JASA ARTICLE

Automated detection of Bornean white-bearded gibbon (*Hylobates albibarbis*) vocalizations using an open-source framework for deep learning

A. F. Owens,[1] Kimberley J. Hockings,[2] Muhammed Ali Imron,[3] Shyam Madhusudhana,[4,5] Mariaty,[6] Tatang Mitra Setia,[7] Manmohan Sharma,[1] Siti Maimunah,[8] F. J. F. Van Veen,[1,a)] and Wendy M. Erb[5]

[1]*Department of Earth and Environmental Science, Faculty of Environment, Science and Economy, University of Exeter, Penryn, TR10 9FE, United Kingdom*

[2]*Centre for Ecology and Conservation, Faculty of Environment, Science and Economy, University of Exeter, Penryn, TR10 9FE, United Kingdom*

[3]*Faculty of Forestry, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia*

[4]*Centre for Marine Science and Technology, Curtin University, Perth, Western Australia, 6102, Australia*

[5]*K. Lisa Yang Center for Conservation Bioacoustics, Cornell Laboratory of Ornithology, Cornell University, Ithaca, New York 14850, USA*

[6]*Fakultas Kehutanan dan Pertanian, Universitas Muhammadiyah Palangka Raya, Palangka Raya, 73111, Indonesia*

[7]*Department of Biology, Faculty of Biology and Agriculture, Universitas Nasional, Jakarta, 12520, Indonesia*

[8]*Fakultas Kehutanan, Instiper Yogyakarta, Yogyakarta, 55281, Indonesia*

ABSTRACT:

Passive acoustic monitoring is a promising tool for monitoring at-risk populations of vocal species, yet, extracting relevant information from large acoustic datasets can be time-consuming, creating a bottleneck at the point of analysis. To address this, an open-source framework for deep learning in bioacoustics to automatically detect Bornean white-bearded gibbon (*Hylobates albibarbis*) "great call" vocalizations in a long-term acoustic dataset from a rainforest location in Borneo is adapted. The steps involved in developing this solution are described, including collecting audio recordings, developing training and testing datasets, training neural network models, and evaluating model performance. The best model performed at a satisfactory level (F score = 0.87), identifying 98% of the highest-quality calls from 90 h of manually annotated audio recordings and greatly reduced analysis times when compared to a human observer. No significant difference was found in the temporal distribution of great call detections between the manual annotations and the model's output. Future work should seek to apply this model to long-term acoustic datasets to understand spatiotemporal variations in *H. albibarbis*' calling activity. Overall, a roadmap is presented for applying deep learning to identify the vocalizations of species of interest, which can be adapted for monitoring other endangered vocalizing species. © *2024 Acoustical Society of America*. . https://doi.org/10.1121/10.0028268

(Received 16 April 2024; revised 24 July 2024; accepted 30 July 2024; published online 9 September 2024)

[Editor: James F. Lynch] Pages: 1623–1632

I. INTRODUCTION

Ever-increasing anthropogenic pressures on the environment, such as habitat loss, have led to widespread population declines in many animal species (Bender *et al.*, 1998). However, for many species, data on population trends are often sparse (Jetz *et al.*, 2019), leading to an increased demand for wildlife population monitoring programs to inform conservation responses (Verma *et al.*, 2016). To help achieve this, conservation scientists and ecologists have turned to developing technologies to automate data collection, enabling the rapid accumulation of large volumes of data (Piel and Wich, 2021). Although this has allowed for unprecedented insight, it can also make practical aspects of ecological inference challenging (Borowiec *et al.*, 2022).

Manual extraction of relevant information from large datasets can be time-consuming, resulting in a bottleneck at the point of analysis (Norouzzadeh *et al.*, 2018). This bottleneck is evident in data generated as part of passive acoustic monitoring (PAM) programs, which involve the use of autonomous acoustic sensors to collect sound recordings in the field (Acevedo *et al.*, 2009). Advancements in recording device design and cost, as well as improved data storage options, have made the task of capturing many hours of acoustic data relatively straightforward (Morgan and Braasch, 2021; Piel and Wich, 2021). Data must then be browsed to identify relevant signals of interest, such as species-specific vocalizations, often by manually listening to each recording in full or visually inspecting the data in spectrogram form (a time-frequency pictorial representation of an audio signal) or both (van Kuijk *et al.*, 2023; Clink *et al.*, 2023). PAM can provide a step-change in

a)Email: f.j.f.van-veen@exeter.ac.uk

standardized population monitoring of vocal species at high temporal resolution and simultaneous large spatial scales, which would be impossible to achieve with "traditional" methods relying on manual data collection (Sugai *et al.*, 2019). However, such PAM programs often capture datasets so large that they cannot be studied manually in full in a reasonable time frame, thus, automating this limiting data-processing step is critical (Morgan and Braasch, 2021; Clink *et al.*, 2023).

Machine learning has proven to be an effective solution for fast and accurate analysis of acoustic data, including the automated detection of signals of interest (Stowell, 2022; Miller *et al.*, 2023). There are many options available for this task, including artificial neural networks (ANNs; Mielke and Zuberbühler, 2013), Gaussian mixture models (GMMs; Heinicke *et al.*, 2015), and support vector machines (SVMs; Noda *et al.*, 2016), among others (reviewed in Knight *et al.*, 2017). These each have associated advantages, and as a result of the diversity of potential signal types and acoustic environments, no single method is optimal in all situations (Clink *et al.*, 2023). However, it is worth noting that ANNs demonstrate comparatively strong adaptability and proficiency in understanding complex patterns in data (Haykin, 2009; Goodfellow *et al.*, 2016). ANNs can be arranged in various architectures, each suited for different tasks. Early work applying ANNs to animal sound made use of the multilayer perception (MLP) architecture, where manually selected summary features, such as syllable duration, peak frequency, etc., are used to inform the network's predictions (Stowell, 2022). Although MLPs have been effective in classifying a wide variety of terrestrial and marine animal calls, the structure of nonspeech acoustic events can be highly variable (Kong *et al.*, 2017), and reducing the data to a series of manually assigned summary features can restrict the wealth of information available to train a network, potentially limiting its effectiveness (Stowell, 2022).

Deep neural networks, such as convolutional neural network (CNN) architectures, rely on feature sets that are not manually selected but are, instead, learned during the training process (Morgan and Braasch, 2021). CNNs are particularly effective for processing visual representations of audio, such as spectrograms, leveraging their ability to learn patterns that occur spatially and temporally in data. This allows CNNs to learn local features regardless of their spatial position within an image (Knight *et al.*, 2017). CNNs are, therefore, ideal candidates for the automated detection of signals within bioacoustic data, where instances of relevant features within a spectrogram are not predefined or readily identifiable (Stowell, 2022). They have been used to analyze vocalizations from a variety of taxa, including insects (Hibino *et al.*, 2021), fish (Guyot *et al.*, 2021), anurans (Colonna *et al.*, 2016), birds (Narasimhan *et al.*, 2017), marine mammals (Miller *et al.*, 2023), bats (Mac Aodha *et al.*, 2018), and other terrestrial mammals (Bjorck *et al.*, 2019), including primates (Wood *et al.*, 2023).

