Bioacoustics and machine learning for automated avian species monitoring in global biodiversity hotspots

Article *in* The Journal of the Acoustical Society of America · October 2020

DOI: 10.1121/1.5146736

Ming Zhong
Microsoft

Naomi Bates
Future Generations University

Rahul Dodhia
Microsoft

Juan Miguel Lavista Ferres
Microsoft

Acoustic Detection of Regionally Rare Bird Species Through Deep Convolutional Neural Networks

Ming Zhong[1], Ruth Taylor[2], Naomi Bates[2], Damian Christey[2], Hari Basnet[2], Jennifer Flippin[2], Shane Palkovitz[2], Rahul Dodhia[1], Juan Lavista Ferres[1]

[1] AI for Good Research Lab, Microsoft
[2] Songs of Adaptation, Future Generations University

Corresponding Author:
Juan Lavista Ferres[1]
AI for Good Research Lab, Microsoft, Redmond, WA 98052, USA
Email address: jlavista@microsoft.com

Abstract: Bioacoustic monitoring with machine learning (ML) models can provide valuable insights for informed decision-making in conservation efforts. In this study, the team built deep convolutional neural networks to analyze field recordings and classify calls of Yellow-vented warbler (*Phylloscopus cantator*) and Rufous-throated wren-babbler (*Spelaeornis caudatus*), both of which are regionally rare in Nepal. Data augmentation techniques for calls of the two bird species were utilized to effectively increase the size of the training set and thus boost model performance. Nepali ornithologists were engaged in iterative data labeling from field recordings, leveraging ML technology in conjunction with expert manual labeling and verification. The model output provides insights of species activity and abundance throughout 2018-2019 in multiple ecosystems along an elevational transect in the Barun River Valley, Nepal. The results of this study may help conservationists better understand species distribution, behavior, diversity, and habitat preference. Additionally, the results provide baseline data to quantify future changes

due to habitat disruption or climate change. This modeling methodology and its framework can be easily adopted by other acoustic classification problems.

Keywords: Deep learning, Convolutional Neural Networks (CNN), Bioacoustic classification, Transfer learning, Species population, Presence survey

I. Introduction

In recent decades, the populations of various animals, including birds, amphibians, insects, and mammals, have exhibited steep declines worldwide. While many decreases are due to habitat loss and overutilization, other unidentified processes threaten 48% of rapidly declining species and are driving species most quickly to extinction [1]. As biodiversity plays a critical role in many aspects, well-designed monitoring programs provide a basis for identifying the species, sites and threats of most significant concern. Such monitoring programs also provide reliable tools when evaluating the integrity of ecosystems and their responses to disturbances, assessing progress in efforts to conserve biodiversity, and measuring the success of actions taken to preserve or recover biodiversity. However, manual observation remains limited and challenging in many scenarios, especially in the areas that are difficult to access physically or when the focus is to study animals' night-time behavior. In such scenarios, passive acoustic monitoring is highly appropriate, as many birds, including rare species, are most readily detectable by their sounds, often more so than by vision. With modern remote monitoring stations, it can continuously monitor large remote areas for avian community composition and tracking migratory and seasonal changes in populations ([2] – [7]).

Earlier applications that have employed such technology either performed automatic recording but relied on manual analysis of sound recordings ([8], [9]) or were based on low-complexity signal processing such as template matching ([10], [11]), feature extraction ([12]), or traditional machine learning methods ([13], [14]).

With the constant increase in computing power and the development of more efficient codes, high-performance computing helps the extremely fast growth of deep learning in recent years, which has been shown to outperform previous state-of-the-art techniques in several tasks. Deep learning has fueled great strides in a variety of computer vision problems, and in particular, Convolutional Neural Networks (CNN) have demonstrated great potential and success in image classification tasks and thus drawn much attention in constructing the automatic bird sound classification systems. Some popular CNN architectures applied to bioacoustics classification include AlexNet [15], LeNet-5 [16], VGG16 [17], ResNet50 [18], among others.

