DSC 291 - BirdWatchers Report

Chirag Agarwal, Varadraj Bartakke, Abhilash Shankarampeta,
Hemanth Bodala, Vignesh Palaniappan
Halıcıoglu Data Science Institute
University of California, San Diego

ABSTRACT

This report documents our end-to-end pipeline for tackling this ecoacoustic classification task, from initial exploration and preprocessing to model experimentation and final submission. We present a detailed analysis of the dataset, highlight key patterns observed through exploratory data analysis (EDA), and discuss both the domain-specific challenges and the technical hurdles faced during development. Our approach leverages advanced audio preprocessing techniques and compares the performance of multiple deep learning architectures, including ConvNeXt, PANNs, and Audio Spectrogram Transformers (AST). Our best-performing model, a **ConvNeXt-based classifier** trained on top-K mel spectrogram chunks, achieved a score of **0.831** on the private leaderboard. We also reflect on unexpected findings, failed experiments, and avenues for future improvements.

1 INTRODUCTION

In recent years, passive acoustic monitoring has become a vital tool for biodiversity research, allowing ecologists to non-invasively study ecosystems by detecting species from audio recordings [10]. However, identifying individual species from hours of noisy, real-world soundscapes remains a challenging machine learning task complicated by overlapping calls, environmental noise, and highly imbalanced class distributions. BirdCLEF+ 2025, part of the LifeCLEF challenge series, pushes the boundaries of ecoacoustic classification by asking participants to detect not only bird species, but also amphibians, mammals, and insects in continuous recordings collected from the biologically rich but under-monitored Magdalena Valley in Colombia.

Our project for the DSC 291 course focused on building a robust, end-to-end pipeline for this competition. In the first phase, we experimented extensively with audio preprocessing techniques to improve signal quality and reduce irrelevant noise. These included denoising, band-pass filtering, silence removal, gain adjustment, padding, chunking, and generating Mel spectrograms. A key innovation in our pipeline was the random application of preprocessing transformations per training clip, encouraging the model to distinguish bird calls from a variety of noisy conditions and background artifacts (a form of data augmentation).

In the second phase, we evaluated multiple deep learning architectures including ConvNeXt, PANNs, and Audio Spectrogram Transformers (AST). Our final submission used a ConvNeXt-based model, which achieved a private leaderboard score of 0.831, placing us competitively among participants. While our current approach shows promise, future work could explore one-vs-all classifiers for rare species and incorporating external audio datasets to improve generalization.

Through this project, we aimed not only to build a high-performing model but also to explore the challenges of real-world audio classification at scale—balancing signal processing, data augmentation, and model selection to solve a complex, ecologically important problem.

2 PROBLEM STATEMENT

The goal of this competition is to accurately identify which species (birds, amphibians, mammals, insects) are calling in recordings made from Magdalena Valley, which involves training models on over 28,000 labeled audio clips in the training data, and then classifying species in test soundscapes.

The core challenge is the domain shift between segmented, high-quality training clips and noisy, real-world test recordings containing overlapping vocals and environmental sounds. The system must perform multi-label classification to detect species presence in each segment while generalizing from controlled training audio to complex natural soundscapes with limited data for rare species.

3 DATASET

3.1 Data Description

The BirdCLEF+ 2025 dataset comprises the following components:

- `train_audio` A directory of short recordings of audio clips of individual birds in OGG format. Each clip contains a vocalizing animal (bird, amphibian, mammal or insect) and is organized by recording ID. The set covers 206 target species sourced from Xeno-Canto, iNaturalist and the Colombian Sound Archive [12, 13].
- `train.csv`
  A table with one row per clip in `train_audio`, including columns for:
  - `recording_id`: Unique identifier matching the file name
  - `primary_label`: Species code (e.g. `species_101`)
  - `secondary_labels`: Any co-occurring species codes
  - `duration`: Length of the clip in seconds
  - Geographic and temporal metadata (e.g. `latitude`, `longitude`, `date`)
- `taxonomy.csv`
  Lookup table which maps numeric label to its biological identifiers and taxonomy.
  - `primary_label` to its biological identifiers and higher-level class. It contains the following columns:
  - `inat_taxon_id`: iNaturalist taxon identifier

  - `scientific_name`, `common_name`: species names
  - `class_name`: higher-level class.
- **`train_soundscapes`**
  A collection of unlabeled audio data usually longer continuous field recordings (typically 60sec each) from same location.
- **`test_soundscapes`**
  Hidden directory of 700 unlabeled audio recordings used for final evaluation of the scores. These files are from different recording sites than train_soundscapes.
- **`sample_submission.csv`**
  The required submission file format. Each row contains:
  - `row_id`: Row Identifier (e.g. `soundscape_12345_10` is the segment from 00:05-00:10 for a soundscape number 12345)
  - `species_id`: 206 species ID columns denoting probability scores of each species in that soundscape audio segment.

3.2 Exploratory Data Analysis

3.2.1 ***Dataset Overview***. The training metadata comprises 28,564 audio clips described by 13 attributes each, including species labels, recording provenance, clip duration, and spatial–temporal coordinates. Training metadata has a shape (28564, 13).

3.2.2 ***Class Imbalance***. Species frequency is highly skewed. The count of recordings per species ranges from 2 to 990, with a mean of approximately 138.7 and a median of 80.5.

