CSE 493S: BirdCLEF+ 2025

Andy Legrand, Ryan Bai, Jacob Chen, Jay Bhateja
Affiliation
{andyleg, ryanbai, jacobc35, jbhateja}@uw.edu

June 7, 2025

Machine Learning Project Summary

Project Scope

The main goal of our project is to create a successful, highly scoring submission for BirdCLEF+ 2025 by analyzing the methods used by previous winning submissions and experimenting with them to produce our own submission.

Methodology

Our methodology revolved around replicating the 2024 solution with our own adjustments. We used pretrained EfficientNet and RegNet backbones and finetuned them solely on the provided BirdCLEF+ data. We then used a grid search to find the best hyperparameters. Audio files were then converted into mel spectrograms and stored on disk to accelerate training. Models were trained with cross-entropy loss with basic data augmentations. Everything was run locally with an RTX 4070, with each model taking around an hour to train.

Results

We had a solid baseline performance with individual EfficientNet and RegNet models, but we achieved notable improvements with ensembling. A 6 model ensemble, with 3x EfficientNet and 3x RegNet and mean pooling yielded the best results, with a test ROC-AUC of 0.769. Some other approaches we tried, such as excluding the loudest 20 percent of data, replacing the softmax layer with a sigmoid layer, and relabeling the dataset yielded minimal improvements. Our results suggest that ensembling and model architecture had the biggest impact, while preprocessing adjustments offered only marginal benefits.

What was Easy

Finetuning models was straightforward thanks to the Hugging Face library. Access to a 4070 also made training runs efficient, and with precomputed mel spectrograms, training loops were even faster. Once our dataset was ready, the process was pretty plug-and-play. The 2024 winning solution was also well-documented and provided clear guidance for getting started.

What was Difficult

Setting up the mel spectrogram and dataset pipeline was more difficult than expected. Since the competition was different this year, some concepts from previous years didn't translate well and factors such as added amphibian and insect noise made data processing and relabeling approaches more difficult. More experimentation was needed than expected to understand what would work well.

1 Introduction

BirdCLEF is an annual recurring competition in which the objective is to develop models for detecting and identifying bird vocalizations in autonomously recorded soundscapes [1]. The competition can be approached as an audio event classification problem, where audio recordings should be classified based on the species heard in the recording. Note that the problem is multi-class and multi-label in nature: there are multiple target species, and since more than one species can be present in a recording, a label of present or not needs to be predicted separately for each species. The most common approach to this audio event classification problem in past iterations of BirdCLEF is to use pretrained image classification models, typically CNNs such as EfficientNet [3] and RegNet [2],to perform classification on the spectrogram images of audio recordings.

The most recent iteration of the competition, BirdCLEF 2024, used soundscapes recorded in the Western Ghats region of India for its train and test dataset, and the classes for prediction includes 108 different bird species. Successful submissions to BirdCLEF 2024 tended to make use of techniques such as data augmentations, model ensembling, and pseudo-labeling of unlabeled data for training [1]. BirdCLEF 2025 features a different dataset recorded in the Magdalena Valley of Columbia, and the classes for prediction includes 206 species, which has been expanded to include amphibians, reptiles, and insects in addition to birds. Since the distribution of data is quite different from BirdCLEF 2024, we would like to investigate whether successful methods from BirdCLEF 2024 generalizes to the BirdCLEF+ 2025 dataset. This would provide an understanding of whether methods from previous submissions are generally applicable to birdcall classification tasks, or if they are specifically fitted for the BirdCLEF 2024 dataset.

1.1 BirdCLEF 2025

The BirdCLEF 2025 dataset, like past datasets, features a large number of soundtracks of varying lengths, and recorded in different locations in the general area. Training data feature a primary label (the most dominant species vocalizing in the recording) and secondary labels (one or more species that may also be present in the recording). The test dataset consists of 1-minute long soundscapes. Submissions should include predictions for each 5-second segment of a soundscape, which should include a probability for each of the 206 target species being present in the 5-second segment. The submissions are evluated by the ROC-AUC metric, which measures the probability that the classifier ranks a positive sample is higher than a negative sample.

2 Scope of the Project

There were several methods which were widely used in successful submissions to BirdCLEF 2024. In particular, many solutions utilized ensembles of diverse

models and using data augmentation such as CutMix and masking [1]. The top ranking solution of BirdCLEF 2024 [1] simplified classification from a multi-class, multi-label problem to simply a multi-class classification problem by only using the primary label of training data. The rationale is that most training soundscapes only include a single species, and simplifying into single-class classification allows models to be trained with Cross-Entropy loss. At inference time, to account for the possibility of multiple species being present in a recording, logits outputted by the model are converted into probabilities using Sigmoid to produce a separate probability for each class. The top solution of 2024 also performed pre-processing by relabeling the training dataset using a classifier and dropping data that the classifier could not accurately predict, to filter out recordings with low quality or where the vocalization is difficult to detect.