In this study, two regionally rare species were chosen: Yellow-vented warbler (*Phylloscopus cantator*) and Rufous-throated wren-babbler (*Spelaeornis caudatus*). The Rufous-throated Wren Babbler is a very rare bird that has an extremely limited range in Nepal. The species is Near Threatened globally; it is listed within Nepal as a Critically Endangered species on a national level ([19], [20]). The nationally endangered Yellow-vented Warbler can be found in the East of Nepal. It is recorded between 75m and 1525m in a few locations, including Makalu Barun National Park [20]. These species provided a proof of concept demonstrating that with limited training samples, deep learning models can classify rare species calls.

As a research project of Future Generations University, this project brings expertise in community development with decades-long global partnerships that ensure long-term data collection and research permissions, data labeling, and collaboration for sustainable, just, and lasting climate action. Protecting 100,000,000 acres of land, Future Generations leadership in community-based conservation established multiple national parks across Asia. Over the past 27 years, the University has employed key indicators (quick, easy-to-use measurements), shaped to fit specific communities' contexts, to empower community members to measure change over time for themselves. This project is unique in its commitment to community engagement.

II. Data collection and pre-processing

Audio data were collected from 8 different sites along an elevational transect in the mountains of Makalu Barun National Park, Nepal, between 2018 and 2019. Audios were recorded into wav format using Song Meter SM4 Acoustic Recorders (Wildlife Acoustics) at a sampling rate of 48kHz and 24-bit rate, and recorders were programmed to record 5-minute audio every 15 minutes 24 hours per day.

The training and test datasets were initially generated using pattern recognition clustering software (Kaleidoscope Pro Analysis Software, Wildlife Acoustics [21]), and local avian experts subsequently analyzed clusters in Nepal to identify calls from the two species of interest, Yellow-vented warbler (*Phylloscopus cantator*) and Rufous-throated wren-babbler (*Spelaeornis caudatus*). A target of 100+ positive detections for each species and 300+ negative detections (ex. rain wind river, insects, other bird species, etc.) were set as the minimum amount of data necessary for model training and development. The positive and negative samples were used as input to train CNN models. Spectrogram images containing a target positive or negative sample were standardized using a 4-second audio clip beginning at each detection's start-time.

Spectrograms were extracted from audio files (with NFFT = 256, Hanning window) using Python 3.6 and then resized to 224 by 224 pixels with RGB channels and stored as color PNG images (see Fig. 1 for example). The color spectrograms were the input for the machine learning model, and the corresponding single-species labels for each image (i.e. species present (positive) or absent (negative)) were used as the ground truth data for training and evaluating the classification model.

