A Deep Learning Approach to the Automated Segmentation of Bird Vocalizations from Weakly Labeled Crowd-sourced Audio

Jacob Glenn Ayers[1], **Sean Perry**[1], **Samantha Prestrelski**[1], **Tianqi Zhang**[1],
Ludwig von Schoenfeldt[1], **Mugen Blue**[2], **Gabriel Steinberg**[3], **Mathias Tobler**[4],
Ian Ingram[4], **Curt Schurgers**[1], **Ryan Kastner**[1]

Abstract

Ecologists interested in monitoring the effects caused by climate change on biodiversity are increasingly turning to passive acoustic monitoring, the practice of placing autonomous audio recording units in ecosystems to monitor species richness and occupancy via species calls. However, identifying species calls in large datasets by hand is an expensive task, leading to the development of machine learning techniques to reduce cost. Due to a lack of annotated datasets of soundscape recordings, these models are often trained on large databases of community created focal recordings. A challenge of training on such data is that clips are given a "weak label," a single label that represents the whole clip. This includes segments with background noise, anthropogenic sounds, and difference species that are labeled as calls in the training data, reducing model performance. Heuristic methods exist to convert clip-level labels to "strong" call-specific labels, where the label tightly bounds the temporal length of the call and better identifies bird vocalizations. Our work improves on the current weakly to strongly labeled method used on the training data for BirdNET, the current most popular model for audio species classification. We utilize an existing RNN-CNN hybrid, resulting in a precision improvement of 12% (going to 90% precision) against our new strongly hand-labeled dataset of Peruvian bird species.

1 Introduction

Climate change threatens to cause devastating biodiversity loss in regions such as the Amazon, which hosts about a quarter of all global biodiversity and more than half of the world's remaining tropical forests. To monitor the biodiversity of such vast and dense regions, ecologists are relying more on passive acoustic monitoring [1, 2]. Compared to traditional techniques such as transects, point counts, and captures, acoustic monitoring scales better with large datasets, is less invasive, and is cheaper thanks to autonomous recording units such as AudioMoths [3–10]. Therefore, acoustic monitoring presents a better way to monitor the effects of climate change.

However, the volume of data produced makes manual labeling of audio data impractical and expensive. In order to greatly reduce the cost for researchers using passive acoustic monitoring to study bird communities, we can use automated techniques such as deep learning [3, 11, 12].

One of the most popular models for automatic bird species classification is BirdNET, a convolutional neural network (CNN) capable of classifying more than 6,000 species and trained on large bird

[1] University of California San Diego, La Jolla, California, USA
[2] University of California Merced, Merced, California, USA
[3] Demining Research Community, New York, New York, USA
[4] Population Sustainability, San Diego Zoo Wildlife Alliance, Escondido, California, USA

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

species focal-recording databases, such as xeno-canto and the Macaulay Library [13–15]. The major downside of these databases is that files are weakly labeled, meaning an entire audio file is labeled with its species. There is no indication of how much of the file contains vocalizations of the species of interest. Thus, training data will frequently include extraneous noise, such as other vocalizations, anthropogenic sounds, or environmental noise. This is in contrast to strongly labeled data, for which the exact start and end times of the call are labeled, allowing the removal of segments where the bird is not calling.

To handle weakly labeled data, it has become common practice in bioacoustics to feed training data through a detection algorithm to convert weak labels into strong labels before presenting them to the classifier [16]. However, most of these detection algorithms (discussed in section 2) are rooted in signal processing, causing reliance on signal-to-noise ratio (SNR) or other static features. This in turn leads to false detections of non-bioacoustic events.

We propose a prepossessing technique using deep learning to convert focal recordings into strongly labeled annotations for training data that improves upon currently used signal processing techniques.

2 Related Work

The weak-to-strong label (WTS) technique used by BirdNET is background foreground separation, a digital signal processing method based on the premise that focal recordings have a high SNR [17–19]. Assuming the peaks of a given signal are when the weakly labeled species vocalizes, identifying signal peaks will output strong labels for the bird calls.

