***Mamethierno Gadiaga***
***Richard Cenedella***
***University Of Rhode Island***
***CSC_561: Neural Networks & Deep Learning***

**Kaggle Competition:** https://www.kaggle.com/competitions/birdclef-2025

**Github:** https://github.com/thiernohgradiagram/fp-561

**Notebooks:**

Restnet 18 CNN: https://colab.research.google.com/drive/1EEwftRHjQhERSNL-Aza7I1FmqjhZPa98

Efficient Net B0: https://colab.research.google.com/drive/17Fw8_Tqxu27tk1sP1ndx0kzjT7CJJUcD?usp=sharing

CNN from Scratch: https://colab.research.google.com/drive/1qtlRctbUl11-pb_djkrg-loRX8VJr--9?usp=sharing

**Title: Comprehensive Strategy for a High-Performing `BirdCLEF+ 2025 Ensemble**

**Introduction:**

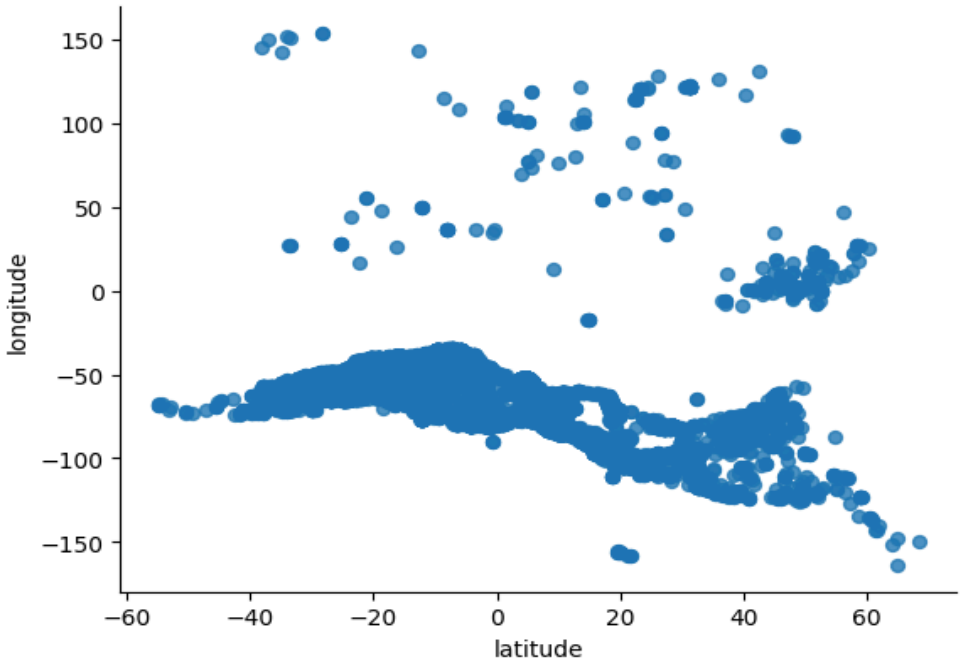
For the final project in this class (CSC_561 Neural Networks & Deep Learning), we decided to join this kaggle competition: https://www.kaggle.com/competitions/birdclef-2025. The main objective of this project is to be able to analyze audio recordings of nature to identify the many different species of animals found in the El Silencio Natural reserve in Colombia. The identification will be based on the unique vocalizations that each animal makes. The types of animals that are represented in this kaggle competition's dataset include various species of birds, amphibians, mammals and insects. This data was captured and collected in a couple different ways. The first way is through user uploaded submissions to animal identification websites. Another way it was collected was by passive acoustic monitoring devices that were placed in the Magdalena valley of Colombia. The end goal of this project will be to enhance the biodiversity monitoring efforts for the researchers in the nature reserve. They want to be able to automate the detection and classification of different taxonomic groups found in the area. This research stems from their ecological restoration efforts in the region. Historically this area was a humid tropical rainforest and it was a significant biodiversity hotspot. While it is still a biodiversity hotspot, the rainforest has been ravaged by illegal logging operations, farming and cattle ranching over time. Today researchers aim to restore the forest and wetlands through community outreach and other biodiversity monitoring programs. Their goals are to increase endangered animal populations back to healthy population numbers. This requires some form of monitoring in order to be able to track population numbers in the wild. Currently, biodiversity surveys are done manually by people on the ground monitoring groups of animals. This is costly and labor intensive for the researchers. By employing machine learning techniques, they hope to be able to automate their analysis of passive acoustic monitoring data and employ greater temporal resolution than ever before. This will provide them with new insights into their restoration efforts and help them to improve the biodiversity in the area and return the forest back to its natural and sustainable state.

To achieve this goal, we propose to use an ensemble model approach to analyze the animal vocalization data. By using an ensemble method approach we can capture different aspects of the data to increase the accuracy of our final

score. All the data sources will go through five different neural networks. These networks will be a CNN on mel spectrogram, CRNN, an audio transformer, MFCC+MLP, and a raw waveform CNN. We will then average the results of all the neural networks together to get the final score. By using the ensemble model approach we hope to get better performance and accuracy then we would get from using just one neural network.

