One Detector per Bird: A Scalable Binary Classification Approach for BirdCLEF+ 2025*

Shreejith Suthraye Gokulnath[1], Chandrima Das[1], Arya Gaikwad[1], Keerthana Senthilnathan[1] and Shruti Prasad Sawant[1]

[1]University of California, San Diego (UCSD), La Jolla, CA 92093, USA

Abstract

Automated bird sound classification has become an essential tool for ecological monitoring and biodiversity research. The BirdCLEF 2025 challenge presents a large-scale, multi-species audio classification task with over 206 target species recorded under highly variable acoustic conditions. To address the challenges of extreme class imbalance, overlapping species, and domain variability, we propose a modular framework that trains one binary classifier per species. This approach allows for targeted feature engineering, interpretability, and scalable parallel training. We develop the system in three iterative stages: starting with a single-species prototype, expanding to a multi-species configuration, and finally scaling to the full species set with metadata integration and targeted data augmentation. Audio recordings are encoded using binary frequency activation patterns and log-mel statistics, with classification performed using Random Forests and XGBoost. Although local cross-validation results showed strong performance, domain shifts in the test data highlighted the complexity of ecologically real-world soundscapes. Our findings underscore the value of species-specific modeling strategies and offer a flexible framework for future bioacoustic monitoring systems. We participated as team `echo` in the BirdCLEF 2025 challenge and achieved a public leaderboard score of **0.568** and a private leaderboard score of **0.561**. These results reflect the generalization difficulty posed by diverse ecological soundscapes and validate the scalability of our binary-classifier approach.

Keywords

BirdCLEF 2025, Bird sound classification, Bird sound classification, Bioacoustics, Species detection, Binary classifiers, Class imbalance, XGBoost, Ecological monitoring,

CEUR-WS.org/Vol-4038/paper_254.pdf

1. Introduction

Bird sound classification plays a critical role in biodiversity monitoring and ecological research. As the global biodiversity crisis intensifies, monitoring wildlife populations has become more critical than ever. Birds, as highly responsive indicators of ecosystem health, are central to many conservation efforts. With the growing availability of passive acoustic monitoring systems, large volumes of field recordings are being collected across diverse habitats offering a unique window into bird communities at scale. However, extracting meaningful insights from these recordings requires automated systems capable of identifying species accurately under real-world conditions, including overlapping calls, background noise, and vast species diversity.Recent work has highlighted the growing potential of passive acoustic monitoring tools enhanced with deep learning, few-shot detection, and metadata-aware models.

The BirdCLEF 2025 challenge addresses this need by tasking participants with detecting bird species in long-form audio recordings. Unlike controlled laboratory datasets, these recordings are heterogeneous in quality, length, and complexity. Moreover, the dataset presents a classic long-tail distribution: a few species are well-represented, while many appear infrequently, making traditional multi-label classification approaches less effective.

To tackle these challenges, we adopt a modular, species-specific modeling strategy. Instead of training a single multi-label classifier, we train a dedicated binary model for each of the 206 target species. This design provides several advantages: it handles extreme class imbalance more effectively, allows for species-specific feature tuning and error analysis, and enables parallel model training.

Our development process unfolded in three stages. We began with a single-species prototype to validate our core pipeline, then expanded to a three-species setup to test generalization, and finally

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

scaled to the full species set using metadata-aware preprocessing, adaptive thresholding, and targeted augmentation for underrepresented classes. Our models rely on interpretable binary frequency activation features derived from short-time Fourier transforms and log-mel spectrograms, combined with ensemble learning methods such as Random Forests and XGBoost.

Despite promising validation results, our final leaderboard performance highlighted persistent challenges in domain generalization, a common issue in ecological AI. Nonetheless, our results demonstrate that modular, per-species classification is a viable and scalable approach for large-scale bioacoustic monitoring. This work contributes a reproducible framework for future research and underscores the value of flexible, interpretable systems in complex environmental tasks.

2. Dataset

The BirdCLEF 2025 challenge dataset encompasses multiple components to facilitate automated bird sound recognition:

1. **Train audio**: The BirdCLEF 2025 training dataset comprises short audio recordings of individual vocalizations from a wide range of species, including birds, amphibians, mammals, and insects. These recordings were contributed by three primary sources:
    - xeno-canto.org
    - iNaturalist
    - The Colombian Sound Archive (CSA), curated by the Humboldt Institute for Biological Resources Research in Colombia

    Resampling of the recordings is done in the dataset to a uniform sampling rate (32kHz) to align with the test set audio. The format of the files is OGG which offers a balance between compression and quality. No additional data was downloaded from xeno-canto or iNaturalist. Each audio file follows the following naming convention: [collection][file_id_in_collection].ogg where the prefix identifies the source collection.
2. **Test Soundscapes**: The test dataset contains approximately 700 soundscape recordings having a duration of exactly one minute. They are also provided in OGG format. These audio files have been resampled to 32 kHz, ensuring consistency with the training data.
   The test_soundscapes directory is automatically populated with the test audio files when the notebook is submitted to the competition platform. No inference of content or ordering is done from the filename alone as they follow a randomized pattern of the form soundscape_xxxxxx.ogg.
3. **Train Soundscapes**: The training dataset includes a collection of unlabeled soundscapes, long-form audio recordings captured at the same general locations as the test soundscapes. These files are intended to help understand environmental background noise, acoustic context, and species co-occurrence patterns in realistic settings. Each file is named using the format: [site]_[date]_[local_time].ogg, which encodes the recording site identifier, date, and local time of the capture. The recordings do not overlap with the hidden test soundscapes inspite of them being recorded at the same geographic region. Therefore overfitting will be avoided and the model can generalize across unseen environments.
4. **Training metadata**: The dataset contains a csv file that provides metadata for each training audio recording. It provides information like rating of an audio, geographic diversity etc. A few columns include:
    - **Primary Label**: It consists of a standardized species code which represents the primary vocalization in the recording. For birds, this corresponds to the eBird species code (e.g., gretin1 for Great Tinamou); for non-bird taxa, the iNaturalist taxon ID is used. These codes can often be appended to URLs for further species information such as https://ebird.org/species/gretin1 for the Great Tinamou.

- **Secondary Label**: It contains a list of other species that are annotated by recordists that are also present in the background of the recording. The field maybe incomplete in some cases and is treated by caution.
- **Latitude and Longitude**: This column contains the latitude and longitude (geographic coordinates) which indicate the location of the recording capture. These are used to understand regional vocal dialects which may be present in some bird species.
- **Author**: This column contains the name of the user who contributed the recording. It can also have a value as "unknown" indicating that the author chose to remain anonymous and was not recorded.
- **Filename**: It contains the name of the associated audio file, which also encodes its source collection.
- **Rating**: Rating varies from 1(low) to 5(high) and is provided by the users of Xeno-canto. A value of 0 implies that no rating is available. In addition to that, iNaturalist and CSA do not provide quality ratings.
- **Collection**: It indicates the origin of the recording: XC (xeno-canto), iNat (iNaturalist), or CSA (Colombian Sound Archive). This field also aligns with the prefix of the filename.

5. **Sample Submission**: This file contains the valid sample submission format which contains a csv of the row_id and species_id. The probability of the presence of each species needs to be predicted for each row.
6. **Taxonomy**: This file consists of data on different species along with taxonomic hierarchy and standardized codes.
7. **Recording Location**: This file contains location metadata associated with each recording, enabling spatially informed analysis.

3. Dataset Analysis

3.1. Species Distribution and Class Imbalance

Comprehensive analysis reveals 206 unique species represented in the dataset, exhibiting typical ecological abundance patterns. The species distribution demonstrates moderate imbalance, with 23 species (11.2%) having fewer than 5 recordings, 39 species (18.9%) having fewer than 10 recordings, and 78 species (37.9%) having fewer than 50 recordings. This long-tail distribution presents significant challenges for machine learning approaches, particularly for rare and endangered species that are primary targets of conservation efforts.

The top 20 most represented species show clear dominance patterns, with several species exceeding 800 recordings while many others fall below 100 recordings. This class imbalance necessitates specialized training strategies, including weighted loss functions, balanced sampling techniques, and data augmentation approaches tailored to minority classes.

3.2. Audio Characteristics and Quality Assessment

Statistical analysis of audio characteristics reveals substantial temporal heterogeneity within the dataset. Duration statistics show a right-skewed distribution with the first quartile at 13.0 seconds, median at 22.6 seconds, third quartile at 45.2 seconds, and maximum extending to 249.8 seconds. This variation reflects the natural diversity of vocalization behaviors across taxonomic groups, from brief insect chirps to extended mammalian calls.

All analyzed recordings maintain consistent technical specifications with 32 kHz sampling rates and OGG compression format. Root mean square energy levels average 0.033 with standard deviation of 0.026, indicating moderate amplitude variability across recordings. Maximum amplitude values range from 0.005 to 1.10, suggesting varying recording conditions and source-to-microphone distances typical of field recordings.