The potential of CNNs is far from fully realized, however (Rammer and Seidl, 2019), and there are relatively few examples of CNNs being used to answer well-defined research questions in ecology, as is so with other deep learning approaches (Dufourq *et al.*, 2021). Additionally, there are few guidelines on how to approach key steps such as model tuning and performance assessments (Knight *et al.*, 2017; Patterson and Gibson, 2017; Stowell, 2022). Further case studies reporting successful applications will advance the development of best practices for overcoming these challenges (Dufourq *et al.*, 2021).

Gibbons (family Hylobatidae) are ideal candidates for the automated detection of species-specific vocalizations. They engage in loud, highly stereotyped song bouts, which are largely confined to a few-hour window before and after sunrise (Cheyne *et al.*, 2008). During a particular calling bout, they usually emit multiple calls, which facilitates the generation of abundant training data (Clink *et al.*, 2023). Because the great call is performed largely by mated females, it is often used as an indicator of a gibbon family group, allowing for group density and spatial distribution estimates to be derived from great call densities, assuming that estimates of female calling rates are available (Cheyne *et al.*, 2016). Furthermore, gibbons reside exclusively in tropical forests, which are often visually challenging and inaccessible, therefore, studying their populations using visual methods, such as line transect and camera trap surveys, is typically very difficult (Vu and Tran, 2019). For these reasons, gibbons are model organisms for developing and testing guidelines for automated detection. So far, this has been performed for a handful of species, including the Hainan gibbon (*Nomascus hainanus*; Dufourq *et al.*, 2021), Western black-crested gibbon (*Nomascus concolor*; Zhou *et al.*, 2023), Northern gray gibbon (*Hylobates funereus*; Clink *et al.*, 2023), and southern yellow-cheeked crested gibbon (*Nomascus gabriellae*; Clink *et al.*, 2024).

Here, we apply a variation of a predefined CNN architecture, DenseNet (Huang *et al.*, 2017), to identify female great call vocalizations of the endangered Bornean white-bearded gibbon (*Hylobates albibarbis*) from a long-term acoustic dataset. We train a detector with high precision, which minimizes false-positive rates (i.e., the rate of detections incorrectly labelled as gibbon great calls), for application to large acoustic datasets recorded by PAM arrays. This can then be used to facilitate accurate population monitoring of wild gibbons on an ever-greater spatiotemporal scale and applied as a case study for developing automated detectors for other endangered vocalizing species.

II. METHODS

A. Data collection

The long-term acoustic dataset used in this study derives from the Mungku Baru Education and Research Forest (MBERF), a $\sim$50 km^2 area of tropical rainforest in Central Kalimantan Province, Indonesia. The MBERF lies in the center of the wider Rungan Landscape, which covers approximately 1500 km^2 between the Kahayan and Rungan rivers north of the provincial capital of Palangka Raya. This

https://doi.org/10.1121/10.0028268

represents the largest area of continuous unprotected lowland rainforest remaining on the island of Borneo (Purnama and Afitah, 2021). There is an estimated population of 4000 white-bearded gibbons in the Rungan Landscape and an estimated density of 2.79 groups per km^2 in the MBERF, making the region significantly important for the conservation of the species (Buckley *et al.*, 2018).

Eight autonomous recording units (ARUs; Song Meter SM4, Wildlife Acoustics, Maynard, MA) were deployed here by W.M.E. in July 2018 (see Fig. 1). These were placed on trees, 5 m above the ground, in a dispersed grid with approximately 1200 m between each device. The MBERF contains a mosaic of different forest types, and the array was designed to capture this heterogeneity with three ARUs situated in "kerangas" (heath) forest, three in "low pole" peat swamp forest, and two in "mixed swamp" forest, where the latter represents a transition habitat between the former two (Buckley *et al.*, 2018).

The ARUs were designated to record from 4 am to 6 pm (local time) daily to capture the full predawn and diurnal period of ape calling and set to default settings [sensitivity of -35 ± 4 dB (0 dB = 1 V/pa at 1 kHz), dynamic range of 14–100 dB sound pressure level (SPL) at 0 dB gain, microphone gain of 16 dB, and inbuilt preamplifier gain of 26 dB] and recorded on two channels with a sampling rate of 24 kHz. Audio was captured in 16-bit waveform audio file format (WAV) and saved as 1-h files. Memory cards and batteries were changed every two weeks.

B. Manual annotation

Manually annotated training and testing datasets were required to develop the automated detector. To create these samples, recordings between 4 and 10 a.m. were selected from a single day every four weeks from a randomly selected device for each habitat (see supplementary material A). This covers the temporal period in which most *H. albibarbis* great calls occur (Cheyne *et al.*, 2008) and ensured that a variety of potential sound environments (i.e., capturing spatial and temporal variation) were included as training inputs to the model, improving its ability to generalize over a wider range of applications. The resultant subset contained 300 h of recordings, covering 50 days spanning from October 2018 to December 2019.

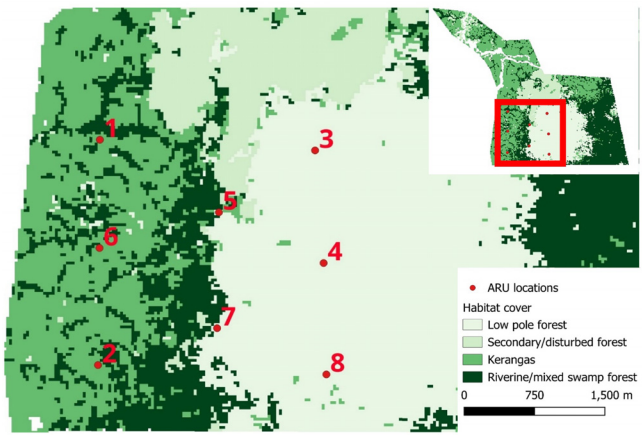

FIG. 1. (Color online) Map of the Mungku Baru Education and Research Forest (MBERF; Buckley *et al.*, 2018) shows the distribution of different habitat types over the survey area and the location of the ARUs.