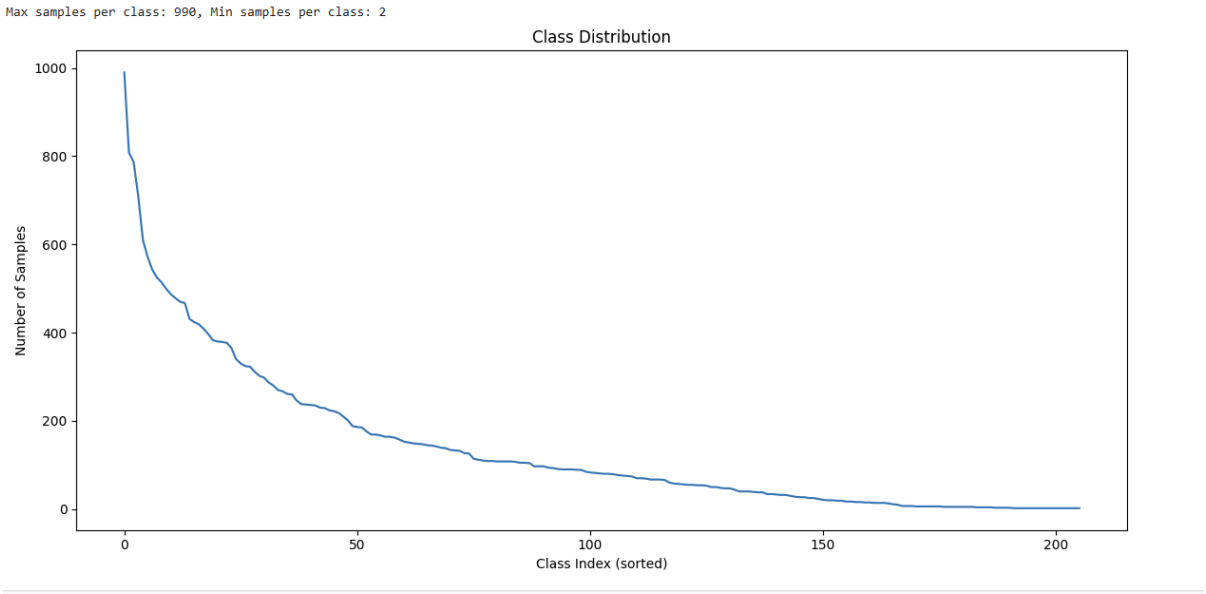
**The Data:** https://www.kaggle.com/competitions/birdclef-2025/data

For this project, we will be using the dataset that is provided to us by the Kaggle competition organizers. It comprises three different sources of animal vocalization data. The sources are xeno-canto.org, iNaturalist and the Colombian Sound Archive of the Humboldt Institute for Biological Resources Research in Columbia. Xeno-canto.org and iNaturalist are websites where end users can upload their own animal recordings. Their primary focus is on animal identification. Each of their datasets consists of user created recordings of animals that they find around them. The dataset for this project consists of a smaller subset of those datasets. This dataset specifically consists of recordings of various birds, amphibians, mammals and insects all found within the Middle Magdalena Valley of Colombia. The data is split up into four different sections. These sections are called train_audio, test_sondscapes, train_soundscapes and hidden test data. For the train_audio, the files consist of the following information. The first feature is the primary_label which specifies the code for the species. In this case it is eBird code for birds and iNaturalist taxon ID for everything else. In total there are 206 species that can be identified in the training data. There are also 28,564 labeled training examples available to us in this dataset. Then they have a feature called type and in some of the samples it explains what type of call an animal is making. This may describe how the sound is meant to scare off other animals or call fellow animals to a particular place. The next feature is the filename. This is the filename of the sound files. This links the sound files to the spreadsheet so that they have some background context to what is being heard in the audio. The next feature is called collection, and it indicates the source of the data and what the recording's id was in that collection. Some recordings also provide a quality rating if the data comes from the Xeno-canto dataset. If there are no ratings, then the number 0 will be used to indicate that no rating is available. The next feature is the url of where the sound file came from. After that the next feature is the geographical location of the recording. These are given as latitudinal and longitudinal coordinates. The chart below shows all the coordinates that were present in the csv for the data.




This shows where the animals are located in relation to each other based off of where the recordings were taken. As one can see, a lot of the animals are in very close proximity to each other with few outliers. Since they are so close, recordings may have other animals that can be identified in the recording but they are not the primary animal to be studied. In the dataset this is called a secondary_label. They may not all have secondary labels. The next feature is called scientific_name and this is just the scientific name of the animal that was heard in the sample. Likewise, the next feature is called common_name and this is the common name of the animal in the recording. Then the next feature is called author. If an author of the file can be specified it will be found in this column. If there is no author then the name “Unknown” will be used in place of the actual author's name. The last feature is just the license for the sound file. In total there are 13 features. For the test_soundscapes data, there are approximately 700 1 minute recordings to process. They are all resampled to 32 kHz. Not all species from the train data occur in this data. Along with test_soundscapes there is also train_soundscapes. This data is unlabeled, but it comes from the same location as test soundscapes. They do not overlap with the hidden test data. In this dataset there are 9,726 samples to analyze. Through the combination of datasets between the three different sources and the ensemble approach we will take to analyze the data, the model should be able to achieve a high accuracy when tested with the hidden test data.