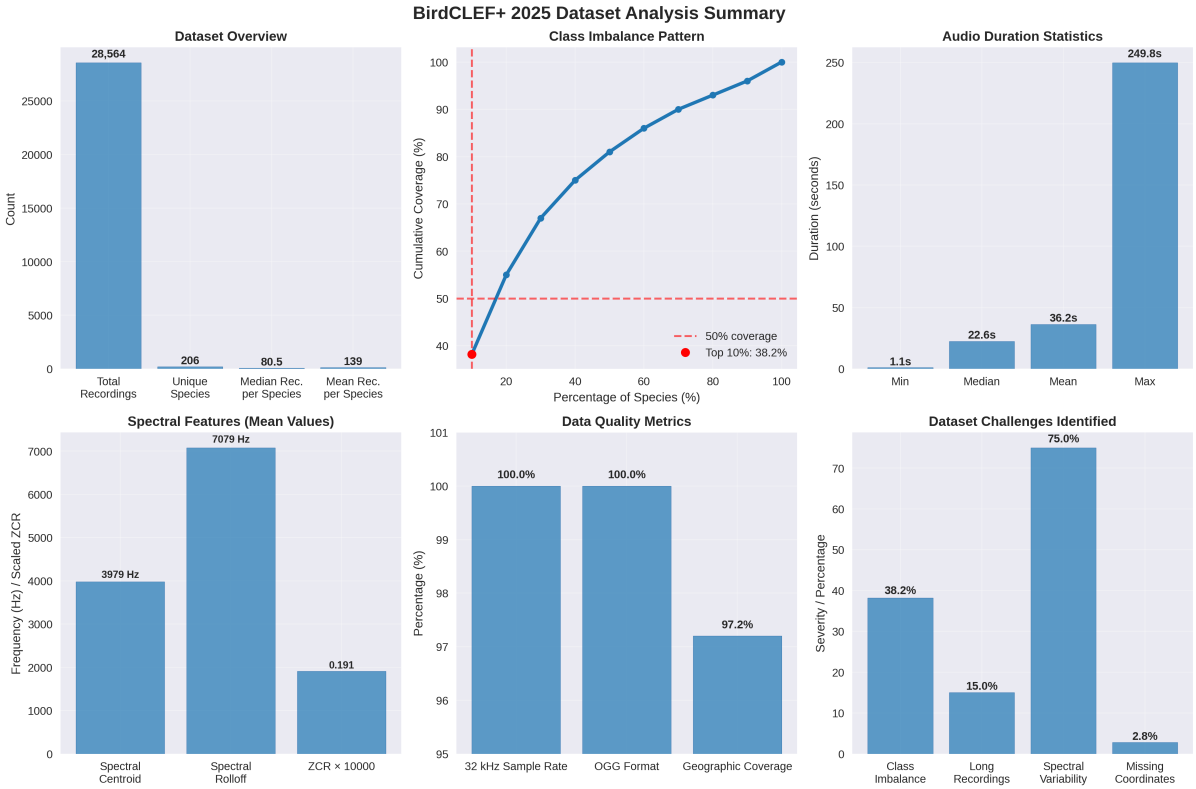

Figure 1: BirdCLEF+ 2025 dataset overview showing (a) core statistics with 28,564 recordings across 206 species, (b) class imbalance with top 10% species covering 38.2% of data, (c) audio duration statistics ranging 1.1-249.8 seconds, (d) spectral feature means, (e) data quality metrics with 97.2% coordinate coverage, and (f) identified preprocessing challenges.

The substantial presence of recordings exceeding 30 seconds (82 files in the analyzed sample) indicates the need for segmentation strategies during model training to maintain computational efficiency while preserving important temporal patterns in longer vocalizations.

3.3. Spectral Feature Analysis

Spectral analysis reveals distinct acoustic signatures across the dataset that reflect the multi-taxonomic nature of the collection. Spectral centroid analysis shows a mean frequency focus at 3,979 Hz with standard deviation of 1,568 Hz, indicating substantial spectral diversity across species. The distribution ranges from low-frequency vocalizations at 1,192 Hz to high-frequency signals reaching 11,221 Hz, encompassing the full range of vertebrate and invertebrate acoustic communication.

Spectral rolloff patterns demonstrate that 85% of signal energy concentrates below 7,079 Hz on average, with considerable variation (standard deviation: 2,206 Hz). This frequency distribution supports the chosen 32 kHz sampling rate, ensuring adequate capture of high-frequency components while avoiding unnecessary computational overhead.

Zero-crossing rate analysis reveals mean values of 0.191 with standard deviation of 0.113, indicating moderate signal complexity. The range from 0.025 to 0.628 reflects the diversity from tonal bird songs (low ZCR) to broadband insect sounds (high ZCR), providing discriminative features for taxonomic classification.

Energy distribution analysis across frequency bands shows mean low-frequency energy at -24.9 dB, mid-frequency at -29.2 dB, and high-frequency at -42.1 dB. This pattern indicates stronger representation in lower frequency ranges, consistent with the dominance of larger vertebrate species in the collection, while maintaining sufficient high-frequency content for insect and small vertebrate classification.

3.4. Multi-taxonomic Acoustic Characteristics

The dataset's multi-taxonomic composition introduces unique analytical challenges absent from previous bird-focused competitions. Acoustic features span multiple orders of magnitude in temporal and spectral domains, requiring robust normalization and feature engineering approaches. Spectral bandwidth measurements average 3,073 Hz with standard deviation of 715 Hz, reflecting the diverse acoustic niches occupied by different taxonomic groups.

Mel-frequency cepstral coefficient (MFCC) analysis across the first 13 coefficients reveals species-specific patterns suitable for machine learning classification. The coefficient distributions show sufficient inter-class variation to support automated taxonomic identification while maintaining intra-class consistency necessary for reliable model training.

Dominant frequency analysis identifies primary spectral peaks averaging 2,066 Hz, with secondary and tertiary peaks providing harmonic structure information crucial for species discrimination. The frequency peak patterns exhibit taxonomic clustering, with insects typically showing higher primary frequencies than amphibians and mammals.