Other techniques also depend on amplitude or other such static features for bioacoustic sound event detection. One approach employs a threshold for amplitude and duration for detecting bird note onsets and offsets [20]. In another, a wavelet peak detector was used to find frames with high energy, which each had a 6-second window drawn around it. The 5 windows with the highest peaks were then extracted [21].

However, given the crowdsourced nature of these focal recordings, such assumptions may not always hold during other loud sound events, such as a tree branch falling, gunshots, or human vocalizations. These cases create false positives in the training data which may lead to downstream errors and inefficiencies during training.

3 Methodology

Following the same assumptions as BirdNET's signal strength estimator, we assume the weakly labeled bird is the most prominent bird in the audio recording. Therefore, any positive bird detection by the WTS technique can be labeled with the corresponding weak label. We propose that TweetyNet, a CNN-RNN hybrid [22], and other CNNs can be retrained using open source bird/no-bird datasets and outperform the current WTS method to create labeled training data. We compare TweetyNet's performance to BirdNET's WTS pipeline as well as preexisting binary bird-no/bird DNNs like Microfaune [12] on hand-labeled xeno-canto clips.

More detailed descriptions for each method are as follows:

Foreground-Background Separation: BirdNET uses signal-to-noise ratio (SNR) thresholds foreground-background separation to isolate bird calls [13, 18]. The algorithm takes a normalized short-time Fourier transform (STFT) and generates a binary mask of the resulting STFT keeping bins in the mask that have a magnitude greater than or equal to three times its column median and row median. The binary mask highlights elements that seem more powerful than the background noise indicating a species of interest. A morphological opening is performed to reduce noise. From there, a temporal indicator vector is created by taking the sum of all the rows and setting each non-zero value in the result to 1. Two successive dilation operations are performed on the vector to further reduce noise. An annotation heuristically contains a bird if the temporal indicator vector is one.

RNN Model: Microfaune is composed of a convolution layer, a recurrent layer, and a final max-pooling layer [12]. The convolution layer takes in audio that has been converted to a mel-spectrogram and extracts features from that audio. It passes this to the recurrent layer, which computes the features at the time step level based on neighboring time steps. The predictions on these steps amount to a

Technique	Time Ratio	Number of 3 Second Segments	Precision	Recall	F1
FG BG Sep	1.000	**21582**	0.7797	**.9831**	0.8697
Microfaune	1.329	13200	0.7767	0.7062	0.7398
TweetyNet	**0.853**	18365	**0.9009**	0.9704	**0.9344**

Table 1: The aggregate metrics for the results of each model.

local score array of elements [0,1] representing bird presence/absence. The model was trained on field [23], mobile device [24], and soundscape recordings [25, 26]. We take the local score arrays and apply an isolation technique that sets a lower bound static threshold of 0.15, and a relative threshold of 3.2 times the median of the local score array. It then loops through all elements in the array that are greater than both the static and relative thresholds and applies a 1.5 s window around said point. All overlapping windows are consolidated into a single strong label.

CNN-RNN Hybrid Model: TweetyNet is a neural network model designed to segment and annotate bird songs at the syllable level [22]. The network is built from a convolution layer which uses two rounds of standard convolution and max-pooling to learn the features of the data, a recurrent layer that computes the features found for each time-bin, and a linear layer that assigns a similarity score for each class to each time-bin based on the recurrent layer's output. We assigned the similarity score of the bird to each time-bin and used a threshold to distinguish between time-bins with and without bird calls. We have retrained the model with field recordings (freefield1010 [23]) and mobile device recordings (warblrb10k [24]) similar to Microfaune.

Each method was evaluated on data from xeno-canto of a list of 1000 provided by an ornithologist familiar with the Madre de Dios region, a biodiversity hotspot in Southeastern Peru [27]. We selected audio files marked "A" or "B" quality [28]. The audio was then uploaded into a browser-based acoustic labeling tool [29]. Eight students were instructed to select a species, listen to all corresponding xeno-canto recordings in order to identify the intersecting sounds most-likely associated with the weak label, and only labeled calls from the specified species ignoring other vocalizations. Students were explicitly told to label 5 audio files of high priority species and aim for about 3 files per species for the rest (see appendix). Post-labeling, the audio was processed into uniform clips by binning the annotations into three second intervals. Following this method, 630 of the selected 1000 species were labeled, resulting in 17,125 three second annotations.

