Developing RailNET: An Exploration into the use of Deep Learning Models for Rare Species Monitoring - Technical Report

Johannes Nelson

April 2023

Abstract

With biodiversity in rapid decline worldwide, efficient wildlife monitoring solutions are increasingly important. Deep learning models offer new opportunities in acoustic monitoring, a field where fully manual annotation of audio recordings in prohibitively labor-intensive. State-of-the-art models for bird monitoring perform well in general, but performance varies with rare, elusive species and generalization from the focal recordings in training data to soundscape data collected in monitoring programs has proved to be a major challenge. This project focuses on the Virginia rail–a cryptic marsh bird that can serve as a stand-in for other elusive monitoring targets for whom species-specific classifiers might be of interest. We explored various data preparation methods such as applying strong labels to weakly labeled public recordings and engineering a negative class that is biologically and regionally relevant. We also tested different data augmentation techniques designed to better mimic the distribution of soundscape data. We experimented with a handful of architectures: a basic CNN, ResNets with 18 and 34 layers, a Wide ResNet, a ResNet with self-attention, and VGGish. Results indicate that the application of strong labels improved model performance compared to automated energy threshold selection techniques, as expected, and also led to models that showed notable improvements in their ability to generalize to real-world, soundscape data. The ResNets performed best overall and were ultimately used to create the final classifier. An ensemble of three ResNets trained with both data augmentation regimes was able to achieve the highest F1 and F-Beta scores of all classifiers and ensemble combinations tested. We then compared this final ensemble, which we call RailNET, to BirdNET, the state-of-the-art model for avian monitoring–assessing performance of each on our soundscape set at various confidence score thresholds. For classification of the Virginia rail, RailNET achieved significantly higher recall at every score cutoff with only minor sacrifices to precision. This demonstrates our workflow's effectiveness at creating a species- and region-specific acoustic monitoring solution for rare, elusive targets.

1 Introduction

In order to design and implement effective wildlife conservation strategies, researchers need efficient ways to collect meaningful data about wildlife distributions. Birds, being conspicuous, vocal, and present in most terrestrial ecosystems are often used as indicators for underlying environmental health. Because bird vocalizations are distinct between species, wildlife surveyors often use their audio cues when gathering data in the field. As recording devices have become smaller and cheaper, and as advances in machine learning have allowed for more efficient processing of audio data, passive acoustic monitoring has emerged as a promising method for monitoring wildlife at larger scales [Bro+17; Gib+19]. State-of-the-art models like BirdNET can classify nearly a thousand species of birds [Kah+21], but performance varies across species. Species who are rare or cryptic and who vocalize infrequently pose a challenge to automated classifiers. Recall is important for these targets, where a small number of vocalizations may be all that are available in a large number of recordings. Precision is important as well, in so far as it determines the amount of manual validation of the final predictions that will be required. Adding to these challenges is that fact that these species are often underrepresented in the public databases of recordings that are used to train these models. Furthermore, audio classification models for biodiversity monitoring struggle with the domain shift from training–where they are trained on relatively high quality, focal recordings–to application where they are being used to analyze often poor quality, unfocused soundscape recordings.

For this project, we explored various data-centric and model-centric approaches to address these problems, such as the application of strong labels to the weakly labeled recordings of public databases, the curation of a regionally and biologically relevant samples to represent the binary classifier's negative class, and data augmentation regimes designed to narrow the distribution gap between the focal recordings available in public databases and the soundscape recordings of monitoring programs. We test a base-CNN model, ResNets with 18 and 34 layers, a Wide ResNet, a ResNet-18 with self-attention, and VGGish. In addition, we analyze how the models perform with various score threshold cutoffs as recommended by [Kni+17], who also recommend reporting on a representative set of performance metrics, advice which we follow by tracking precision, recall, F-score, and F-beta score (with a beta value of 2). Our results indicate that an ensemble of ResNets trained with a strongly labeled positive class, a regionally relevant negative class, and with various data augmentation regimes performs best and outperforms BirdNET in its ability to classify Virginia rail vocalizations with satisfactory recall and precision. This workflow could potentially be followed to create any species-specific classifier where higher recall is desired while still maintaining acceptable levels of precision. It could also feasibly be adapted to multi-class classification for a suite of target species.

2 Related Work

The decline in biodiversity around the globe is alarming. The impact on bird populations is well documented, with a staggering estimate from [**rose**] of nearly one-third of all birds in North America having been lost since 1970. Passive acoustic monitoring has been used to successfully improve distribution data for rare species, to monitor occupancy at landscape scales, and to estimate population density[CA16; WGP19; Woo+19; PT21]. For monitoring efforts at large spatial and temporal scales, the need to manually validate recordings severely restricts the scope of monitoring efforts, necessitating automated methods for processing the audio data in search of target signals. The classification of bird sounds using machine learning is not new. Machine- and deep-learning practitioners who are not even ornithologists have been attracted to this particular use-case because of the fact that there exist large, labeled audio databases of bird calls that are publicly available–making this area rife with opportunities for experimentaion and exploration. The classification task is often approached as a multi-class problem, with a focus on general avian diversity monitoring in order to help ecologists estimate state variables such as species richness. For example, [Kah+21] created BirdNET, a model that is capable of classifying vocalizations from over 900 bird species. BirdNET has become the state-of-the-art model for avian diversity monitoring. It is updated regularly with more species (including non-avian species) and increasingly user-friendly interfaces to enable conservationists without programming backgrounds to use it. BirdNET's performance across species varies, however, and the model might not always be suitable for species-specific monitoring efforts, such as those targeting sensitive or endangered species. [Col+22] found for their study of a group of focal species that acceptable BirdNET confidence cutoffs for each species ranged from 0.15 to 0.95, illustrating the inconsistent performance across different species–and these were fairly common, vocal species. For rare, elusive species that vocalize infrequently, challenges in automated detection are amplified.