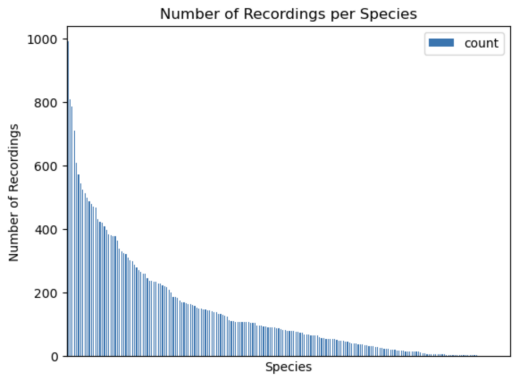

Figure 1: Number of Recordings per Species

3.2.3 ***Data Summary***. Numeric fields most notably clip duration exhibit wide variation, from just a few seconds to several minutes. On average, each species appears as a primary label in about 138 clips and as a secondary label in about 19 clips.

Table 1: Descriptive Statistics for Species Clip Counts

Statistic	Primary Label	Secondary Label	Total
Count	206	206	206
Mean	138.66	19.13	157.79
Std Dev	169.18	47.23	203.95
Min	2	0	2
25^{th} percentile	19	0	20
50^{th} percentile	80.5	3	86
75^{th} percentile	182.75	17.5	202.25
Max	990	489	1479

3.2.4 ***Species Clip Counts***. Combining primary and secondary labels shows total clip counts peaking at 1,479 for the most common species and declining sharply thereafter. The top five species by total clip occurrences are *grekis*, *trokin*, *compau*, *banana* and *roahaw*, underscoring which classes dominate the dataset.

Table 2: Primary, Secondary, and Total Audio Counts for Selected Species

Species ID	Primary count	Secondary count	Total count
grekis	990	489	1479
compau	808	44	852
trokin	787	188	975
roahaw	709	2	711
banana	610	104	714
	...		
66578	2	0	2
66531	2	0	2
66016	2	0	2
64862	2	0	2

3.2.5 ***Total Audio Time per Species***. Aggregated durations reveal that the leading species amass over 32,000s ($\approx$ 540min) of audio, while many rare species remain under 100s. This temporal imbalance further accentuates the need for careful balancing during model training.

Table 3: Total Audio Time per Species (Top 5 and Bottom 5)

Species ID	Total Seconds	Total Minutes
compau	32493.26	541.55
grekis	29803.54	496.73
roahaw	28062.40	467.71
whtdov	26327.30	438.79
laufal1	25430.51	423.84
	...	
868458	23.56	0.39
42113	22.65	0.38
42087	21.13	0.35
21116	13.06	0.22
1564122	11.33	0.19

4 DATA TRANSFORMS AND AUGMENTATION PIPELINE

4.1 Audio Preprocessing and Feature Extraction

4.1.1 DefaultFeatureExtractor. Our custom feature extractor implemented intelligent audio preprocessing with the following key features:

Centered Padding and Truncation: Unlike traditional padding approaches that add zeros at the end, our implementation uses centered padding/truncation to preserve the temporal structure of vocalizations. For truncation, we perform center-cropping by randomly selecting a start position within the valid range. For padding, we split the required padding equally between the beginning and end of the sequence, ensuring the original audio content remains centered.

Attention Mask Generation: The extractor generates attention masks to help the model focus on actual audio content rather than padded regions, improving training efficiency and model performance.

Standardization: All audio was processed at 32 kHz sampling rate, chosen to capture the frequency range of most bird vocalizations while maintaining computational efficiency.

4.2 Advanced Augmentation Strategies

4.2.1 Multi-Label Mixer. Our custom implementation extends traditional Mixup [14] to handle multi-label scenarios, addressing the task shift between single-species training clips and multi-species soundscapes:

SNR-Controlled Mixing: Samples are mixed with controlled signal-to-noise ratios (0-5 dB) to simulate realistic overlapping vocalizations. The background signal amplitude is adjusted based on the desired SNR using RMS normalization and logarithmic scaling.

Weighted Sample Selection: The mixer uses species-specific weights to ensure balanced representation during mixing, preventing dominant species from overwhelming rare ones. This addresses the extreme class imbalance where some species had only 2 recordings while others had nearly 1000.

Label Union Strategy: When mixing samples, labels are combined using a union approach, creating synthetic multi-label training examples that better reflect real-world soundscape conditions.

Temporal Randomization: Mixed samples undergo random circular shifts to prevent position-dependent learning, ensuring the model doesn't rely on specific temporal positions of vocalizations.

4.2.2 Background Noise Addition (AddBackgroundNoise). This transform addresses the domain gap between clean focal recordings and noisy field recordings:

Dynamic Background Construction: The system randomly selects and concatenates background audio files to match the target length, ensuring diverse noise profiles. When background samples are shorter than needed, multiple files are concatenated; when longer, random segments are extracted.

SNR-Based Mixing: Background noise is added with SNR values ranging from 3-30 dB, simulating various recording conditions from quiet forests to noisy environments.

RMS Normalization: Both foreground and background signals are RMS-normalized before mixing, ensuring consistent energy levels across the augmented dataset.

4.2.3 No-Call Sample Integration (NoCallMixer). To prepare the model for segments without bird vocalizations, this transform introduces negative examples:

Probabilistic Replacement: Training samples are probabilistically replaced with background/silent segments according to a specified probability threshold.

Label Zeroing: When no-call samples are introduced, the corresponding labels are set to zero vectors, teaching the model to recognize the absence of target species.

Temporal Alignment: No-call segments are carefully length-matched and resampled to maintain consistency with the target audio format.

4.3 Spectrogram-Level Transformations

Beyond the waveform-level augmentations, our pipeline included complementary spectrogram-level transforms:

- **Frequency Masking:** Random frequency bands were masked in mel spectrograms to improve robustness against spectral artifacts and environmental noise.
- **Time Masking:** Temporal segments were randomly masked to encourage the model to rely on distributed temporal features rather than specific time windows.
- **Spectral Normalization:** Mel spectrograms were normalized using AudioSet statistics to leverage pre-training knowledge from the BirdSet dataset [3, 11].