All of the annotations from the automated labeling techniques as well as the manual annotations were converted into left-aligned uniform three second segments. We compute True Positives (TP), False Positives (FP), and False Negatives (FN) for each automated labeling technique by comparing their strong labels to the human labeled ground-truth. Over a given three second interval, an interval is TP if both the automated label and human label agree and FP If the automated label identifies a bird where the human does not. The converse is labeled a FN. From these descriptions of TP, FP and FN, we define precision, recall, and F1 as seen in appendix (Equation 1).

4 Results

The foreground-background separation technique produced the most annotations (21582) and had the highest recall (98.31 %), as shown in Table 1. However, TweetyNet had the highest precision (90.0 %) out of the three techniques, as well as the highest F1 score (93.4 %), in addition to being the fastest method of the three, taking 85.3 % of the time as foreground-background separation. Microfaune was outperformed in every metric. In addition, before three-second preprocessing, we collected the statistics for the annotation duration's can be found in Table 2 in the appendix.

When examining the scores for each species (seen in Figure 1), we observe TweetyNet's precision and F1 scores were higher than the other models for more species, implying TweetyNet performed well across the different species studied. The foreground-background separation technique was higher with respect to recall than the other models, implying for most species there were few false negatives at the cost of a higher rate of false positive.

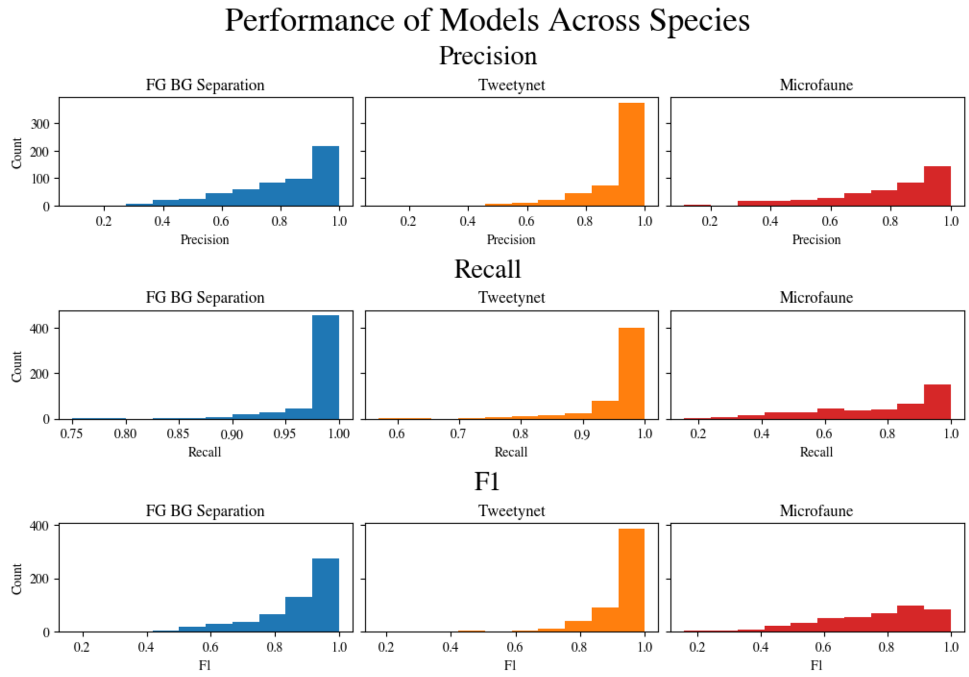

Figure 1: The distribution for each performance metric across the 630 species of interest.

5 Discussion

The experimental results favor TweetyNet. Notably, TweetyNet has higher precision, implying lower false positives and fewer sections of audio incorrectly labeled as a region of interest. This leads to less noisy data in training, which is our ultimate goal. TweetyNet's higher F1 also means that TweetyNet is not increasing precision by being overly restrictive, implying it produces enough annotations for a sufficiently large training dataset. Lastly, according to Figure 1, we can see that the TweetyNet scores for precision and F1 metrics are more skewed left, implying TweetyNet can work well across a wide range of species. While BirdNET and more recent approaches [30] continue to use a signal-to-noise ratio method, these results imply that TweetyNet is well suited for distilling higher-quality training data from a dataset of focal recordings.

