Deep Learning For Ecology Project 2024

Josiah K. Kimani [1]

Abstract

Understanding soundscapes of the environment we live in enables us to perceive the changing ecological niches with time. Soundscapes ecology attempts to visualize our environment to assess patterns and interactions of biophony, anthrophony, and geophony. For this visualization to happen, acoustic data from these three categories of sounds in a soundscape is analyzed and processed. This process can be daunting and time-consuming considering the ever-growing size of acoustic data from ecologists. In this paper, I attempt to solve this analysis of acoustic data by the use of Artificial Intelligence, particularly deep learning, to process, visualize, and classify sounds in any soundscape.

1. Introduction

The application of Artificial Intelligence (AI) and Machine Learning (ML) is currently revolutionizing how global ecological research is done (Ryo, 2024). The fast turn around time in data analysis, processing and high precision prediction is at the core of this revolution. Making sense of different sounds in the ecosystem has proven to be more effective with the use of deep learning models.

In this paper, I develop a deep learning model that predicts the presence of anthrophony, biophony, wind(in place of geophony) and also their absence. This prediction is vital towards understanding the soundscape of any environment which enables the making of informed decisions on conservation measures as well as the interaction of human and animal sounds with that of nature. (Pijanowski et al., 2011) assesses the interaction of human and animal sounds which shows the effect of human activities on the mating vocalizations of birds. (To et al., 2021) espouses a novel observation on the changing vocalization frequencies of birds in the urban environments. Interestingly, it is observed that an urban bird vocalizes in higher frequencies to avoid sound masking by the anthropogenic noise that exists.

[1]African Institute for Mathematical Sciences (AIMS) South Africa, 6 Melrose Road, Muizenberg 7975, Cape Town, South Africa. Correspondence to: Josiah K. Kimani <josiah@aims.ac.za>.

2. Literature Review

Deep Learning has been applied to many different tasks including but not limited to computer vision, speech recognition, Natural Language Processing(NLP) and audio processing. Classification is at the core of these tasks which therefore supports the exploration of the Deep Learning in computational bioacoustics. In computational bioacoustics, we attempt to study the classification different sounds of in different environments over time either for conservation, understanding biodiversity health, impact of urban development, climate change, noise pollution among others.

Despite the power that comes with the use of deep learning, there exists challenges unique to bioacoustic data, such as overlapping of different sounds, lack of precise labeling for supervised and semi-supervised models and unbalanced datasets for training and validation. (Stowell, 2022) suggests data augmentation approaches to mitigate unbalanced datasets. However, there need to be a careful choice of the augmentation technique so that we don't introduce artifacts on our data or change the correct frequency of the vocalizations.

3. Methodology

The audio file was converted to frequency amplitudes which were then used to generate mel spectrograms for each segment of the audio. The mel spectrograms after normalization and adding an extra dimension for channeling formed the input of the model.

For the multiclass classification of anthrophony, biophony, wind and absence/silence, I considered a Convolutional Recurrent Neural Network(CRNN) architecture which involved two additional 2D recurrent(Long Short-Term Memory (LSTM)) layers after alternating three 2D convolution and max-pooling layers. The motivation for this combination is as a result of the spatial and temporal features of bioacoustic signals. Spatial as a result of varying frequencies

and temporal because of the time-domain characteristics. (Gupta et al., 2021) demonstrated the effectiveness of this architecture based on the classification of 100 bird species. CRNN architecture outperformed standalone CNNs and other hybrid models.

Layer (type)	Output Shape	Param #
input_layer_3 (InputLayer)	(None, 64, 157, 1)	0
conv2d_9 (Conv2D)	(None, 61, 154, 32)	544
max_pooling2d_9 (MaxPooling2D)	(None, 30, 77, 32)	0
conv2d_10 (Conv2D)	(None, 27, 74, 64)	32,832
max_pooling2d_10 (MaxPooling2D)	(None, 13, 37, 64)	0
conv2d_11 (Conv2D)	(None, 10, 34, 128)	131,200
max_pooling2d_11 (MaxPooling2D)	(None, 5, 17, 128)	0
reshape_1 (Reshape)	(None, 5, 17, 128, 1)	0
conv_lstm2d_2 (ConvLSTM2D)	(None, 5, 14, 125, 16)	17,472
conv_lstm2d_3 (ConvLSTM2D)	(None, 11, 122, 32)	98,432
flatten_1 (Flatten)	(None, 42944)	0
dense_1 (Dense)	(None, 4)	171,780

Total params: 452,260 (1.73 MB)
Trainable params: 452,260 (1.73 MB)
Non-trainable params: 0 (0.00 B)

Figure 1. This is the **CRNN** architecture used in this paper(3 Conv2D, 3 MaxPooling2D, 2 ConvLSTM2D and 1 Dense layer)

As shown in figure 1, the total number of parameters is **452,260** which are all trainable since the model is built from scratch.