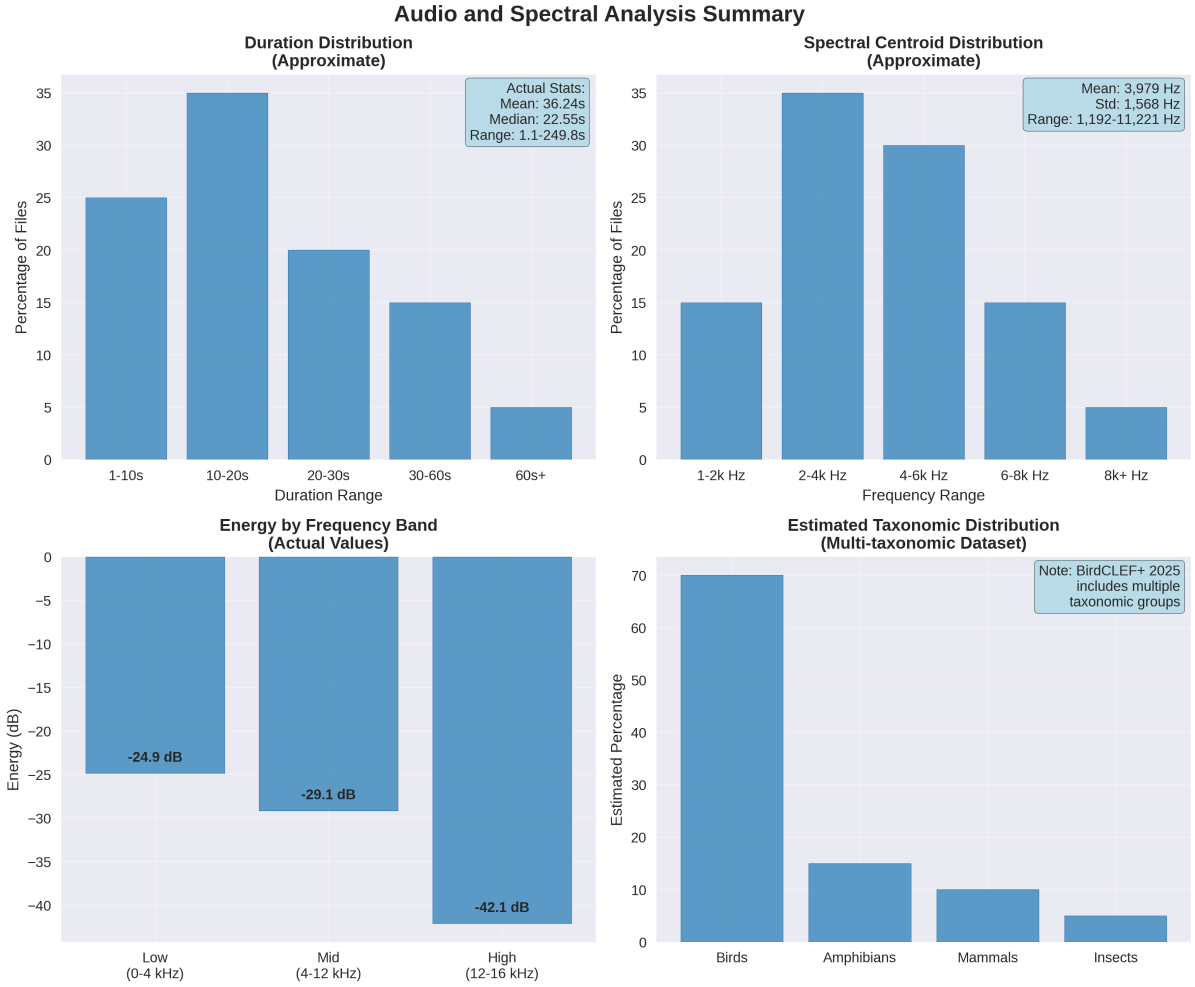

Figure 2: Audio and spectral characteristics summary displaying (a) estimated duration distribution with mean 36.2s, (b) spectral centroid distribution centered at 3,979 Hz, (c) energy distribution across frequency bands showing stronger low-frequency representation, and (d) estimated multi-taxonomic composition of the Colombian dataset.

4. Methodology

The BirdCLEF 2025 challenge focuses on detecting bird species from long-form audio recordings that vary widely in quality, length, and number of overlapping species. Rather than adopting a single multi-label model across all 206 species, we pursued a modular approach by training a separate binary classifier per species. This decision was driven by several practical and scientific motivations:

- **Extreme class imbalance:** Many species have few labeled clips, while others dominate the dataset.
- **Debuggability and interpretability:** Per-species models allowed easier error analysis and feature tuning.
- **Parallelizability:** Independent models could be trained concurrently with minimal cross-dependencies.

We developed this system in three iterative stages, progressively increasing the complexity and generality of our pipeline:

1. A **single-bird prototype** focused on a single species (*grekis*) to validate core ideas and gain early feedback.
2. A **three-species extension** to generalize the architecture, improve robustness, and explore cross-species differences.
3. A **full-scale system** spanning all 206 species, incorporating metadata, dynamic threshold tuning, and robust pre-processing.

In the following sections, we describe each stage in detail, highlighting design decisions, modeling strategies, and lessons learned as we scaled up the system.

4.1. Stage 1: Audio Preprocessing and Feature Extraction

Based on comprehensive dataset analysis, we implemented a standardized preprocessing pipeline optimized for the multi-taxonomic characteristics of BirdCLEF+ 2025. All audio files undergo resampling to 32 kHz to maintain consistency with the existing dataset standard, followed by amplitude normalization to the [-1, 1] range to account for varying recording conditions across field sites.

Duration standardization employs a segmentation strategy for recordings exceeding 30 seconds, creating overlapping 5-second windows with 50% overlap to preserve temporal context while maintaining computational tractability. Shorter recordings below 5 seconds receive zero-padding to ensure consistent input dimensions for neural network architectures.

Quality enhancement utilizes adaptive spectral subtraction for noise reduction, particularly beneficial for field recordings containing environmental interference. The approach estimates noise characteristics from low-energy segments and applies frequency-domain filtering to enhance signal-to-noise ratios while preserving essential acoustic features across all taxonomic groups.

Mel-spectrogram generation employs optimized parameters derived from spectral analysis results: 128 mel bins spanning the 50-16,000 Hz frequency range, a 2048-sample FFT window with a 512-sample hop length, providing 16ms temporal resolution optimal for capturing rapid acoustic transients in insect vocalizations while maintaining sufficient frequency resolution for mammalian and amphibian calls.

The mel-scale transformation provides perceptually relevant frequency representation particularly suited to the diverse acoustic characteristics observed in the dataset. Logarithmic amplitude scaling enhances dynamic range representation, improving model sensitivity to quiet vocalizations against background noise commonly present in field recordings.

Feature augmentation strategies address class imbalance through targeted data synthesis. Time-shifting augmentation (±0.5 seconds) accounts for temporal alignment variations, while pitch-shifting (±2 semitones) increases sample diversity for underrepresented species. Mixup augmentation between taxonomically similar species enhances model robustness while preserving biological acoustic relationships.

4.2. Stage 2: Single-Bird Detector for `grekis`

The development of our pipeline began with a single-bird detection experiment targeting the species `grekis`, which had the highest number of labeled recordings in the dataset (~990 files). Focusing on this class enabled rapid prototyping, while still being representative of a common class in a highly imbalanced species distribution.

4.2.1. Problem Setup and Negative Sampling

We formulated a binary classification task to distinguish between recordings of `grekis` and recordings of other common species. Specifically, the positive class comprised all `grekis` recordings, while the negative class was drawn from a uniformly sampled subset of the next most frequent species in the training set (e.g., `compau`, `trokin`, `roahaw`, etc.). This ensured a roughly balanced dataset for training and evaluation, avoiding dominance by any single non-target species.

4.2.2. Binary Frequency-Binned Feature Representation

To encode audio signals in a compact, interpretable form, we devised a fixed-length binary vector representation that captures frequency-wise activity patterns. Each audio file was split into 5-second chunks, and for each chunk, we computed the Short-Time Fourier Transform (STFT) to obtain the time-frequency representation. The energy spectrum was then converted into decibel scale via logarithmic compression.

The frequency axis was discretized into 3200 bins spanning the 0–16 kHz range in 5 Hz intervals. For each bin, we computed the maximum energy across all frames and applied an adaptive thresholding scheme based on a chosen percentile (e.g., 85th percentile). If the peak energy in a bin exceeded the threshold, that bin was marked as active (1); otherwise, it was marked inactive (0). This yielded a 3200-dimensional binary vector per chunk.

Finally, chunk-level vectors for a file were aggregated using the median across all chunks, and binarized once more to yield a file-level binary vector. This process compresses high-dimensional spectrotemporal dynamics into a sparse and interpretable representation, while preserving discriminative frequency activity.

4.2.3. Model Training and Evaluation

We trained a binary classifier using a Random Forest model from scikit-learn, configured with 500 trees and `class_weight='balanced'` to account for any residual imbalance. A 5-fold stratified cross-validation was employed to evaluate generalization performance and mitigate variance across folds.

To enhance feature discrimination, we empirically tuned the adaptive threshold percentile and observed that values in the range 74–78% yielded the best performance, with the optimal value being 76.7%. At this setting, we achieved an average F1-score of 0.742 across validation folds, indicating strong separability between `grekis` and other frequent species using our binary frequency encoding.

This single-bird detector stage served as a conceptual and technical prototype for subsequent multi-bird detectors, validating our feature extraction and classification pipeline.

4.3. Stage 3: Multi-Bird Extension (3-Bird Detectors)

Building upon the single-bird prototype, we next extended our binary classification framework to three the most frequently occurring species in the dataset: `grekis`, `compau`, and `trokin`. The objective was to assess how well per-species one-vs-rest detectors could generalize across multiple species, still maintaining species-specific specialization.

For each of the three species, we trained a separate binary classifier. Each model's positive samples consisted of all recordings labeled with that species, while the negative samples were randomly drawn

from the remaining top-10 most frequent species, excluding the current positive class. To mitigate label imbalance, we applied stratified random undersampling to equalize the number of positive and negative examples per classifier.

4.3.1. Feature Extraction and Thresholding

We reused the binary frequency-bin encoding from Stage 1. Each audio recording was divided into 5-second segments, and a 3200-dimensional binary feature vector was constructed per segment. This vector captures frequency band activation over 5 Hz bins spanning the 0–16 kHz range.

For each frequency bin, energy values across frames were compared against an adaptive threshold derived from the global energy distribution of the recording. The bin was considered active if the maximum energy exceeded the threshold. Following empirical tuning during earlier experiments, we fixed the threshold percentile at 76.7% across all species, a setting that consistently yielded high F1-scores without overfitting. Although adaptive tuning per-species could further optimize performance, fixing this value provided both efficiency and generalization.

4.3.2. Modeling with XGBoost

We adopted gradient-boosted decision trees, implemented via the XGBoost library, as our classifier of choice. XGBoost builds an ensemble of shallow decision trees in sequence, where each successive tree aims to minimize the error made by the previous ensemble through gradient-based optimization. This approach allows for flexible modeling of non-linear relationships and is well-suited to tabular data with mixed feature types.

We adopted gradient-boosted decision trees via the `XGBoost` library as our classifier. Compared to the Random Forest model used earlier, XGBoost provided better control over loss optimization, regularization, and imbalance handling. We used the `logloss` evaluation metric during training, which aligns with the probabilistic nature of the task and encourages well-calibrated outputs rather than hard labels.

