Overview of BirdCLEF 2025: Under-studied species recognition in soundscape recordings

Taha H. Ababou[1,*,†], Vajinder Kaur, Paul Moon and Reese Mullen

Masters in Statistical Practice, Boston University

Abstract

Humid tropical rainforests are home to incredible biodiversity, but tracking mobile and elusive species in these environments is tough—traditional surveys take time, labor, and money. Passive acoustic monitoring (PAM) using autonomous recorders is a promising alternative, yet most approaches still rely on humans manually labeling hours of recordings. While deep learning has made it easier to identify common bird species from audio, rare and lesser-known species are still a major challenge due to class imbalance and limited data. BirdCLEF 2025 tackled this issue by asking participants to detect and rank 206 rare bird species from long-form audio collected in Colombia's El Silencio Natural Reserve. Our solution used mel-spectrograms, an ensemble of four EfficientNet v2 models trained with a mix of Focal and BCE loss, and temporal smoothing. The approach achieved a leaderboard accuracy of 0.835.

Keywords

LifeCLEF, BirdCLEF, bioacoustics, passive acoustic monitoring, endangered species, acoustic classification, machine learning, mel-spectrograms

---

*GitHub repository: https://github.com/VajinderKaur/BirdCLEF-2025/

†These authors contributed equally.

hababou@bu.edu (T. H. Ababou); vajinder@bu.edu (V. Kaur); phmoon01@bu.edu (P. Moon); mullenrd@bu.edu (R. Mullen)

https://tahaababou.com/ (T. H. Ababou); https://github.com/vajinderkaur (V. Kaur); https://github.com/phmoon01 (P. Moon); https://github.com/mullenrd (R. Mullen)

1. Introduction

Passive acoustic monitoring (PAM), employing autonomous recorders to survey wildlife across extensive spatial and temporal scales, has emerged as a transformative approach for biodiversity assessment [1]. Unlike traditional observer-based surveys, which are costly and logistically constrained, PAM allows continuous sampling of mobile and habitat-diverse species whose vocalizations serve as key indicators of ecosystem health and restoration success [2]. However, the inherent complexity of tropical soundscapes—with high species richness, overlapping vocal activity, and intense ambient noise—continues to hinder reliable automated detection, especially for under-studied taxa characterized by very limited labeled data [3].

Colombia's lowland Magdalena Valley exemplifies these challenges. As a megadiverse region where over 70% of original rainforest has given way to cattle pasture and fragmented remnants face illegal logging, efficient biodiversity monitoring is imperative [4,5]. El Silencio Natural Reserve, encompassing 5,407 acres of restored lowland forest and wetlands, hosts an extraordinary assemblage—including 295 bird species, 34 amphibians, 69 mammals, 50 reptiles, and nearly 500 plant taxa—many of which are rare or endangered [6]. Traditional field-based surveys here demand substantial human effort and remain constrained by access and cost.

To advance scalable monitoring within this critical landscape, BirdCLEF 2025 challenged participants to develop machine-learning pipelines capable of identifying under-studied species directly from continuous audio collected at El Silencio. By leveraging mel-spectrogram transforms, deep convolutional backbones, and tailored loss functions, the competition sought robust few-shot classification methods that generalize across complex tropical soundscapes and contribute actionable insights for conservation practitioners.

2. BirdCLEF 2025 Competition Overview

In recent years, bioacoustic monitoring has benefited enormously from deep neural networks capable of recognizing species calls in focused, noise-controlled recordings [7,8]. Yet, applying these models to authentic tropical soundscapes—with overlapping vocalizations, shifting ambient noise, and scarce labels for many taxa—remains an open challenge. BirdCLEF 2025 sought to close this gap by providing two complementary datasets: 14,500 labeled focal recordings covering 206 species and a hidden test corpus of 5,356 one-minute soundscapes (roughly 90 hours) captured year-round across seven sites in El Silencio Natural Reserve. Each focal clip included metadata on location, date, and recording quality to support strategies like domain adaptation and few-shot learning.

In this project, our team designed a complete inference pipeline that divides continuous audio into 5-second segments, converts each segment to a mel-spectrogram, and passes it through an ensemble of EfficientNet V2 S models trained with a custom Focal+BCE loss. We then apply temporal smoothing across successive segments to enhance detection consistency. Performance is measured by mean Average Precision at 5 (mAP@5), which equally weights all species—thereby rewarding accurate ranking for rare and common taxa alike.