Future work might consider the performance of these models on non-focal recordings with lower SNR as the data we studied were primarily focal recordings. Furthermore, as we relied on student-labeled annotations, any future replication should consider using expert annotators.

Our Python code encapsulating these methods as well as our student-hand-labeled, strongly-labeled evaluation dataset can be accessed at `https://github.com/UCSD-E4E/AID_NeurIPS_2024`.

6 Conclusion

We have demonstrated that through deep learning, we can create better WTS pipelines than that currently being used by the largest passive acoustic monitoring analysis tool available to ecologists worldwide. The new tool can therefore be used to better parse through training datasets to reduce noise and better identify species of interest. Future work should consider further investigations into WTS pipelines (such as ensembling these weakly to strongly label pipelines to further improve precision) as they can continue to produce richer datasets. That way improvements can be made to multispecies models without having to modify the underlying architectures. Towards these ends, we hope our work can make it easier to identify species in passive recordings before they can no longer be heard.

Acknowledgments

This work was funded in part by the REU Site Engineers for Exploration, supported by the National Science Foundation under grant number 1852403. Additional support for some students was provided by the Summer Research Internship Program and Halıcıoğlu Data Science Institute (HDSI) Undergraduate Scholarship Program at the University of California, San Diego.

We would like to thank our friends and collaborators from the San Diego Zoo Wildlife Alliance and the World Wide Fund for Nature Peru for collecting the Audiomoth recordings in the Peruvian Amazon. We would also like to thank the Computer Science and Engineering and Electrical and Computing Engineering departments of UC San Diego. We would like to thank Nathan Hui, the Staff Engineer at Engineers for Exploration, for his technical guidance throughout the project. We would like to thank Vanessa Salgado, Nishant Balaji, Shreyas Anantha Ramaprasad, and Pranjal Chatterjee for significant contributions to the automated segmentation techniques. We would also like to thank Alan Arce, Dylan Nelson, Edmundo Zamora, Harsha Jagarlamudi, Jinsong Yang, Kelly Kong, Qiaochen Sun, and Wesley Zhen who volunteered their time by labeling hours of audio through the Data Science Capstone class at UC San Diego. Finally, we would like to thank Letty Salinas Sanchez from the Universidad Nacional Mayor de San Marcos Natural History Museum in Lima, Peru who provided us with a list of bird species known in Madre de Dios, Peru.

References

[1] R. A. Betts, Y. Malhi, and J. T. Roberts, "The future of the amazon: new perspectives from climate, ecosystem and social sciences," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1498, pp. 1729–1735, 2008.

[2] L. S. M. Sugai, T. S. F. Silva, J. Ribeiro, José Wagner, and D. Llusia, "Terrestrial passive acoustic monitoring: Review and perspectives," vol. 69, no. 1, pp. 15–25, _eprint: https://academic.oup.com/bioscience/article-pdf/69/1/15/27503065/biy147.pdf. [Online]. Available: https://doi.org/10.1093/biosci/biy147

[3] S. R. P.-J. Ross, D. P. O'Connell, J. L. Deichmann, C. Desjonquères, A. Gasc, J. N. Phillips, S. S. Sethi, C. M. Wood, and Z. Burivalova, "Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions," *Functional Ecology*, vol. 37, no. 4, pp. 959–975, 2023. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.14275

[4] A. P. Hill, P. Prince, E. Piña Covarrubias, C. P. Doncaster, J. L. Snaddon, and A. Rogers, "Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," *Methods in Ecology and Evolution*, vol. 9, no. 5, pp. 1199–1211, 2018. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12955

[5] P. Stahlschmidt and C. A. Brühl, "Bats as bioindicators – the need of a standardized method for acoustic bat activity surveys," *Methods in Ecology and Evolution*, vol. 3, no. 3, pp. 503–508, 2012. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2012.00188.x