4.4 Pipeline Integration and Impact

The complete transformation pipeline was applied with carefully tuned probabilities based on the BirdSet methodology and our experimental validation:

Transform	Probability	Rationale
MultilabelMix	0.7	High probability to simulate multi-species scenarios
AddBackgroundNoise	0.5	Moderate probability to maintain signal quality
NoCallMixer	0.075	Low probability to avoid overwhelming positive samples
Frequency Masking	0.5	Improve spectral robustness
Time Masking	0.3	Encourage temporal feature distribution

Table 4: Augmentation probabilities and rationales

Domain Adaptation: The augmentation strategy effectively bridged the gap between focal training recordings from Xeno-Canto and the complex soundscape recordings in the test set from the Magdalena Valley.

Class Balance: The weighted sampling in MultilabelMix helped address the extreme class imbalance, where some species had only 2 recordings (e.g., species 66578, 66531) while others had nearly 1000 recordings (e.g., grekis with 990 recordings).

Generalization: The diverse augmentations improved model robustness to various acoustic conditions, contributing to our competitive performance of 0.831 on the private leaderboard.

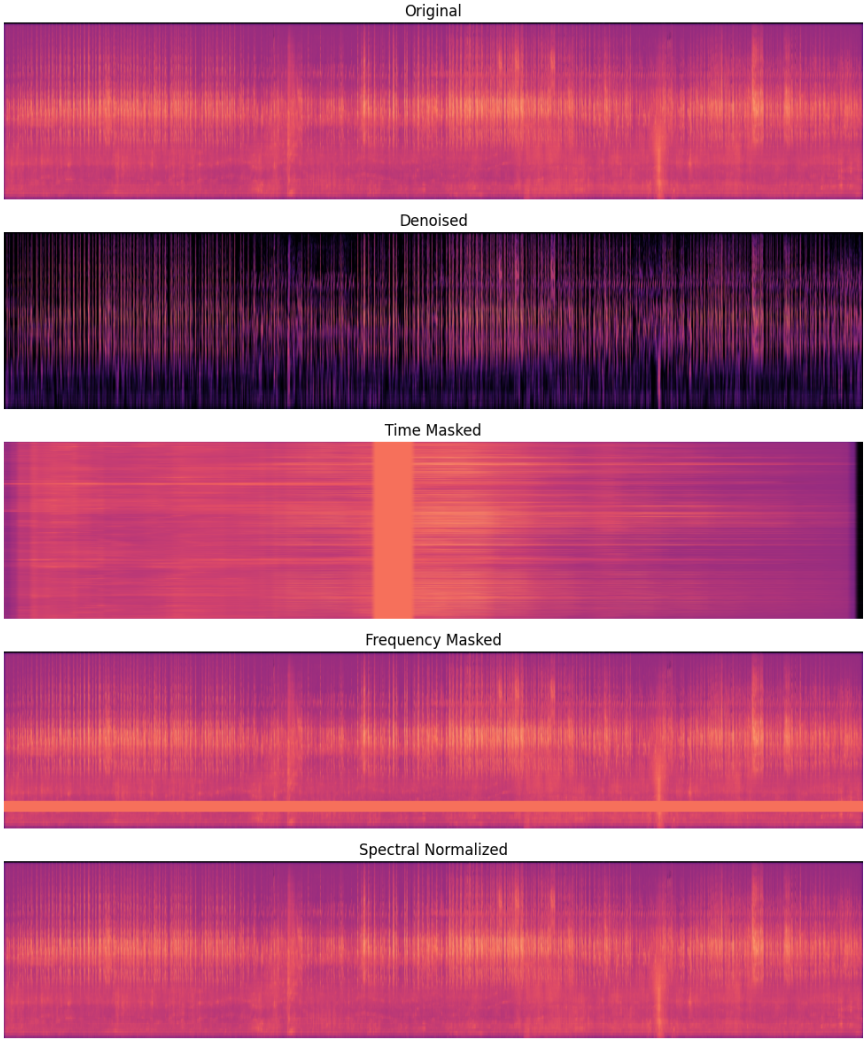

Figure 2: Spectrogram Transformations

The mathematical formulation for our SNR-controlled mixing can be expressed as:

$$x_{mixed} = x_{original} \frac{RMS x_{original}}{10^{SNR20}} \cdot x_{background} \quad (1)$$

where x_{mixed} is the resulting mixed signal, $x_{original}$ is the primary audio signal, $x_{background}$ is the background noise or secondary species audio, and SNR is the signal-to-noise ratio in decibels.

This comprehensive augmentation pipeline was crucial for achieving strong generalization performance, enabling our ConvNext model to effectively classify bird species in the challenging real-world conditions of the Magdalena Valley soundscapes, transitioning from clean, single-species focal recordings to noisy, multi-species environmental recordings.

5 MODEL SELECTION

5.1 PANN Model Architecture and Training

In the early stages of our project, we explored the PANN architecture (CNN14 variant) [5], which has previously demonstrated strong performance in environmental sound classification tasks. However, unlike the full PANN architecture that processes raw audio waveforms internally, we adapted the model to accept pre-computed mel spectrograms as input, effectively isolating and utilizing only the convolutional backbone of the network.

5.1.1 CNN14 Spectrogram Architecture Overview. CNN14's architecture consists of multiple convolutional blocks followed by global pooling layers designed for audio event detection. In our modified version, we removed the waveform-based frontend and provided log-mel spectrograms (224×224) directly as input to the convolutional layers.