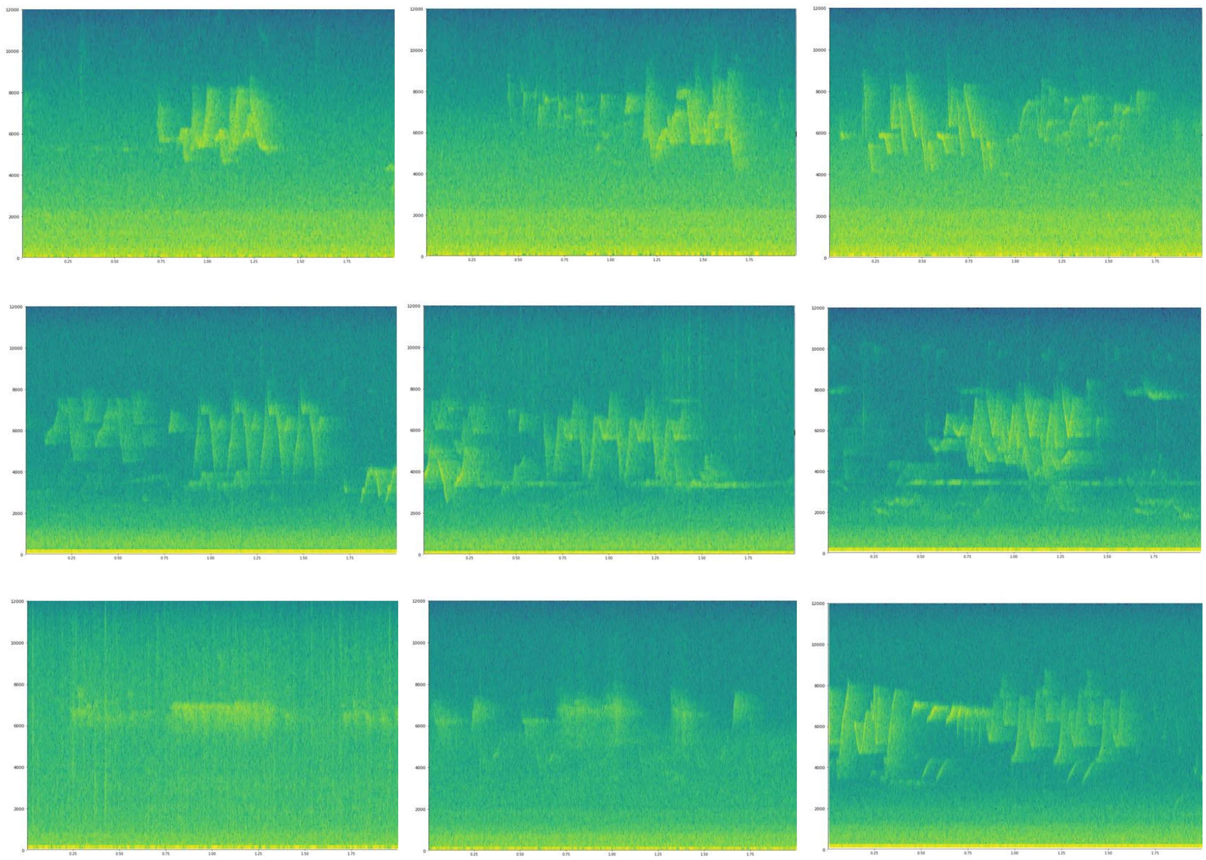

Fig. 1. Sample spectrograms for 4-second audio recordings. First row: calls from the species *Phylloscopus cantator*; Second row: calls from the species *Spelaeornis caudatus*; Third row from left to right: rain background, river background, unidentified species.

III. Approaches

A. Transfer learning and fine-tuning with a pre-trained CNN model

Here the neural network model ResNet50 was applied to classify the calls of the 2 bird species. This ResNet50 CNN architecture is a variant of ResNet model which has 48 Convolution layers along with 1 Max Pooling and 1 Average Pooling layer. It begins with the RGB images (size 224

× 224 × 3) as input and performs the initial convolution and max-pooling using 7×7 and 3×3 kernel sizes, respectively. Afterward, it stacks a series of residual blocks. With the skip connection of residual blocks, it allows the model to propagate larger gradients to initial layers. These layers are able to learn as fast as the final layers, in order to train deeper networks. Finally, the network has an average pooling layer, followed by a fully connected layer. When training the ResNet50 model, the Adam optimizer algorithm was applied, and an initial learning rate of 1e-4 with a decay factor of 1e-7.

In the context of deep learning, most models include millions of parameters. ResNet50, for example, has 23 million parameters. To train such complex models, it typically requires an extensive dataset to achieve an optimal parameter configuration. However, in practice it may be very difficult to collect large amounts of labeled data, especially if a species rarely calls or if the species is endangered and there are few individuals. Besides, using experts to obtain a large number of labeled samples in acoustics is an expensive and time-consuming endeavor. Given this scenario, transfer learning with fine-tuning [22] is a useful technique when there is only a small number of labeled data available.

Transfer learning is a machine learning technique where a model trained on one task (or domain) is re-purposed on a second related task (or domain). Pre-trained models are usually shared in the form of the millions of parameters/weights the model achieved while being trained to an optimal state. In this study, the model weights were initially trained on ImageNet [23] dataset with 1000 classes of objects, but their pre-trained weights can be leveraged by a different task or domain [24]. This approach is effective because the source model was trained on a large number of images and made predictions on a relatively large number of classes. In turn, it required the model to extract distinct features from images in order to perform well. With fine-tuning, some layers are frozen from the pre-trained model, and it is sufficient to train the last several layers only, instead of having to train the whole model with random initialization of all parameters.