[6] S. H. Borges, "Bird assemblages in secondary forests developing after slash-and-burn agriculture in the brazilian amazon," *Journal of Tropical Ecology*, vol. 23, no. 4, pp. 469–477, 2007. [Online]. Available: http://www.jstor.org/stable/4499120

[7] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[8] H. H. Welsh Jr. and L. M. Ollivier, "Stream amphibians as indicators of ecosystem stress:a case study from california's redwoods," *Ecological Applications*, vol. 8, no. 4, pp. 1118–1132, 1998. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/1051-0761%281998%29008%5B1118%3ASAAIOE%5D2.0.CO%3B2

[9] J. Ayers, S. Perry, V. Tiwari, M. Blue, N. Balaji, C. Schurgers, R. Kastner, M. Tobler, and I. Ingram, "Reducing the barriers of acquiring ground-truth from biodiversity rich audio datasets using intelligent sampling techniques," in *Tackling Climate Change with AI Workshop, Conference on Neural Information Processing Systems*, 2021.

[10] L. S. M. Sugai, T. S. F. Silva, J. Ribeiro, José Wagner, and D. Llusia, "Terrestrial Passive Acoustic Monitoring: Review and Perspectives," *BioScience*, vol. 69, no. 1, pp. 15–25, 11 2018. [Online]. Available: https://doi.org/10.1093/biosci/biy147

[11] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000273

[12] V. Morfi and D. Stowell, "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, no. 8, p. 1397, 2018.

[13] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[14] J. Martinsson and S. Kahl, "Architecture description of birdnet v2.4," https://github.com/kahst/BirdNET-Analyzer/issues/177#issuecomment-1772538736, 2023.

[15] G. Bota, R. Manzano-Rubio, L. Catalán, J. Gómez-Catasús, and C. Pérez-Granados, "Hearing to the unseen: Audiomoth and birdnet as a cheap and easy method for monitoring cryptic bird species," *Sensors*, vol. 23, no. 16, p. 7176, 2023.

[16] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar. 2022.

[17] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks." *CLEF (working notes)*, vol. 1866, 2017.

[18] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," *LifeCLEF 2016*, pp. 547–559, 2016.

[19] J. C. C.-H. To, "Cocktail party problem for bird sounds."

[20] T. Koumura and K. Okanoya, "Automatic recognition of element classes and boundaries in the birdsong with variable sequences," *PloS one*, vol. 11, no. 7, p. e0159188, 2016.

[21] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 636–640.

[22] Y. Cohen, D. A. Nicholson, A. Sanchioni, E. K. Mallaber, V. Skidanova, and T. J. Gardner, "Automated annotation of birdsong with a neural network that segments spectrograms," *eLife*.

[23] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," 2013.

[24] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018. [Online]. Available: https://arxiv.org/abs/1807.05812

[25] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: a dataset and benchmark for avian flight call detection," in *Proc. IEEE ICASSP*, April 2018.

[26] V. Morfi and D. Stowell, "microfaune: Bird detection and identification using deep learning," https://github.com/microfaune/microfaune, 2018.

[27] L. Brotto, J. Murray, D. Pettenella, L. Secco, and M. Masiro, "4.2 biodiversity in the peruvian amazon," *Biodiversity conservation in certified forests*, p. 112, 2010.

[28] W. Vellinga, "Xeno-canto-bird sounds from around the world," in *Xeno-canto Foundation for Nature Sounds. Occurrence dataset*. GBIF. org, 2020.

[29] S. Perry, V. Tiwari, N. Balaji, E. Joun, J. Ayers, M. Tobler, I. Ingram, R. Kastner, and C. Schurgers, "Pyrenote: a web-based, manual annotation tool for passive acoustic monitoring," in *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, 2021, pp. 633–638.

[30] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, 2023.