While this approach offered simplicity and reduced preprocessing complexity, it came with notable tradeoffs: (1) Limited

Temporal Representation: By bypassing the raw waveform frontend, the model lost access to phase and fine temporal features that could aid species differentiation. (2) Mismatch with Original Design: The PANN architecture was originally optimized for joint learning from waveform and spectrogram features; disabling part of this pipeline degraded its representational capacity. (3)Reduced Robustness to Noise: Without explicit waveform-based denoising or augmentation layers, the model struggled to handle noisy and overlapping acoustic scenarios.

Table 5: PANN Spectrogram-Only Model Configuration

Parameter	Value
Input Resolution	224×224 (mel spectrogram)
Total Parameters	∼80M
Loss Function	BCEWithLogitsLoss
Optimizer	Adam
Learning Rate	1e-4
Batch Size	64
Epochs	100

5.1.2 Training Observations. Despite extensive training and hyperparameter tuning, the spectrogram-only PANN model failed to match the performance of other architectures we experimented with. On both internal validation and public leaderboard scores, the model consistently underperformed, particularly in multi-species detection and rare class generalization.

Its limited ability to capture complex spectral-temporal dependencies made it less competitive for the BirdCLEF+ 2025 task, where overlapping calls and domain shift play a significant role.

5.1.3 Conclusion from PANN Experiments. While the CNN14 backbone remains a strong audio feature extractor under certain conditions, our simplification of the architecture led to a substantial performance drop in this particular setting. These findings motivated our subsequent focus on fully pre-trained architectures such as HTSAT and ConvNeXt, which demonstrated superior robustness and generalization capabilities across diverse acoustic environments.

5.2 AST and HTSAT Model Architecture and Training

5.2.1 Audio Spectrogram Transformer (AST). As part of our model exploration, we first experimented with transformer-based architectures for audio spectrogram classification. The Audio Spectrogram Transformer (AST) [4], inspired by the success of Vision Transformers (ViT), adapts self-attention mechanisms to 2D spectrogram inputs, enabling global modeling of spectral and temporal dependencies.

In our implementation, we trained AST directly on log-mel spectrogram inputs resized to 128×1024 resolution, following the original AST design which processes spectrogram patches as 16×16 tokens. The model consisted of 12 transformer encoder layers with 12 attention heads and an embedding dimension of 768.

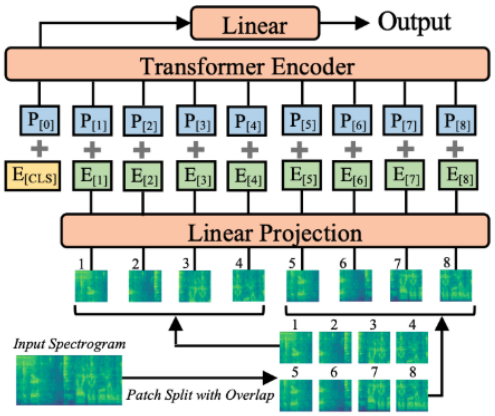

Figure 3: AST architecture

Unlike our ConvNeXt pipeline, the AST model was trained from scratch without any pretrained weights, due to the lack of large-scale domain-specific audio pretrained models available for AST in this setting. Despite this, AST demonstrated better performance compared to our earlier PANN-based spectrogram model, successfully capturing complex spectro-temporal interactions across the full input sequence.

However, the absence of pretraining limited its ability to generalize to rare species and highly noisy segments. While AST served as a useful bridge between convolutional and transformer-based architectures, its performance remained slightly behind our best-performing models.

Table 6: AST Model Configuration

Parameter	Value
Input Resolution	128×1024 (mel spectrogram)
Total Parameters	∼88M
Loss Function	BCEWithLogitsLoss
Optimizer	AdamW
Learning Rate	1e-4
Batch Size	8
Epochs	10

5.2.2 HTSAT: Hierarchical Token-Semantic Audio Transformer. Building on the insights from AST, we next explored Hierarchical Token-Semantic Audio Transformer (HTSAT) [2], a state-of-the-art pretrained audio transformer architecture. Unlike AST, HTSAT incorporates both local patch-level features and hierarchical global context modeling, making it particularly suitable for bioacoustic signals where fine-grained local patterns and global scene context both play essential roles.

We leveraged pretrained HTSAT weights from large-scale ESC and AudioSet training, providing a rich initialization

that captured broad acoustic patterns prior to BirdCLEF fine-tuning. To further enhance temporal modeling capacity, we extended HTSAT with an additional bidirectional LSTM layer and multi-head attention mechanism on top of the extracted transformer features, allowing improved aggregation of temporal information across the full audio sequence.

Additionally, the HTSAT training pipeline incorporated an aggressive rare species augmentation framework, featuring:

- Adaptive SpecAugment
- Scene simulation mixing
- Advanced noise injection (Gaussian, colored, impulse noise)
- Conservative species-aware Mixup and Cutmix
- Rare-species-focused oversampling and weighted sampling

These augmentations were designed to address the extreme class imbalance of BirdCLEF and improve model robustness for low-sample species.

Table 7: HTSAT Model Configuration

Parameter	Value
Input Resolution	256×256 (log-mel spectrogram)
Base Model	HTSAT-Swin Transformer pretrained
Total Parameters	∼92M
Temporal Enhancement	BiLSTM + Multi-head Attention
Loss Function	Combined BCE + Focal Loss with Label Smoothing
Optimizer	AdamW
Learning Rate	1e-4
Batch Size	16
Epochs	15
Augmentation Focus	Aggressive Rare Species Curriculum

5.2.3 Performance Summary. Both AST and HTSAT outperformed our earlier CNN14 spectrogram-only model by effectively capturing global context and richer spectral-temporal representations. HTSAT, with its pretrained initialization and hierarchical token representation, achieved slightly superior performance compared to AST, particularly in detecting rare and overlapping species.