In this study, the model design included pre-trained weights of ResNet50 and fine-tuned parameters, adding a fully connected layer, a dropout layer and an output layer.

B. K-Fold Cross-Validation

In this dataset, there are only a few hundreds of detected calls for the two target species, *Phylloscopus cantator* and *Spelaeornis caudatus*, that include different stereotypes of calls from each species. By partitioning the available data into three sets (training, validation and testing), we drastically reduce the number of samples which can be used for learning the model, and the results that depend on a particular random choice for the three sets are not stable. A solution to this problem is a procedure called K-fold cross-validation, which generally results in a less biased model compared to other methods. With this procedure, it ensures every observation from the original dataset has the chance of appearing in the training and test set. This is one of the best approaches if we have limited input data. This method follows the below steps:

Step 1: Split the entire data randomly into K folds (here, we use K=5).

Step 2: Fit the model (training and validation) using the K - 1 (K minus 1) folds and test the model using the remaining Kth fold. Note down the scores/errors.

Step 3: Repeat this process until every K-fold serves as the test set. Then take the average of all recorded scores. That will be the performance metric for the model.

C. Data Augmentation

While many deep neural network models have parameters in the order of millions, they are heavily reliant on big data to avoid overfitting. Unfortunately, in many real-world applications, the amount of data that can be used for training is rather limited, either due to the huge manual efforts required to collect data, or due to the fact that it is almost impossible to acquire large amounts of data in some cases. As an effective data-space solution to the problem of limited data, data augmentation encompasses a suite of techniques that enhance the size and quality of training datasets such that better deep learning models can be built using them.

Among various data augmentation methods for image processing, some basic ones include flips, rotations, shifts, noise injections, color space transformations, sharpening or blurring, and random erasing or cropping. Specifically, for audio recordings, there are methods such as time-stretching, pitch shifting, and mixing multiple audios [25]. Beyond them, there are more advanced techniques, for example, generative adversarial network (GAN)-based methods [26], which can be used to generate synthetic images.

For this model implementation, basic techniques were applied to increase the size of data that can be used for model training: rotation (up to 5 degrees), shifting (width and height shifting up to 10% of the original spectrogram), and cropping.

Another effective method we adopted to boost the training data size is to use spectrograms with smaller time-windows. While the detected calls for the two regionally rare species, *Phylloscopus cantator* and *Spelaeornis caudatus*, usually last for 2 seconds or longer (see Fig. 1. as an example), our baseline model was fit based on spectrograms generated from a 4-second time window. In order to boost the size of training data, we break down each 4-second detection into 3 shorter detections, where each detection lasts for 2 seconds (that is, to create 3 spectrograms starting at second 0, 1, and 2, respectively, from each original 4-second detection). Even though breaking down spectrograms into 2-second windows may not include one complete call within each spectrogram and may bring some noisy labels during model training, but with this implementation, the size of data available for training tripled.

D. Model Training

Our manually validated dataset consists of 195 positive detections for *Phylloscopus cantator*, 320 positive detections for *Spelaeornis caudatus*, and 1060 negative detections composed of various types of noises (rain, wind, river, bugs, other bird species, etc.), where each detection lasts for 4 seconds.

Finding sufficient clips of exemplar training data from the field recordings was challenging, because of call volume variations, overlapping calls with other species, and background noises. In addition, to distinguish a species with multiple and varying calls, it was also essential and challenging to determine other species that had similar calls to the target species and label these close calls as negative training data. Particularly, for the target species *Spelaeornis caudatus*, there are two other species that have acoustically similar calls (Fig. 2).