In our project, we aim to apply these methods from 2024 solutions to the BirdCLEF 2025 dataset, and investigate which of these methods leads to an increase in performance. This would allow us to understand where the methods proposed by 2024 solution generalizes to a different distribution of data.

2.1 Addressed hypotheses

We will apply the methods we identified from 2024 solutions to see if they lead to improved performance when training models on the BirdCLEF 2025 dataset. We will be testing the following hypotheses:

- Using an ensemble of multiple models leads to improved performance.
- A model trained on the entire dataset will perform worse than a model trained on the quietest 80 percent of the dataset.
- Using sigmoid during inference instead of softmax yields a higher score.
- Relabeling the dataset with a previously trained model ensemble and training a new ensemble on this will increase our score.

3 Methodology

To evaluate our hypotheses, we implemented a general pipeline for training models on audio classification with the BirdCLEF 2025 dataset. We would then experiment with adding various data augmentations and pre-processing steps to evaluate their effect on performance. The predominant approach among previous submissions is to convert soundscapes into mel spectrograms, a visualization of audio signals that is more closely aligned to human perception of sound. The spectrograms are then classified using visual models, typically CNNs. Popular model choices are EfficientNet and RegNet pretrained on ImageNet classification. We used the model checkpoints EfficientNet_b0 and RegNetY_008, pretrained on ImageNet, and finetuned using our training pipeline.

[1] https://www.kaggle.com/competitions/birdclef-2024/discussion/512197

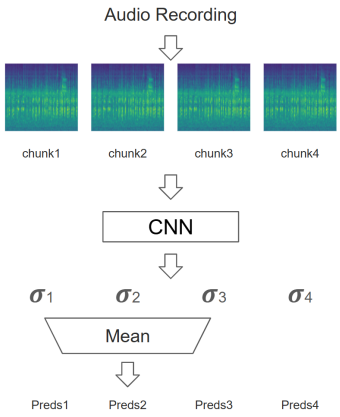

Figure 1: The inference strategy we used to create predictions with fine-tuned models. Input recordings are converted into spectrograms and split into chunks. Probabilities outputted by the CNN are averaged with those of adjacent chunks to create final predictions. Windows of 5 predictions used during averaging in practice, unlike 3 as shown in figure

3.1 Training and inference pipeline

We trained the models using 10-second soundscapes from the training dataset. Similar to the 2024 winning solution, we used the primary label of each recording as labels for training, which simplifies the problem into multi-class classification. Models are trained using Cross-Entropy loss.

The conversion of training soundscapes into mel spectrograms was relatively compute-expensive. To reduce training time, we opted to create mel spectrograms and load them from the disk ahead of model training.

We finetuned our models on the 2025 BirdCLEF dataset. We did a grid search to find the best hyperparameters to use for all of our tests.

Hyperparameter	**Value**
Learning Rate (lr)	5×10^{-4}
Dropout	0.2
Batch Size	32
Weight Decay	5×10^{-5}

Finetuning the CNN based models was fairly efficient in terms of compute, as we had predicted. We were able to run everything locally on an NVIDIA RTX 4070 12GB GPU, with each model taking ∼1hr to train.

To produce predictions on test dataset and submit them for evaluation, we adopted the inference pipeline shown in figure 1.

Test soundscapes are 1 minute in length, and predictions need to be made for each 5-second segment of the soundscape. For each 5-second segment, we

would include 2.5 seconds before and after the segment to obtain a 10-second segment of the soundscape, and input into the fine-tuned model to produce logits. Using 10-second inputs instead of 5-second inputs is inspired by the 2024 winning solution, and provides the model with more context and captures more information regarding the period of repeating bird vocalizations. Raw logits for each 10-second chunk are converted into probabilities (using either softmax or sigmoid, depending on the experiment setup), and averaged with probabilities of the 2 preceding and subsequent chunks to produce final predictions.

3.2 Experiment setup

We first trained EfficientNet and RegNet models without ensembling or data augmentations, and recorded the validation accuracy and f1-score achieved during training. We submitted predictions with these models to the competition and recorded the ROC-AUC score on the test set.

To test the hypothesis that ensembling improves performance, we experimented with different methods of combining models into ensmebles and recorded the ROC-AUC score that ensembles achieved on the test set. To evaluate the hypothesis that converting logits into probabilities with sigmoid is more effective than softmax at inference time, we produced predictions using both methods and compared the ROC-AUC score that they achieved on the test set.

We then tested cleaning the dataset by removing duplicate files and using only the 80% of files with lowest power, as suggested by the 2024 winning solution. Finally, we used our most successful ensemble to re-classify the training dataset, and dropped files where the classification does not match the label, then re-trained the models and evaluated performance.

