A.F., Gill, F.B. (Eds.), Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, USA.

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.

Santos, M.S., Abreu, P.H., Japkowicz, N., Fernández, A., Soares, C., Wilk, S., Santos, J., 2022. On the joint-effect of class imbalance and overlap: a critical review. Artif. Intell. Rev. 55, 6207–6275. https://doi.org/10.1007/s10462-022-10150-3.

Schlüter, J., 2021. Learning to monitor birdcalls from weakly-labeled focused recordings. CLEF (Working Notes) 1627–1638.

Schuster, G.E., Walston, L.J., Little, A.R., 2024. Evaluation of an autonomous acoustic surveying technique for grassland bird communities in Nebraska. PLoS One 19, e0306580. https://doi.org/10.1371/journal.pone.0306580.

Sethi, S.S., Bick, A., Chen, M.-Y., Crouzeilles, R., Hillier, B.V., Lawson, J., Lee, C.-Y., Liu, S.-H., Parruco, C.H.F., Rosten, C.M., Somveille, M., Tuanmu, M.-N., Banks-Leite, C., 2024. Large-scale avian vocalization detection delivers reliable global biodiversity insights. Proc. Natl. Acad. Sci. 121, e2315933121. https://doi.org/10.1073/pnas.2315933121.

Settles, B., 2011. From theories to queries: active learning in practice. In: Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010. JMLR Workshop and Conference Proceedings, pp. 1–18.

Shonfield, J., Bayne, E.M., 2017. Autonomous recording units in avian ecological research: current use and future applications, 12, p. 14. https://doi.org/10.5751/ace-00974-120114.

Shugaev, M.V., Tanahashi, N., Dhingra, P., Patel, U., 2021. BirdCLEF 2021: building a birdcall segmentation model based on weak labels. CLEF (Working Notes) 1649–1658.

Singer, D., Hagge, J., Kamp, J., Hondong, H., Schuldt, A., 2024. Aggregated time-series features boost species-specific differentiation of true and false positives in passive acoustic monitoring of bird assemblages. Remote Sens. Ecol. Conserv. https://doi.org/10.1002/rse2.385.

Singhal, A., 2001. Modern information retrieval: a brief overview. IEEE Data Eng. Bull. 24, 35–43.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152. https://doi.org/10.7717/peerj.13152.

Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D., 2018. Terrestrial passive acoustic monitoring: review and perspectives. Bioscience 69, 15–25. https://doi.org/10.1093/biosci/biy147.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: a citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006.

Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller Models and Faster Training. arXiv. https://doi.org/10.48550/arxiv.2104.00298.

Team, R.C., 2023. R: a language and environment for statistical computing.

Team, A., 2024. Audacity (R).

Tolkova, I., Chu, B., Hedman, M., Kahl, S., Klinck, H., 2021. Parsing Birdsong with Deep Audio Embeddings. arXiv. https://doi.org/10.48550/arxiv.2108.09203.

Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) 53, 1–34. https://doi.org/10.1145/3386252.

Wang, L., Yu, C., Salian, S., Kierat, S., Migacz, S., Florea, A.F., 2022. GPUNet: Searching the Deployable Convolution Neural Networks for GPUs. arXiv. https://doi.org/10.48550/arxiv.2205.00841.

Ware, L., Mahon, C.L., McLeod, L., Jetté, J.-F., 2023. Artificial intelligence (BirdNET) supplements manual methods to maximize bird species richness from acoustic data sets generated from regional monitoring. Can. J. Zoöl. 101, 1031–1051. https://doi.org/10.1139/cjz-2023-0044.

Wei, Y.-C., Chen, W.-L., Tuanmu, M.-N., Lu, S.-S., Shiao, M.-T., 2024. Advanced montane bird monitoring using self-supervised learning and transformer on passive acoustic data. Ecol. Inform. 84, 102927. https://doi.org/10.1016/j.ecoinf.2024.102927.

Wightman, R., 2019. PyTorch image models. GitHub Reposit. https://doi.org/10.5281/zenodo.4414861.

Wilgenburg, S.L.V., Beck, E.M., Obermayer, B., Joyce, T., Weddle, B., 2015. Biased representation of disturbance rates in the roadside sampling frame in boreal forests: implications for monitoring design. Avian Conserv. Ecol. 10. https://doi.org/10.5751/ace-00777-100205 art5-14.

Wilgenburg, S.L.V., Mahon, C.L., Campbell, G., McLeod, L., Campbell, M., Evans, D., Easton, W., Francis, C.M., Haché, S., Machtans, C.S., Mader, C., Pankratz, R.F., Russell, R., Smith, A.C., Thomas, P., Toms, J.D., Tremblay, J.A., 2020. A cost efficient spatially balanced hierarchical sampling design for monitoring boreal birds incorporating access costs and habitat stratification. PLoS One 15, e0234494. https://doi.org/10.1371/journal.pone.0234494.

Wood, C.M., Kahl, S., 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J. Ornithol. 1–6. https://doi.org/10.1007/s10336-024-02144-5.

Wood, C.M., Kahl, S., Chaon, P., Peery, M.Z., Klinck, H., 2021. Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. Methods Ecol. Evol. 12, 885–896. https://doi.org/10.1111/2041-210x.13571.

Wood, C.M., Socolar, J., Kahl, S., Peery, M.Z., Chaon, P., Kelly, K., Koch, R.A., Sawyer, S. C., Klinck, H., 2024. A scalable and transferable approach to combining emerging conservation technologies to identify biodiversity change after large disturbances. J. Appl. Ecol. 61, 797–808. https://doi.org/10.1111/1365-2664.14579.

Xie, J., Zhong, Y., Zhang, J., Liu, S., Ding, C., Triantafyllopoulos, A., 2023. A review of automatic recognition technology for bird vocalizations in the deep learning era. Ecol. Inform. 73, 101927. https://doi.org/10.1016/j.ecoinf.2022.101927.

Yu, F., Wang, D., Shelhamer, E., Darrell, T., 2018. Deep layer aggregation. In: 2018 IEEECVF Conf. Comput. Vis. Pattern Recognit, pp. 2403–2412. https://doi.org/10.1109/cvpr.2018.00255.