For rare species monitoring, recall becomes very important, because missing even a small handful of vocalizations might significantly impact the resulting analyses. Validating presence in a recording for a highly vocal bird involves only picking up on one of many vocalizations that are available on the recordings, whereas validating presence for an elusive target involves picking up on at least one of what is often a small number of signals. This creates the need to balance very high recall (the ability to detect as many vocalizations as possible) with enough precision that sifting through the false positives of a high-sensitivity detection process does not become prohibitively time-consuming. Since monitoring programs can easily collect thousands of hours of audio, both of these metrics need to be taken into careful consideration when training models.

Using convolutional neural networks has proven to be effective in audio classification. [TPS21] compared transfer learning methods using a variety of pretrained CNNs, including GoogLeNet, SqueezeNet, ShuffleNet, VGGish, YAMNet, finding that the latter two models, which were specifically geared towards audio classification rather than image classification, performed best. BirdNET, by [Kah+21], is based on a Wide ResNet architecture and includes 157 layers and roughly 27 million trainable parameters. It uses 64 x 384 single-channel mel-spectrograms as input. The training data was made up of 1.5 million of these samples and represented nearly 1,000 species. They oversampled underrepresented classes in order to address class imbalance, but they only do this to a maximum of 350 samples. They used the ADAM optimizer, starting with a learning rate of 0.001, and a batch size of 32, reducing the learning rate by a factor of 0.5 whenever validation loss stalled. They did not explore ensemble techniques because of computational demands.

Following in the footsteps of the above studies, this project will explore transfer learning with pretrained ResNets and VGGish models. We differ in our approach to data curation guided by the belief that more effort in creating a cleaner training dataset would yield better performance. We also experimented with self-attention layers in the ResNets to see if they might help the models to track dependencies across more distant parts of the input spectrograms. The fact that we are focused on single-species monitoring means that we have a smaller dataset, which allowed us to more easily experiment with ensemble classifiers. When tracking performance metrics during training, we prioritize recall, while keeping in mind that it must be paired with high enough precision to make the model viable to use in real applications. For this reason, we also closely track the F1 score and the F-Beta score (with a beta value of 2). We then perform an analysis of various score cutoffs and how they impact final performance.

3 Dataset and Features

Our data comes from two sources: Xeno-canto and the Macaulay Library, two online databases of labelled bird recordings. Xeno-canto was our primary source, and we only used Macaulay Library when more recordings for specific vocalization types were needed. The recordings in these databases vary in format, length, sampling rate, and number of channels. We standardized the inputs so that all data samples had equal sampling rates, lengths, and number of channels prior to model training in order to avoid the need for additional preprocessing steps within the data-loading pipeline that might have slowed down training and our ability to iterate efficiently.

3.1 Standardizing Training Data

This process began by converting all .mp3 files to .wav files. Then, we identified the best sampling rate for our inputs. Since most recordings of Virginia rails in the databases were sampled at 44.1 or 48 kHz, we decided on 44.1 kHz as the sampling rate for our inputs so that we could downsample rather than upsample, which would risk introducing interpolation artifacts into our inputs. The few recordings sampled at rates lower than 44.1 kHz were discarded. We then averaged all multi-channel recordings down to a single channel. These data preparation steps were done using R with the help of packages warbleR, which includes functions to interact with the Xeno-canto database, and tuneR, which includes functions to handle and manipulate audio data.

3.2 Defining Positive and Negative Classes

To create equally sized audio segments, we decided on three-second windows, which match the window length that [Kah+21] used for BirdNET. The recordings on Xeno-canto are weakly labeled, meaning that they indicate that the species is present somewhere within them, but do not give timestamps. While some include information about background species vocalizations, these also do not have timestamps. Because many approaches to bird sound classification are focused on a large number of species (and because domain knowledge in identifying birds by vocalization might not be available to these projects), they need to rely on automated means for extracting spectrograms of audio segments. This process, however, results in potentially significant data integrity issues, with the signals of the target class regularly mixed in with other signals–a fact illustrated in Figure 1. We use scripted processes in R to page through the recordings and accurately label target signals with timestamps, which we then extract as distinct three-second audio files. In addition to this application of strong labels, we also track subclasses of vocalization types to ensure that the Virginia rail's vocal repertoire is adequately represented and that the model does not only learn to distinguish a subset of vocalizations.

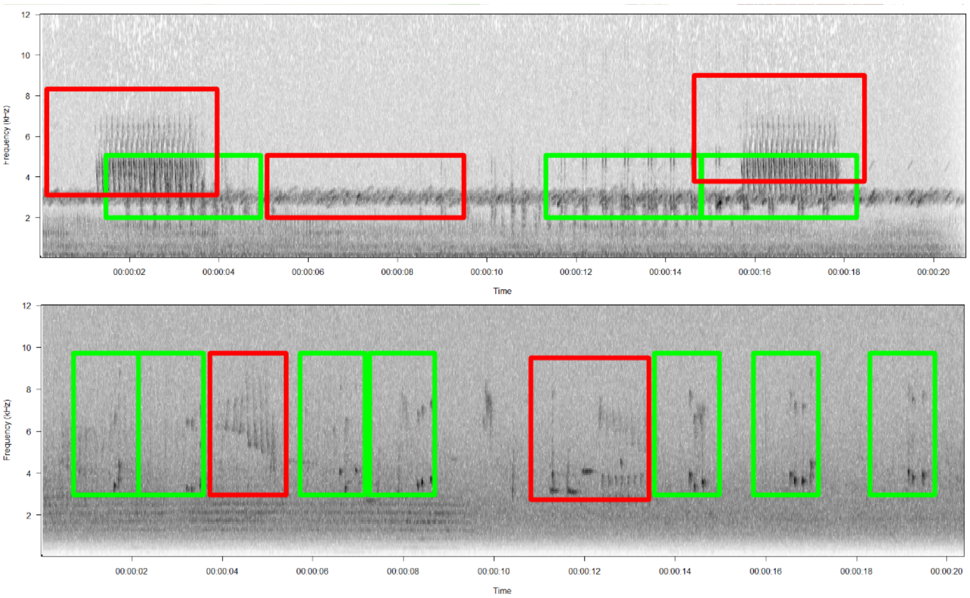


Figure 1: Two, 20-second spectrograms from weakly labeled Virginia rail recordings illustrating the challenge in automatically extracting desired signals for training data preparation. Energy threshold based techniques lead to the regular inclusion of undesired signals. Green boxes indicate rail vocalizations, red boxes indicate vocalizations of other species.