7 Appendix

7.1 Precision, Recall, and F1

In our paper, we reference the evaluation metrics of precision, recall, and F1, which are calculated using True Positives (TPs), False Positives (FPs), and False Negatives (FNs). The exact formulas are as follows:

$$\begin{aligned}\text{precision} &= \frac{\text{total\_TP}}{\text{total\_TP} + \text{ total\_FP}}\\[2ex] \text{recall} &= \frac{\text{total\_TP}}{\text{total\_TP} + \text{ total\_FN}}\\[2ex] \text{F1} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}\end{aligned} \tag{1}$$

7.2 Annotation Durations

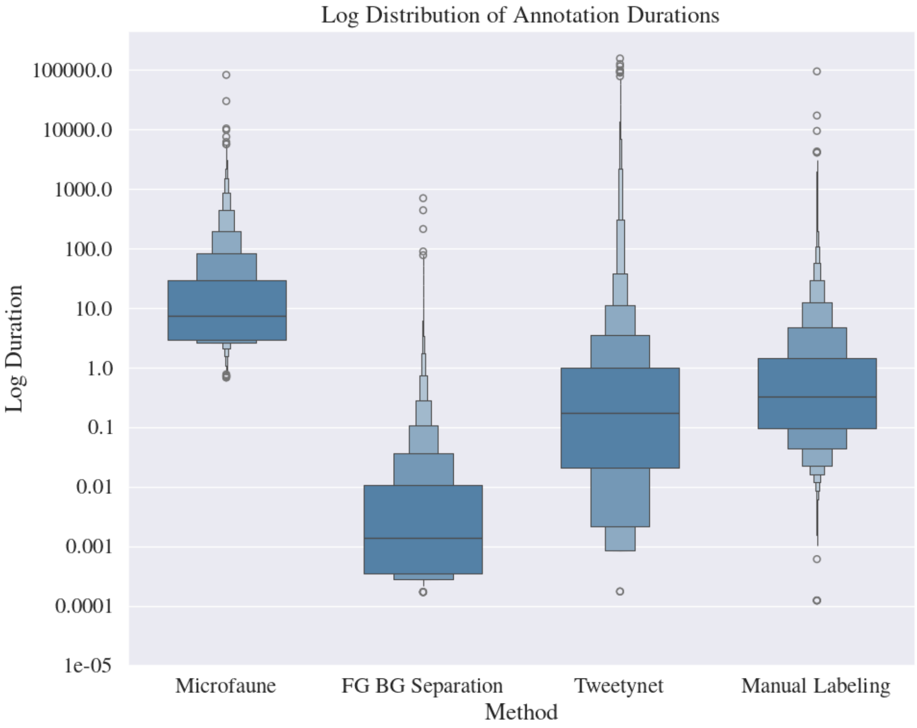

Figure 2: The distribution of annotation lengths for each method.

Technique	Mode	Mean	STD	Min	Q1	Median	Q3	Max
Manual	0.3	1.07	1.85	0.00	0.36	0.61	1.17	143.32
FG BG Sep	0.03	0.13	0.24	0.02	0.03	0.06	0.14	17.07
Microfaune	1.52	3.85	4.44	0.84	1.59	2.36	4.28	135.10
TweetyNet	0.05	1.38	5.70	0.02	0.19	0.47	1.00	178.00

Table 2: Statistics for the duration of annotations produced by each technique.

Prior to performing chunking to produce uniform, 3-second annotations, we collected data on the lengths of annotations generated by each method. As can be seen in Figure 2 and Table 2, the lengths of annotations produced by TweetyNet had the most similar distribution to the lengths of the ground truth (Manual Labeling) annotations. This supports that using DL techniques can generate annotations that are closer to those made by humans than previous techniques.

7.3 Species Label Distribution

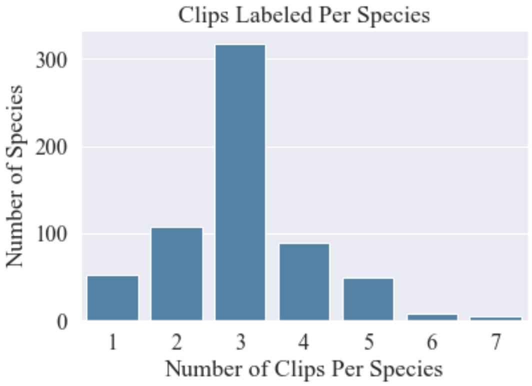

Figure 3: The distribution of labels per species. Most species had 3 labeled clips each. Due to the crowdsourced nature of the labeling process, species had differing number of labels.