We have made our codebase for training the models [2] and submitting inferences [3] publicly available.

4 Results/Summary

The results of our experiments are presented in table 1.

4.1 Ensembles

Our aim for this hypothesis was to test if using an ensemble of models (3x RegNet, 3x EfficientNet) would perform better than a single RegNet or EfficientNet. We experimented with 3 methods of ensembling: taking the mean of probabilities outputted by different models, taking the minimum, and taking the maximum. Intuitively, taking the minimum would predict that a species is not present in the recording if any model predicts that it is not present. Taking the maximum would predict that a species is present in the recording if any model believes that it is present.

[2] https://github.com/andyllegrand/birdclef25

[3] https://www.kaggle.com/code/hanchenbai/birdclef2025-inference-notebook

Test	Val Accuracy	Val F1	ROC-AUC (test set)
EfficientNet_b0	.661	.665	.732
RegNet	.669	.662	.726
Ensemble (mean)	.732 (avg)	.721 (avg)	.769
Ensemble (mean, sigmoid)	.732	.721	.670
Ensemble (min)	.732	.721	.744
Ensemble (max)	.732	.721	.752
Ensemble (cleaned data)	.718	.712	.760
Ensemble (relabeled data)	.761	.752	.753

Table 1: Model performance comparisons on validation and test sets.

Our data supports the effective of ensembling, as using an ensemble brought our test score from .732 up to .769. Interestingly, we found that mean ensembling was the most effective method, whereas the 2024 winning solution suggested that taking the minimum was most effective in their case.

4.2 Dataset cleaning

To test this hypothesis we used an identical ensemble training pipeline to test 1, but excluded the loudest 20 percent of the data (calculated by taking the average power over the entire sound file). We saw similar results with this method to using the entire dataset, however the test score dropped by 1 percent.

4.3 Softmax and Sigmoid

To test this we simply replaced the softmax layer in our inference notebook with a sigmoid layer. This dropped our score considerably. This seems reasonable, since the models were trained on a cross-entropy loss objective. However, the 2024 winning solution suggested that training on cross-entropy loss and using sigmoid worked for them. It would be interesting to inverstigate the reason that their method does not generalize to new datasets.

4.4 Dataset Relabeling

For this test we used our highest scoring model ensemble to relabel the dataset. To do this, we ran the entire dataset through our highest scoring ensemble, then relabeled individual mel spectrograms the model could not accurately predict to a 207th no call class. This approach gave us slightly worse results than using the base ensemble.

5 Discussion

5.1 Larger implications

Our experiment suggests that many of the methods which were successful in BirdCLEF 2024 solutions were not replicable using the BirdCLEF 2025 dataset. This shows that preprocessing and inference techniques developed by successful BirdCLEF submissions may be specific to the dataset of the competition, and not easily applicable to birdcall classification or audio event classification in general.

5.2 What went well

The general model architecture of converting to mel spectrograms then processing with RegNet and EfficientNet seemed to work quite well. We expected this to be the case as most solutions used this method instead of processing raw audio. Ensembling also worked quite well, which we expected, however we found that using mean pooling instead of min pooling or max pooling was most effective. This could come down to training instability, and one poor model skewing the results of min and max pooling more than mean pooling.

Sigmoid activation worked quite poorly, which we did not expect. Sigmoid should work better for if the clips in the test set contain multiple calls at a time, while softmax should perform better if there is only one call present. Perhaps this suggests that the test data this year is more suitable for a single class problem.

Additionally, dropping the loudest 20 percent of the data did not yield the increase in score that we hoped it would. This again could come down to differences in the 2024 and 2025 datasets. Perhaps the 2024 dataset contained lots of noise which the 2025 dataset did not.

Lastly, relabeling the dataset was unsuccessful for us as well. Due to this years dataset containing insects and amphibians in addition to birds using googles bird classifier would not have been effective, so instead we used a previously trained model ensemble. This could just come down to our approach of using a previously trained ensemble instead of the google model being ineffective.

5.3 What we would do differently

If given more time, we would have liked to compare the success of different models, especially comparing transformer architectures to CNNs. While we originally intended this to be a focus of our project, we were unable to set up transformer architectures for experiments in time.

References

[1] Stefan Kahl, Tom Denton, Holger Klinck, Vijay Ramesh, Viral Joshi, Meghana Srivathsa, Akshay Anand, Chiti Arvind, Harikrishnan Cp, Suyash Sawant, Robin Vv, Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Robert

Planqué, and Alexis Joly. Overview of birdclef 2024: Acoustic identification of under-studied bird species in the western ghats. In *Conference and Labs of the Evaluation Forum*, 2024.

[2] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces, 2020.

[3] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.