The selected sound files were loaded into the sound analysis software Raven Pro 1.6 (K Lisa Yang Center for Conservation Bioacoustics, 2024) and visualized as spectrograms using a 3462-sample Hann window with a 90% overlap and a 4096-sample discrete Fourier transformation. With assistance from a team of undergraduate interns (see the Acknowledgments), each recording was listened to in full and visually scanned to identify instances of great calls. These were defined as vocalizations containing introductory, climax, and descending phrases (see Fig. 2). A selection was created for each instance by drawing a box around the call in the spectrogram, providing information about its time-frequency boundaries.

Each selection was then annotated based on its completeness and quality. A selection was marked as "clear" when the entirety of the call could be heard and was visually clear in the spectrogram, "faint" when the whole call could be heard but was not fully revealed in the spectrogram (and vice versa), and very faint when the call was only partially seen and heard (i.e., part of the great call was not captured in the recording). Each selection was reviewed and edited where needed by A.F.O. to prevent interobserver bias. In total, 1611 great calls were annotated.

The manually annotated data were then randomly split with 70% allocated for training (210 h and 1089 calls) and 30% allocated for testing (90 h and 522 calls). Due to the ambiguous nature of "very faint" calls, they were removed from the training process and used solely for testing thereafter. This prevented misleading information, i.e., nontarget events, from being fed into the positive class weightings. Models were trained using clear and faint instances (729 calls) as well as only clear instances (522 calls).

C. Automated detection

The development and testing of the automated detector used Koogu (version 0.7.2), an open-source framework for deep learning from bioacoustic data (Madhusudhana, 2023). Koogu offers a variety of functions for deep learning, including (1) preparing audio for use as input to machine learning models, (2) training models, (3) assessing model performance, and (4) using trained models for the automated analysis of large datasets. For a full workflow describing how the following steps were implemented within Koogu, see supplementary material B.

1. Data preparation

All audio files were first down-sampled to 4500 Hz to reduce the overall file size and improve the efficiency of downstream computations (cf. Miller *et al.*, 2023). The resulting Nyquist frequency (2250 Hz) is above the highest frequency within the great call selections (2077 Hz), and,

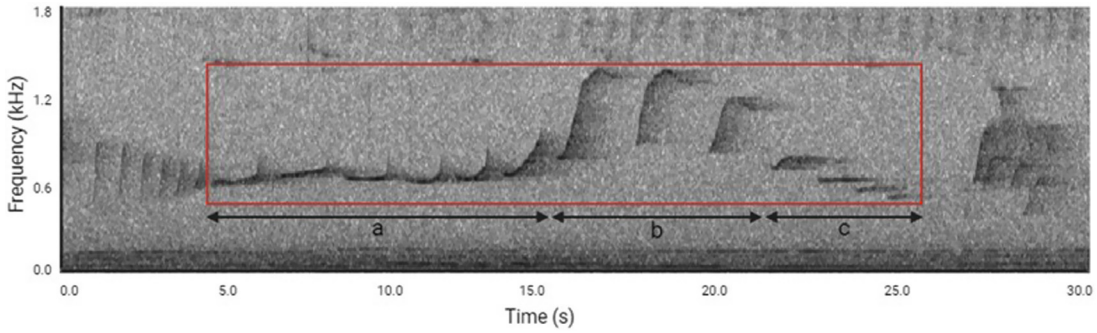

FIG. 2. (Color online) A spectrogram image of a female *H. albibarbis* great call, created using Raven Pro 1.6. The box represents the time and frequency boundaries of a manually annotated selection. The time boundaries span from the start of the first introductory (a) note to the end of the last descending (c) note. The frequency boundaries span from the lowest frequency descending (c) note to the highest frequency climax (b) note.

hence, no relevant information was lost in this process. The recordings were then split into consecutive 28 s segments (longer than the longest manually annotated great call at 27.7 s) with a hop size of 1 s, leaving an overlap of 27 s between clips. The waveform of each segment was normalized by scaling the amplitudes to occur in the range –1.0–1.0.

The resulting start and end times of each segment were then compared to those from manual annotations. Segments that fully contained the temporal extents of an annotated great call were considered as positive inputs while segments with partial overlap were excluded from training as these could resemble non-great call events and, therefore, lead to uncertainties during the training process. Segments without temporal overlap were considered as negative inputs (i.e., background noise).

Spectrograms for the positive and negative classes were then computed with an analysis window of 0.192 s and a 75% overlap. The bandwidth was also restricted to between 200 and 2200 Hz to exclude noise outside of the target frequency range. This resulted in input spectrograms with a shape of 384×580 (height $\times$ width) pixels. To address the imbalance between positive and negative classes, the maximum number of training inputs for each class was reduced to 10 000. This applied all the positive class inputs while randomly subsampling inputs from the negative class. Following this, there were 5763 clear and 1253 faint positive class spectrograms as well as 10 000 negative class spectrograms.

2. Data augmentation

Despite manually annotated calls showing a high variance in background noise, call duration and note length, for example, data augmentation was applied to further improve input variance. To do this, several predefined augmentations supported by Koogu were applied on waveforms before conversion into spectrograms and the spectrograms themselves. These were performed at each epoch, i.e., each time the model passed through the entire training dataset, during the feeding of inputs into the model. This meant that the same original sample could have different levels of augmentations between epochs.

First, Gaussian noise (Schlüter and Grill, 2015) was added to 25% of the training input's waveforms at each epoch to simulate varying levels of background noise. The amount of noise added randomly varied from –20 dB to –30 dB below the peak dB of the input signal.

The spectrogram was then smeared and squished along the time axis (cf. Madhusudhana, 2023). These augmentations were independently added to 20% of the input spectrograms each epoch at a magnitude of –1,1 (smearing backward and forward by one frame of the spectrogram) and –2,2 (stretching and squishing over up to two frames of the spectrogram). This process essentially blurred inputs along the time axis because while the target calls were contained within a standard frequency range, the duration of the signals was highly variable.

Finally, Koogu's "AlterDistance" augmentation was applied to 25% of the spectrogram inputs. This aimed to mimic the effect of increasing or reducing the distance between the calling gibbon and the receiver by attenuating or amplifying higher frequencies while keeping lower frequencies relatively unchanged. This was applied by a random factor between −5 dB (attenuation) and 5 dB (amplification).

3. Network parameters and training

The DenseNet architecture was chosen as the base CNN architecture for this study as it has been shown to achieve comparatively high accuracy with fewer parameters, making it efficient in terms of computational resources (Huang *et al.*, 2017). Early variations of the model suffered from overfitting, occurring when the model learns noise or random fluctuations in the training data rather than the underlying pattern itself. This occurs when the model is too complex relative to its intended task. Bearing this in mind, the standard DenseNet architecture was adapted to a "quasi-DenseNet" architecture (Madhusudhana *et al.*, 2021), which reduces the number of connections within each dense block, limiting model size and complexity. To limit the model's complexity further and improve computational efficiency, bottleneck layers were also added (cf. Huang *et al.*, 2017). Finally, batch normalization was enabled to improve model convergence. For the final model architecture, see Fig. 3.

https://doi.org/10.1121/10.0028268

Training inputs were then divided further with 15% randomly selected as a validation set to evaluate the model's performance throughout the training process. Dropout layers were added (Srivastava *et al.*, 2014) at a rate of 5% to further reduce overfitting and improve generalization. The models were then trained over 80 epochs using the Adam optimizer (Kingma and Ba, 2017) with a minibatch size of 24. The learning rate was initially set at 0.01 and then reduced successively by a factor of 10 at epochs 20 and 40.