However, despite these improvements, ConvNeXt ultimately delivered the best overall performance in our experiments. The combination of large-scale BirdSet pretraining, convolutional inductive biases, and robust augmentation strategies made ConvNeXt particularly well-suited for the complex, noisy, and highly imbalanced BirdCLEF+ 2025 task.

5.3 ConvNext Model Architecture and Training

Our approach leveraged the ConvNext architecture [7], a modern convolutional neural network that combines the efficiency of traditional CNNs with design principles inspired by Vision Transformers. This section details our modeling strategy, from pre-training on the BirdSet dataset to fine-tuning for the BirdCLEF+ 2025 competition.

5.3.1 ConvNext Architecture Overview. ConvNext represents a significant advancement in convolutional neural network design, incorporating several key innovations that make it particularly suitable for audio classification tasks:

Modernized Building Blocks: ConvNext employs depthwise separable convolutions, larger kernel sizes (7×7), and fewer activation functions, creating a more efficient and effective feature extraction pipeline compared to traditional ResNet-style architectures.

Layer Normalization: Unlike traditional CNNs that use Batch Normalization, ConvNext utilizes Layer Normalization, which provides more stable training dynamics and better performance on varying input distributions—crucial for handling the diverse acoustic conditions in our dataset.

Inverted Bottleneck Design: The architecture uses an inverted bottleneck structure where the hidden dimension is expanded in the intermediate layers, similar to Transformer feed-forward networks, allowing for more expressive feature representations.

GELU Activation: ConvNext employs GELU (Gaussian Error Linear Unit) activation functions instead of ReLU, providing smoother gradients and improved training stability.

5.3.2 Model Configuration. For our implementation, we utilized the ConvNext-Base architecture with the following specifications:

Parameter	Value
Model Variant	ConvNext-Base
Input Resolution	224×224 (mel spectrogram)
Total Parameters	86M
Embedding Dimension	1024
Depth Configuration	[3, 3, 27, 3]

Table 8: ConvNext-Base model configuration

5.4 ConvNext Pre-training Strategy on BirdSet

Our modeling approach followed a two-stage training paradigm, beginning with large-scale pre-training on the BirdSet dataset:

Large-Scale Representation Learning: We pre-trained our ConvNext model on the BirdSet XCL dataset, containing over 520,000 focal recordings from nearly 10,000 bird species worldwide. This extensive pre-training provided our model with rich acoustic representations spanning diverse geographic regions and species.

Multi-Label Pre-training: Unlike traditional image classification pre-training, our approach used multi-label classification from the start, training the model to handle the inherent

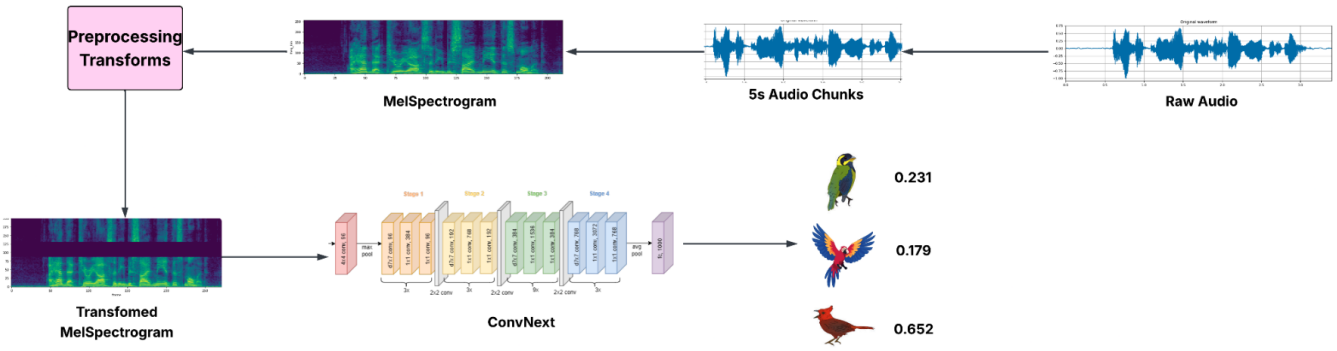

Figure 4: ConvNeXt architecture

complexity of acoustic environments where multiple species may be present.

Transfer Learning Benefits: The BirdSet pre-training provided several advantages:

- Robust feature extractors adapted to avian vocalizations
- Understanding of acoustic patterns across diverse species
- Improved handling of noisy and variable recording conditions
- Better generalization to unseen species and environments

5.5 ConvNext Fine-tuning for BirdCLEF+ 2025

Building upon the BirdSet pre-trained weights, we fine-tuned the model specifically for the Magdalena Valley dataset:

5.5.1 Target Adaptation. **Species-Specific Fine-tuning:** The final classification layer was adapted to predict the 206 target species present in the BirdCLEF+ 2025 competition, focusing the model's attention on the specific taxonomic groups found in the Colombian Magdalena Valley.

Domain-Specific Adaptation: Fine-tuning allowed the model to adapt from the global species distribution in BirdSet to the regional ecosystem characteristics of tropical lowland forests and wetlands.

5.5.2 Training Configuration. Our fine-tuning process employed the following hyperparameters:

Hyperparameter	Value
Learning Rate	5e-4
Optimizer	AdamW
Weight Decay	1e-5
Batch Size	128
Scheduler	Cosine Annealing
Warmup Ratio	5e-2
Training Epochs	10
Loss Function	Asymmetric loss

Table 9: Fine-tuning hyperparameters

Learning Rate Scheduling: A cosine annealing schedule with warm-up was used to ensure stable convergence and optimal performance, starting with a warm-up period to gradually increase the learning rate before following the cosine decay.