Target species	Close call Species #1	Close call Species #2
Spelaeornis caudatus	*Phylloscopus reguloides*	*Pnoepyga albiventer*

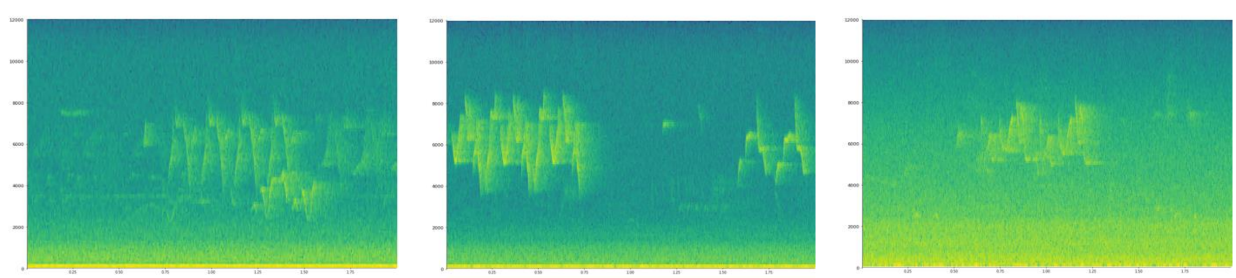

Fig. 2. Spectrogram of *Spelaeornis caudatus* and two other species (*Phylloscopus reguloides* and *Pnoepyga albiventer*) with acoustically similar calls.

After scoring, all the false positives in the training data were verified by experts and correctly labeled to retrain the model after the first round of model training. External training data (exemplar calls manually verified from xeno-canto.org) were added to supplement the project's data.

IV. Results

A. Model Performance

Three key metrics are reported to evaluate and compare the performance of the model on the testing data set: 1) sensitivity (true positive rate, recall); 2) specificity (true negative rate), and 3) area under a curve (AUC). Sensitivity measures the proportion of true positives that were identified correctly; and specificity measures the proportion of true negatives that were identified correctly . While sensitivity and specificity are dependent on the choice of threshold score, the area under a curve (AUC) provides an aggregate measure of performance across all possible classification thresholds. It is not affected by the class imbalance.

TABLE I: Classification results (sensitivity, specificity, and AUC) for both target species by each CNN model. The results are based on the average score of conducting 5-fold cross-validation, with a neutral threshold score 0.5.

Species	CNN Model Description	Sensitivity (%)	Specificity (%)	AUC (%)
Phylloscopus cantator	based on 4-second spectrograms, no data augmentation	86.15	98.91	98.62
	based on 2-second spectrograms, no data augmentation	94.92	99.92	99.02
	based on 2-second spectrograms, with data augmentation	**95.94**	**99.92**	**99.58**
Spelaeornis caudatus	based on 4-second spectrograms, no data augmentation	53.12	94.82	91.50
	based on 2-second spectrograms, no data augmentation	78.46	**95.96**	97.05
	based on 2-second spectrograms, with data augmentation	**90.15**	93.92	**97.85**

For both species *Phylloscopus cantator* and *Spelaeornis caudatus,* the model based on 2-second spectrograms performed significantly better, especially sensitivity, compared to the model based on 4-second spectrograms. Using data augmentation made further improvement for the model based on 2-second spectrograms (Table I). The sensitivity for classifying the species *Spelaeornis* *caudatus* was not as good as that of the model classifying the species *Phylloscopus cantator*, and resulted in about 10% of detections that were misclassified as negative. A closer investigation of the data revealed that the labeled calls for *Spelaeornis caudatus* included detections with various levels of clarity, different call stereotypes, and maybe some incorrectly labeled detections. It appears that the neural network model did not find enough commonalities among these detected calls to make correct classification. Some examples of spectrograms are shown in Fig. 3.