The input size of the model is **(None, 64, 157, 1)** where *(None)* is a placeholder for the sample size, *(64,157)* is the size and *(1)* is channel of the spectrogram. The size of the spectrogram stems from the choice of the model's hyperparameters:

- n fft = 1024 *(number of points for Fast Fourier Transform)*,
- hop length = 256 *(number of samples the window moves between successive frames)*,
- n mels = 64 *(number of melody frequency bins used for the spectrogram)*

The output of the model is 4 mutliclass labels representing anthrophony, biophony, wind and absence/nothing probabilities. Therefore, I used the **sigmoid** function which resulted into four probabilities of the corresponding labels. Note that there exists a possibility of having permutations of anthrophony, biophony and wind per event or spectrogram.

The loss function chosen was the **binary-cross entropy** since we have binary labels in addition to having multi-label classification.

Adam(Adaptive Momentum Estimation) optimizer was used for accelerating gradient descent while incorporating adaptive learning rates.

4. Dataset Analysis and Plots

Dataset used in this model was comprised of two sets:

- Initial Data : 18 audio files and corresponding annotations provided before hand
- My Data: 4 audio files and their corresponding annotations which I recorded using my phone and annotated using the sonic visualizer software.

4.1. Initial Data

The distribution of the training and validation dataset after 80-20 split is shown in the figure 2 below:

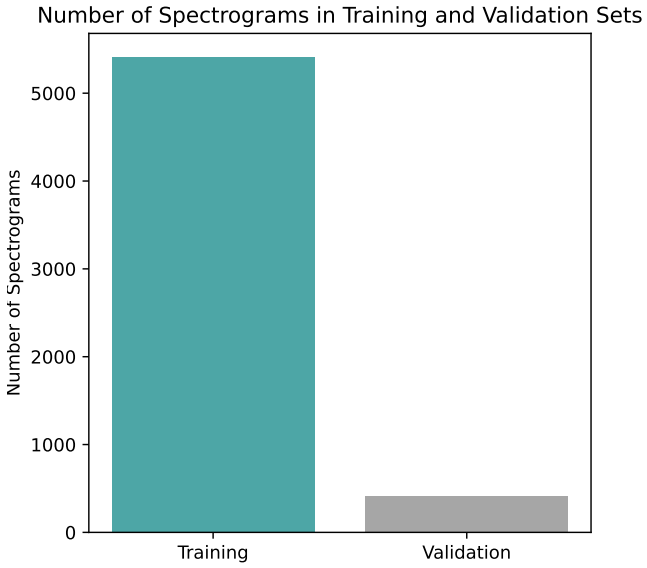

Figure 2. Training and validation data distribution before augmentation

The distribution of the unique events before augmentation is shown in figure 3 below:

To avoid the above unbalanced distribution of unique events in the training data, I augmented the data using the rolling technique. This resulted in a distribution as shown in figure 4 below:

4.2. My Data

The distribution of the training and validation dataset after 80-20 split is shown in the figure 5 below:

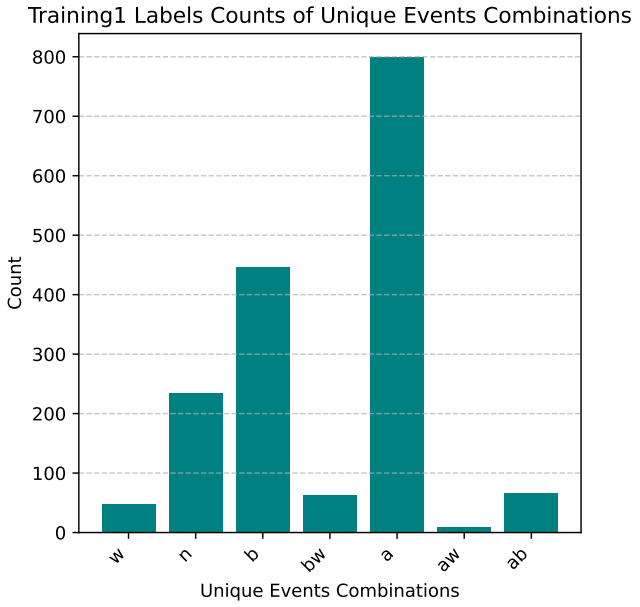

Figure 3. Unique events training data distribution before augmentation

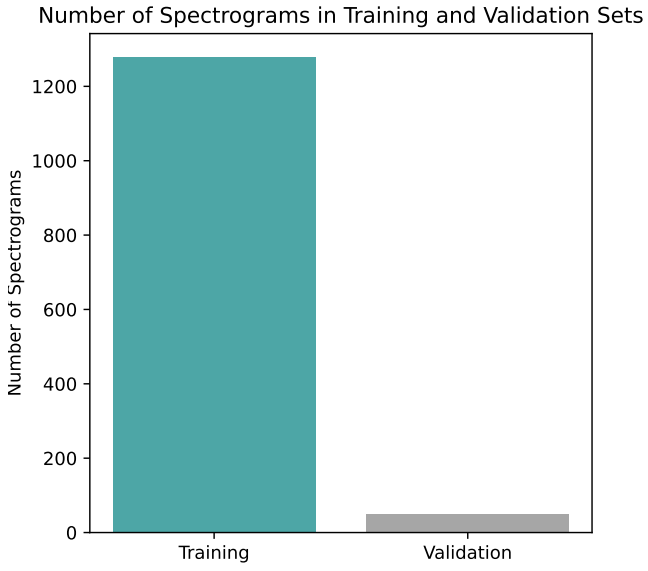

Figure 5. Training and validation data distribution before augmentation

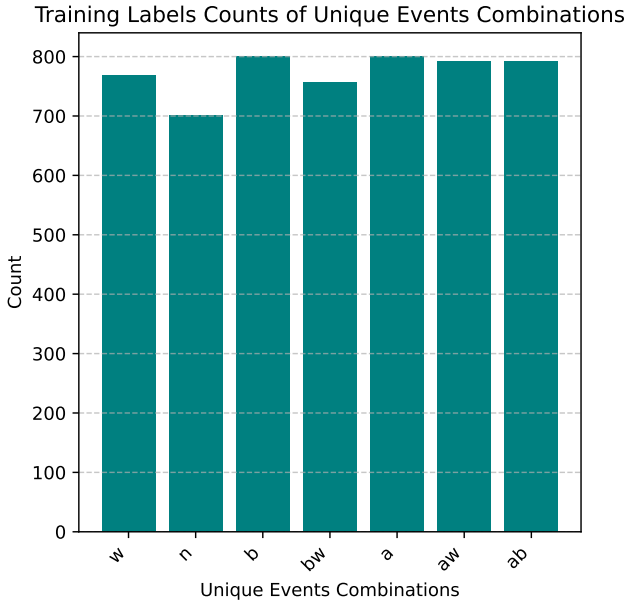

Figure 4. Unique events training data distribution after augmentation

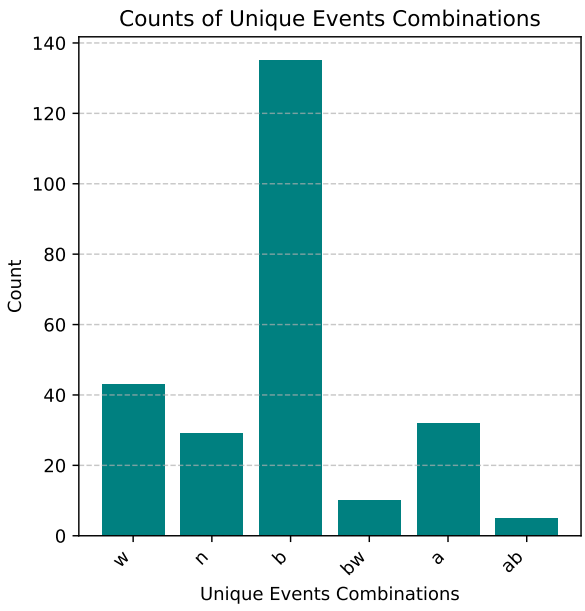

Figure 6. Unique events training data distribution before augmentation

The distribution of the unique events before augmentation is shown in figure 6 below:

To avoid the above unbalanced distribution of unique events in the training data, I augmented the data using the rolling technique. This resulted in a distribution as shown in figure 7 below:

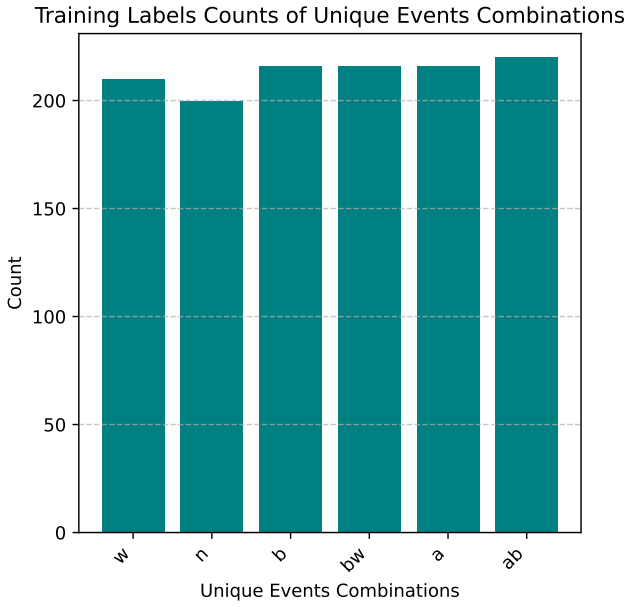

Figure 7. Unique events training data distribution after augmentation

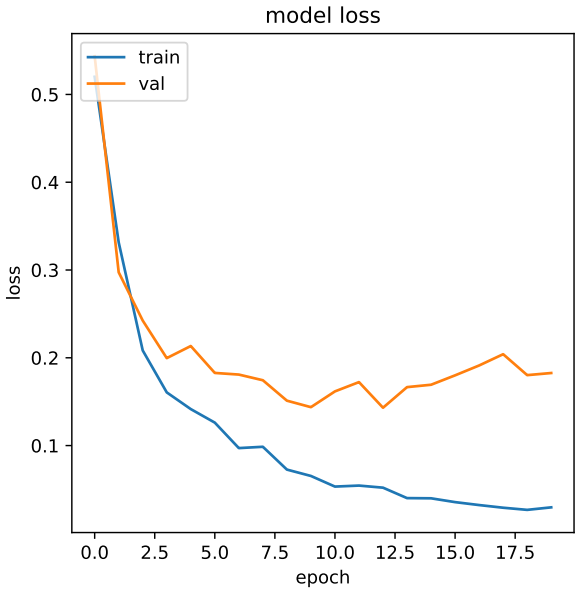

Figure 8. Initial data model loss of training vs validation

5. Results and Discussion

5.1. Training

Using the same model architecture, I trained a model on the initial data set and achieved a training accuracy of 85%.

The plots of the loss figure 8 and accuracy figure 9 for training vs validation for the initial data as shown below:

Using my data I trained another model achieving a training accuracy of 70%. Similarly, the plots of the loss figure 10 and accuracy figure 11 for training vs validation for the initial data as shown below:

For the model trained on the initial we can observe that the loss for the training and validation datasets decreases monotonically downwards while their accuracies increase monotonically upwards. However, it can be observed that the curves are not smooth for both training and validation datasets pointing to instability of the model probably caused by the complexity in the model's architecture.

The model trained on my data is very unstable as seen by both the loss and accuracy plots. This also supports the hypothesis of complexity in the model's architecture.

5.2. Prediction

Predictions were made on some test files which the model was not exposed to. An acoustic plot over time for the initial data model figure 12 and my data model figure 13 was discovered as below: The difference in the two predictions validates the accuracies achieved in each training and validation sets. Accuracies for both models can be improved with availability of more data, better augmentation techniques or