4. Testing

Trained models were then applied to the test dataset to provide a preliminary assessment of model performance and establish a desirable detector threshold value. To do this, each test segment was assigned a confidence score by the model between zero (lowest) and one (highest), indicating how likely it was to contain a great call. Performance scores were outputted for thresholds at an interval of 0.01, calculating the number of true positives (TPs), false positives (FPs), and false negatives (FNs). If there was a 100% overlap between a segment and an annotated great call and the confidence score was above the threshold, it was marked as a TP. If there was no or partial overlap and the score was above the threshold, it was marked as a FP. If there was full overlap but the confidence score was below the threshold, then it would be marked as a FN.

These quantities were used to compute recall [Eq. (1)], precision [Eq. (2)], and F score [Eq. (3)] at each threshold (P, precision; R, recall):

$$R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}), \tag{1}$$

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}), \tag{2}$$

$$F = 2(P \times R)/(P + R). \tag{3}$$

The optimal threshold was then selected using maximum F score as it is a good indicator of overall model performance (Clink *et al.*, 2023).

Once a threshold had been decided, the model was rerun on the test dataset audio files to produce detections and analyze the models' output. Neighboring segments for which scores were above the identified threshold were grouped to form a single detection. The score of each detection was then set as the maximum of its component segment scores, and its start and end times were set to the start time of the first segment and the end time of the last segment, respectively. These detections were then outputted in the format of Raven Pro selection tables.

D. Post-processing

Initial inspection of the models' outputs showed that they produced clusters of detections for each great call, overestimating the number of calls within the data. Detections that overlapped with the first detection in each cluster were grouped together (see supplementary material C). These groups were then filtered to retain only the highest-scoring detection(s) within each group to minimize duplicate detections for single great call events. When the start times of the remaining selections were the same, only one detection was retained. There was no further judication between remaining detections when their scores were tied as this could indicate that two great calls overlapped or occurred close in time.

Finally, the non-post-processed and post-processed model outputs of the best performing model were evaluated against the manually annotated test dataset. To do this, we compared the distribution of total detections every 5 min across the 4–10 a.m. period using a Kolmogorov–Smirnov test. By comparing the non-post-processed and post-processed model outputs, this also served to validate the effectiveness of the post-processing stage.

III. RESULTS

A. Preliminary assessment

The preliminary assessment of the best performing model, trained on only clear calls, showed precision ranging from 0.19 to 0.97 and recall ranging from 0.93 to 0.23 at thresholds from 0.01 to 0.99 (Fig. 4). This gave a maximum F score of 0.75 at a threshold of 0.36 (precision, 0.80; recall, 0.70).

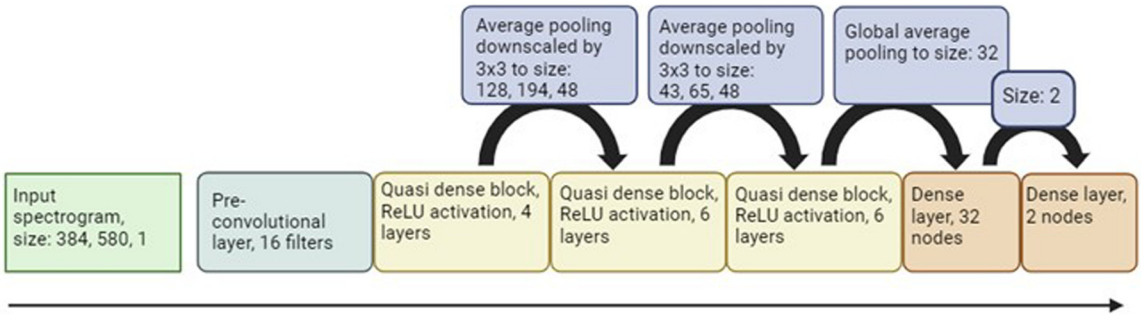

FIG. 3. (Color online) Flowchart showing the final model architecture. The final model had a growth rate of 12, began with a 16-filter pre-convolutional layer, contained 3 quasi-dense blocks with 4, 6, and 6 layers, respectively, and finished on a 32-node dense layer. Average pooling layers downscaled the inputs by a factor of 3×3 (height $\times$ width) in the transition blocks between quasi-dense blocks. Global average pooling was used to reduce the spatial dimensions of outputs of the final block to the 32-node feature vectors.

As the aim was to optimize the model for clear and faint calls, the testing was rerun excluding very faint calls. In this case, precision ranged from 0.12 to 0.94 and recall ranged from 0.99 to 0.33 at thresholds from 0.01 to 0.99 (Fig. 4). The maximum F score was improved to 0.78 at a threshold of 0.78 (precision, 0.80; recall, 0.76). To maximize performance for clear and faint calls while minimizing FPs, a threshold of 0.78 was, therefore, chosen to evaluate the model's performance on the test dataset.

B. Comparison with manual annotations

After rerunning the model selected in the preliminary assessment on the test dataset at the desired threshold and processing the output, it produced 535 detections. These were then visually analyzed in Raven Pro and compared to the manually annotated dataset to discern the occurrences of TPs, FPs, and FNs. In this case, a TP instance was defined as any model detection that overlapped with a great call. The model was found to have produced 511 TPs and 24 FPs, missing a further 133 calls (FNs). This gives a precision of 0.96, a recall of 0.79, and an F score of 0.87.

On closer inspection of the post-processed model output, 86 of the TP detections were the result of repeated detections for singular great call events. Additionally, 22 FNs were classed as TPs before the post-processing stage. These were created as a result of inadvertently grouping detections in which two great calls overlapped in time or were adjacent to one another.

Overall, the non-post-processed model identified 409 (78%) of the 522 manually annotated calls. This included 98% of all clear calls, 73% of faint calls, and 44% of very faint calls. Out of the FN instances before post-processing, 77% were very faint calls with only 0.05% representing six missed clear annotations. A further 38 great calls were identified, which had been missed in the manual annotation stage, including 6 clear, 8 faint, and 24 very faint calls.