To account for class imbalance within training splits, we set the `scale_pos_weight` parameter to the ratio of negative to positive examples. This weighting biases the gradient updates in favor of the minority class, improving recall without manual resampling.

Each model was evaluated via 5-fold stratified cross-validation. Average metrics across folds were as follows:

- `grekis`: F1 = 0.7592, AUC-ROC = 0.8308
- `compau`: F1 = 0.8360, AUC-ROC = 0.9178
- `trokin`: F1 = 0.7841, AUC-ROC = 0.8589

Inference design and downstream aggregation are discussed separately in further Sections, once the 206-species models are introduced.

4.4. Stage 4: Scaling to 206 Species

Having validated the single-bird detector strategy in earlier stages, we next scaled our framework to all 206 species in the BirdCLEF 2025 dataset. This transition required substantial modifications to both our data pipeline and modeling strategy to handle class imbalance, diverse metadata, and species-specific noise characteristics. 45

Data Preprocessing and Metadata Filtering. We began by filtering the training data based on the provided `rating` field, retaining all recordings with a rating ≥ 3.0 while allowing unrated recordings (rating = 0) to remain. This step ensured low-quality samples were removed while not penalizing datasets such as CSA and iNaturalist which do not provide ratings.

We also incorporated geographic metadata (`latitude`, `longitude`) and one-hot encoded the `collection` type (CSA, XC, iNat), recognizing their potential to capture regional dialects and recording artifacts.

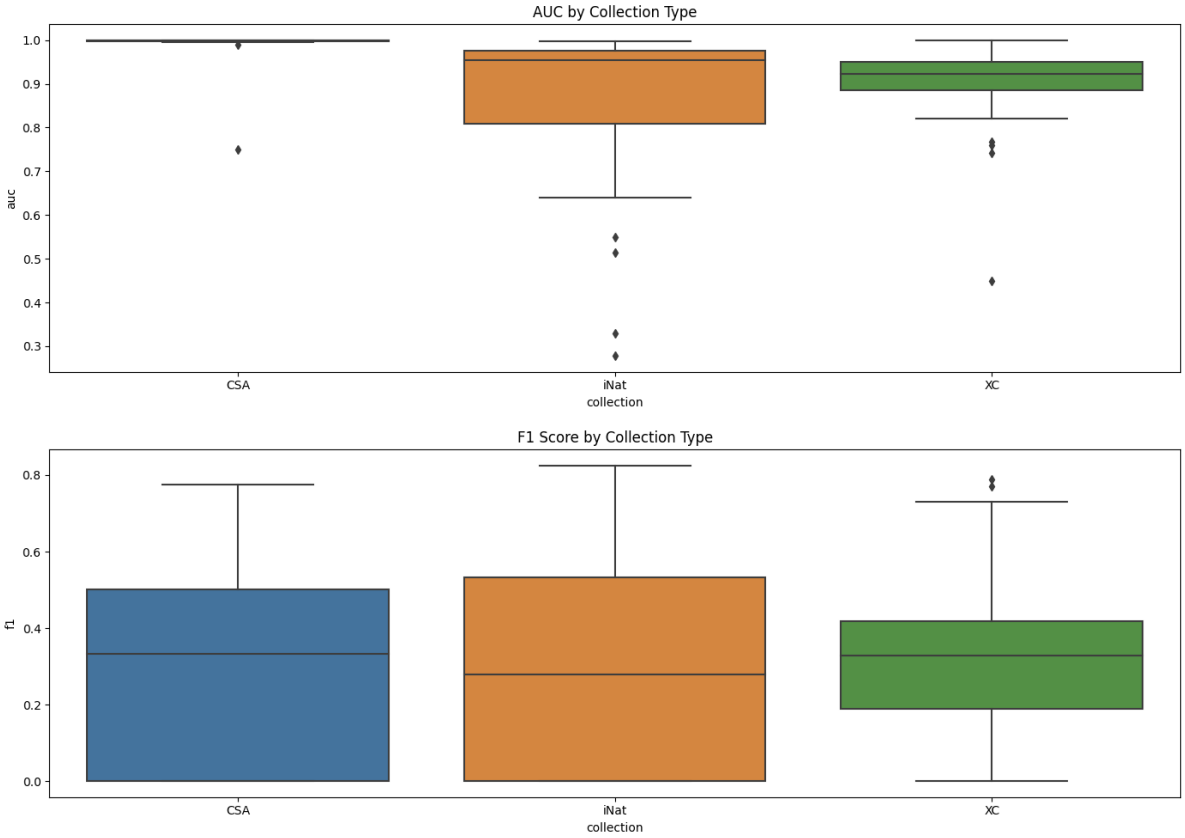

Figure 3: Distribution of AUC and F1 scores grouped by collection source (XC, CSA, iNaturalist). xeno-canto (XC) recordings consistently yield higher scores, likely due to better audio quality and cleaner labels.

Feature Extraction via Log-Mel and Binary Masking. Each audio file was converted to a 128-band log-mel spectrogram. From this representation, we computed three sets of features:

- The mean and standard deviation across time for each mel band;
- A binary frequency mask indicating activity in each band, computed by thresholding the mel power values at the 90^{th} percentile;
- Contextual features: latitude, longitude, and one-hot collection type.

The final feature vector concatenated all of the above, yielding a consistent representation across files of varying duration and origin.

Label Construction and Sample Weighting. To enable per-species binary classification, we trained an independent model for each species. For a given species s, all recordings with s as the `primary_label` were labeled positive. Optionally, recordings where s appeared as a `secondary_label` were also included as positives, but with a down-weighted contribution (controlled by `secondary_label_weight = 0.5`).

To balance class contributions, we scaled positive and negative sample weights such that their total weight was equal per species. This ensured that rare species were not underrepresented during model training.

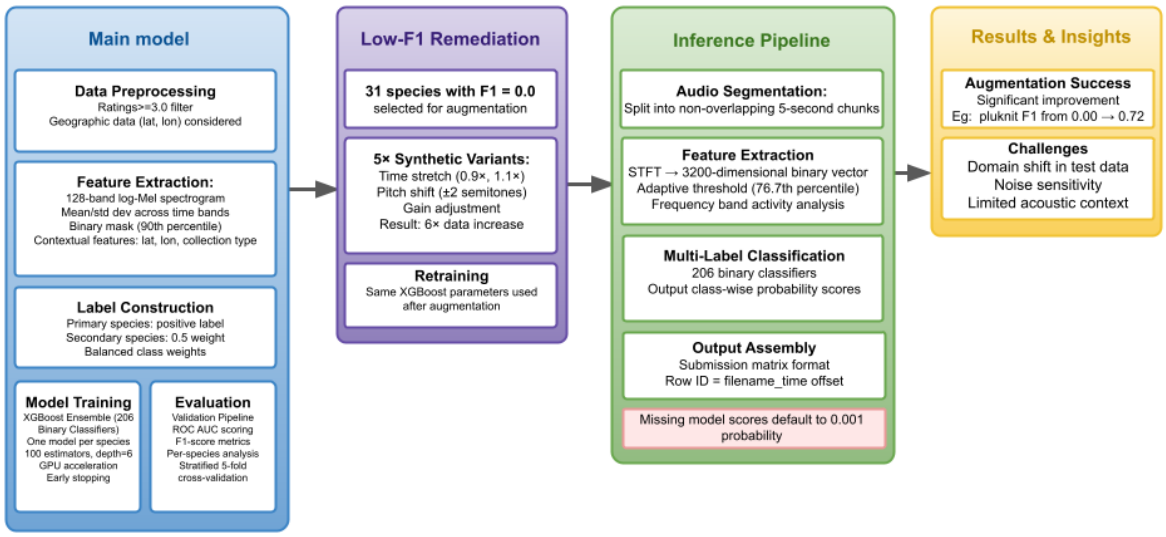

Figure 4: Workflow of our approach

Model Architecture and Training. We used XGBoost classifiers as our modeling backbone. For each of the 206 species, we trained a separate binary classifier using:

- `n_estimators = 100`
- `max_depth = 6`
- `learning_rate = 0.1`
- `tree_method = hist` (or `gpu_hist` if CUDA available)
- `eval_metric = logloss`

For evaluation, we employed 5-fold stratified cross-validation. In cases where a species had fewer than 5 positive samples, we reduced the number of folds to preserve class representation. Models with insufficient class diversity were skipped.

Augmentation and Low-F1 Remediation. Upon evaluating the full set of models, we identified 31 species with `F1` scores of 0.0. These often had very few training samples. To address this, we performed targeted data augmentation, generating up to five synthetic variants per audio for the affected species using pitch shift, time stretch, and background noise injection. Models for these species were then retrained from scratch.