2.1. Goal and Evaluation Protocol

Participants were tasked with generating ranked lists of the top 5 species per 5-second segment within continuous one-minute recordings. The primary metric, mean Average Precision at 5 (mAP@5), equally weighted rare and common species, incentivizing accurate ranking over mere presence/absence classification. Entrants could tune confidence thresholds to balance precision and recall, mimicking real-world deployment scenarios.

2.2. Dataset

This challenge requires identifying which species—spanning birds, amphibians, mammals, and insects—are calling in recordings made at El Silencio Natural Reserve, Colombia. More accurate automated solutions will enable conservationists to monitor animal populations at scale, informing ecological restoration efforts.

Dataset Files Provided:

- **train_audio/**: Short recordings of individual vocalizations uploaded by xeno-canto, iNaturalist, and the Colombian Sound Archive (CSA). All files are in OGG format at 32 kHz. Filenames follow `[collection][file_id_in_collection].ogg`. `[collection]` is the folder under `train_audio/` directory which contains `[file_id_in_collection].ogg` audio files. There are a total of 28564 audio files.
- **test_soundscapes/**: Approximately 700 one-minute recordings (hidden during development) used for scoring. Filenames are randomized (e.g., `soundscape_xxxxxx.ogg`). These are only available after the notebook is submitted.
- **train_soundscapes/**: Unlabeled recordings from the same general locations as the test set. Filenames use `[site][date][local_time].ogg`.
- **train.csv**: The training data used in this study consists of a variety of metadata that helps describe the recordings. The most relevant fields for our analysis are as follows:
  - **primary_label**: This is a unique code for each species. For birds, it uses the eBird code, while for non-bird species, it uses the iNaturalist taxon ID. If you are curious about the species, you can easily access detailed information by appending the code to the eBird or iNaturalist URLs.
  - **secondary_labels**: This field contains a list of additional species that were noted by recordists as potentially present in the recording. However, this list can be incomplete, and not all species might be accurately marked.
  - **latitude & longitude**: The geographical coordinates where the recording was made. These are crucial as some species may exhibit regional variations in their calls, so ensuring diversity in the geographical locations represented in the training data can improve the model's performance.
  - **author**: The person who contributed the recording. If the author is not provided, the entry will be marked as unknown.
  - **filename**: The name of the audio file associated with the recording. This helps in linking each metadata entry to its corresponding audio data.
  - **rating**: This field contains a quality rating of the recording, ranging from 1 to 5, where 1 indicates a low-quality recording and 5 indicates a high-quality one. If no

rating is available, the value is recorded as 0. Note that iNaturalist and the CSA (Community Science Association) do not provide quality ratings, so their entries will not have values for this field.

  - **collection**: This indicates the source of the recording. The collection can be either "XC" (for Xeno-canto), "iNat" (for iNaturalist), or "CSA" (for the Community Science Association). The filenames also contain references to the collection and the unique ID within that collection.

- **sample_submission.csv**: Template showing the required submission format with `row_id` and one column per species (206 total).
- **taxonomy.csv**: Provides species taxonomy, including eBird codes or iNaturalist taxon IDs and class names (Aves, Amphibia, Mammalia, Insecta).
- **recording_location.txt**: High-level description of the El Silencio Natural Reserve recording sites.

3. EDA

3.1. Class Imbalance in Species Audios Distribution

The training dataset includes 206 distinct species, identified using primary labels. To investigate class imbalance, we examined the top 30 most represented species based on the number of audio recordings. A clear imbalance was observed: the most frequent species had close to 1,000 recordings, while the 30th species had approximately 300 recordings. This substantial gap indicates that the dataset is skewed toward a few species, which may bias model performance if not properly addressed.