Even though we were not creating a classifier for multiple species, we wanted our negative class to accurately represent the sounds that the model might be exposed to in wetland ecosystems. So, we chose six common, highly vocal marsh birds that we believed would make up a significant portion of other signals in our target soundscapes. These choices were made both with domain knowledge and with exploratory analysis of the soundscape recordings we have available. We downloaded recordings of these species from Xeno-canto, filtering by latitude and longitude to a broad but targeted region to ensure relevance. Rather than manually extracting three second segments around true signals, we randomly sliced these recordings at three-second intervals, knowing that by doing this we were also sampling segments of relevant background noise and additional background species–both of which we believed would be valuable in defining a comprehensive negative class that would improve our model's precision.

3.3 Soundscape Test Set

Because performance on the validation set during training does not always indicate clearly how the model will generalize to soundscape data, we curated a soundscape dataset from the author's own monitoring projects. It is a small subset of a larger monitoring effort, made of roughly four hours of recordings with known rail vocalizations. The rail sounds within these recordings are rare and sporadic, with the exception of one recording which has over 300 vocalizations alone. In total, when the soundscape recordings are broken into three-second segments for inference, there are roughly 4200 audio samples, of which just under 400 contain Virginia rail vocalizations. Our hope was that by testing on this dataset we would get a clearer idea for real-world performance. We acknowledge, however, that with such a limited soundscape test set from a single monitoring program, further testing is needed to better estimate true generalizability.

4 Methodology

4.1 Preprocessing

Since the data samples were standardized prior to training, no resampling or resizing was necessary during data loading. We chose grayscale, mel spectrograms because of their widely reported success in other audio classification tasks. To compute them we chose a FFT window size of 1024, a hop length of 517, and 128 mel frequency bands. This resulted in input spectrograms that were 128x256 in size. As a final preprocessing step, we apply a transformation to the spectrogram which converts the amplitude values to a logarithmic scale, effectively compressing the dynamic range of the signal and emphasizing important components.

4.2 Data Augmentation

One of the major challenges in this domain has been generalizing from the focal recordings to the soundscape recordings generated by acoustic monitoring programs. This challenge is highlighted by [Kah+21], who note significant performance decreases when testing on their soundscape dataset. To address these challenges, we employed data augmentation techniques that applied realistic transformations to the focal training recordings. One transform included the addition of Gaussian noise in order to emulate ambient noise conditions and lower signal to noise ratios. Another technique involved the mixing of training samples with actual recordings that were full of other species and relevant environmental sounds. These augmentation techniques are illustrated in Figure 2. In addition to these, we also applied frequency masking, time masking, and horizontal shifts.

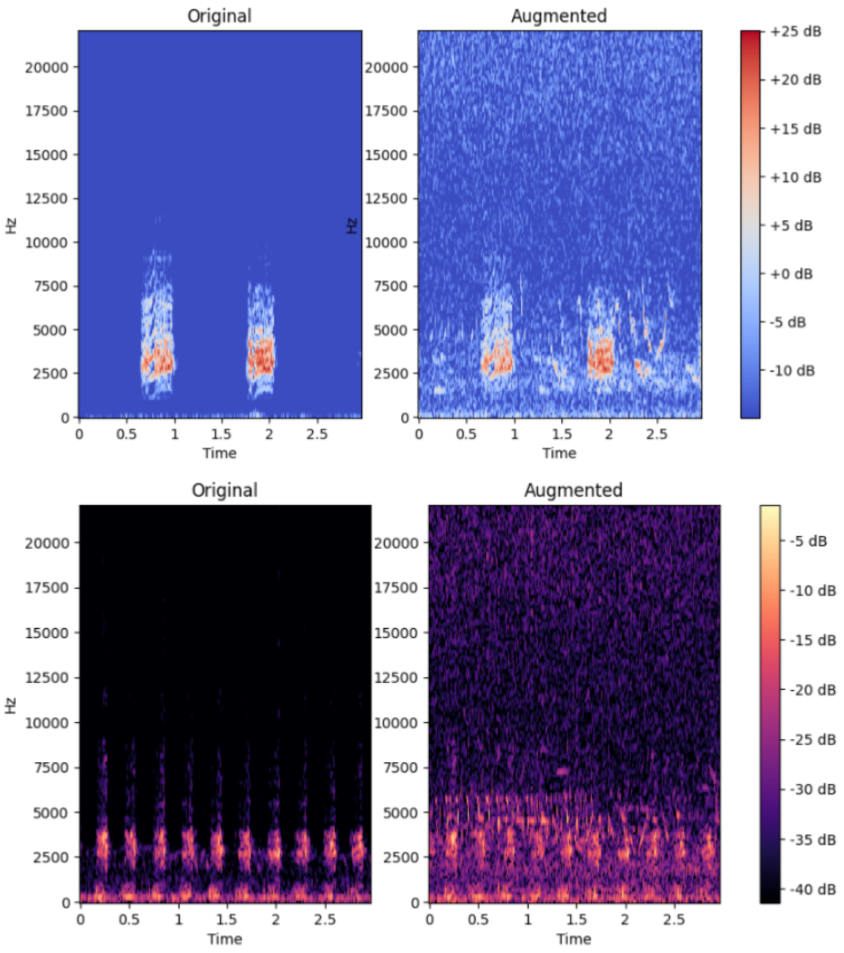


Figure 2: Illustration of data augmentation technique where clear, focal training samples (left) are mixed with genuine soundscape recordings to emulate target signals being obscured by other species and noisy ambient conditions