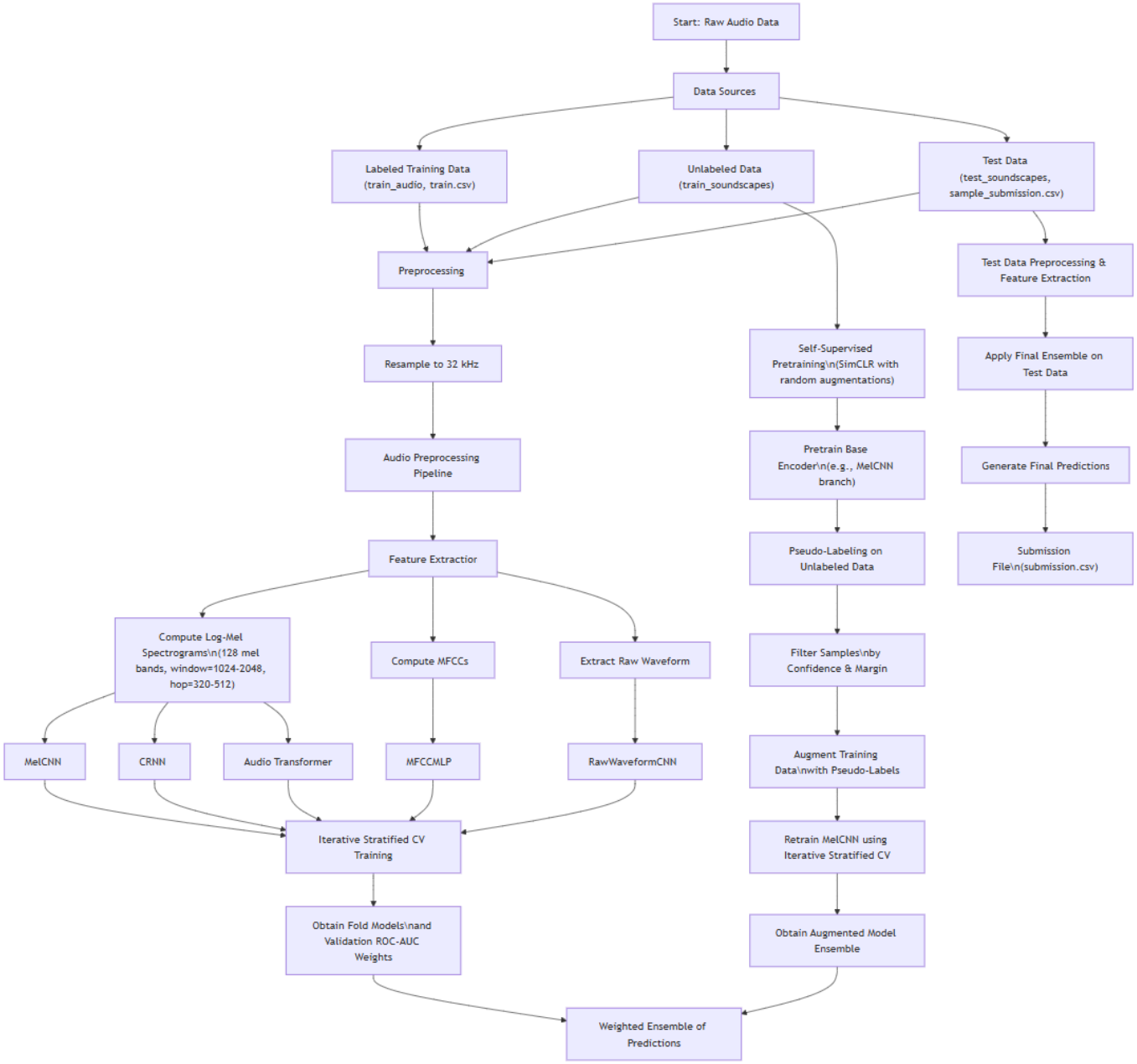
**Dataset Challenges:**




While there are a lot of audio samples to analyze in the dataset, one challenge we face is the issue of class balance. In the diagram above, this chart shows how many example audio files there are per species of animals. What this shows is that some animals have a lot of example audio files that the model can learn from while others have very few examples. This will make identifying rare animals more difficult for the model as there are just fewer examples of how they sound. One way we can counteract this is by introducing various data augmentation techniques. Some of these methods include time shifts, gaussian noise, pitch adjustments without changing tempo, and speed changes. For our project we decided that only up to 2 different augmentations may be applied randomly to the data to help balance the classes. Along with different augmentation techniques, we will employ ensemble learning techniques to further alleviate the issues with this unbalanced dataset.

**Methods: Ensemble Model Architectures for Species Classification:**

We propose an ensemble of **multiple neural network architectures**, each tailored to different feature types and aspects of the bioacoustic data. This diversity will help the model generalize from limited samples imageclef.org.




The ensemble will include:

- **CNN on Mel-Spectrograms:** A convolutional neural network that treats the mel-spectrogram as an image. For example, an **EfficientNet** or **ResNet** backbone (pretrained on ImageNet, to leverage transfer learning) can be adapted to spectrogram inputs by using 1-channel inputs. Such CNNs will learn frequency-temporal filters

(e.g. to detect bird chirp patterns or frog croaks) and are typically very strong classifiers for audio spectrograms. We will fine-tune these models on the BirdCLEF data; pretrained image models give a head-start on feature extraction even with few labels. In past bird audio competitions, ensembles of CNNs with different backbones have achieved top performance blog.csdn.net, so we will include multiple CNN variants (e.g. EfficientNet-B3, ResNet34, etc.) each trained on spectrograms with slightly different settings (different time crop lengths, frequency ranges, or augmentation settings) to provide complementary predictions.

- **CRNN (Convolutional Recurrent Network):** To capture temporal dynamics beyond the CNN's receptive field, we include a model combining CNN layers (for feature extraction from the spectrogram or MFCC sequence) with a **Recurrent layer (LSTM/GRU)** or **Temporal Attention** mechanism. The CNN learns local spectral features, and the RNN/attention summarizes how calls evolve over time, which is useful for species with distinctive rhythm or periodic calls. This CRNN architecture can operate on mel-spectrogram patches or MFCC sequences. It is optimized for sequences, helping to detect a species call that might be a few seconds long within a longer clip. With limited data, we will constrain model size (e.g. a small GRU) and use dropout to prevent overfitting.