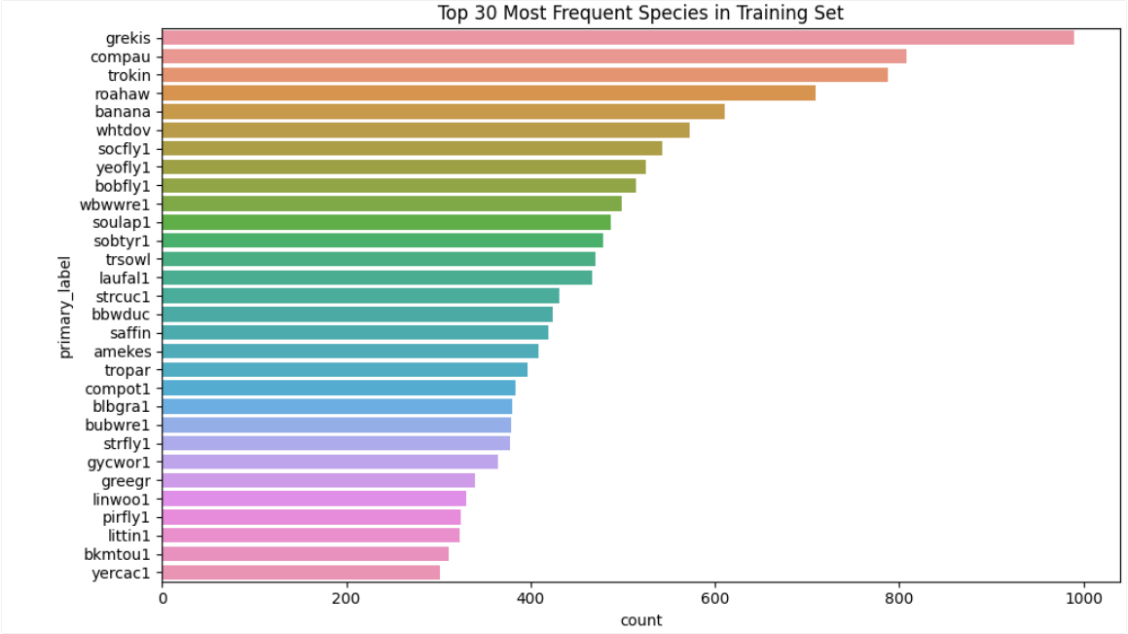

Figure 1: Top 30 Most Frequent Species in the Training set

3.2. Duration of Recordings

Beyond count, recording duration (a column generated by us for `train.csv`) provides another layer of complexity. We generated a box plot of recording durations, which revealed a high number of outliers—especially for species with lower recording counts. These outliers represent recordings with exceptionally long durations, suggesting that some low-frequency species may still carry valuable temporal information. However, such outliers can also distort summary statistics.

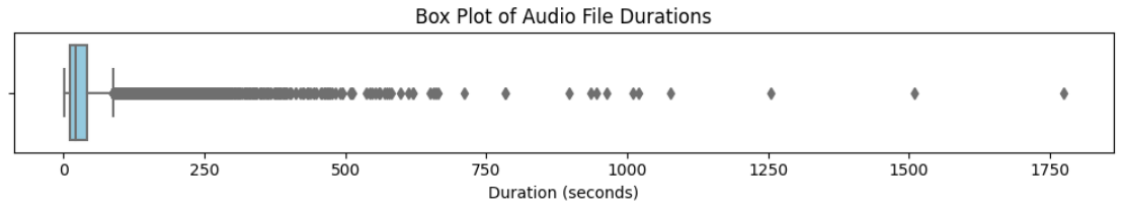

Figure 2: Outliers for the Audio Files Durations

There was another question in our mind after observing the box plot ; *what if a species that has fewer recordings might be having significantly longer recordings, potentially compensating for the low count in terms of available data.* To explore this, we analyzed the relationship between the number of recordings per species and their average durations (derived from audio files in the train_audio directory).