The distribution of total detections every 5 min across the 4–10 a.m. period for the non-post-processed and post-processed model outputs were compared against the manually annotated test dataset (Fig. 5). Kolmogorov–Smirnov tests indicated no significant difference ($D = 0.036$, $p > 0.05$ and $D = 0.045$, $p > 0.05$, respectively). Also, we found no significant difference between preprocessed and post-processed data ($D = 0.020$, $p > 0.05$).

IV. DISCUSSION

Our best performing model was effective at detecting high-quality *H. albibarbis* great calls with a low rate of FPs. The best performing model (F score, 0.87) exceeded previously reported SVM models for detecting gibbon vocalizations e.g., *Hylobates funereus* detector, F score, 0.78 (Clink *et al.*, 2023), and was largely comparable to other CNN models for gibbon great calls, e.g., *Nomascus hainanus* detector, F score, 0.91 (Dufourq *et al.*, 2021) and *Nomascus concolor* DenseNet detector, F score, 0.92 (Zhou *et al.*, 2023).

We found that the best performing model detected 38 great calls that had been missed during the manual annotation stage, amounting to 6.8% of the total number of target events. Human listening is subject to error (Brauer *et al.*, 2016; Knight *et al.*, 2017), and this study relied on multiple human observers with differing levels of training to construct the manually annotated dataset. Although it has been shown that human observers with less experience may perform worse than some automated detectors (Jennings *et al.*, 2008), the consensus from multiple observers may have reduced the level of human error (Drake *et al.*, 2016). Furthermore, signals with low signal-to-noise ratio (SNR) are difficult for humans and machine learning algorithms to detect (Knight *et al.*, 2017). Precision and recall are often measured relative to manually annotated datasets, but these are not always perfect. With this in mind, it has been recommended to view the manually annotated dataset as the output from an alternative detector rather than a ground-truth set (Knight *et al.*, 2017).

As a result of the apparent likelihood of great calls to be missed by manual annotation (Knight *et al.*, 2017), it is

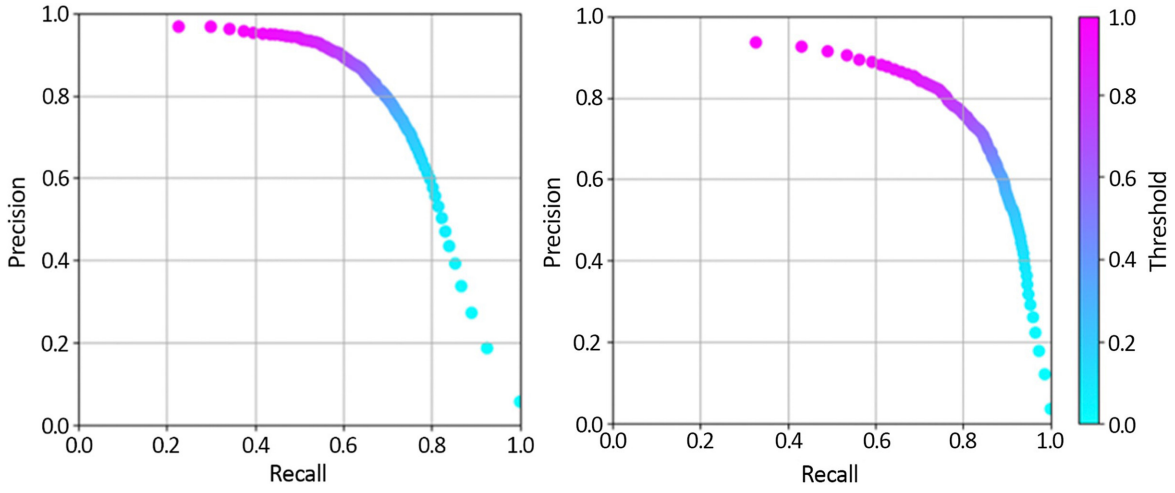

FIG. 4. (Color online) Precision-recall curves of the best performing model are tested against all calls (left) and clear and faint calls (right).

https://doi.org/10.1121/10.0028268

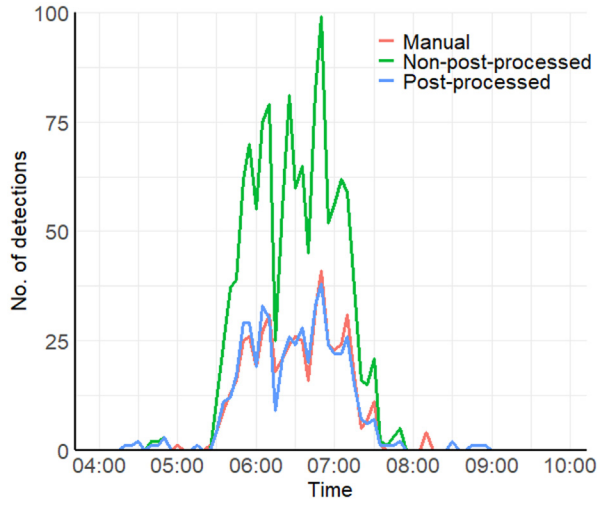

FIG. 5. (Color online) Histogram shows the number of great calls detected every 5 min between 04:00 and 10:00 for the manually annotated test dataset, as well as the output for the non-post-processed model and the post-processed model when run on the test dataset.

unrealistic to presume that all target calls were identified in the 210-h training dataset. This could pose a problem when training the model if any of the randomly selected unannotated time periods for the negative class contained target events, potentially increasing the rate of FNs. Dufourq *et al.* (2021) found that better results were obtained by specifically including negative class segments with typical ambient noise, such as other species' vocalizations, which could potentially confuse the classifier. This method of "handpicking" the negative class could reduce the FN rate and FP rate by negatively labelling potentially confusing information. The low false-positive rate reported in this study, as well as the low false-negative rate for clear calls, suggests that our method of randomly selecting the negative class was sufficient for our purposes. Where reducing the false-negative rate has greater importance, such as when identifying infrequent vocalizations, a more thorough approach could be preferable. Nonetheless, it is important to recognize the trade-off between ensuring that the negative class contains as little erroneous information as possible and the time required to construct an adequate training dataset.

In this study, we categorized selected calls according to their quality (cf. Cañas *et al.*, 2023), although we acknowledge that the distinction of clear, faint, and very faint great call selections was not based on p measurements. These categories cannot be translated directly to distance; however, in most cases, it is likely that clear calls were recorded from gibbons singing closer to the ARUs. Overall, our best performing model identified 98% of all clear calls and only missed six from the manually annotated dataset. This was comparable to the human observers, who also missed six clear calls. The model performed worse than the human detector at detecting faint and very faint calls, however, picking up 73% and 44% of the manually annotated instances, respectively. With this in mind, the results suggest that detection likelihood is affected by the caller-ARU distance (Spillmann *et al.*, 2015). This supports the suggestion by Jahn *et al.* (2017) that the difference in recall between an automated detector and human listener is caused by the former having a smaller detection radius. Future studies should aim to apply relationships between signal strength and the distance of the source from the receiver to estimate call detection probability over distance. This will help to determine the effective area being sampled by PAM studies.