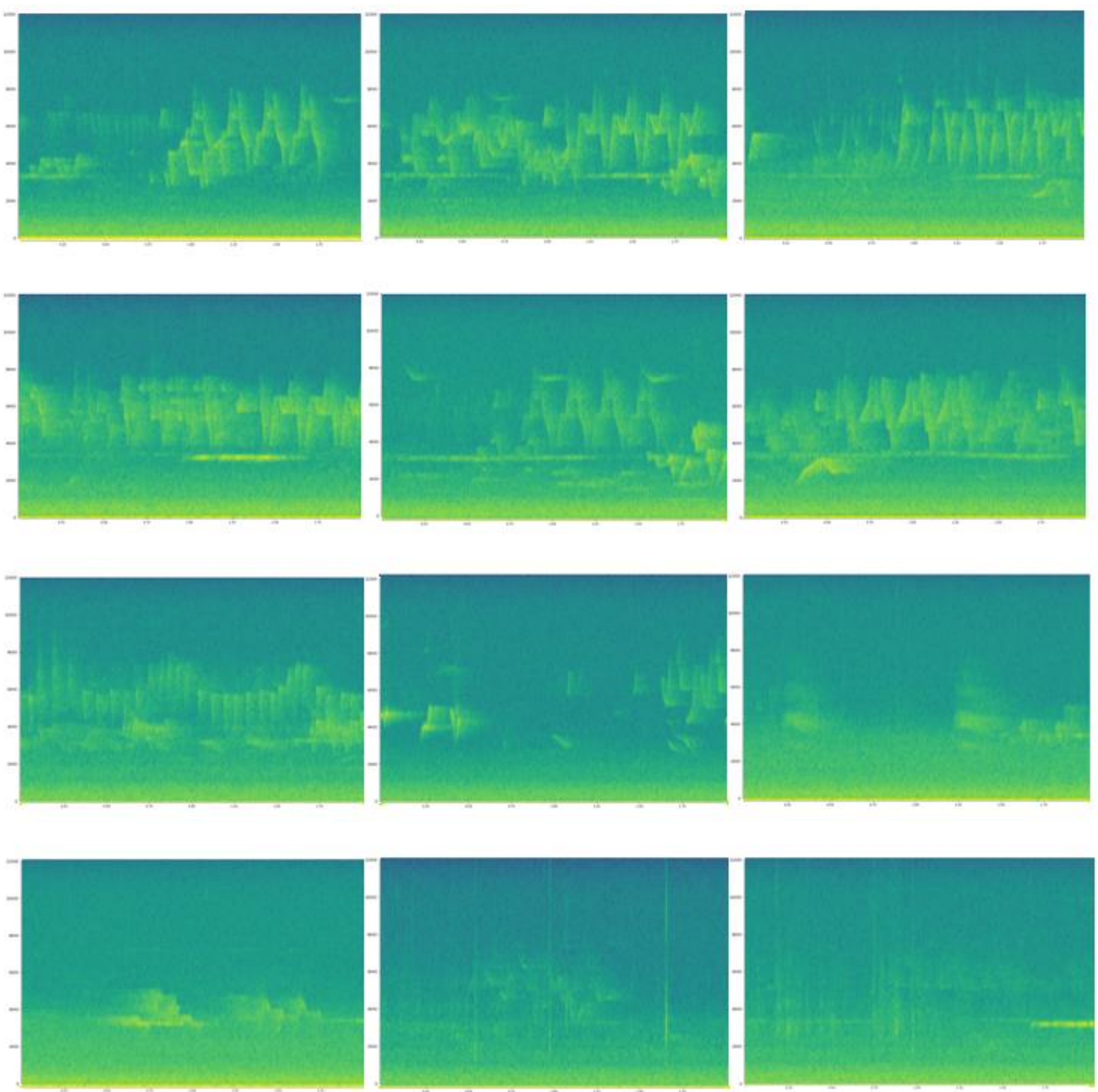

Fig. 3. Examples of spectrograms for 4-second audio recordings with detected calls from the species *Spelaeornis caudatus*. Row 1-2: examples of detections that the model can correctly classify; Row 3-4: examples of detections that the model wrongly classified as "no call".

B. Scoring on Unlabeled Data

The model was run using over one year of data with dates ranging from 3/2018 to 7/2019 using data from three stations in Makalu Barun National Park around the elevations where the target species were expected - Hinju Camp (elevation 1820 m), Deurali danda (elevation 2100 m), and Tutin Camp (elevation 2300 m).

In order for results to be analyzed, a threshold needs to be chosen for what probability will be counted as the presence of the species. Table II shows the number of detected calls for three sample threshold probability ranges (clip numbers rounded to the nearest ten). While the model predicts the probability of target species calls for each extracted spectrogram from the corresponding audio clip, the probability itself does not give a definite answer of presence/absence of species calls. As our next step, we will send those results to the local ecologists and conduct output validation by sampling spectrograms with different predicted probability ranges and then choosing the optimal threshold.