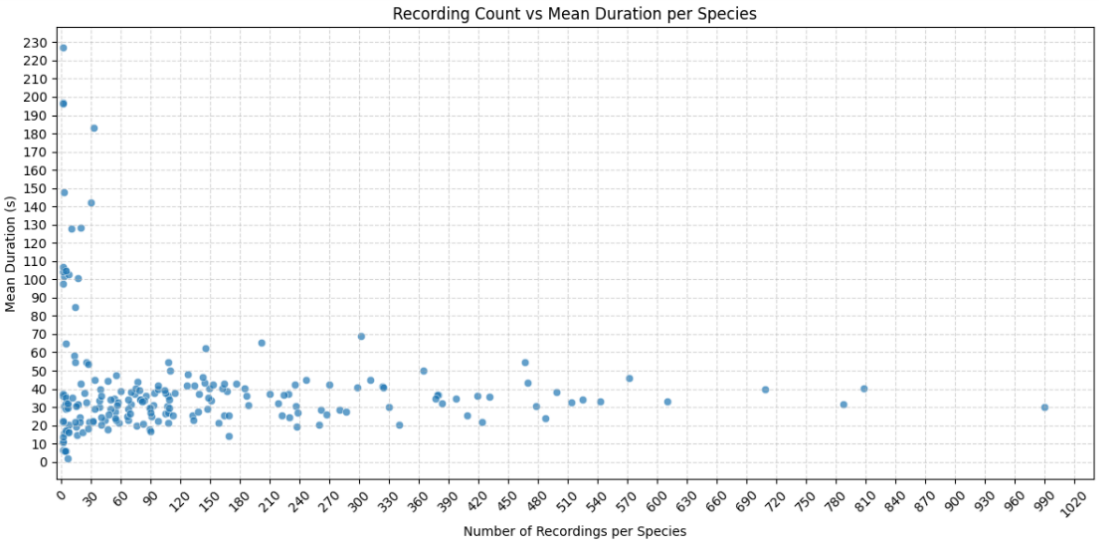

Figure 3: Number of Recordings vs Mean Duration of recording per species

A scatterplot of species count vs. mean duration confirmed our hypothesis: species with fewer recordings often had longer individual recordings. In contrast, species with higher recording counts tended to have more consistent, shorter durations.

To investigate species with particularly sparse data, we zoomed into those with fewer than 40 recordings. Here, we opted to analyze the median duration rather than the mean. The median is

less sensitive to extreme values and provides a more robust measure of central tendency when dealing with skewed data or outliers. This gives a clearer view of the typical recording length, particularly for underrepresented species.

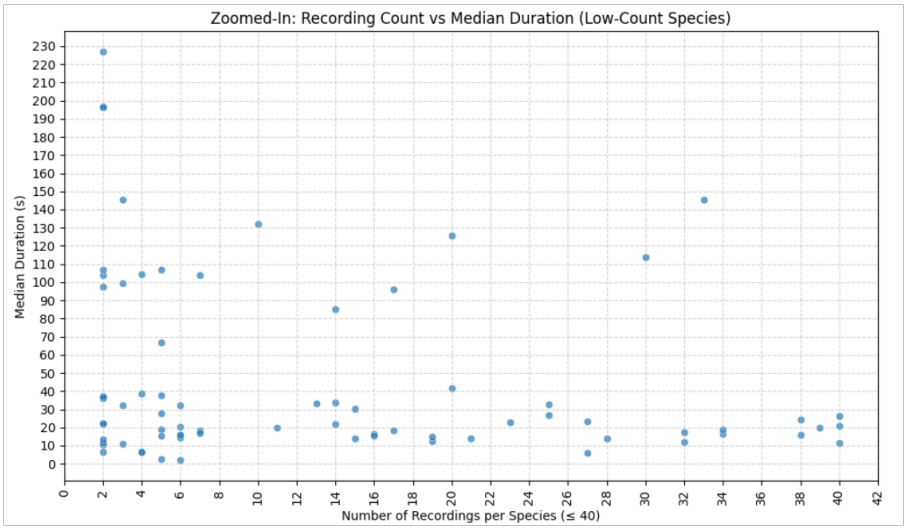

Figure 4: Zooming in for the species recording less than 40

3.3. Summary Table of Sparse Species

We generated a table listing species with fewer than 40 recordings. Many of these had no rating information (i.e., a rating value of 0), which doesn't necessarily imply poor quality but simply that no user ratings were available. Additionally, the highlighted species—those with as few as 2 or 3 recordings—tend to have both lower durations and no rating data, raising concerns about data sparsity and quality for these classes. (The table (Table 1) doesn't contain all under 40, we chose some of them for the report since it was a long list)