5.6 ConvNext Architecture Advantages for Audio Classification

ConvNext's design principles provide several specific advantages for bird sound classification:

Spectral-Temporal Feature Extraction: The larger kernel sizes (7×7) in ConvNext are particularly effective at capturing both spectral and temporal patterns in mel spectrograms, essential for distinguishing between different bird vocalizations.

Computational Efficiency: Despite its large parameter count, ConvNext's design provides excellent efficiency during both training and inference, crucial for processing the extensive BirdCLEF dataset.

Robustness to Input Variations: The Layer Normalization and modern architectural choices make ConvNext more robust to the varying acoustic conditions present in our augmented training data.

Gradient Flow: The improved gradient flow in ConvNext enables effective training of deep networks, allowing the model to learn complex hierarchical representations of acoustic features.

5.7 Performance Analysis

Our ConvNext implementation demonstrated superior performance compared to alternative architectures tested in our experiments:

Comparison with Other Architectures: Based on the BirdSet benchmark results, ConvNext consistently outperformed other models including EfficientNet, Audio Spectrogram Transformer (AST), and waveform-based models like EAT and Wav2Vec2 across multiple evaluation metrics.

AUROC Performance: ConvNext achieved the highest Area Under the Receiver Operating Characteristic curve (AUROC) scores, indicating superior discrimination ability between

positive bird vocalizations and background noise or other species.

Generalization Capability: The model's performance on diverse test datasets from different geographic regions demonstrated strong generalization capabilities, crucial for the varied acoustic conditions in the Magdalena Valley.

Final Competition Performance: Our ConvNext-based approach achieved a score of 0.831 on the private leaderboard, demonstrating the effectiveness of combining large-scale pre-training with targeted fine-tuning and comprehensive data augmentation.

The success of our ConvNext implementation highlights the importance of modern architectural design principles in audio classification tasks, particularly when combined with domain-specific pre-training and carefully designed augmentation strategies. The model's ability to effectively transfer knowledge from the diverse BirdSet dataset to the specific ecological conditions of the Colombian Magdalena Valley demonstrates the value of large-scale pre-training in bioacoustic applications.

6 EXPERIMENTS & RESULTS

6.1 ConvNext Experiments

6.1.1 Experimental Setup. Our ConvNext [7] experiments were conducted using a systematic approach to evaluate different loss functions and training strategies for the BirdCLEF+ 2025 multi-label classification task. All experiments utilized the ConvNext-Base architecture with 86M parameters, pre-trained on the BirdSet dataset and fine-tuned on the competition data.

The experimental configuration remained consistent across all runs:

- **Model Architecture:** ConvNext-Base (86M parameters)
- **Input Resolution:** 224×224 mel spectrograms
- **Batch Size:** 128
- **Optimizer:** AdamW with learning rate 5e-4
- **Training Epochs:** 12
- **Scheduler:** Cosine Annealing with 5% warmup

All models were evaluated using the same comprehensive data augmentation pipeline, including MultilabelMix (p=0.7), AddBackgroundNoise (p=0.5), NoCallMixer (p=0.075), and spectrogram-level transformations.

6.1.2 Loss Function Ablation Study. We conducted a systematic evaluation of different loss functions to address the extreme class imbalance and multi-label nature of the dataset. The experiments progressively incorporated more sophisticated loss functions and sampling strategies.

Binary Cross-Entropy (BCE) Baseline

Our initial experiment used standard Binary Cross-Entropy loss as a baseline:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\hat{y}_{ij} \;\; 1 - y_{ij}\log 1 - \hat{y}_{ij} \qquad (2)$$

where N is the batch size, C is the number of classes (206), y_{ij} is the ground truth label, and $\hat{y}_{ij}$ is the predicted probability.

Cross-Entropy (CE) Loss

We then evaluated standard Cross-Entropy loss, which is commonly used for multi-class problems but can be adapted for multi-label scenarios: $\sum \int \alpha \beta \gamma \mathcal{L} \hat{y} \log$

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\hat{y}_{ij} \qquad (3)$$

The improvement in public score suggests better discrimination capability, though with slightly lower private score performance.

Weighted Cross-Entropy Loss

To address class imbalance, we incorporated label frequency-based weights:

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j \cdot y_{ij}\log\hat{y}_{ij} \qquad (4)$$

where $w_j = \sqrt{\frac{N_{total}}{N_j \cdot C}}$ and N_j is the number of samples for class j.

The weighted approach achieved the highest private score among the simpler loss functions, demonstrating the importance of addressing class imbalance.

Focal Loss with Multilabel Mixer

We implemented Focal Loss [6] to focus learning on hard examples, combined with frequency-based resampling in the multilabel mixer:

$$\mathcal{L}_{Focal} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \alpha_j 1 - \hat{y}_{ij}{}^{\gamma} y_{ij}\log\hat{y}_{ij} \qquad (5)$$

where α_j represents class weights and $\gamma = 2$ is the focusing parameter. We resampled labels based on $w_j = \sqrt{\frac{N_{total}}{N_j \cdot C}}$ and N_j is the number of samples for class j.

Asymmetric Loss with frequency-based Sampling

Finally, we evaluated Asymmetric Loss [9], specifically designed for multi-label classification with severe class imbalance, combined with frequency-based resampling in the multilabel mixer:

$$\mathcal{L}_{ASL} = \sum_{j=1}^{C}\left[y_j \cdot 1 - p_j{}^{\gamma}\log p_j \;\; 1 - y_j \cdot p_j^{\gamma_-}\log 1 - p_j\right] \qquad (6)$$

where γ and γ_- are positive and negative focusing parameters, respectively.