Although the model did not have perfect recall, the importance of detecting all great calls within a recording will depend on the research question. With regard to calling activity over time, the distribution of calling frequency detected by the model was not significantly different from that of the manual annotations. Therefore, our model can be used to reliably estimate spatiotemporal variations in *H. albibarbis* calling activity. Through analyzing recordings from multiple habitats, our model could be applied to understand the relative importance of forest subtypes for the species as singing behavior is density dependent (Cheyne *et al.*, 2008) with less singing activity at lower group densities. This could operate on a continuous long-term time frame, which would be hard to achieve when relying on human observers in the field.

For an in-depth understanding of gibbon group abundance and density, further information is necessary. One method is to use individuals as the sampling unit (Buckland, 2006) by analyzing their call structure (Clink *et al.*, 2023) or localizing vocalizations using estimates of direction to the source from multiple ARUs (Stevenson *et al.*, 2015). This may prove highly effective at estimating gibbon abundance and density over the short term, yet, could prove too complex over the course of hundreds or thousands of hours of audio. An alternative method is to estimate vocalization density per unit time, apply an estimation of vocalization rate, and then convert vocalization density into group density (Marques *et al.*, 2013). This does require a knowledge of the area covered by each ARU, though, for this task, Marques *et al.* (2013) notes that automated detectors need not perform extraordinarily well so long as TP and FP rates are characterized accurately. This method does not require for the effective area of ARUs to overlap, therefore, in theory, could monitor a larger area with the same resources.

Whereas post-processing greatly reduced the number of repeated detections for single great call events, these would still account for many FPs if only one TP per great call is allowed. The post-processing protocol did not seek to adjudicate between detections if there was a tie in the highest-scoring instances within a group as in some cases, this represented two great call events close in time. In fact, 22 FNs derived from instances when the grouping of detections failed to take this into account, and limiting the number of detections per group to one would have increased the FN rate further. An alternative approach would be an improved post-processing stage to better interpret the output of the model. In audio classification, CNNs inspect audio recordings as image-like segments and, hence, are unable to use broader-scale contextual information such as whether the

current point in a recording is preceded by a target call (Wang *et al*., 2022). It has, therefore, been proposed to combine CNNs with other machine learning techniques, such as hidden Markov models (HMMs), or deep learning techniques, such as recurrent neural networks (RNNs). Postprocessing the output of a CNN with a HMM or a RNN has been shown to improve the *F* score (Madhusudhana *et al*., 2021; Wang *et al*., 2022). For instance, Wang *et al*. (2022) showed that the application of a combined convolutional recurrent neural network (CRNN) reduced error rates arising from overestimation of gibbon calls (49%–54%) to 0.5%. It was noted, however, that while post-processing the model output with a HMM performed second best, it required much less computational power. Future work should, thus, consider both of these options when attempting to improve on our post-processing method.

A key aim of this study was to help address the analysis bottleneck evident in PAM data by improving on the time taken to manually analyze recordings. During the batch processing stage, it took 19 s for our trained model to process each hour of test recordings. This greatly improved on a human processing speed of minimum 1 h per hour of audio for this study, varying depending on the level of observer experience. One caveat is the significant amount of time required to construct a manually annotated dataset to train and test a CNN when compared to other machine learning approaches (Stowell, 2022). Despite our study depicting how data augmentation can be effective where training data is limited, careful consideration should be taken when under time pressure if no training datasets are already available. In some cases, it may be better to adopt an approach that requires less data to develop such as a SVM or GMM (Clink *et al*., 2023).

Finally, Stowell (2022) noted that it was increasingly common to evaluate deep learning models on test sets specifically designed to differ in some respects from the training data such as location, SNR, or by season. In this case, the model was designed with application to the MBERF bioacoustic dataset in mind and, hence, it is appropriate that the test data came from the same location as the training data. Test inputs spanned across all times of the year and were recorded from three different habitats. This ensured that a variety of potential sound environments were included in the testing stage. However, for applications in other locations, especially outside the Rungan Forest Landscape, it would be advantageous to first test the model on recordings captured elsewhere within *H. albibarbis*' range.

V. CONCLUSION

Our study demonstrates how an open-source deep learning framework can be adapted to produce a CNN capable of detecting *H. albibarbis* great calls, performing at a comparable level to similar CNN approaches for gibbon great calls. Our model performed best on the highest-quality calls and yielded a low false-positive rate, meeting the objectives of this study. There was a much lower likelihood of successful detection for the lowest-quality calls, however, and future studies should aim to estimate call detection probability over distance to determine the effective area being sampled.

Further development of the post-processing stage could help to reduce the number of repeat detections for each call. However, the current output can be used to estimate calling rate over time. Further work should seek to apply this model to long-term acoustic datasets over a variety of habitats to study spatial and temporal variation in gibbon calling activity. Furthermore, in combination with future studies on sound propagation of gibbon vocalizations, this represents an opportunity to monitor *H. albibarbis*' populations on an ever-greater spatiotemporal scale. Our work presents some key considerations to inform decision-making for such projects and a full workflow script to visualize how these can be implemented in developing an automated detector.

SUPPLEMENTARY MATERIAL

See the supplementary material for the composition of the manually annotated dataset, including dates, see supplementary material A (SuppPub1.pdf). For the full Koogu workflow, see supplementary material B (SuppPub2.tex). For the post-processing script, see supplementary material C (SuppPub3.tex).

ACKNOWLEDGMENTS

We acknowledge the Universitas Muhammadiyah Palangkaraya and Borneo Nature Foundation for access to and support in the MBERF, as well as the Indonesian government for permission to perform this research [RISTEK-DIKTI Permit Nos. 189/SIP/FRP/E5/Dit.KI/VI/2018 and 43/E5/E5.4/SIP.EXT/2019 (W.M.E.), Badan Riset dan Inovasi Nasional foreign research Permit Nos. 223/SIP/IV/FR/5/2023 (F.J.F.V.), and 225/SIP/IV/FR/5/2023 (A.F.O)]. We thank Erik Estrada and Rido for their invaluable support in deploying and maintaining the ARUs and data in the field. We also thank Georgia Allen, Amy Barron, Sophie Carpenter, and Elena Gough for their contributions in manually annotating the dataset. In addition, we acknowledge the contributions of the Indigenous Dayak Ngaju community of Mungku Baru, including Pak Edo, Pak Yuli, Pak Viktor, and Rico, who supported the data collection and shared their knowledge together with local conservation practitioners from Yayasan Borneo Nature. This project has been a collaboration between international and Indonesian researchers from the start and is recognized in joint authorship here. Finally, please note the following sponsors of this research, awarded to W.M.E.: Primate Conservation, Incorporated; British Academy; Conservation International; American Association of Physical Anthropologists; International Primatological Society; and American Institute for Indonesian Studies. A.F.O., F.J.F.V., W.M.E., M.A.I., and K.J.H. developed the project and concept; M., T.M.S., and Si.M. provided permissions and guidance for field study design and execution; A.F.O., Sh.M., and M.S. carried out

building and analysis of the model; A.F.O. drafted the manuscript to which all other authors then contributed to produce the final version.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

Ethical approval was provided by the University of Exeter (Application ID 1845574), BRIN (Application No. 22 022 023 000 026), and Institutional Animal Care and Use Committee of Rutgers, the State University of New Jersey Protocol No. PROTO201800073.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.10926304 (Owens *et al.*, 2024).