Table II. Number of model results returned for three selected probability ranges.

Species	Predicted Probability Range	# of 2-sec clips ML results show species presence
Spelaeornis caudatus	0.99-1	240,300 clips
	0.7-1	982,230 clips

	0.5-1	1,247,170 clips
Phylloscopus cantator	0.99-1	51,550 clips
	0.7-1	189,380 clips
	0.5-1	237,260 clips

Visualization of these big data results is a helpful tool for data analysis, as well as further verification and spot checking of results. Utilizing Plotly Dash (https://plotly.com/dash/), a web interface was created to visualize daily and hourly count (Fig. 4), with interactive options to filter results by species, predicted probability range (threshold), model iteration, and station.

Fig. 4. Web interface that can visualize the number of detected calls in multiple monitoring stations over time for certain targeted species. The interface allows the users to choose different probability ranges from model predictions.

V. Discussion

In this study, we demonstrate how deep convolutional neural networks (CNN) and transfer learning can achieve higher accuracy for the classification of calls from the targeted rare species with limited training data. We provide both methodological and practical contributions by testing the performance of a machine learning approach to augment the manual validation process, which is time-consuming and labor-intensive.

With limited labeled data, especially for rare species, the CNN model performs reasonably well. While transfer learning leverages the learning from one task which is generally trained on a large size dataset, it does not require learning from scratch for the new task, which is motivated by the observation that the earlier features of a CNN model contain more generic features (e.g. edge detectors or color blob detectors) that should be useful for many tasks. In this study, we used a pre-trained ResNet50 model to implement transfer learning with fine-tunings, and there are other options of pre-trained CNN models, such as VGG16 or DenseNet ([27]), that can be used to achieve comparable results. Except for these pre-trained models, which are based on ImageNet, transferring learned knowledge from networks trained on audio data (for example, SoundNet ([28]) or SincNet ([29])) is another reasonable choice.

Data augmentation is another effective way to increase the training sample size in order to achieve better classification performance. Beyond the ones that we used in our model, there are more complicated data augmentation methods such as adding or removing noises, image sharpening or masking, changing audio loudness, and audio mixing,

Finally, the methodology and implementation framework presented in this study can be easily adopted by other similar bioacoustics applications, where target signals require manual validation. This study sets initial steps for placing deep learning CNN analysis as the natural evolution of analysis methods for passive acoustic monitoring data.

VI. Further Research

In order for the results to be accurately used for species presence survey data, more iterations of label verification and model retraining are needed. The next step for this research is to establish a pipeline for verifying the ML results, determining when to re-run the model with additional verified training data, and ultimately choosing a threshold per species that represents accurate species presence survey data.

One tool that will aid this verification is being added to the interface and will be tested with further research. 10% stratified sample of the results will be returned for experts to spot check and compare with model analysis in order to determine if the model needs to be retrained or if the results are accurate for species presence research. A framework for this verification is essential because each species call will require different amounts of training data and/or a different threshold that returns accurate results.

Acknowledgements

The authors would like to thank everybody who participated in the experiment for their support. This work was supported by AI for Earth grants at Microsoft. Our appreciation to Dan Morris for connecting different parties for fruitful discussions and useful online materials. Our gratitude to the Nepal Government Department of National Parks and Wildlife Conservation, Makalu Barun National Park, The East Foundation (TEF), and the Barun Bachaon ("Save the Barun") Taskforce for their partnership.

References

[1] Stuart, S.N., J.S. Chanson, N.A. Cox, B.E. Young, A.S.L. Rodrigues, D.L. Fischman, and R.W. Waller. Status and trends of amphibian declines and extinctions worldwide. *Science*, 2004, 306, 1783–1786.