We defined two principal data augmentation regimes to test: a moderate regime and a heavy regime. The moderate regime mixed Gaussian noise and soundscape noise with a probability of being applied to each sample of 0.75. The strength of the original signal relative to the noise added was randomly selected for each sample from a predefined range of potential signal-to-noise ratios. For the moderate regime, we set this range at (-10, 15). We then applied a horizontal shift with a 0.5 probability (to mimic the signal occurring at a different time-frame in the window). We applied no time or frequency masking.

The heavy regime was simply an intensification of the moderate. The probability of applying each noise transformation (Gaussian and soundscape) was increased to 1.0, meaning that both transformations were applied to every sample. We restricted the signal-to-noise ratio range further to (-10, 10) to decrease the chances of clearer signals. We kept the horizontal shift transformation the same, but we added time and frequency masking at probabilities of 0.25 each, transforms which artificially obscure time or frequency bands. We tested each of our chosen model architectures with both data augmentation regimes.

4.3 Model Architectures

To establish baseline measures of performance, we tested a support vector machine as well as a simple CNN model with two convolutional layers, a max pooling layer, and three fully-connected layers.

For our primary experiments, we tested tested VGGish and ResNet architectures. We initialized them with pretrained weights and after exploratory analysis of results and convergence times decided that unfreezing all layers improved performance without incurring severe costs in convergence time. We prioritized ResNets, because of their success in similar classification tasks. Specifically, we tested a Wide ResNet, an 18-layer ResNet, a 34-layer ResNet, as well as a ResNet with a self-attention layer after the second residual block.

In addition to the above architectures, we tested the performance of a single ResNet with 18 layers and moderate augmentation that was trained with data that was automatically curated from the public database to serve as a baseline against which we could compare models trained on strongly labeled, hand-picked samples. For the automated curation process, we calculated the mean energy of the entire recording and automatically extracted three second segments that had higher energy than the overall mean.

4.4 Hyperparameters and Experimentation

Prior to running full experiments, we explored various optimizers and learning rate schedules, starting with combinations that we derived from recommendations in the literature by [Kah+21]. We found convergence to be quickest with a decaying learning rate and the ADAM optimizer, with a starting learning rate of 0.001, decaying by a factor of 0.5 if validation loss stalled for more than three epochs. We chose the cross-entropy loss function in the event that future modifications led us to include multiple classes for call type or for other species.

For each combination of architecture and data augmentation regime, we performed five-fold cross validation, training on each fold for 60 epochs. During training, we tracked loss, accuracy, recall, F1 score, and F-Beta score with a beta value of 2. Top performing model/augmentation combinations were then retrained on the full training dataset for final evaluation.

5 Results

5.1 Validation and Soundscape Set Performance

Performance on the validation set during five-fold cross validation was generally promising (Table 1), even with baseline models. The poorest performers were the SVM, the simple CNN, and the model trained with automatically curated data. While it might seem self-evident that a more carefully curated dataset with strong labels would lead to better performance, what was less obvious was the more significant ability for it to generalize to the soundscape test set, a comparison which can be seen in Tables 1 and 2. Since generalization is a common issue in bird sound classification problems, and since many models are trained on data that is somehow automatically extracted from weakly labeled recordings, this result might point towards data integrity as a primary culprit.

As expected, the high performance on the validation set was not reflected in the soundscape tests, where all models suffered. In general, though, the models which performed best on the validation set also performed best on the soundscape set. Tables 1 and 2 show performance comparisons between the validation set and the soundscape set after 60 epochs of training. VGGish models, the basic CNN, and the model trained on automatically curated data showed the most severe drops in recall and F1 score. The ResNet with an attention layer and the wide ResNet also showed more unexpected drops in performance that we believe might be due to inadequate hyperparameter tuning since our exploratory analysis of hyperparameters was focused on regular ResNet architectures and might not work as well on these modified versions. Validation set performance metrics for the top performing models relative to baseline implementations are visualized in Figure 3.

Model	Test Loss	Test Accuracy	Test F1 Score	Test Recall
VGGish (M)	0.138	0.96	0.96	0.96
VGGish (H)	0.224	0.93	0.93	0.93
ResNet18 (M)	0.073	0.97	0.97	0.98
ResNet18 (H)	0.115	0.96	0.96	0.96
ResNet34 (M)	0.072	0.98	0.98	0.97
ResNet34 (H)	0.124	0.96	0.96	0.96
ResNet18 (M + attn)	0.039	0.99	0.99	0.98
ResNet (wide)	0.084	0.97	0.97	0.97
ResNet18 (M + auto)	0.210	0.93	0.93	0.92
BaseCNN (M)	0.236	0.92	0.92	0.92
SVM	—	0.76	0.80	0.78

Table 1: Average performance metrics at 60 epochs across all folds after five-fold cross validation. M = moderate augmentation, H = heavy augmentation.

Model	Test Loss	Test Accuracy	Test F1 Score	Test Recall
VGGish (M)	0.332	0.90	0.39	0.33
VGGish (H)	0.286	0.92	0.44	0.35
ResNet18 (M)	0.202	0.94	0.69	0.65
ResNet18 (H)	0.152	0.96	0.77	0.71
ResNet34 (M)	0.200	0.95	0.70	0.66
ResNet34 (H)	0.178	0.95	0.72	0.67
ResNet18 (M + attn)	0.208	0.93	0.60	0.51
ResNet (wide)	0.207	0.94	0.64	0.54
ResNet18 (M + auto)	0.519	0.78	0.43	0.85
BaseCNN (M)	0.388	0.85	0.20	0.19

Table 2: Average performance metrics at 60 epochs across all folds after five-fold cross validation on the soundscape test set. M = moderate augmentation, H = heavy augmentation.