Table 1
Species Data with Recordings less than 40, Mean Duration, and Rating

Scientific Name	Recordings	Mean Duration (s)	Rating
Allobates niputidea	**2**	**292.49**	**0.0**
Alouatta seniculus	23	218.78	4.0
Andinobates opisthomelas	10	212.25	0.0
Ardea cocoi	40	112.77	5.0
Arundinicola leucocephala	34	105.46	5.0
Boana boans	25	389.77	5.0
Boana pugnax	6	25.34	0.0
Bradypus variegatus	**2**	**51.37**	**0.0**
Cathartes aura	11	134.95	5.0
Cerdocyon thous	**2**	**11.96**	**0.0**
Chauna chavaria	14	147.70	5.0
Chloroceryle aenea	28	68.68	5.0
Chrysuronia goudoti	15	57.99	5.0
Cicadidae	30	338.67	0.0

Scientific Name	Recordings	Mean Duration (s)	Rating
Cocconotus aratifrons	3	107.82	0.0
Cochlearius cochlearius	39	278.23	5.0
Colostethus inguinalis	**2**	**382.30**	**0.0**
Copiphora colombiae	3	46.04	0.0
Copiphora gracilis	**5**	**12.24**	**0.0**
Copris susanae	17	129.77	0.0
Crax alberti	20	96.31	5.0
Daedadelus waehnerorum	6	2.99	0.0
Dendrobates truncatus	2	8.86	0.0
Dendropsophus bogerti	6	48.23	0.0
Elachistocleis panamensis	5	38.96	4.5
Elachistocleis pearsei	2	296.62	0.0
Eschatoceras bipunctatus	**4**	**9.38**	**0.0**
Espadarana prosoblepon	4	48.27	4.0
Fluvicola pica	19	108.69	5.0
Gryllidae	33	896.58	0.0
Hyalinobatrachium tatayoi	7	39.32	4.0
Hypnelus ruficollis	17	84.24	5.0

3.4. Spectrogram Examples

To understand the audio content visually, we converted audio files into mel spectrograms. These transformations illustrate the variation in acoustic patterns between species and provide a basis for image-based modeling approaches. Spectrograms also help identify background noise, overlapping calls, or silence, which are important factors for training robust models.