6.1.3 Results Analysis. Table 10 summarizes the performance of all ConvNext experiments:

Configuration	Public Score	Private Score
ConvNext + BCE	0.800	0.830
ConvNext + CE	0.814	0.825
ConvNext + CE + Label Weights	0.809	0.831
ConvNext + Focal Loss + Resampling	0.807	0.828
ConvNext + Asymmetric Loss + Resampling	0.807	**0.832**

Table 10: ConvNext Experiments Results

6.1.4 Key Findings.

(1) **Loss Function Impact:** The choice of loss function significantly affected performance, with specialized multi-label losses (Asymmetric Loss) achieving the best private score (0.832).
(2) **Class Imbalance Handling:** Incorporating label weights and frequency-based resampling consistently improved private leaderboard performance, indicating better generalization to the test distribution.
(3) **Public vs. Private Performance:** The gap between public and private scores varied across configurations, with simpler approaches (CE) showing larger gaps, suggesting potential overfitting to the public test set.
(4) **Robust Performance:** All configurations achieved competitive scores above 0.82 on the private leaderboard, demonstrating the effectiveness of our pre-training strategy and data augmentation pipeline.

6.1.5 Discussion. The experimental results reveal several important insights about loss function selection for multi-label ecoacoustic classification. While standard BCE provided a solid baseline, the incorporation of class-aware strategies (weighted CE, Asymmetric Loss) yielded superior performance on the private leaderboard, which better reflects real-world generalization.

The Asymmetric Loss configuration achieved the highest private score (0.832), likely due to its explicit design for handling positive-negative sample imbalance in multi-label scenarios. The moderate public scores (0.807) combined with strong private performance suggest that this approach successfully avoided overfitting to the public test distribution.

The consistency of results across different loss functions (private scores ranging from 0.825 to 0.832) demonstrates the robustness of our overall pipeline, including the BirdSet pre-training strategy and comprehensive data augmentation framework.

6.2 PANN Experiments

6.2.1 Experimental Setup. We first evaluated the PANN-14 architecture [5], which is a convolutional model pretrained on AudioSet. PANN models have been widely used for general-purpose audio tagging due to their strong performance on large-scale weakly labeled audio datasets. Since BirdCLEF presents a highly multi-label and long-tail classification task, PANN served as a useful baseline for CNN-based architectures.

We fine-tuned PANN-14 using log-mel spectrograms of resolution 320×320 with 64 mel bands. The model was trained for 15 epochs with a batch size of 64, using the Adam optimizer with a learning rate of 3e-4 and StepLR scheduler. We applied basic augmentations such as Mixup [14], SpecAugment [8], and background noise injection to simulate realistic recording conditions.

6.2.2 Results and Observations. The model achieved a public leaderboard score of 0.512 and a private leaderboard score of 0.507. These relatively low scores reflect the limitations of CNN-based architectures in modeling the highly overlapping bird vocalizations and species co-occurrences present in the BirdCLEF dataset.

PANN struggled particularly with rare classes, as no explicit class balancing, label weighting, or curriculum sampling was applied. Although stable, its lower scores highlight the necessity for more complex temporal modeling when handling fine-grained ecoacoustic data.

6.2.3 Key Findings.

- PANN provided a stable CNN-based benchmark for comparison.
- The model struggled with rare species due to lack of class balancing.
- CNNs have limited capacity for learning long-range dependencies in ecoacoustic signals.

6.2.4 Discussion. While PANN offered a valuable starting point, its performance plateaued when directly applied to BirdCLEF data. The results underline that while CNNs extract local features well, they are insufficient for modeling complex sequential structures in bird vocalizations that span longer time periods and exhibit hierarchical acoustic patterns.

6.3 AST Experiments

6.3.1 Experimental Setup. We next evaluated the Audio Spectrogram Transformer (AST) [4], a transformer-based model that applies vision-style attention mechanisms to audio spectrogram patches. Unlike PANN, AST models global dependencies in both time and frequency domains through self-attention, making it better suited for multi-label ecoacoustic problems.

Two experimental settings were explored. In both, AST was fine-tuned using 224×224 log-mel spectrograms with 128 mel bands. The model was trained with a batch size of 64, AdamW optimizer (learning rate 3e-4), and cosine annealing scheduler with warmup. Standard data augmentations such as Mixup, SpecAugment, time/frequency masking, and noise injection were applied.

In Setting 2, we introduced curriculum learning with targeted resampling for rare species to explicitly address the extreme class imbalance.

6.3.2 Results and Observations. In Setting 1 (vanilla AST), the model achieved a public score of 0.608 and private score of 0.623. After introducing curriculum learning in Setting 2, performance improved to 0.622 (public) and 0.657 (private), representing a meaningful improvement in model generalization.

The curriculum sampling strategy effectively exposed the model to underrepresented species during training, helping mitigate label sparsity while maintaining stability across species distributions.

6.3.3 Key Findings.

- Transformer-based modeling outperformed CNNs, capturing long-range acoustic dependencies.
- Curriculum sampling improved rare species detection, yielding +0.034 private score gain.
- The model generalized well to unseen data with minimal public-private gap.

6.3.4 Discussion. The AST results demonstrate that transformer based models, even when directly pretrained on AudioSet, outperform classical CNNs on BirdCLEF data. Incorporating curriculum sampling proved essential for rare class generalization. This experiment highlights the importance of both architectural design and data sampling strategy for multi-label biodiversity monitoring.