- **Audio Transformer:** Leverage a transformer-based architecture (such as an **Audio Spectrogram Transformer** or a **Perceiver**) that can model long-range context in the audio. Transformers have shown success in audio classification by attending to relevant time-frequency regions. We could use a pretrained AudioSet model (e.g. AST or PANNs CNN14 transformer variant) if allowed, or train a transformer from scratch on our data augmented with unlabeled examples. This model will be adept at capturing complex patterns like overlapping calls from multiple species. It provides a different inductive bias compared to CNNs (global self-attention vs. local convolution), enriching the ensemble. Given the data limits, we might freeze parts of a pre-trained transformer and fine-tune just the later layers to reduce the risk of overfitting.

- **Raw Waveform CNN:** A 1D convolutional network operating on raw wave input (with possible initial SincNet or wavelet-based filters). This model will learn frequency-selective filters automatically. To make it feasible with limited labels, we will initialize it from a self-supervised pre-training on unlabeled audio (detailed in the next section). The architecture could resemble an encoder from a known model (e.g. a smaller WaveNet-like stack or the front-end of wav2vec). This provides the ensemble an *end-to-end* learned feature extractor that might capture subtle cues (like precise pitch or timbre cues) that fixed feature transforms might miss. We will ensure this model remains lightweight (fewer layers) so that it can train with the data available and run efficiently.

- **MFCC+MLP Classifier (Lightweight model):** In addition to deep models, we can include a simple **machine learning model** on MFCC features or other handcrafted features. For instance, a small **multi-layer perceptron** or even a **Gradient Boosted Trees (XGBoost)** model can take as input aggregated MFCC statistics or species-specific acoustic indices. This model is less powerful than the neural nets but can excel when data are very scarce, since MFCCs distill the audio and the model has fewer parameters. It acts as a

safety net in the ensemble – because it's less likely to overfit, it may perform more consistently on rare classes. Its predictions can slightly boost ensemble performance especially on species where deep nets struggle due to extremely few examples.

**Ensemble Method / Technique to Maximize ROC-AUC**

Prior competitions have shown that an ensemble of diverse models can significantly outperform any single model blog.csdn.net. Once we have multiple trained models, we need to combine their predictions effectively to maximize the ROC-AUC metric. We will evaluate several ensemble techniques and use a **weighted averaging (blending)** approach with optimized weights. **Weighted Averaging (Blending) is t**he simplest and often most robust method to average the prediction probabilities of each model in the ensemble. Rather than a plain average, we will assign higher weights to models that performed better on validation (higher AUC) and to models that complement each other. We can determine the weights by analyzing out-of-fold prediction performance: for example, solve an optimization (even a simple linear regression or grid search) to maximize AUC on the combined OOF predictions. In one audio competition, a weighted blending based on out-of-fold performance was used to boost ensemble score huggingface.co. We will do similar – using our CV results to find, say, model A should get 0.4, B 0.3, C 0.3 in the average (or even use a geometric mean for stability huggingface.co). This blending will be done at the probability level for each species. Weighted averaging is fast at inference and usually avoids overfitting, making it suitable for the final Kaggle submission.

**Audio Feature Extraction Techniques**

To robustly capture bird, amphibian, mammal, and insect sounds, we will combine multiple audio feature representations:

- **Mel-Spectrograms:** Convert audio into mel-frequency spectrogram images (time-frequency representations). These preserve rich frequency patterns (e.g. bird chirps, insect stridulations) and are well-suited for CNN-based classification. Mel-spectrogram features have proven highly effective for deep learning in audio tasks ieeexplore.ieee.org. We will use log-mel spectrograms with appropriate parameters (e.g. 32 kHz sample rate, 128 mel bands, window ~1024-2048, hop ~320-512) to capture relevant frequencies of target species. We may also experiment with Per-Channel Energy Normalization (PCEN) or different time-frequency resolutions to optimize for various call types (short chirps vs. long calls).

- **MFCCs (Mel-Frequency Cepstral Coefficients):** Extract MFCC features as a compact summary of the audio's spectral shape. MFCCs capture perceptually relevant timbre information and are relatively robust to noise and channel distortions ideas2it.com, which is useful in diverse outdoor recordings. We can compute 20-40 MFCC coefficients over time frames and use them in two ways: (1) feed sequences of MFCCs into a recurrent or convolutional model, or (2) aggregate statistics of MFCCs as features for a lightweight classifier. Including MFCC-based models adds diversity to the ensemble, as they may capture features (e.g. broad

spectral envelope) that complement spectrogram-based models.

- **Raw Waveform Processing:** Incorporate models that learn directly from raw audio waveforms. End-to-end waveform models can learn custom filters (e.g. via an initial convolutional layer or SincNet/learnable filterbank) to extract features optimal for the data. This avoids hand-crafted feature biases and could capture phase or temporal details lost in spectrograms. However, training from raw audio typically requires more data to outperform spectrograms
  reddit.com. To mitigate this with limited labeled data, we will **pretrain** the raw-wave model on unlabeled audio (self-supervised, see below) so it can learn useful filters. By including a raw waveform branch (e.g. a 1D CNN or WaveNet-style model), the ensemble gains another perspective that might detect subtle wave patterns or temporal transients that spectrograms smooth out. In practice, spectrogram-based approaches are expected to dominate with small data reddit.com, but a pre-trained waveform model could contribute unique signals for certain species.