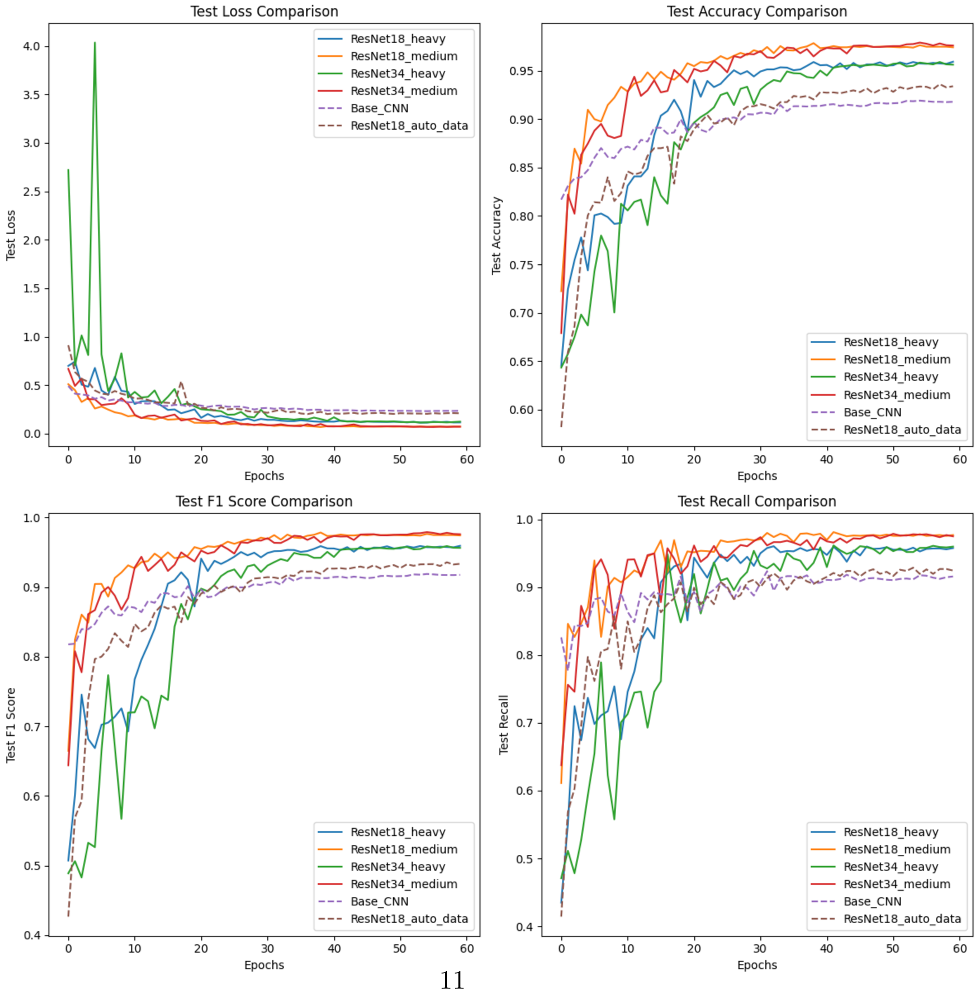


Figure 3: Validation set performance metrics for top four performing models compared to two baseline models–the simple CNN and the ResNet18 trained with training samples automatically extracted with an energy threshold selection technique.

Another interesting result was that while the models trained with moderate augmentation did better on the validation set than those with heavier augmentation, the models with heavier augmentation performed better on the soundscape set, indicating that the intensity of augmentation impacts generalization ability. It also demonstrates the important point that when training these models with focal recordings, optimizing to the validation set–which is also made of focal recordings–should not be the ultimate aim, but that efforts should rather focus on augmenting the training data in a way that realistically represents true use-case circumstances.

5.2 Final Training

In light of our results, we moved ahead to a final round of training using the full dataset with the ResNet-18 and -34 models, which we trained with both data augmentation regimes. With a training dataset of only around 1600 samples, the difference between training on four folds or the entire dataset was significant. During this training, we tracked performance against the soundscape set, which you can see in Figure 4 where we show loss on both datasets over the course of training. We set the scheduler to decay by a factor of 0.5 every 15 epochs, roughly mirroring the distribution of decay steps we saw when the scheduler depended on validation loss during experimentation. We also trained for 80 epochs rather than 60 to see if further training would improve performance, which it did not. After every epoch, we saved separate checkpoints for models that performed best in terms of recall and F1-score.

Figure 3 shows the relatively erratic performance in terms of test loss on the soundscape set during training, which was expected due to the soundscape set's limited size and severe class imbalance. Nonetheless, loss decreases over time seeming to hit a low for all four models at around the 40 epoch mark, a fact that was reflected in that the best performing models for both F1 score and recall were also saved from around this time.

The models trained on the full dataset showed significant improvements on the soundscape set. During the final tests on the soundscape set, we assessed the performance of each model individually, as well as every combination of ensembles where we took the average of the softmax scores as outputs. The ensemble predictions generally showed improved performance, and the best performer in terms of both F1 and F-Beta score was an ensemble of both ResNet-34 models as well as the ResNet-18 model with heavy augmentation applied during training–achieving a recall of 0.91 at a high precision of 0.87. We decided to call this final ensemble RailNET. Results for the four base models as well as two ensemble examples are shown in Table 3.