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., and Aide, T. M. (**2009**). "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," Ecol. Inf. **4**(4), 206–214.

Bender, D. J., Contreras, T. A., and Fahrig, L. (**1998**). "Habitat loss and population decline: A meta-analysis of the patch size effect," Ecology **79**(2), 517–533.

Bjorck, J., Rappazzo, B. H., Chen, D., Bernstein, R., Wrege, P. H., and Gomes, C. P. (**2019**). "Automatic detection and compression for passive acoustic monitoring of the African forest elephant," Proc. AAAI Conf. Artif. Intell. **33**(01), 476–484.

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., and White, A. E. (**2022**). "Deep learning as a tool for ecology and evolution," Methods Ecol. Evol. **13**(8), 1640–1660.

Brauer, C. L., Donovan, T. M., Mickey, R. M., Katz, J., and Mitchell, B. R. (**2016**). "A comparison of acoustic monitoring methods for common anurans of the northeastern United States," Wildl. Soc. Bull. **40**(1), 140–149.

Buckland, S. T. (**2006**). "Point-transect surveys for songbirds: Robust methodologies," Auk **123**(2), 345–357.

Buckley, B. J. W., Capilla, B. P., Maimunah, S., Adul, Armadyanto, Boyd, N., Cheyne, S. M., Iwan, Husson, S. J., Santiano, Salahudin, Ferisa, A., Namaskari, N., van Veen, F., and Harrison, M. E. (**2018**). "Biodiversity, Forest Structure and Conservation Importance of the Mungku Baru Education Forest, Rungan, Central Kalimantan, Indonesia," BNF Reports (Borneo Nature Foundation, Palangka Raya, Indonesia), https://www.borneonaturefoundation.org/wp-content/uploads/2019/03/KHDTK-Report-2016-2017.pdf.

Cañas, J. S., Toro-Gómez, M. P., Sugai, L. S. M., Restrepo, H. D. B., Rudas, J., Bautista, B. P., Toledo, L. F., Dena, S., Domingos, A. H. R., de Souza, F. L., Neckel-Oliveira, S., da Rosa, A., Carvalho-Rocha, V., Bernardy, J. V., Sugai, J. L. M. M., dos Santos, C. E., Bastos, R. P., Llusia, D., and Ulloa, J. S. (**2023**). "A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring," Sci. Data **10**(1), 771.

Cheyne, S., Gilhooly, L., Hamard, M. C., Höing, A., Houlihan, P. R., Kursani, Loken, B., Phillips, A., Rayadin, Y., Ripoll Capilla, B., Rowland, D., Sastramidjaja, W. J., Stephanie, S., Thompson, C. J. H., and Zrust, M. (**2016**). "Population mapping of gibbons in Kalimantan, Indonesia: Correlates of gibbon density and vegetation across the species' range," Endang. Species Res. **30**, 133–143.

Cheyne, S. M., Thompson, C. J. H., Phillips, A. C., Hill, R. M. C., and Limin, S. H. (**2008**). "Density and population estimate of gibbons (*Hylobates albibarbis*) in the Sabangau catchment, Central Kalimantan, Indonesia," Primates **49**(1), 50–56.

Clink, D. J., Kier, I., Ahmad, A. H., and Klinck, H. (**2023**). "A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings," Front. Ecol. Evol. **11**, 1071640.

Clink, D. J., Kim, J., Cross-Jaya, H., Ahmad, A. H., Hong, M., Sala, R., Birot, H., Agger, C., Vu, T. T., Thi, H. N., and Chi, T. N. (**2024**). Automated detection of gibbon calls from passive acoustic monitoring data using convolutional neural networks in the 'torch for R' ecosystem. arXiv:2407.09976.

Colonna, J., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., and Gama, J. (**2016**). "Automatic classification of anuran sounds using convolutional neural networks," in *Proceedings of the Ninth International C* Conference on Computer Science and Software Engineering—C3S2E '16*, July 20–22, Porto, Portugal (ACM, New York), pp. 73–78.

Drake, K. L., Frey, M., Hogan, D., and Hedley, R. (**2016**). "Using digital recordings and sonogram analysis to obtain counts of yellow rails," Wildl. Soc. Bull. **40**(2), 346–354.

Dufourq, E., Durbach, I., Hansford, J. P., Hoepfner, A., Ma, H., Bryant, J. V., Stender, C. S., Li, W., Liu, Z., Chen, Q., Zhou, Z., and Turvey, S. T. (**2021**). "Automated detection of Hainan gibbon calls for passive acoustic monitoring," Remote Sens. Ecol. Conserv. **7**(3), 475–487.

Goodfellow, I., Bengio, Y., and Courville, A. (**2016**). *Deep Learning* (The MIT Press, Cambridge, MA).

Guyot, P., Alix, F., Guerin, T., Lambeaux, E., and Rotureau, A. (**2021**). "Fish migration monitoring from audio detection with CNNs," in *Proceedings of the 16th International Audio Mostly Conference*, September 1–3, Trento, Italy (ACM, New York), pp. 244–247.

Haykin, S. S. (**2009**). *Neural Networks: A Comprehensive Foundation*, 3rd ed. (Pearson Education, McMaster University, Hamilton, Ontario, Canada).

Heinicke, S., Kalan, A. K., Wagner, O. J. J., Mundry, R., Lukashevich, H., and Kühl, H. S. (**2015**). "Assessing the performance of a semi-automated acoustic monitoring system for primates," Methods Ecol. Evol. **6**(7), 753–763.

Hibino, S., Suzuki, C., and Nishino, T. (**2021**). "Classification of singing insect sounds with convolutional neural network," Acoust. Sci. Tech. **42**(6), 354–356.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 2261–2269.