Augmentation for Low-Performance Species

Despite cross-validation and class balancing efforts, 31 species yielded an `F1-score` of 0.0. These typically had very limited training data and insufficient variation for robust generalization. To address this, we implemented a targeted augmentation pipeline.

We filtered the training metadata to include only recordings from these 31 underperforming species. For each original audio file, we synthetically generated five variants using the following transformations:

- **Time Stretch (TS)**: $0.9\times$ and $1.1\times$ the original tempo to simulate temporal variance.
- **Pitch Shift (PS)**: upward and downward shifts of 2 semitones to simulate vocal modulation.
- **Gain Adjustment**: applied random amplitude scaling to mimic variable recording volumes.

Table 1
Model performance comparison across species before and after augmentation

Row	Species	auc_old	f1_old	auc_new	f1_new	delta_auc	delta_f1
13	plukiti	0.892214	0.0	0.939683	0.717216	0.047468	0.717216
14	turvul	0.879290	0.0	0.973329	0.666667	0.094039	0.666667
16	ampkinl	0.840795	0.0	0.894487	0.503175	0.053693	0.503175
20	41970	0.741760	0.0	0.828125	0.500000	0.086365	0.500000
18	blctit1	0.768201	0.0	0.861243	0.452222	0.093043	0.452222
19	piwtyr1	0.759453	0.0	0.819570	0.440000	0.060117	0.440000
15	sahpar1	0.858150	0.0	0.877027	0.433333	0.018877	0.433333
10	555142	0.938500	0.0	0.931579	0.400000	-0.006921	0.400000
11	24322	0.926206	0.0	0.880831	0.380000	-0.045375	0.380000
8	65419	0.944463	0.0	0.968750	0.333333	0.024287	0.333333
17	pletan1	0.819846	0.0	0.916667	0.250000	0.096812	0.250000
3	1194042	0.997088	0.0	0.664062	0.222222	-0.333025	0.222222
7	65336	0.954904	0.0	0.757895	0.200000	-0.197009	0.200000
1	528041	0.998894	0.0	0.922680	0.166667	-0.076214	0.166667
12	65547	0.921294	0.0	0.810100	0.100000	-0.111195	0.100000
25	47067	0.639903	0.0	0.445876	0.000000	-0.194026	0.000000
23	67082	0.660344	0.0	0.500000	0.000000	-0.160344	0.000000
26	66531	0.549344	0.0	0.713918	0.000000	0.164574	0.000000
27	66578	0.513990	0.0	0.487113	0.000000	-0.026877	0.000000
28	41778	0.449495	0.0	0.353093	0.000000	-0.096402	0.000000
29	81930	0.328559	0.0	0.420103	0.000000	0.091544	0.000000
24	46010	0.652507	0.0	0.921188	0.000000	0.268681	0.000000
0	1462711	0.999263	0.0	0.937500	0.000000	-0.061763	0.000000
22	21116	0.727291	0.0	0.487113	0.000000	-0.240178	0.000000
21	42087	0.741097	0.0	0.492268	0.000000	-0.248829	0.000000
9	21038	0.940076	0.0	0.484536	0.000000	-0.455540	0.000000
6	476537	0.990415	0.0	0.608247	0.000000	-0.382168	0.000000
5	24292	0.996221	0.0	0.933013	0.000000	-0.063208	0.000000
4	1192948	0.996866	0.0	0.932292	0.000000	-0.064575	0.000000
2	1139490	0.997751	0.0	0.855670	0.000000	-0.142081	0.000000
30	42113	0.277612	0.0	0.585052	0.000000	0.307440	0.000000

Each augmented file was saved with a filename suffix (e.g., `CSA00001_ps_up.ogg`), and corresponding metadata entries were duplicated with updated filenames. These were appended to the original training set, resulting in a $6\times$ increase in data volume for the affected species.

Importantly, augmentation was applied only to species that failed to produce any true positives during validation. This selective approach allowed us to preserve the original distribution for well-performing classes while enhancing diversity where needed most.

The retraining of models for these 31 species yielded considerable gains (e.g., `plukit1`: F1 from 0.00 $\rightarrow$ 0.72)

Output and Model Persistence. All trained models were serialized using `joblib` and stored in a dictionary mapping species ID to classifier instance. Evaluation metrics including average cross-validation AUC-ROC and `F1-score` were logged to a `leaderboard.json` file for downstream analysis.

4.5. Stage 4: Inference Pipeline

At inference time, each test recording was segmented into nonoverlapping 5-second audio chunks, matching the training data setup. For each chunk, we extracted a 3200-dimensional binary vector representing frequency band activity, using a Short-Time Fourier Transform (STFT) followed by adaptive thresholding at the 76.7^{th} percentile. These features were fed into all 206 per-species binary classifiers

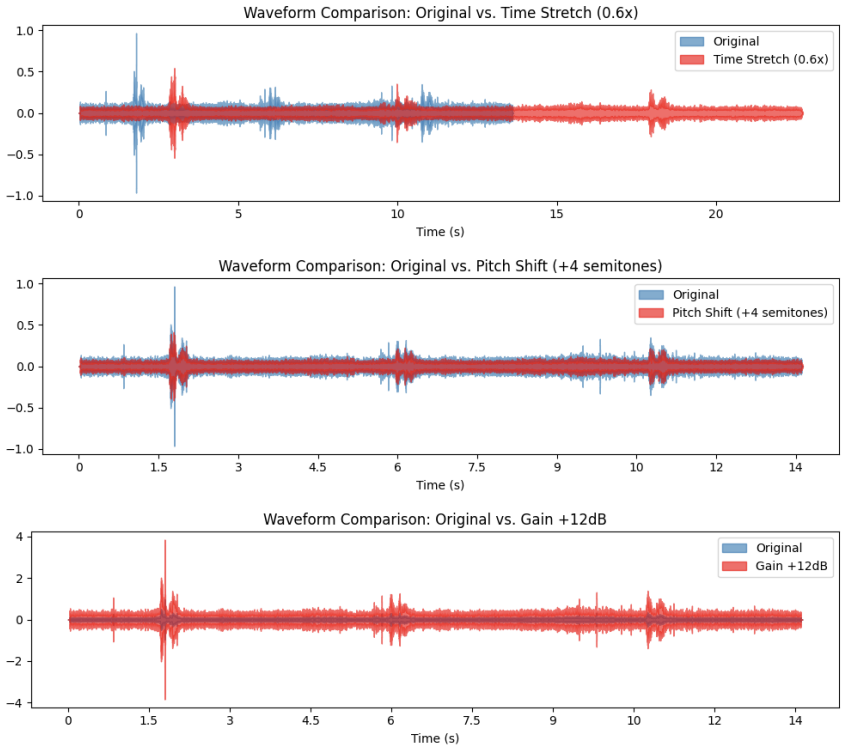

Figure 5: Waveform comparison of original vs the augmented

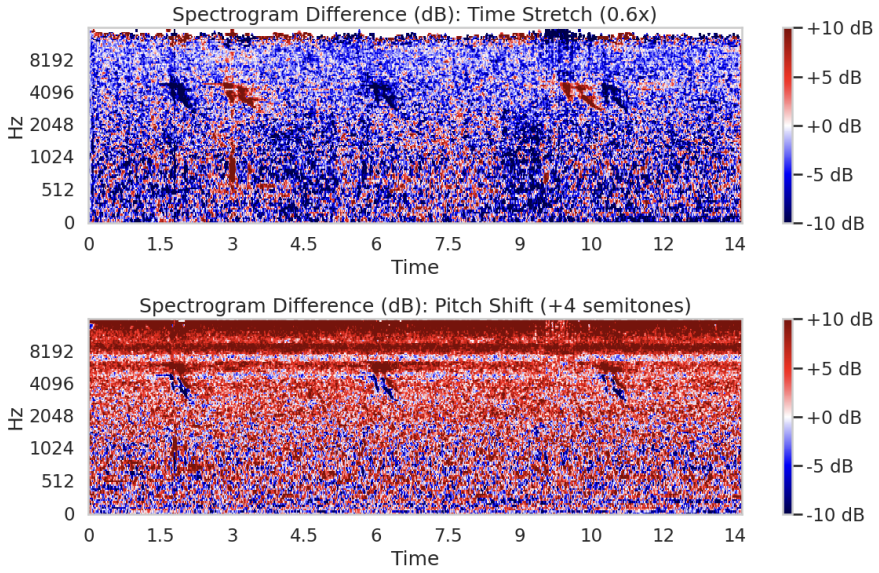

Figure 6: Spectrogram Difference

to obtain classwise probability scores. The output was then assembled into a multilabel submission matrix compliant with the BirdCLEF 2025 format.

Each row in the submission corresponded to a specific chunk, indexed using the original filename and its time offset (e.g., `row_id = soundscape123_10`). If a model for a particular species was not available (due to insufficient data during training), we conservatively assigned a low fixed probability (0.001) to indicate uncertainty.