**Self-Supervised and Semi-Supervised Learning Integration**

To fully exploit the unlabeled audio provided (as encouraged by the competition) imageclef.org, our strategy integrates both self-supervised pre-training and semi-supervised training:

- **Self-Supervised Pre-training:** We will train a neural **audio encoder** on the unlabeled audio recordings using self-supervised learning (SSL). The encoder could be a CNN or transformer that processes either raw wave or spectrogram inputs. We will adopt contrastive learning approaches such as **SimCLR or BYOL** on audio snippets: the model will be trained to produce similar embeddings for two augmented versions of the same audio clip and dissimilar embeddings for different clips. This way, the model learns to recognize inherent acoustic patterns without labels. In bioacoustics, contrastive SSL has been shown to yield representations that improve downstream classification, especially when labeled data are scarce arxiv.org. For example, we can take 5-second crops from the unlabeled soundscapes, apply augmentations (random time shift, filter, noise), and train the network to maximize agreement between embeddings of the same clip (positive pair) and minimize it for others (negative pairs) arxiv.org. This training will teach the model to encode general bird/insect/mammal sound characteristics (such as tonal vs. broadband calls, ambient noise patterns, etc.). We will use the learned encoder weights to initialize our spectrogram CNNs and/or the raw waveform model. This initialization gives a head start, requiring the model to only fine-tune to the specific species classification task rather than learning from scratch.

- **Unsupervised Pretrained Models:** In addition to our own SSL, if the competition permits, we will leverage existing pre-trained models as a form of self-supervised feature extraction. For instance, **BirdNET embeddings** (a model trained on thousands of bird audio hours) are known to provide powerful features for bird calls arxiv.org. We could run BirdNET (or a Google *Perch* model) on our audio to get embedding vectors, and feed those into a new classifier trained on the BirdCLEF labels. This is effectively transfer learning from a massive external dataset, which aligns with self-supervision goals. **Note:** We will only do this if allowed by

competition rules (using external models/data); otherwise, our own SSL pretraining on provided data will be the focus.

- **Semi-Supervised Learning (using unlabeled data):** Beyond pretraining, we will directly incorporate unlabeled data into the training loop via **pseudo-labeling** and **consistency training**:

    - *Pseudo-Labeling:* We will train an initial version of our model (or ensemble) on the labeled data, then use it to predict labels on the unlabeled audio. The highest-confidence predictions for each species will be treated as "pseudo-labels." We then combine these pseudo-labeled samples with the real labeled data to retrain/refine the models. This effectively increases the training set size. We will do this iteratively: each round, add new confident predictions to the training set and retrain, which can gradually improve the model's accuracy as it learns from more data. This technique has improved performance in past audio competitions; for example, in an audio tagging challenge, soft pseudo-labeling boosted validation scores significantly (e.g. CV score from 0.849 to 0.870 in one case) huggingface.co. We will be cautious to avoid reinforcing errors – only very confident predictions (or using ensemble agreement) will be added, and we'll monitor that the public LB doesn't degrade (to avoid overfitting to noise huggingface.co).

    - *Consistency / Teacher-Student Training:* We will also explore a mean-teacher approach or FixMatch-like strategy, where the model learns to be consistent on unlabeled data under perturbations. For example, for an unlabeled audio clip, we obtain a "teacher" prediction using the current model (with no augmentation), then train the model ("student") to predict the same output when the clip is augmented (e.g. adding noise or time-shift). This encourages the model to develop stable predictions and can leverage unlabeled data without explicitly assigning hard labels. Such semi-supervised consistency training can refine decision boundaries using the distribution of unlabeled examples.

By integrating these methods, the limited labeled data can be augmented with **information gleaned from the abundance of unlabeled recordings**, improving the model's ability to recognize rare species. Our final training pipeline will likely intermix these approaches – e.g. first self-supervised pretrain on all unlabeled, then supervised train on labeled, then a round of pseudo-label fine-tuning. This addresses the core challenge of BirdCLEF+ 2025: achieving high accuracy with very few labeled examples by *teaching the model from the data itself* imageclef.org.

**Cross-Validation Strategy for Robust Generalization**

A careful cross-validation (CV) scheme is essential to reliably gauge performance and to maximize the use of limited data. We will design a CV that is stratified and group-aware, while mindful of Kaggle's computational constraints:

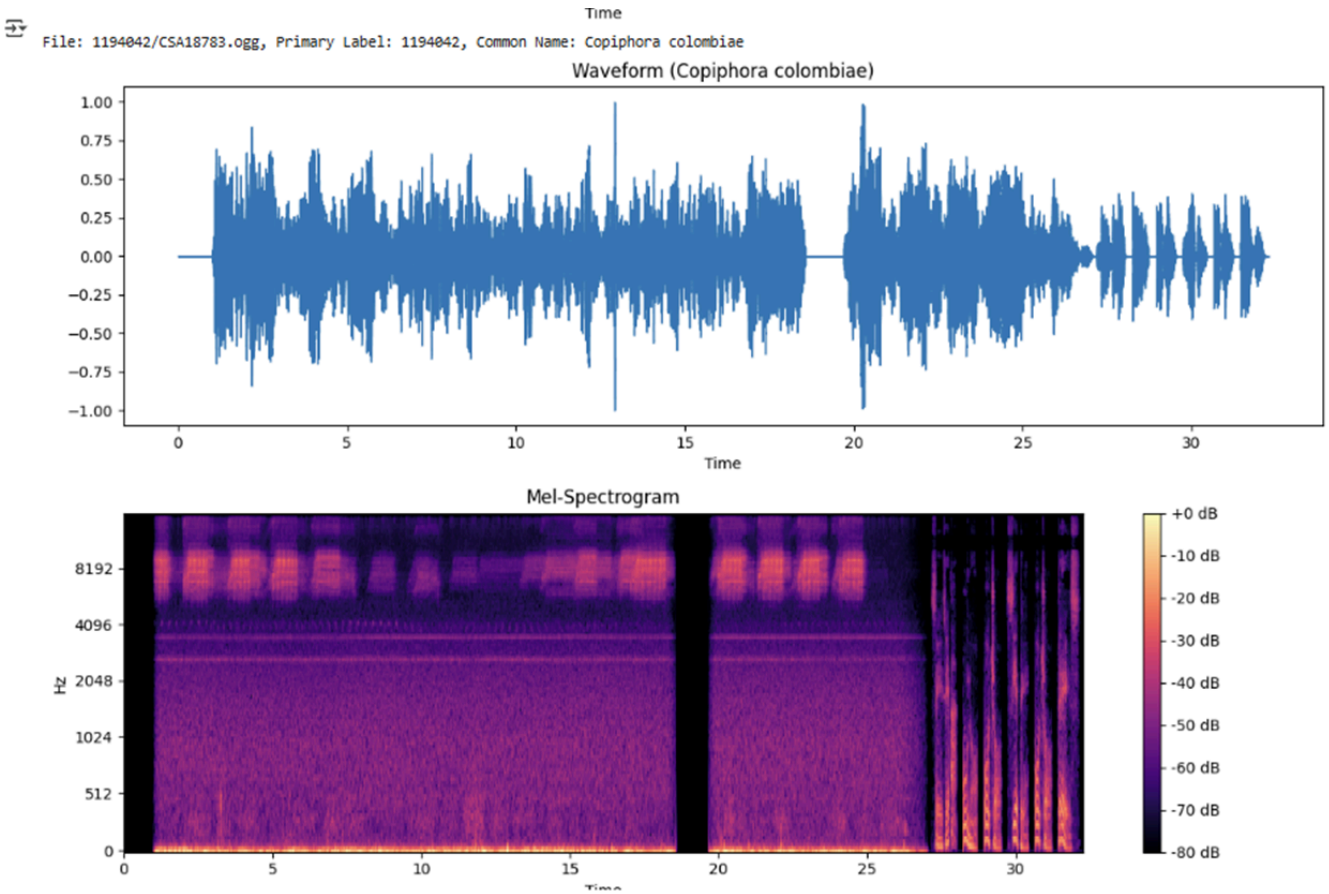
- **Stratified K-Folds:** We plan to use a 5-fold stratified cross-validation by species. Given the class imbalance (rare species with few recordings), stratification ensures each fold's training set has at least some examples of each species (where possible). If the task is multi-label (multiple species in one clip), we will use iterative stratification to maintain label distribution across folds. In BirdCLEF 2023, it was noted that a naive fold split

could omit certain classes in a fold; a solution was to drop that fold from validation blog.csdn.net. We will verify that each fold covers all or most species – if one fold is missing a rare species entirely, we might adjust (e.g. use 4-fold CV instead, as one team did blog.csdn.net). Our splits will be fixed and used consistently for model development and blending.

By employing stratified group k-fold, we maximize generalization and resilience to overfitting. The ensemble will be built on models that have all been validated on unseen folds, giving us confidence in their generalization. This CV strategy balances rigor with practicality under Kaggle's constraints by possibly limiting the number of folds or parallelizing fold training within the allowed timeframe.

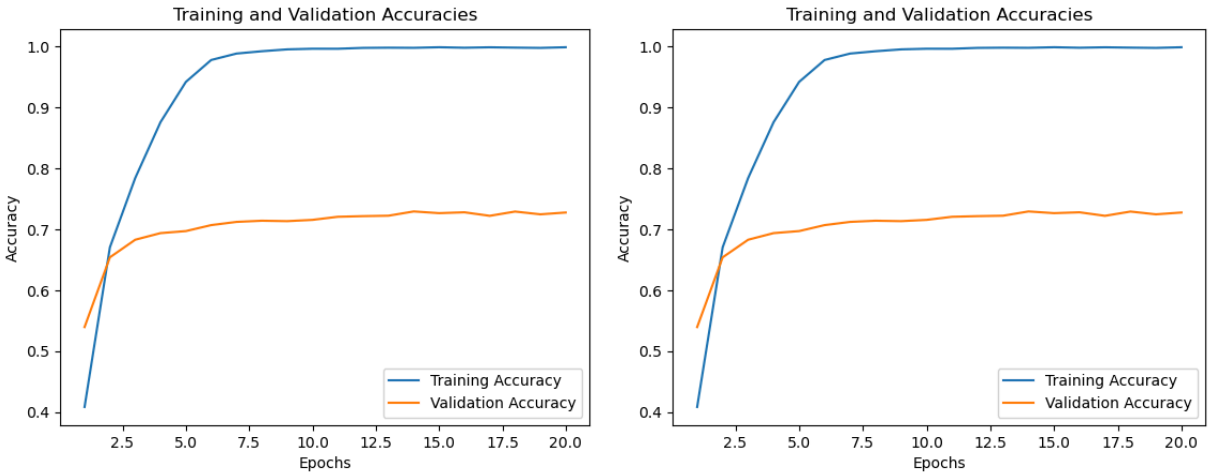
**Experiments and Analysis:**

The first machine learning architecture we decided to implement was a CNN on Mel-Spectrogram. The reason for this was that it was most familiar to us and we knew we could get good results from it. It may not be the most accurate model but it would provide us with a good starting point. To do this we used librosa to convert the sound files to a Mel-Spectrogram that could then be used during the training process. An example of this can be seen in the image below.