Jahn, O., Ganchev, T. D., Marques, M. I., and Schuchmann, K.-L. (**2017**). "Automated sound recognition provides insights into the behavioral ecology of a tropical bird," PLoS One **12**(1), e0169041.

Jennings, N., Parsons, S., and Pocock, M. J. O. (**2008**). "Human vs. machine: Identification of bat species from their echolocation calls by humans and by artificial neural networks," Can. J. Zool. **86**(5), 371–377.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Kiel, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., and Turak, E. (**2019**). "Essential biodiversity variables for mapping and monitoring species populations," Nat. Ecol. Evol. **3**(4), 539–551.

K Lisa Yang Center for Conservation Bioacoustics (**2024**). "Raven Pro: Interactive sound analysis software," (The Cornell Lab of Ornithology, Ithaca, NY), available at http://ravensoundsoftware.com (Last viewed April 1, 2024).

Kingma, D. P., and Ba, J. (**2017**). "Adam: A method for stochastic optimization," arXiv:1412.6980v9.

Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M., and Bayne, E. (**2017**). "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs," Avian Conserv. Ecol. **12**(2), 14.

Kong, Q., Xu, Y., and Plumbley, M. D. (**2017**). "Joint detection and classification convolutional neural network on weakly labelled bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, August 28–September 2, Kos, Greece (IEEE, New York), pp. 1749–1753.

Mac Aodha, O., Gibb, R., Barlow, K., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., and Jones, K. (**2018**). "Bat detective—Deep learning tools for bat acoustic signal detection," PLoS Comput. Biol. **14**(3), e1005995.

Madhusudhana, S. (**2023**). "shyamblast/Koogu: v0.7.2," Zenodo, v0.7.2 https://doi.org/10.5281/zenodo.8254287.
Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E., Helble, T., Cholewiak, D., Gillespie, D., Širović, A., and Roch, M. A. (**2021**). "Improve automatic detection of animal call sequences with temporal context," J. R. Soc. Interface **18**(180), 20210297.
Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. (**2013**). "Estimating animal population density using passive acoustics," Biol. Rev. **88**(2), 287–309.
Mielke, A., and Zuberbühler, K. (**2013**). "A method for automated individual, species and call type recognition in free-ranging animals," Anim. Behav. **86**(2), 475–482.
Miller, B. S., Madhusudhana, S., Aulich, M. G., and Kelly, N. (**2023**). "Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: A double-observer analysis," Remote Sens. Ecol. Conserv. **9**(1), 104–116.
Morgan, M. M., and Braasch, J. (**2021**). "Long-term deep learning-facilitated environmental acoustic monitoring in the Capital Region of New York State," Ecol. Inf. **61**, 101242.
Narasimhan, R., Fern, X. Z., and Raich, R. (**2017**). "Simultaneous segmentation and classification of bird song using CNN," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5–9, New Orleans, LA (IEEE, New York), pp. 146–150.
Noda, J. J., Travieso, C. M., Sanchez-Rodriguez, D., Dutta, M. K., and Singh, A. (**2016**). "Using bioacoustic signals and support vector machine for automatic classification of insects," in *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, February 11–12, Noida, India (IEEE, New York), pp. 656–659.
Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (**2018**). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," Proc. Natl. Acad. Sci. U.S.A. **115**(25), E5716–E5725.
Owens, A. F., Estrada, E., Mariaty, and Erb, W. M. (**2024**). "A collection of Bornean white-bearded gibbon (*Hylobates albibarbis*) great call vocalisations for training and testing an automated detector," Zenodo, v.1.0, Dataset. https://doi.org/10.5281/zenodo.10926304
Patterson, J., and Gibson, A. (**2017**). *Deep Learning: A Practitioners Approach*, edited by M. Loukides and T. McGovern (O'Reilly Media, Inc., Sebastopol, CA).
Piel, A. K., and Wich, S. A., eds. (**2021**). *Conservation Technology* (Oxford University Press, Oxford, UK).
Purnama, A., and Afitah, I. (**2021**). "Motivasi Masyarakat Terhadap Pengelolaan Khdtk Mungku Baru, Palangka Raya" "Community motivation for the management of Khdtk Mungku Baru, Palangka Raya", Anterior J. **20**(2), 43–49.
Rammer, W., and Seidl, R. (**2019**). "Harnessing deep learning in ecology: An example predicting bark beetle outbreaks," Front. Plant Sci. **10**, 1327.
Schlüter, J., and Grill, T. (**2015**). "Exploring data augmentation for improved singing voice detection with neural networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, October 26–30, Malaga, Spain (ISMIR, Canada), pp. 121–126.
Spillmann, B., van Noordwijk, M. A., Willems, E. P., Mitra Setia, T., Wipfli, U., and van Schaik, C. P. (**2015**). "Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls," Am. J. Primatol. **77**(7), 767–776.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (**2014**). "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. **15**, 1929–1958.
Stevenson, B. C., Borchers, D. L., Altwegg, R., Swift, R. J., Gillespie, D. M., and Measey, G. J. (**2015**). "A general framework for animal density estimation from acoustic detections across a fixed microphone array," Methods Ecol. Evol. **6**(1), 38–48.
Stowell, D. (**2022**). "Computational bioacoustics with deep learning: A review and roadmap," PeerJ **10**, e13152.
Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W., and Llusia, D. (**2019**). "Terrestrial passive acoustic monitoring: Review and perspectives," BioScience **69**(1), 15–25.
van Kuijk, S. M., O'Brien, S., Clink, D. J., Blake, J. G., and Di Fiore, A. (**2023**). "Automated detection and detection range of primate duets: A case study of the red titi monkey (*Plecturocebus discolor*) using passive acoustic monitoring," Front. Ecol. Evol. **11**, 1173722.
Verma, A., van der Wal, R., and Fischer, A. (**2016**). "Imagining wildlife: New technologies and animal censuses, maps and museums," Geoforum **75**, 75–86.
Vu, T. T., and Tran, L. M. (**2019**). "An application of autonomous recorders for gibbon monitoring," Int. J. Primatol. **40**(2), 169–186.
Wang, Y., Ye, J., and Borchers, D. L. (**2022**). "Automated call detection for acoustic surveys with structured calls of varying length," Methods Ecol. Evol. **13**(7), 1552–1567.
Wood, C. M., Barceinas Cruz, A., and Kahl, S. (**2023**). "Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modeling of an endangered primate," Am. J. Primatol. **85**(8), e23507.
Zhou, X., Hu, K., Guan, Z., Yu, C., Wang, S., Fan, M., Sun, Y., Cao, Y., Wang, Y., and Miao, G. (**2023**). "Methods for processing and analyzing passive acoustic monitoring data: An example of song recognition in western black-crested gibbons," Ecol. Indic. **155**, 110908.