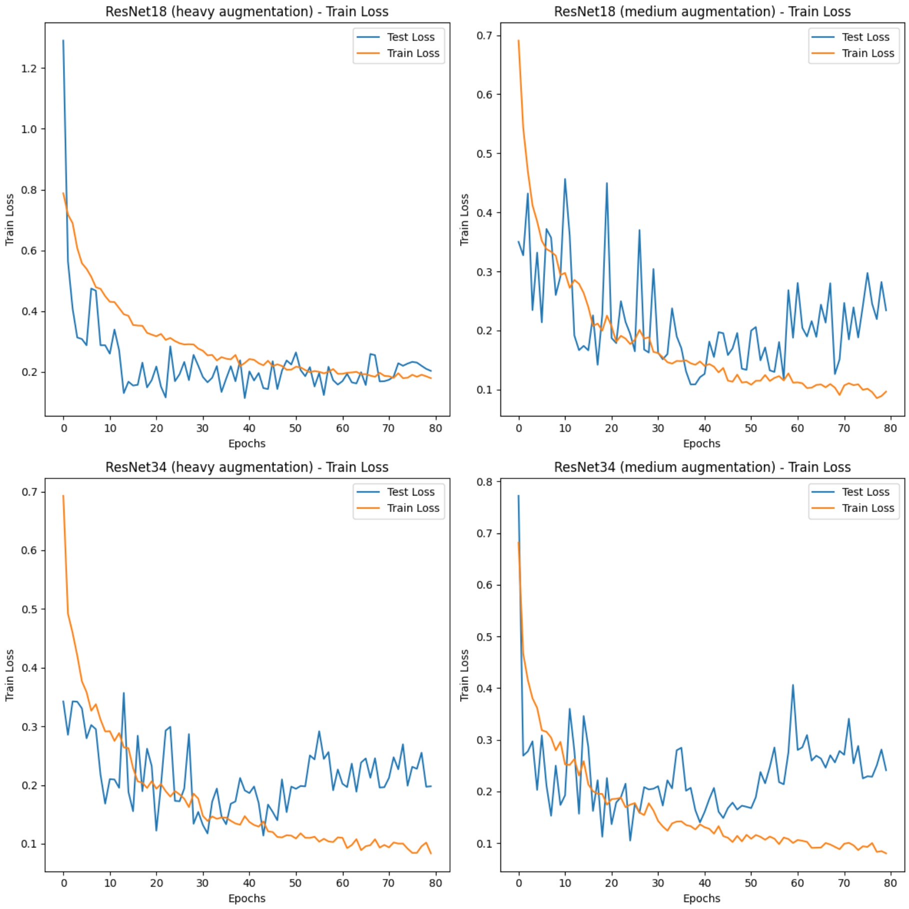


Figure 4: Training and test loss for the final four models, measured during training on the entire training dataset and testing on the soundscape set.

Model	Recall	Precision	F1 Score	F-Beta Score
ResNet18 (M)	0.90	0.61	0.72	0.82
ResNet18 (H)	0.89	0.78	0.83	0.87
ResNet34 (M)	0.93	0.66	0.77	0.86
ResNet34 (H)	0.89	0.85	0.87	0.88
Ensemble (all)	0.91	0.82	0.86	0.89
ResNet18 (H) + ResNet34 (H + M)	0.91	0.87	0.89	0.90

Table 3: Performance metrics for different models and model ensembles measured on the soundscape test set

After analyzing ROC curves for the final models–and following the recommendations of [Kni+17]to incorporate an assessment of confidence thresholds in analyses–we decided to assess performance at different softmax score cutoffs. Tables 4 and 5 show a performance comparison between BirdNET and RailNET at increasing cutoff levels. Overall, RailNET achieves significantly higher recall at every confidence cutoff while maintaining high levels of precision even at cutoffs resulting in 96% recall. For comparison, for BirdNET to detect 96% of the available vocalizations, precision must be reduced to 35%, a score that would result in a significant manual validation effort for researchers to sort through false positives. Furthermore, at even a strict confidence cutoff of 0.9, RailNET achieves a recall of 77% with 99% precision. Figures 4 and 5 show the distribution of each model's confidence scores on positive and negative samples, clearly showing the higher degree of certainty with which RailNET makes its predictions on this dataset. Adding to these promising results was the fact that during a manual validation of some of RailNETs higher confidence, false positive predictions, we actually discovered three additional Virginia rail vocalizations that had fallen through the cracks during the annotation of ground truth labels on the soundscape test set. While three signals might seem insignificant, these were extremely faint signals of a vocalization type that rails usually often only issue once or twice in a sitting as opposed to the "chatter" vocalizations that made up a majority of the soundscape set's signals. This means that these signals were relatively high value, since detecting or missing them could mean the difference between labeling an entire recording site occupied or unoccupied.

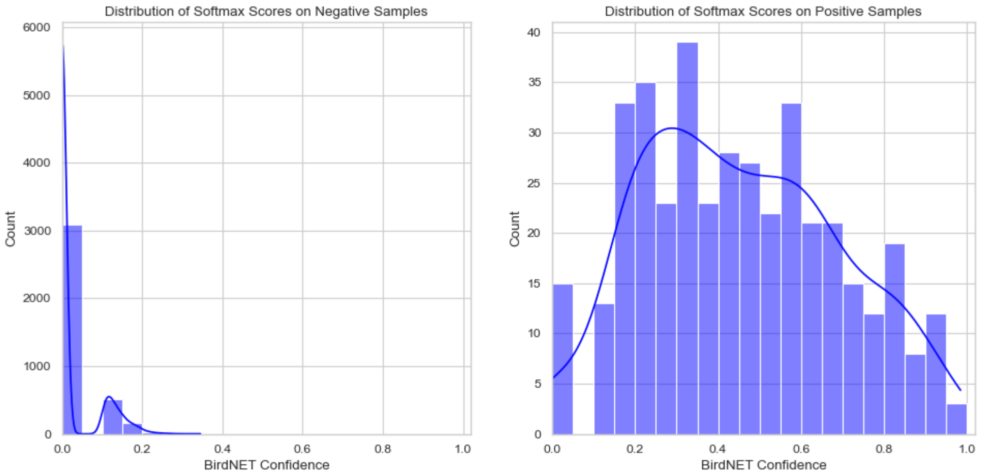
If this performance holds when applied to other soundscape recordings, it points towards a promising workflow to train and develop species-specific models for other rare, elusive targets, with plenty of room for modification and experimentation. We imagine, though, that genuine generalization will still be an issue with regard to the negative class. We built our negative class from samples of bird species specific to our study area, a fact which could account for our high precision. To work as well in other programs, it is likely that the model as it exists now would need to be retrained with a negative class that better reflects the signals the model is likely to encounter.