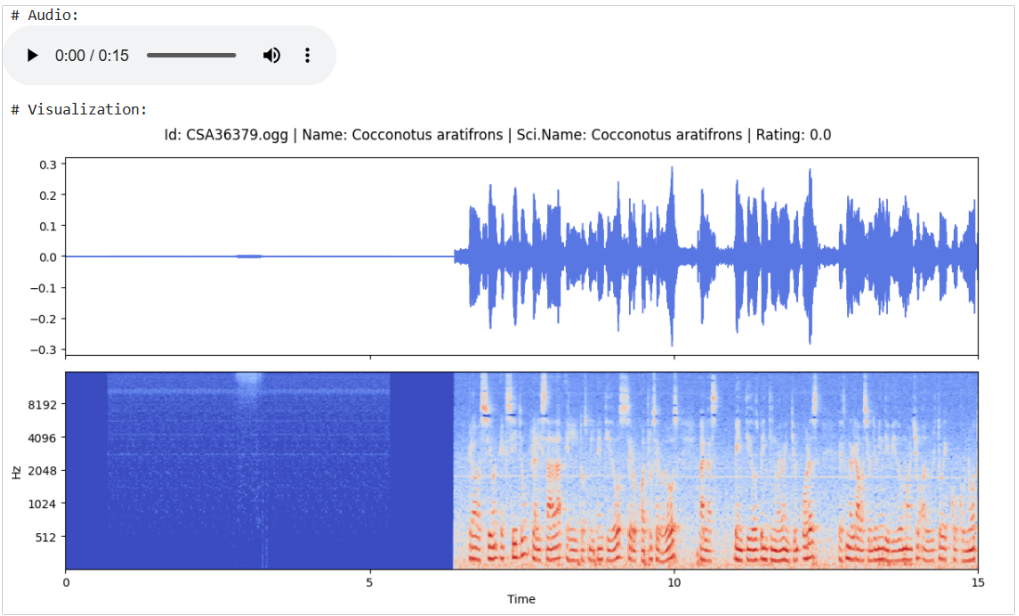

Figure 5: Example 1 of how recording is converted to a Spectrogram

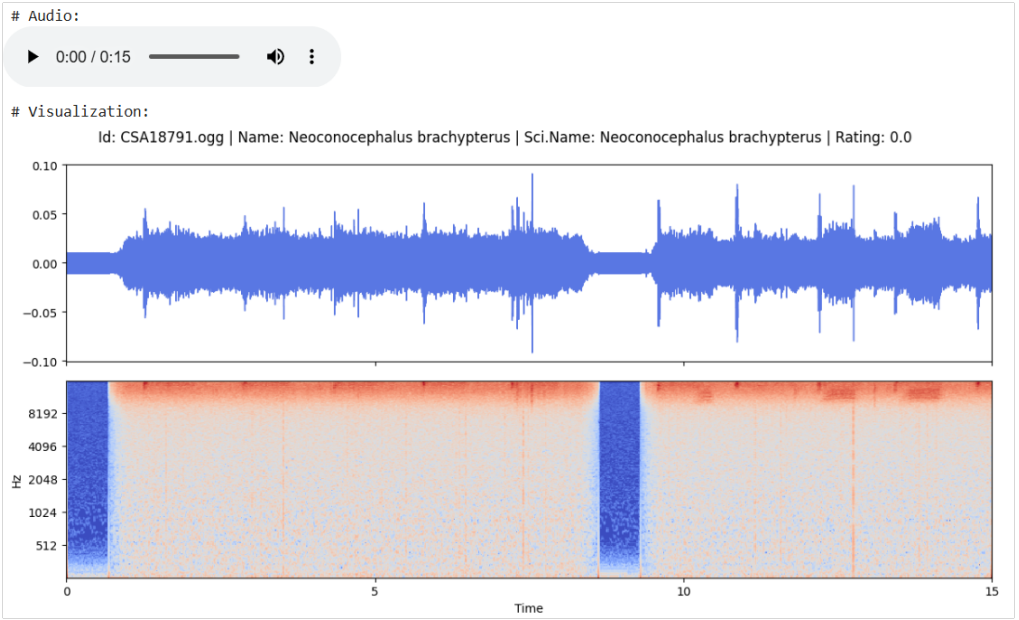

Figure 6: Example 2 of how recording is converted to a Spectrogram

4. Methods

Our pipeline was designed to handle real-world audio from tropical soundscapes while addressing the practical challenges of class imbalance, background noise, and sparse training labels. The overall strategy included spectrogram-based preprocessing, a transfer learning backbone, a composite loss function, ensemble inference, and temporal smoothing. Below we describe each step and its motivation.

`Note:` We begin by looking at discussions and the current code of participants. We proceeded with a baseline model built by a participant, which is shown on Kaggle, and made further changes by incorporating a random selection of 5-second segments rather than just starting 5-second segments, removing raw audio augmentation, and hyperparameter tuning. In addition to this, we tried different models like B0 and B1 for the same.

4.1. Preprocessing and Segmentation

Each one-minute test soundscape was divided into non-overlapping 5-second segments—long enough to capture complete vocalizations, yet short enough for localized predictions. These audio windows were resampled to 32 kHz and converted into 256×256 mel-spectrograms using a 2,048-point FFT, 512-sample hop length, and 256 mel bands spanning 20–16 kHz. The resulting spectrograms highlight vocal patterns from multiple taxonomic groups while normalizing signal magnitudes to a [0,1] range for training stability.