[2] Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 2013, 1, e103. https://doi.org/10.7717/peerj.103

[3] Furnas, B. J., & Callas, R. L. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *The Journal of Wildlife Management*, 2015, 79, 325–337. https://doi.org/10.1002/jwmg.821

[4] Frommolt, K.-H. Information obtained from long-term acoustic recordings: Applying bioacoustic techniques for monitoring wetland birds during breeding season. *Journal of Ornithology*, 2017, 158, 1–10. https://doi.org/10.1007/s10336-016-1426-3

[5] Hill, A. P., Prince, P., Piña Covarrubias, E., Patrick Doncaster, C., Snaddon, J. L., & Rogers, A. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution*, 2017, 9, 1199–1211. https://doi.org/10.1111%2F2041-210x.12955

[6] Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., & Bayne, E. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology*, 2017, 12(2), 14. https://doi.org/10.5751/ACE-01114-120214

[7] Matsubayashi, S., Suzuki, R., Saito, F., Murate, T., Masuda, T., Yamamoto, K., & Okuno, H. G. Acoustic monitoring of the great reed warbler using multiple microphone arrays and robot audition. *Journal of Robotics and Mechatronics*, 2017, 29, 224–235.

[8] Furnas, B. J., & Callas, R. L. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *The Journal of Wildlife Management*, 2015, 79, 325–337. https://doi.org/10.1002/jwmg.821

[9] Frommolt, K.-H. Information obtained from long-term acoustic recordings: Applying bioacoustic techniques for monitoring wetland birds during breeding season. *Journal of Ornithology*, 2017, 158, 1–10. https://doi.org/10.1007/s10336-016-1426-3

[10] Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. A toolbox for animal call recognition. *Bioacoustics*, 2012, 21, 107–125. https://doi.org/10.1080/09524622.2011.648753

[11] Colonna, J. G., Cristo, M., Júnior, M. S., & Nakamura, E. F. An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications*, 2015, 42, 7367–7374. https://doi.org/10.1016/j.eswa.2015.05.030

[12] Mesaros A, Heittola T, Benetos E, Foster P, Lagrange M, Virtanen T, Plumbley MD.

Detection and classification of acoustic scenes and events: outcome of the DCASE. 2016 challenge. IEEE/ACM Trans. *Audio Speech Lang. Process*, 2018, 26, 379–393. (doi:10.1109/TASLP.2017.2778423)

[13] Keen, S., Ross, J. C., Griffiths, E. T., Lanzone, M., and Farnsworth, A. A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). *Ecological Informatics*, 2014, 21, 25–33.

[14] Zhao, Z., S.-h. Zhang, Z.-y. Xu, K. Bellisario, N.-h. Dai, H. Omrani, and B. C. Pijanowski. Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, 2017, 39, 99–108.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[16] Y. LeCun. Lenet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 2015.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016, pp. 770-778.

[19] BirdLife International. IUCN Red List for birds. Downloaded from http://www.birdlife.org on 4/23/2020. 2020.

[20] C. Inskipp, H.S. Baral, S. Phuyal, T.R. Bhatt, M. Khatiwada, T. Inskipp, A. Khatiwada, S. Gurung, P.B. Singh, L. Murray, L. Poudyal. & R. Amin. The status of Nepal's Birds: The national red list series. *Zoological Society of London*, UK, 2016.

[21] Wildlife Acoustics, Kaleidoscope Pro Analysis Software. Boston, MA. Available at: https://www.wildlifeacoustics.com/products/kaleidoscope-pro, 2019.

[22] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 2016, 35, pp. 1285-1298.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[24] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv:1608.08614*, 2016.

[25] J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, Signal Processing Letters (SPL) 24 (3) (2017) 279–283.

[26] Shorten, C., Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019, 6, 60.

[27] Huang, G., Liu, Z., van der Maaten, L. and K. Q. Weinberger, K. Q. Densely Connected Convolutional Networks. In *CVPR*, 2017.

[28] Y. Aytar, C. Vondrick, and A. Torralba. SoundNet: Learning sound representations from unlabeled video. In *NIPS*, 2016.

[29] M. Ravanelli and Y. Bengio, Speaker recognition from raw waveform with sincnet. arXiv:1808.00158, 2018.