Threshold	Precision	Recall	F1 Score	F-Beta Score
0.1	0.35	0.96	0.72	0.72
0.2	0.91	0.85	0.86	0.86
0.3	0.99	0.70	0.75	0.75
0.4	1.00	0.55	0.60	0.60
0.5	1.00	0.41	0.47	0.47
0.6	1.00	0.28	0.33	0.33
0.7	1.00	0.17	0.21	0.21
0.8	1.00	0.10	0.13	0.13
0.9	1.00	0.04	0.05	0.05

Table 4: Performance metrics for BirdNET on the soundscape test set at various confidence score cutoffs

Threshold	Precision	Recall	F1 Score	F-Beta Score
0.1	0.32	0.98	0.48	0.69
0.2	0.56	0.96	0.71	0.84
0.3	0.69	0.96	0.80	0.89
0.4	0.81	0.94	0.87	0.91
0.5	0.88	0.91	0.90	0.91
0.6	0.94	0.89	0.91	0.90
0.7	0.97	0.86	0.91	0.88
0.8	0.98	0.84	0.90	0.86
0.9	0.99	0.77	0.88	0.81

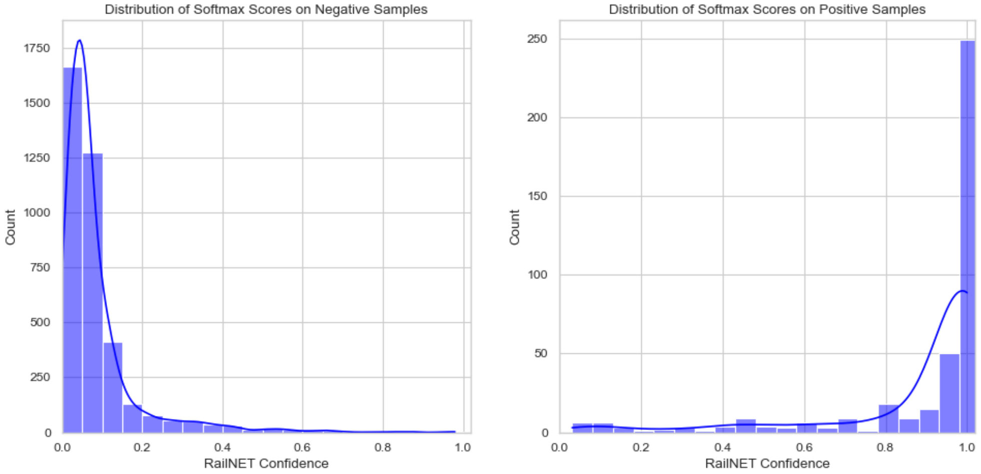
Table 5: Performance metrics for RailNET on the soundscape test set at various confidence score cutoffs



(a) Negative sample confidence

(b) Positive sample confidence

Figure 5: Distribution of BirdNET confidence scores of Virginia rail detections on the soundscape test set. (a) Confidence scores on negative samples, and (b) Confidence scores on positive samples.



(a) Negative sample confidence

(b) Positive sample confidence

Figure 6: Distribution of RailNET confidence scores of Virginia rail detections on the soundscape test set. (a) Confidence scores on negative samples, and (b) Confidence scores on positive samples.

6 Discussion

While somewhat unsurprising that a model trained for a specific species like RailNET outperforms a state-of-the-art model for whom that species is just one of nearly 1000 classes, these results nonetheless point toward an important shortcoming in the ability for large multi-class models to deliver satisfactory performance on particular species. BirdNET is unrivalled in its ability to help ecologists monitor the overall avian community to estimate variables like species richness and diversity. For projects aimed at community level studies, this is the obvious choice. For studies seeking to monitor rare, elusive targets, our results indicate that it might be worthwhile to train a more focused model using a similar workflow to the one defined above. Further, some of the lessons learned here might also be of value to the multi-class approach, such as the importance of focusing on data-integrity and realistic data augmentation.

6.1 Findings

By comparing our models to a baseline ResNet that was trained with automatically curated data, we illustrate the important fact that data quality matters, especially when there is a known challenge with generalization. When training large multi-class models, it might not matter enough to go through the laborious process of applying strong labels, but for a use-case such as this one–and if high recall and precision are necessary–the effort might be worth it. With scripted processes to locate and apply the labels, it took one person a few hours to search through 350 recordings and create a dataset of 1600 samples.

Our negative class being composed of audio relevant to our study area also probably plays a large part in RailNET's high precision, which may mean that to use RailNET in other studies, retraining on an a recomposed negative class would be necessary.

The scores on the validation and soundscape sets indicate that heavier levels of data augmentation might lead to improved generalization. Because of this, we believe that data augmentation choices should not be superfluous, but rather should reflect a genuine understanding of the domain shift from training to application, leading to realistic transformations on focal recordings.