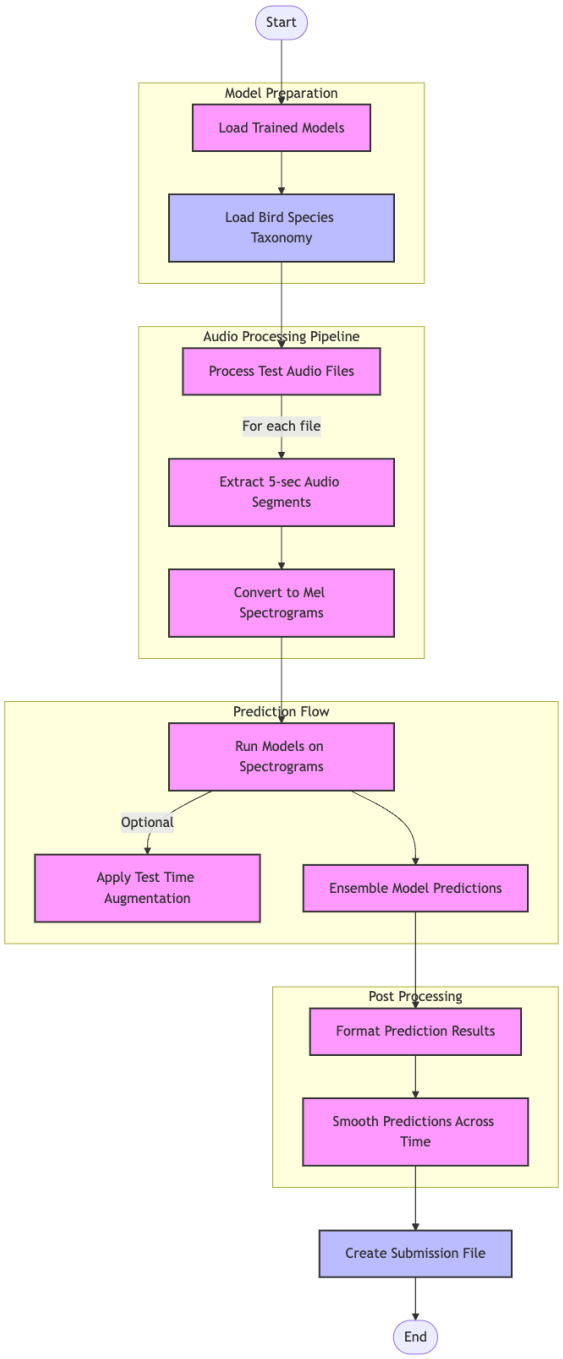

Figure 7: Pipeline architecture

4.2. Model Architecture and Training Strategy

We adopted the `tf_efficientnetv2_s.in21k_ft_in1k` architecture from the TIMM library, selected for its favorable accuracy-to-computation tradeoff. The model was initialized with ImageNet-21k weights and modified to accept single-channel spectrograms. The final classification layer was replaced with a 206-unit linear head to reflect the multi-label nature of the task.

Training employed a custom composite loss function combining binary cross-entropy (BCE) and sigmoid focal loss. BCE stabilized learning on dominant species, while focal loss prioritized harder, under-represented examples using parameters $\alpha = 0.25$, $\gamma = 2$, and weights 0.6 and 1.4 for BCE and focal loss, respectively.

4.3. Ensembling and Inference

To improve robustness, we trained four models using different validation folds and averaged their predictions. This ensemble strategy reduced variance and yielded smoother outputs, especially on rare classes. During inference, each 5-second segment was independently passed through the ensemble without test-time augmentation.

4.4. Temporal Post-processing

While deep models can classify segments effectively, their outputs may exhibit jitter due to transient background noise or clipped calls. To address this, we smoothed predictions across time using a weighted moving average: [0.2, 0.6, 0.2] for internal segments and [0.9, 0.1] at the edges. This filter suppressed spurious activations and improved consistency without sacrificing responsiveness to real events.

5. Results

A total of 1,247 participants (1,556 teams) submitted 37,029 entries. Our approach, featuring an ensemble of four EfficientNet models trained with Focal+BCE loss, achieved a competitive private mAP@5 score of 0.835 (rank 116) surpassing the baseline we used..

CNN ensembles remained dominant; however, innovative post-processing and tailored data augmentation were essential for top performance. Few-shot methods were surprisingly underutilized despite limited training examples.

5.1. Per-Species Analysis

Performance varied considerably among species. Class imbalance and acoustic complexity significantly impacted detection accuracy, highlighting the need for species-specific modeling strategies.

6. Next Steps

Based on insights from the Kaggle competition discussions, we outline the following steps for improving the model pipeline:

1. **Preprocessing and Augmentation**:
    - Convert the Mel spectrograms into image format.
    - Apply image normalization.
    - Use advanced augmentation techniques including:
        - RandAugment
        - Random Erasing
        - Time masking
        - Frequency masking
    - Implement Mixup augmentation with a probability of 1.0 to enhance generalization.
2. **Model Architecture**:
    - Use EfficientNet-B0 as the backbone CNN.
    - Extract intermediate features and combine them with the output of the global pooling layer to enrich representation.
3. **Training Strategy**:
    - Train and validate using a single fold to establish a baseline.
    - Extend to a multi-fold training setup and ensemble the results for improved robustness and accuracy.

All competition data, code, and models are publicly accessible at:
https://www.kaggle.com/competitions/birdclef-2025

Our team's source code submission is accessible at:
https://github.com/VajinderKaur/BirdCLEF-2025/