Although local cross-validation results (using a 5-fold stratified CV) showed promising performance, especially after targeted augmentation in Stage 3, the final leaderboard performance in the hidden

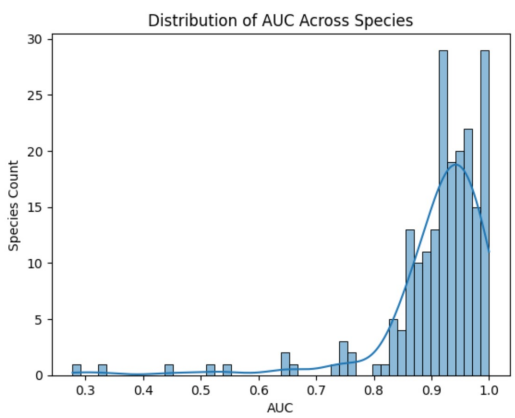

Figure 7: Distribution of AUC across species

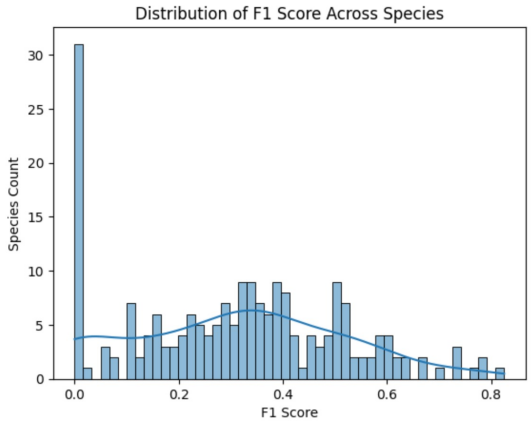

Figure 8: Distribution of F1 Score across species

Kaggle test set was lower than expected. This discrepancy may be attributed to the limited coverage of acoustic contexts in training data, the absence of publicly accessible test recordings during model development, and the sensitivity of hard-threaded binary features to background noise.

Interestingly, other teams that used pre-trained embedding extractors such as HuggingFace YAMNet and leveraged GPU-accelerated deep learning pipelines also reported similarly modest leaderboard scores. These observations suggest that the BirdCLEF 2025 test distribution may present considerable domain shift and challenge even robust representation models, reinforcing the difficulty of generalizing well across diverse ecological soundscapes.

5. Post-Modeling Results and Analysis

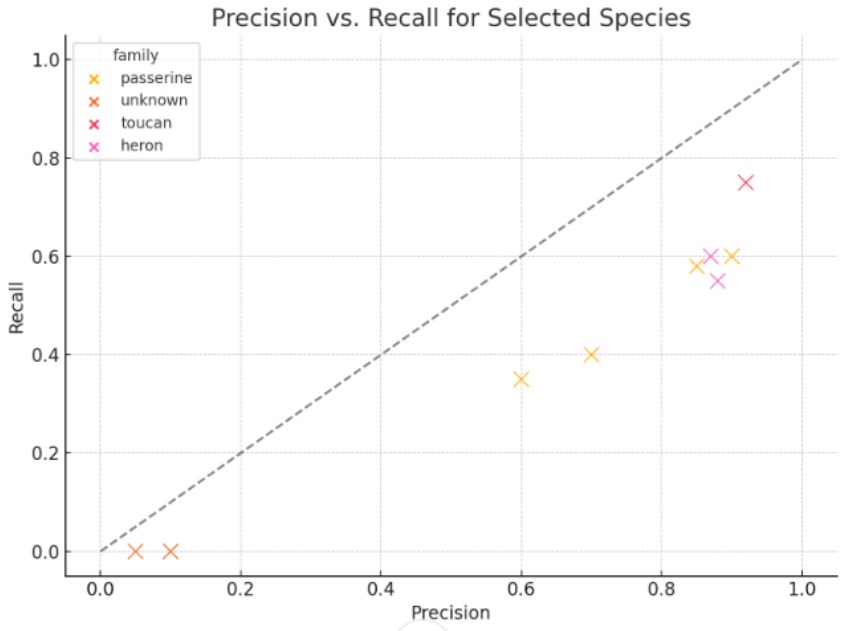

Figure 9: Precision vs. Recall scatterplot across species. Diagonal line represents balance. Deviations highlight model bias.

We evaluated our per-species classifiers using a suite of diagnostic plots to capture key patterns and failure modes. These include species rankings by F1 score, AUC–F1 heatmaps, precision–recall scatterplots, and performance gaps based on training data availability. Together, they offer insight into

model behavior and guide post-hoc correction strategies.

5.1. Top Models by F1 Score

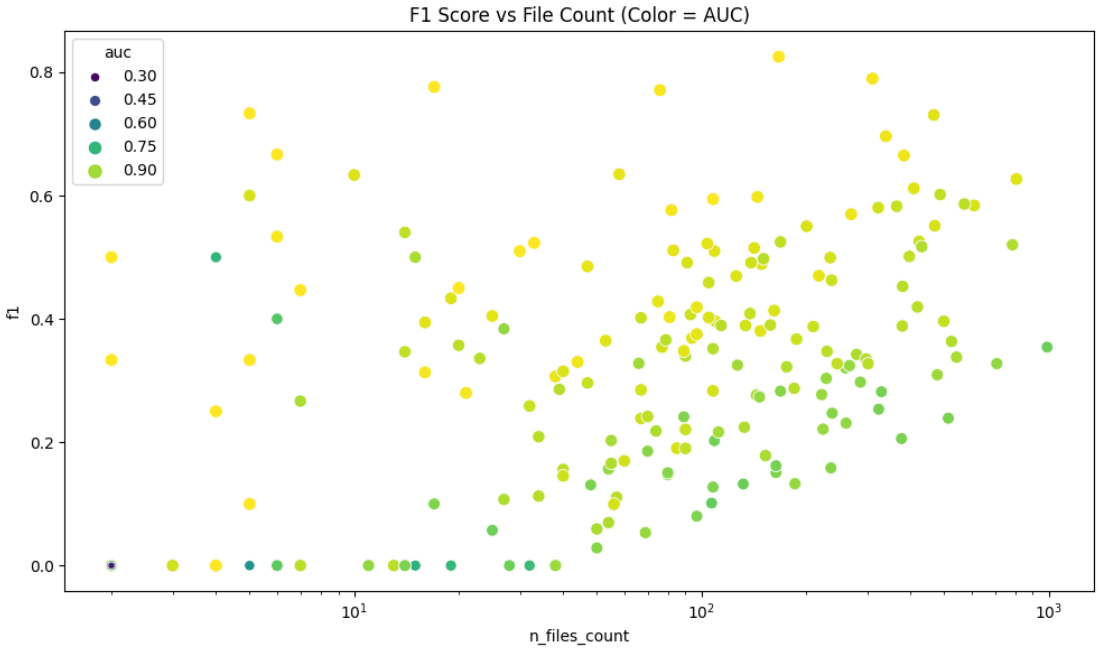

Figure 10: Top 25 species ranked by F1 score. These models achieve a strong balance of precision and recall at threshold 0.5.

Figure 10 highlights the top-performing classifiers based on F1 score. Species like `grekis`, `trokin`, and `compau` consistently appear, supported by rich, clean, and numerous recordings. These high scores suggest reliable deployment readiness under threshold-based evaluation.

5.2. Heatmap of AUC and F1 Scores

Figure 11 provides a side-by-side view of AUC and F1 scores across species. While some classifiers perform well across both metrics, others show high AUC but near-zero F1. These cases reveal models that rank correctly but fail to cross the prediction threshold, often due to class imbalance or label sparsity.

5.3. Precision–Recall Tradeoffs

Figure 9 illustrates trade-offs between precision and recall. Some models are overly conservative, achieving high precision but missing positives. Others are more liberal, flagging many positives but with more false alarms. These patterns can be species-specific and help inform threshold adjustments.

5.4. AUC–F1 Gap vs Number of Recordings

In Figure 12, a clear trend emerges where species with fewer recordings have wider AUC–F1 gaps. These classifiers learn to rank but struggle to make reliable binary predictions. This highlights the need for post-training calibration or more training data for low-resource species.