Our assessment of performances at different score cutoffs also points towards the importance of keeping confidence thresholds in mind when looking at training metrics, rather than just taking the performance scores generated by a default 0.5 cutoff to be wholly indicative of the models ability to perform well.

6.2 Limitations

There is still a lingering question with regard to true generalizability of our model. While it is promising that it performs well on the unseen data that was collected from a genuine monitoring effort which made up our soundscape test set, this set was limited to a few hours of audio, did not include the full range of vocalization types of the Virginia rail, and was made up of recordings that all came from the same region at roughly the same time. Being from a single monitoring program means the background sounds and other species present were fairly constant throughout the whole soundscape set. Monitoring programs in other regions are likely to have other sources of anthropogenic and animal sound present which could impede performance. Further, we made this model regionally relevant by creating a negative class from recordings of species likely to co-occur with the Virginia rail in this specific context, meaning that it may need to be retrained with regionally relevant data if it were to maintain performance in other geographic regions. Whether or not RailNET can serve as an all-purpose Virginia rail detector in a wide range of contexts, we hoped to create an adaptable workflow for the creation a species-specific classifier.

6.3 Future Research Directions

RailNET will be implemented in conjunction with BirdNET during an acoustic monitoring program this spring, where upwards of 2,000 hours of audio will be collected from a variety of sites in a federal wildlife refuge. This will hopefully help to increase detection rates of the Virginia rail, but will also offer the opportunity to test the model's ability to truly generalize to a new set of real-world data, allowing us to make modifications if necessary. We imagine that we will have to recompose a relevant negative class and retrain if we want to keep precision as high as it was during these tests. We also plan to eventually increase the number of classes to encompass a larger group of cryptic marsh birds and to test these methods on other elusive avian targets such as owls.

In terms of model experimentation, since it seemed to be the main factor in promoting generalizability, focusing on more sophisticated data augmentation techniques and testing increasing levels of intensity might prove fruitful. Furthermore, we were somewhat surprised by the relatively poor performance of the Wide ResNet and the ResNet with an attention layer, especially since the latter showed such strong performance during certain folds of cross-validation. We believe that it is possible that more time spent tuning hyperparameters for these modified architectures may lead to more promising results.

7 Conclusions

We hope this work can serve as a starting base for further experimentation in rare-species acoustic monitoring. Initial assessments show that it successfully addressed a widespread challenge in bioacoustics: the shift from focal recordings during training to soundscape recordings during application. We believe that curating a clean, high-quality dataset, applying rigorous, domain-relevant data augmentation, and composing an intentional negative class is responsible for this. We are encouraged by RailNET's ability not only to to capture the rare signals it was trained to find, but to do so with very high precision. If its ability to generalize to new data holds, this will result in a marked decrease in the amount of manual validation required by wildlife professionals. We hope that these experiments offer other acoustic monitoring programs guidance for training deep learning models to detect vocalizations of elusive targets–ultimately leading to better-informed conservation action for species in need.

8 References

References

[Bro+17] Ella Browning et al. "Passive acoustic monitoring in ecology and conservation". en. In: (2017), p. 75.

[CA16] Marconi Campos-Cerqueira and T. Mitchell Aide. "Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling". en. In: *Methods in Ecology and Evolution* 7.11 (Nov. 2016). Ed. by Kate Jones, pp. 1340–1348. ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.12599. URL: https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12599 (visited on 11/08/2021).

[Col+22] Jerry S Cole et al. "Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data". en. In: *Ornithological Applications* (Apr. 2022), duac003. ISSN: 0010-5422, 2732-4621. DOI: 10.1093/ornithapp/duac003. URL: https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duac003/6572065 (visited on 04/28/2022).

[Gib+19] Rory Gibb et al. "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring". en. In: *Methods in Ecology and Evolution* 10.2 (Feb. 2019). Ed. by Luca Börger, pp. 169–185. ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.13101. URL: https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13101 (visited on 11/09/2021).

[Kah+21] Stefan Kahl et al. "BirdNET: A deep learning solution for avian diversity monitoring". en. In: *Ecological Informatics* 61 (Mar. 2021), p. 101236. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2021.101236. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574954121000273 (visited on 11/01/2021).

[Kni+17] Elly C. Knight et al. "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs". en. In: *Avian Conservation and Ecology* 12.2 (2017), art14. ISSN: 1712-6568. DOI: 10.5751/ACE-01114-120214. URL: http://www.ace-eco.org/vol12/iss2/art14/ (visited on 02/14/2022).

[PT21] Cristian Pérez-Granados and Juan Traba. "Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research". en. In: *Ibis* 163.3 (July 2021), pp. 765–783. ISSN: 0019-1019, 1474-919X. DOI: 10.1111/ibi.12944. URL: https://onlinelibrary.wiley.com/doi/10.1111/ibi.12944 (visited on 11/01/2021).

[TPS21] Eleni Tsalera, Andreas Papadakis, and Maria Samarakou. "Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning". en. In: *Journal of Sensor and Actuator Networks* 10.4 (Dec. 2021), p. 72. ISSN: 2224-2708. DOI: 10.3390/jsan10040072. URL: https://www.mdpi.com/2224-2708/10/4/72 (visited on 03/01/2023).

[WGP19] Connor M. Wood, Ralph J. Gutiérrez, and M. Zachariah Peery. "Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology". en. In: *Ecology* 100.9 (Sept. 2019). ISSN: 0012-9658, 1939-9170. DOI: 10.1002/ecy.2764. URL: https://onlinelibrary.wiley.com/doi/10.1002/ecy.2764 (visited on 11/08/2021).

[Woo+19] Connor M. Wood et al. “Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework”. en. In: *Ecological Indicators* 98 (Mar. 2019), pp. 492–507. ISSN: 1470160X. DOI: `10.1016/j.ecolind.2018.11.018`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S1470160X18308793` (visited on 11/10/2021).