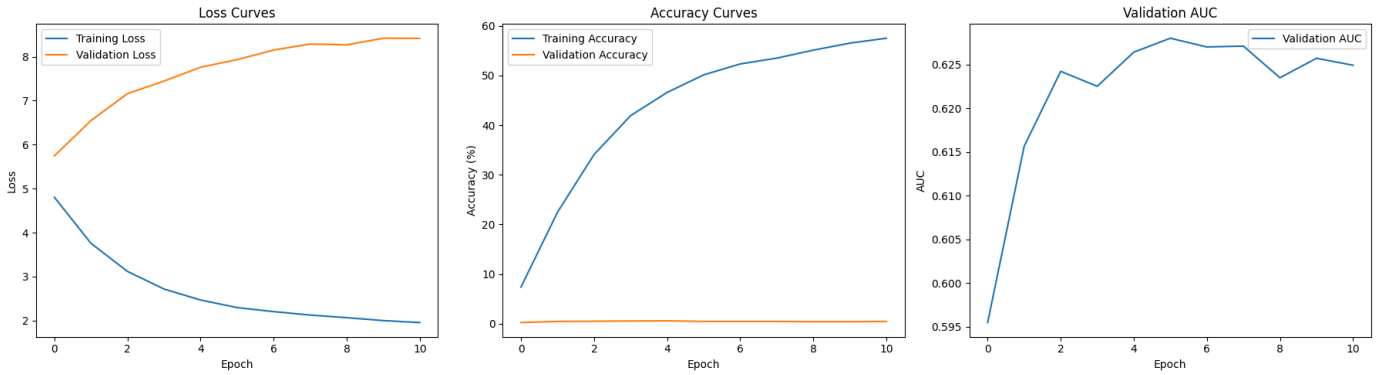
In this case the Mel-Spectrogram represents the vocalization data of a Copiphora Colombiae or a type of cricket. For this project we decided to limit the shape of the image to 128 by 256 with a 32 kHz sample rate. We did try other sizes like 128 by 128 and 150 by 150 but the size we ultimately decided on worked the best for this project. We did keep the hop size and window size as their default values and never made any changes to them during our experimentation. While each audio clip in the training dataset may be a different length, we decided to limit each one to just 5 seconds so that they are all uniform. We did try different clip lengths but it did not have any significant effect on the training and validation results.

After the Mel-Spectrograms were created from the audio files in the training dataset, we began to work on implementing the CNNss. The first CNN we tried was Resnet18. Initially we did not use any pretrained weights on the model, dropout was set at 0.02, batch size was set to 64, the learning rate was set at 0.002 and weight decay was set to 0.0001. Other than reducing the clip length to five seconds for every clip in the training dataset, data augmentation techniques were not used when using this model. Whatever was in the training dataset was sent through the model with no significant alterations. For the initial test, our training accuracy was 99% and our validation accuracy was 60%. This shows clear overfitting of the dataset with this model so we continued the hyperparameter search. We ended up improving the validation accuracy up to 72% but training accuracy was still at 99% which indicates that overfitting was still occurring. The charts below show how well the model was able to learn the dataset over 20 epochs.

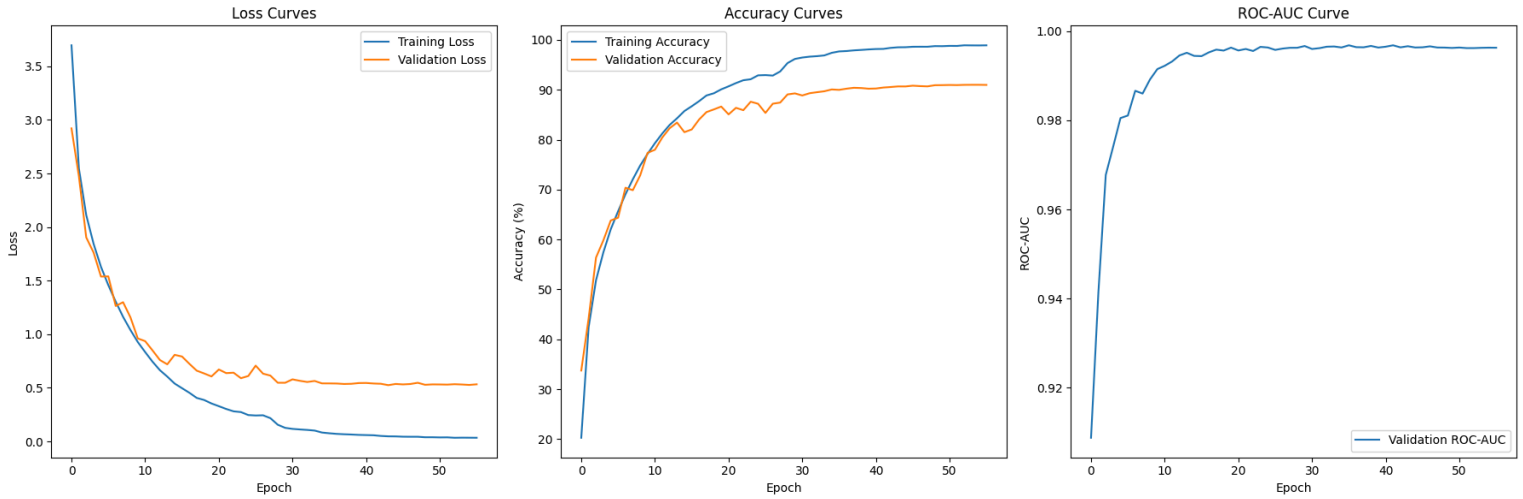



To achieve these results, one change that was made was that we incorporated pretrained weights from Imagenet1k_v1. Also we included label smoothing to help the model be able to better generalize the data and to combat the overfitting issue. Another change that occurred was the amount of dropout was adjusted to 0.06, we changed the learning rate to 0.004, and weight decay was adjusted to 0.0005. Also we made a change to the shape of the Mel-Spectrogram size. Initially the shape of the data was set to 150 by 150 but we found that 128 by 256 worked a lot better for the model. For this model we stuck to using some common architectures and employed no data augmentation techniques and we were still able to improve our validation score from the initial value of 60% all the way up to 72%. With some data augmentation techniques and additional parameter fine tuning, more progress on the overfitting problem could have potentially been made but due to time constraints, this is the best we could achieve with it.