6.4 HTSAT Experiments

6.4.1 Experimental Setup. Finally, we evaluated the Hierarchical Token Semantic Audio Transformer (HTSAT) [2], which extends standard audio transformers with hierarchical attention blocks. HTSAT captures both fine-grained local patterns and longer-range structures, making it highly applicable for the multi-scale nature of bird vocalizations.

We fine-tuned HTSAT using 256×256 log-mel spectrograms with 128 mel bands. The model was trained with a batch size of 32, AdamW optimizer (learning rate 1e-4), and cosine annealing scheduler. A slightly more aggressive augmentation pipeline was applied, including SpecAugment, Mixup, random cropping, and noise injection. Label smoothing was incorporated into the binary cross-entropy loss to stabilize training under noisy labels.

6.4.2 Results and Observations. HTSAT achieved the highest public leaderboard score among all models, with 0.684 on public and 0.663 on private leaderboard. Its strong public score demonstrates its ability to fit in-domain data efficiently. However, the slightly larger public-private gap suggests mild overfitting on validation data compared to AST Setting 2.

Hierarchical attention helped the model capture both short-term syllables and long-term call structures, improving its ability to distinguish subtle acoustic signatures across species.

6.4.3 Key Findings.

- HTSAT achieved the best public leaderboard score (0.684), demonstrating strong in-domain fitting.
- Hierarchical attention improved both fine-grained and global temporal modeling.
- Slight overfitting was observed when generalizing to private leaderboard samples.

6.4.4 Discussion. HTSAT represents the most advanced architecture tested in our experiments. While its public score was highest, private score lagged slightly behind AST Setting 2, suggesting potential room for improvement through more specialized training strategies, such as additional curriculum learning [1] or hard negative mining for rare species.

Model Configuration	Public Score	Private Score
PANN-14 (BCE)	0.512	0.507
AST Setting 1 (Vanilla)	0.608	0.623
AST Setting 2 (Rare Species Curriculum)	0.622	0.657
HTSAT (Label Smoothing)	0.684	0.663

Table 11: PANN, AST, and HTSAT Experimental Results.

7 CONCLUSION AND FUTURE WORK

In this work, we developed and systematically evaluated an end-to-end ecoacoustic classification pipeline for BirdCLEF+ 2025. Starting with detailed data preprocessing and augmentation strategies, we progressively built and compared multiple deep learning architectures, including convolutional (PANN, ConvNext) and transformer-based models (AST, HTSAT). Our experiments demonstrated that modern architectures, when combined with targeted data augmentation and curriculum sampling, substantially improve performance on highly imbalanced multi-label bioacoustic data.

Among all models, ConvNext achieved the best private leaderboard score (0.832), benefiting from BirdSet pretraining, advanced augmentation, and loss function tuning. HTSAT achieved the highest public leaderboard score (0.684), highlighting the strength of hierarchical attention in capturing fine-grained acoustic patterns. AST models further demonstrated the importance of rare-species-aware sampling for improving generalization on unseen data. In contrast, CNN-based PANN models underperformed, indicating their limited ability to model long-range spectral-temporal dependencies in complex soundscapes.

While our approach achieved competitive results, several limitations remain. First, rare species with limited training samples remain difficult to classify, even with curriculum sampling. Second, noisy labels and domain shift between training and test distributions continue to challenge generalization. Lastly, our pipeline primarily used supervised learning with limited utilization of the large unlabeled soundscape data available.

In future work, several extensions are promising:

- **Semi-Supervised Learning:** Leveraging large-scale unlabeled soundscape data through self-training or contrastive pretraining could improve representation learning.
- **Multi-Stage Curriculum:** Designing multi-phase curriculum learning schedules could further improve rare species detection.
- **Hard Negative Mining:** Actively mining difficult negative samples may improve species discrimination under heavy class imbalance.
- **Long-Context Transformers:** Incorporating transformer models that process longer time spans (e.g., Audio Spectrogram Transformer XL or conformer architectures) could better capture complex call sequences.

- **External Data Integration:** Incorporating additional training data from other bioacoustic datasets may improve robustness across ecosystems.

Overall, this work demonstrates the potential of combining modern deep learning architectures with domain-specific augmentation pipelines to tackle real-world ecoacoustic monitoring tasks at scale.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML 2009*. 41–48.

[2] Szu-Yu Chen, Yu Tsao Wang, and Yu Wang. 2022. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2131–2145.

[3] Jort F Gemmeke, Daniel P W Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, Ron C Moore, Manoj Plakal, and Marvin Ritter. 2020. Audioset: An ontology and human-labeled dataset for audio events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 879–890.

[4] Y Gong, Y-A Chung, and J Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech 2021*. 571–575.

[5] Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, Ron Moore, Manoj Plakal, Deepak Platt, Rif A Saurous, Brian Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP 2017*. IEEE, 131–135.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV 2017*. 2980–2988.

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *CVPR 2022*. 11976–11986.

[8] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*. 2613–2617.

[9] Tal Ridnik, Elad Ben-Baruch, Amir Zamir, Asaf Noy, Inbal Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss For Multi-Label Classification. In *ICCV 2021*. 82–91.

[10] Dan Stowell, Mathew D Wood, H Pamuła, Yannis Stylianou, and Hervé Glotin. 2019. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. In *Methods in Ecology and Evolution*.

[11] BirdSet Team. 2023. BirdSet: A global large-scale bird audio dataset. (2023). Available at https://birdset.org.

[12] Universidad del Rosario. Colombian Sound Archive - Universidad del Rosario. (????). https://bibliotecadigital.urosario.edu.co/handle/10336/29950 Accessed 2025.

[13] Xeno-Canto Foundation. Xeno-Canto: Sharing bird sounds from around the world. (????). https://www.xeno-canto.org Accessed 2025.

[14] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR 2018*.