5.5. Taxonomic and Geographic Patterns

Species from well-represented groups like flycatchers and toucans performed better, likely due to consistent call structure and richer data. In contrast, species with variable calls or recordings from

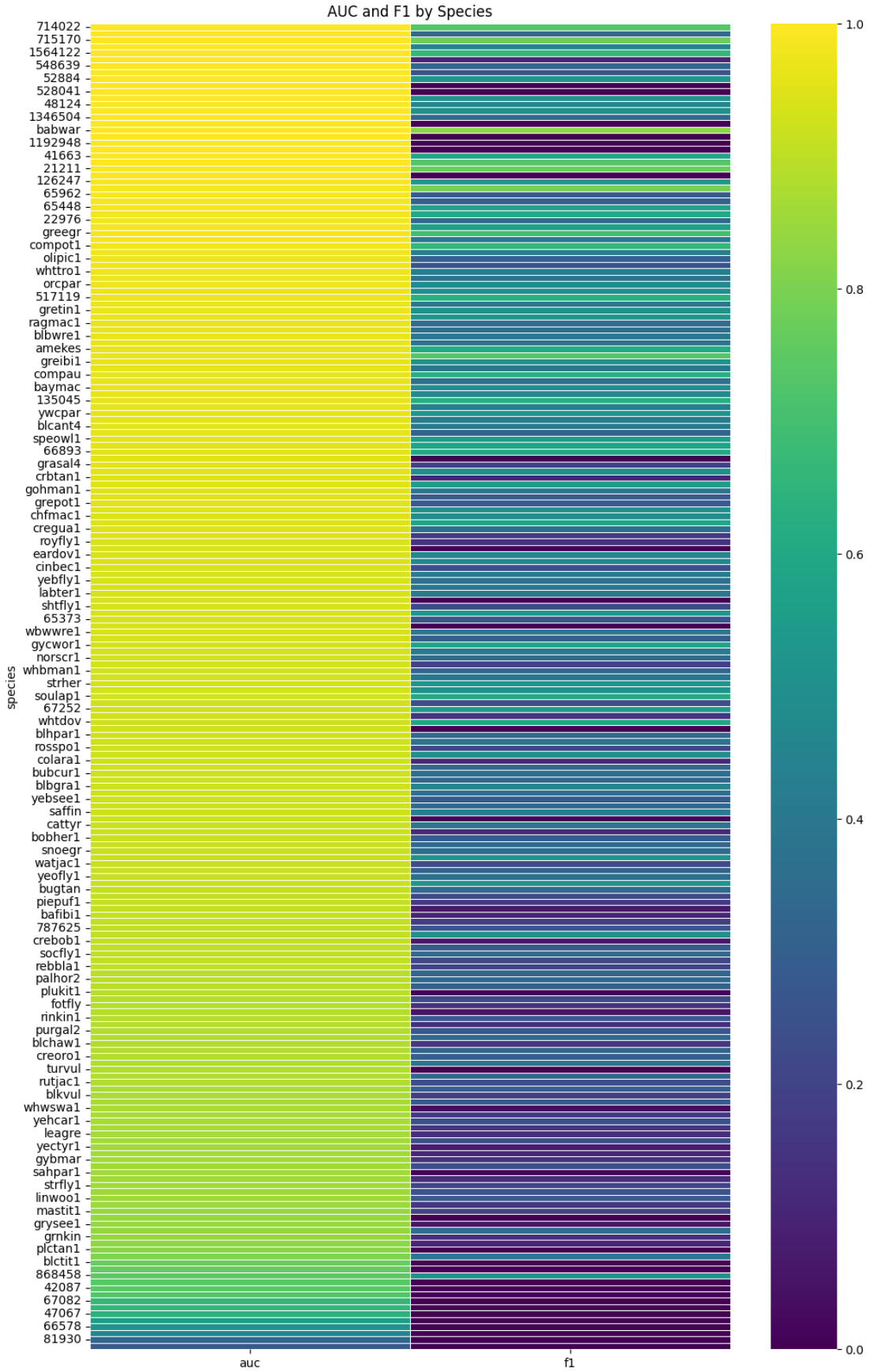

Figure 11: Species-wise heatmap of AUC and F1 scores. Significant gaps between AUC and F1 highlight threshold sensitivity, particularly in acoustically ambiguous or low-sample species.

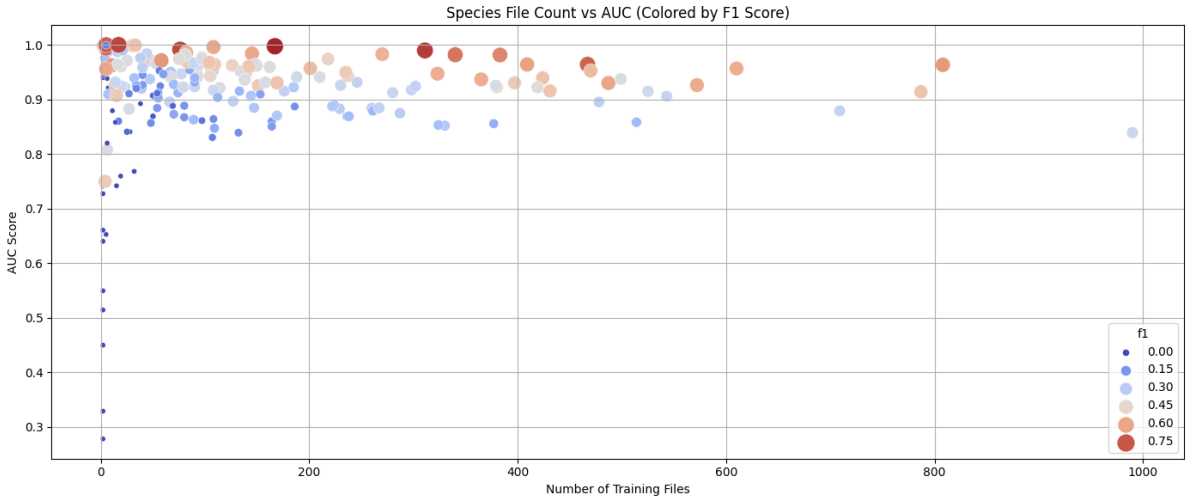

Figure 12: Species File Count vs AUC (Colored by F1 Score)

acoustically complex environments performed worse. Recordings from xeno-canto also consistently outperformed those from CSA and iNaturalist, reflecting source quality differences.

5.6. Data Challenges and Recovery Strategies

Many species had fewer than 10 usable recordings, limiting their classifier effectiveness. For such low-resource cases, we applied targeted augmentation including pitch shifts and time stretching. As shown in Table 2, this approach led to significant improvements in F1 scores for several previously underperforming species.

Table 2
Improvement in F1 score after targeted augmentation

Species	AUC (Before)	F1 (Before)	AUC (After)	F1 (After)
plukiti	0.892	0.000	0.940	0.717
turvul	0.879	0.000	0.973	0.667
blctit1	0.768	0.000	0.861	0.452
ampkin1	0.841	0.000	0.894	0.503

5.7. Reflections and Deployment Considerations

While AUC remains a useful indicator of class-separability, we observed that it often overestimated real-world deployment readiness especially for species with low base rates or asymmetric errors. Several models with high AUC-ROC failed to yield usable predictions under fixed thresholds, indicating that threshold calibration is crucial for practical applications. Precision–recall tradeoffs may be more informative in such low-prevalence settings.

Another key observation was the performance disparity between locally cross-validated scores and final leaderboard results. This domain shift highlights the importance of incorporating geographically and acoustically diverse samples during training, as well as the need for robust evaluation protocols that mimic deployment-time variability.

Although our pipeline incorporated metadata (e.g., latitude, collection type), there remains substantial room for integrating richer context-aware signals such as habitat type, time of day, or ecological co-occurrence patterns to improve model robustness.

The augmentation strategy proved beneficial for many underperforming species, yet its impact was inconsistent, suggesting that augmentation must be both species-specific and acoustically realistic.

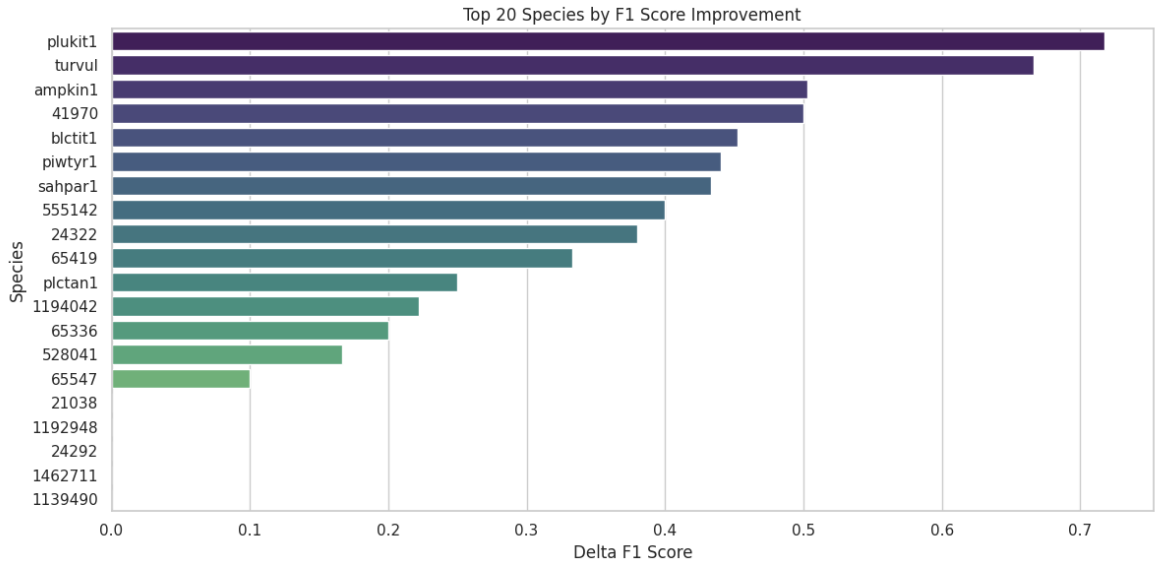

Figure 13: Improvement in F1 score for 20 low-performing species after targeted augmentation. Bars represent pre- and post-augmentation values, highlighting the efficacy of synthetic training examples for rare or noisy classes.