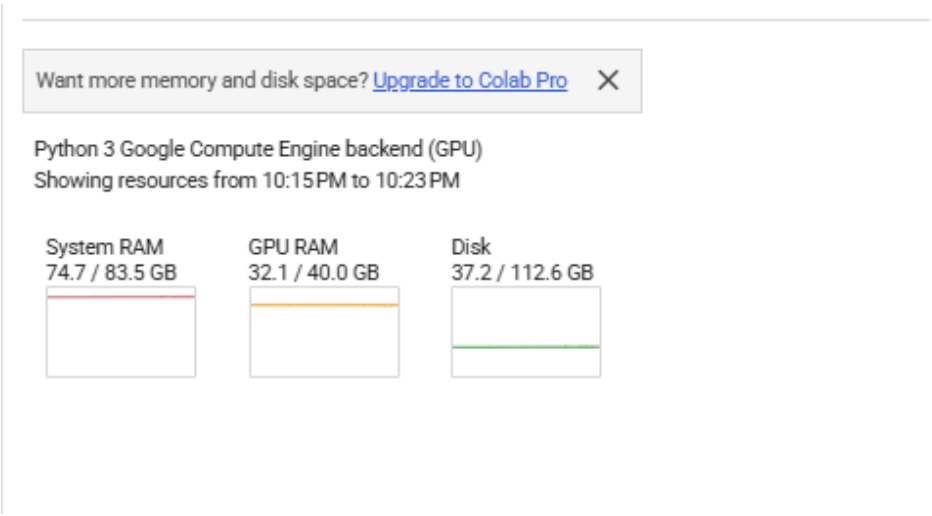
Along with resnet we also tried implementing Efficient Net B0. For this implementation, we did include data augmentation techniques this time. When applying them to the dataset we included no more than two random augmentations at a time to the specific sound files as they were being processed.




While these two models have shown some promise, we decided that a more custom model was needed to be able effectively compete in the competition.The model we came up with is a CNN neural network that we trained from scratch:




Even though BirdCNN received really good results, we did try to continue with the project plan and implement the other neural networks so that we could leverage the power of ensemble learning but we ran into a few issues that prevented us from being able to train those models in time for the presentations. One major issue was memory usage. This can be seen in the image below where system ram and gpu ram is almost maxed out. The usage charts were captured during the training of our BirdCNN architecture.




These models required a lot of memory in order to train them and when we went to implement the other more expensive models in the plan we easily ran out of available memory. In some cases within seconds we could see system ram usage climb to the point where it had to use swap memory which slowed us down significantly. Other methods of data extraction will need to be implemented to overcome this issue. Another issue we faced was related to how long it took to be able to train our models. Loading the data and processing it took up a lot of our time so we had a lot less time to experiment since changes to the data would require us to reload all the data. Also the amount of time it took our models to process each epoch took up a lot of time. Using lighter weight models such as resnet18 helped but it still limited how much time we could devote to fine tuning the parameters. It will also be an issue for the future models that we implement. Luckily the competition is still open until early June so there is still time to overcome these issues.

**Conclusion:**

In conclusion, we have begun to implement this strategy that was specified above. This strategy brings together **state-of-the-art audio processing techniques, an ensemble of complementary models, and advanced semi-supervised learning** to tackle BirdCLEF+ 2025. By extracting multiple feature types and leveraging unlabeled data, our models can recognize species with very few examples [imageclef.org](). A carefully crafted cross-validation and ensembling approach will ensure the solution is robust and achieves high ROC-AUC. All components are designed with Kaggle's practical constraints in mind, so the final solution is not only high-performing but also reproducible and efficient within the competition environment. While we currently are in the process of building the other machine learning models, our current custom CNN model that analyzes mel spectrograms achieves good results when tested on the dataset. By continuing this plan, we aim to maximize performance while fully complying with competition requirements, paving the way for a top-tier BirdCLEF+ 2025 submission.

**Sources:**


1. BirdCLEF+ 2025 Competition Description – Limited labels and unlabeled data emphasis
   imageclef.org

2. Effectiveness of mel-spectrogram features with CNN models
   ieeexplore.ieee.org

3. Advantages of MFCCs (noise and channel robustness)
   ideas2it.com

4. Discussion on raw waveform vs. spectrogram for deep learning
   reddit.com

5. Success of diverse CNN ensemble in BirdCLEF (2023)
   blog.csdn.net

6. BirdCLEF 2023 4th place – cross-validation and fold coverage detail
   blog.csdn.net

7. Self-supervised contrastive learning benefits for audio representations
   arxiv.org

8. Pre-trained bird audio embeddings aiding few-shot classification
   arxiv.org

9. Pseudo-labeling improving audio classification performance
   huggingface.co

10. Weighted ensemble blending using OOF optimization
    huggingface.co

11. Example of loading multiple model checkpoints for ensemble (Kaggle code)
    blog.csdn.net