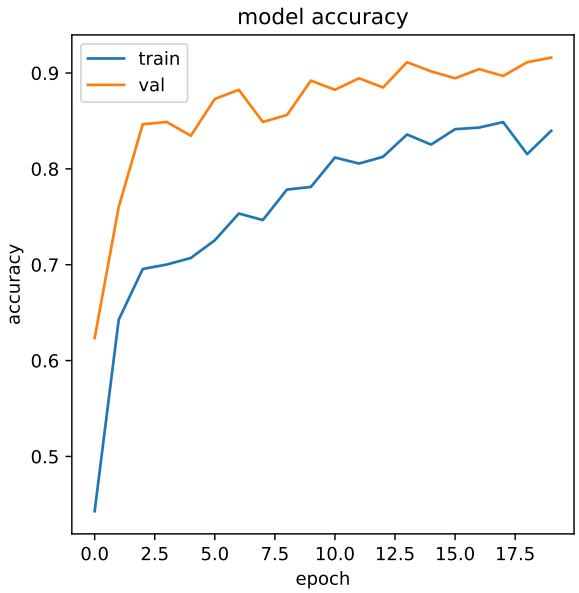

Figure 9. Initial data model accuracy of training vs validation

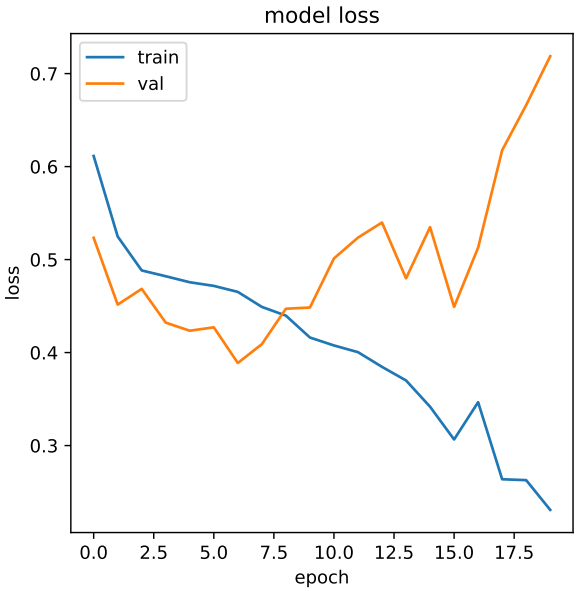

Figure 10. My data model loss of training vs validation

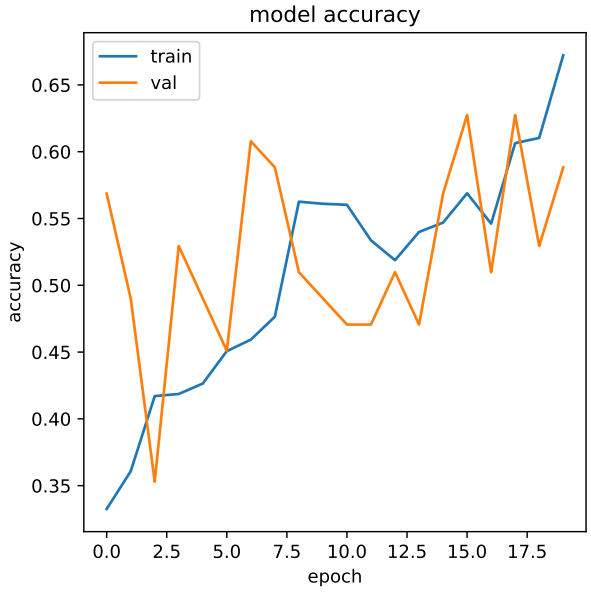

Figure 11. My data model accuracy of training vs validation

high-precision annotation of labels.

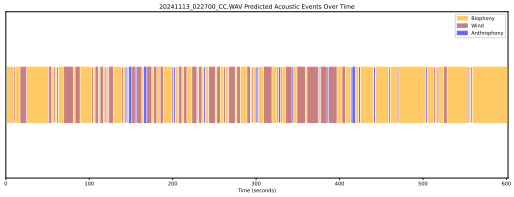

Figure 12. My data model's acoustic plot over time prediction on sample test file

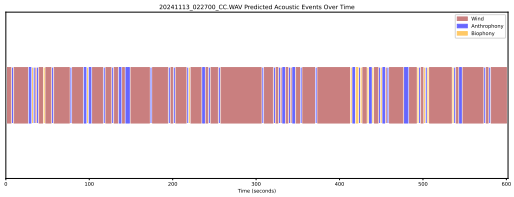

Figure 13. My data model's acoustic plot over time prediction on sample test file

6. Conclusion

There exists room for improvement of this model's architecture given more training examples and compute time. Bioacoustic ecology needs more innovation and use of deep learning models to aid in data processing, analysis and real-time and reliable prediction. Most deep learning models still suffer from lack of generalization as a result of limited data of different biophony, anthropophony and geophony. These models most often times are niche-specific.

7. Source Code and Data Folder

The following is the link to the Google Drive root folder containing the files and folders: Google Drive Link.

References

Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., and Lavista Ferres, J. Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific Reports*, 11, 08 2021. doi: 10.1038/s41598-021-96446-w.

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., and Pieretti, N. Soundscape ecology: The science of sound in the landscape. *BioScience*, 61(3):203–216, 03 2011. ISSN 0006-3568. doi: 10.1525/bio.2011.61.3.6. URL https://doi.org/10.1525/bio.2011.61.3.6.

Ryo, M. Ecology with artificial intelligence and machine learning in asia: A historical perspective and emerging trends. *Ecological Research*, 39(1):5–14, 2024.

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

To, A. W. Y., Dingle, C., and Collins, S. A. Multiple constraints on urban bird communication: both abiotic and biotic noise shape songs in cities. *Behavioral Ecology*, 32(5):1042–1053, 07 2021. ISSN 1045-2249. doi: 10.1093/beheco/arab058. URL https://doi.org/10.1093/beheco/arab058.