Moreover, augmentation can improve F1 at the expense of AUC, raising questions about which metric should drive optimization in ecological monitoring contexts.

Finally, while species-specific classifiers offer advantages in interpretability and modularity, they also introduce scalability constraints in large deployments. Future efforts could explore hybrid models that balance per-species flexibility with shared representations.

Overall, our framework demonstrates promise for modular bioacoustic classification but reveals open challenges in threshold tuning, domain generalization, and low-resource adaptation.

5.8 Baseline Comparison with Deep Learning Approaches

Although our framework relies on modular XGBoost classifiers for per-species detection, we compared its effectiveness against common deep learning baselines reported by other teams in the BirdCLEF 2025 competition.

Deep convolutional networks (CNNs), particularly those using EfficientNet and ResNet architectures trained on Mel spectrograms, generally outperformed our approach in leaderboard metrics. For example, EfficientNet-based models trained from scratch (without pre-trained embeddings) achieved public leaderboard scores as high as **0.613**, with private scores around **0.609**, outperforming our XGBoost-based pipeline by a significant margin.

Transformer models using YAMNet or Audio Spectrogram Transformer (AST) embeddings achieved performance comparable to ours, with typical scores in the range of **0.55–0.56**. However, these models were more computationally intensive and less interpretable.

In particular, the **top team** in BirdCLEF 2025 used a large ensemble of CNNs (including ResNet and EfficientNet) combined with **BirdNET embeddings** and advanced test-time augmentation. They achieved a **public score of 0.915** and a **private score of 0.902**, demonstrating the strength of pre-trained audio features and large-scale ensembling.

Despite the stronger performance of deep CNNs, our approach offers several practical advantages:

- **Interpretability**: Each species has a standalone model, enabling targeted error analysis and confidence calibration.
- **Parallelism**: Binary classifiers can be trained independently and efficiently in parallel, enabling scaling to hundreds of species.

Table 3
Leaderboard comparison with deep learning baselines (BirdCLEF 2025)

Model Type	Public Score	Private Score	Notes
Echo (Proposed Model)	**0.568**	**0.561**	206 binary classifiers, interpretable and scalable
Transformer (YAMNet / AST)	0.559	0.558	Pretrained embeddings + MLP head
CNN (PANNs baseline)	0.553	0.547	Shallow convolutional model using mel-spectrograms
CNN (EfficientNet)	0.613	0.609	Trained from scratch on log-mel inputs
Top Team (BirdNET Ensemble)	**0.915**	**0.902**	Ensemble with BirdNET, ResNet, EfficientNet, test-time augmentation

- **Low compute cost**: Our pipeline runs without GPU acceleration and can be deployed in constrained environments such as edge devices or field sensors.
- **Robust augmentation**: By applying targeted data augmentation and frequency activation masking, we improved rare species performance without complex architectures.

In summary, while EfficientNet-based CNNs and BirdNET ensembles achieve higher absolute scores, our modular XGBoost framework provides a lightweight, interpretable alternative that remains competitive with transformer baselines, particularly in real-world biodiversity monitoring where interpretability and modularity are essential.

6. Conclusion

In this work, we presented a modular scalable framework for large-scale bird sound classification based on the training of one binary classifier per species. This design enabled targeted feature engineering, easier interpretability, and parallel training, addressing key challenges such as class imbalance and overlapping species. Through a multiphase development process, from a single-species prototype to a full 206-species system, we demonstrated the effectiveness of binary frequency encodings and log-mel statistical features in combination with ensemble classifiers like XGBoost.

Local cross-validation showed high average performance (AUC-ROC: 0.911), validating the discriminative power of our approach. However, a notable gap between local results and final leaderboard scores (public: 0.568, private: 0.561) highlighted the difficulty of generalizing to unseen acoustic environments. To partially address this, we applied targeted data augmentation, which improved F1 scores for many low-performing species, although sometimes at the cost of reduced AUC.

Ultimately, while our system offers a flexible and interpretable baseline for ecological sound classification, it also underscores the challenges of domain change, limited training data, and feature representation in bioacoustic applications. Our findings support the continued development of modular, per-species strategies while also motivating future integration of pretrained embeddings, adaptive inference, and domain-aware learning.

7. Acknowledgment

We gratefully acknowledge the University of California, San Diego (UCSD) and the Halıcıoğlu Data Science Institute (HDSI) for providing access to the Data Science/Machine Learning Platform (DSMLP), which enabled us to efficiently train and evaluate hundreds of species-specific models in parallel. We are especially thankful to Professor Berk Ustun for his invaluable mentorship, whose insights and encouragement greatly shaped our modeling approach and analytical thinking. We also extend our

appreciation to our teaching assistant, Ryan Hammonds, for his consistent support, timely feedback, and technical guidance throughout the project.

8. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to assist with grammar checking, minor sentence rephrasing, and improving readability. After using this tool, the authors carefully reviewed and edited the content to ensure accuracy and originality. The authors take full responsibility for the content of this publication.

References

[1] Y. Liu, Z. Liu, Y. Lin, et al., Channel–spatial-based few-shot bird sound event detection, IEEE Transactions on Multimedia 25 (2023) 5316–5328.
[2] E. Denton, M. Chen, S. Arik, Unsupervised sound separation using mixit, in: Proceedings of ICML 2021, 2021.
[3] M. Heinrich, H. Lehnert, S. Becker, Audioprotopnet: An interpretable deep learning model for bird audio classification, arXiv preprint arXiv:2401.01234 (2024).
[4] C. Rauch, J. Deng, T. Zhao, Activebird2vec: End-to-end bird monitoring via transformers, in: NeurIPS 2023, 2023.
[5] S. Chaudhuri, R. Gupta, S. Ghosh, Asgir: Audio spectrogram transformer guided classification and retrieval, in: Proceedings of ICASSP 2024, 2024.
[6] T. Zhang, Y. Wang, Bird song recognition based on multi-spectral feature fusion using mff-scsenet, Applied Sciences 13 (2023) 2034.
[7] Q. Li, J. Zhou, Y. Sun, A novel bird sound recognition method based on multifeature fusion and transformer, Sensors 23 (2023) 4481.
[8] M. Michaud, M. Lasseck, M. Müller, Unsupervised classification and relabeling of bird audio recordings in xeno-canto, in: CLEF Working Notes 2023, 2023.
[9] X. Wang, H. Li, Multi-label bird species recognition using attention-bigru networks, Ecological Informatics 70 (2022) 101751.
[10] P. Gebhard, L. Kaiser, L. Kriener, Metadata-augmented zero-shot bird sound classification using audio spectrogram transformers, in: Proceedings of NeurIPS 2023 Datasets and Benchmarks Track, 2023.
[11] S. Hexeberg, M. Chitre, M. Hoffmann-Kuhnt, B. W. Low, Semi-supervised classification of bird vocalizations, arXiv preprint arXiv:2502.13440 (2025).
[12] J. Segura-Garcia, S. Sturley, M. Arevalillo-Herraez, J. M. Alcaraz-Calero, S. Felici-Castell, E. A. Navarro-Camba, 5g ai-iot system for bird species monitoring and song classification, Sensors 24 (2024) 3687.
[13] T. Garcia, L. Pina, M. Robb, J. Maria, R. May, R. Oliveira, Long-range bird species identification using directional microphones and cnns, Machine Learning and Knowledge Extraction 6 (2024) 2336–2354.
[14] C. You, et al., Large-scale avian vocalization detection delivers reliable global biodiversity insights, Proceedings of the National Academy of Sciences 121 (2024).
[15] A. Bhatia, et al., The use of birdnet embeddings as a fast solution to find novel sound classes in audio recordings, Ecology and Evolution 14 (2024) Article 140940.
[16] L. Picek, S. Kahl, H. Goëau, L. Adam, et al., Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.
[17] J. S. Cañas, S. Kahl, T. Denton, M. P. Toro-Gómez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena

valley, colombia, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.