Bird Sound Classification Using Deep Learning

A Deep Ensemble Approach for BirdCLEF 2025

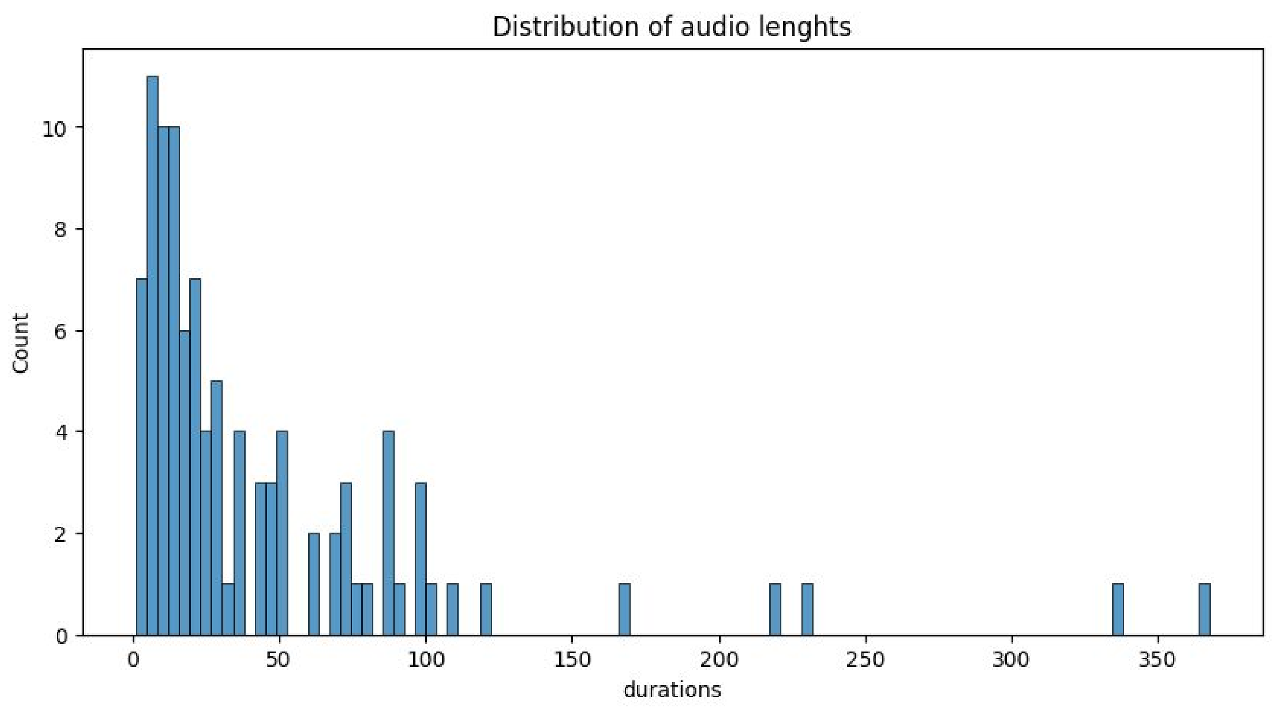
Authors: Malek, Wael, Amjad, Abdullah, Ebru

Introduction

BirdCLEF 2025 challenges participants to detect bird species from complex soundscape recordings. The task involves classifying multiple species present in 60-second audio clips using machine learning models. Our solution converts audio signals into mel spectrograms and uses an ensemble of CNN-based models to perform multi-label classification for 207 bird species.

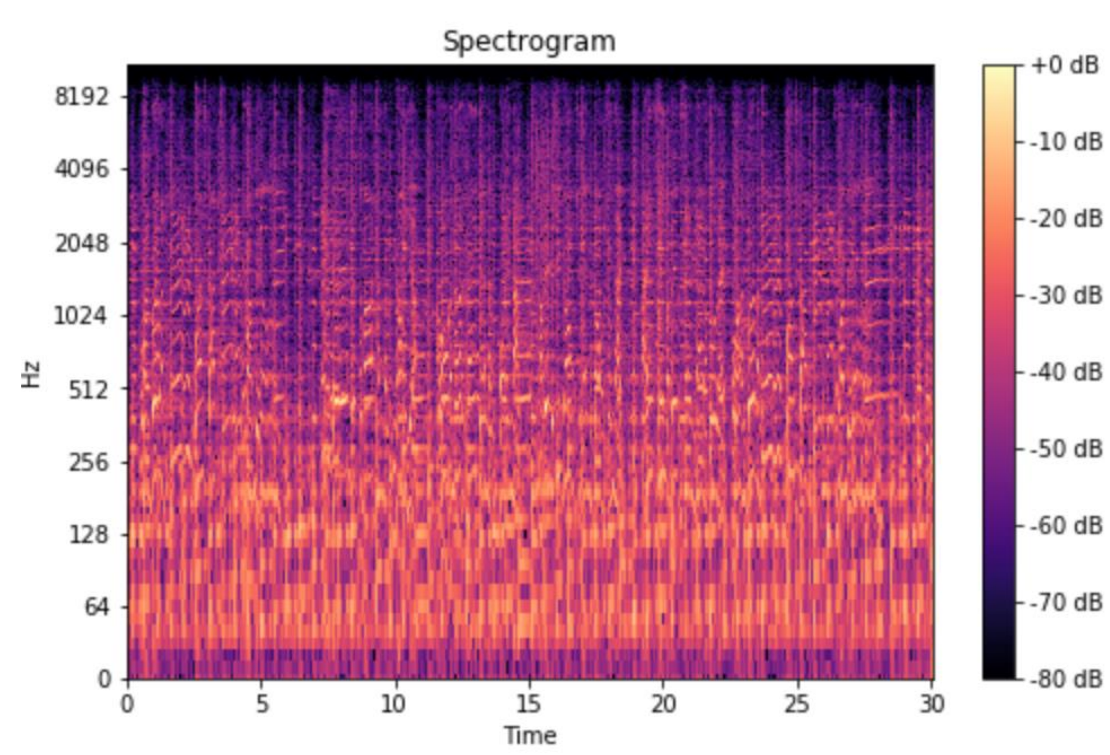
This poster summarizes the technical pipeline — from audio preprocessing to deep learning inference and postprocessing — and presents our key findings.

Dataset



- Source: BirdCLEF 2025 dataset
- Training data: Labeled bird vocalizations categorized by species
- Test data: Unlabeled 60s soundscapes (.ogg format)
- Sampling rate: 32,000 Hz
- Chunking: Each test file is split into twelve 5-second windows
- Label space: 207 bird species for multi-label prediction

Preprocessing



- Mel Spectrogram Extraction: Audio → mel spectrograms using torchaudio.transforms.MelSpectrogram with the following parameters:
  - FFT size: 1024, Window length: 1024, Hop length: 512, Mel bands: 128, Frequency range: 50–16,000 Hz
- Chunk-wise Processing:
  - Each 5-second window → one mel spectrogram.
  - Each 60s clip → 12 spectrograms stacked along batch dimension.

Inference and Postprocessing

- Model Ensemble:

We use an ensemble of 3 trained TimmSED models (sed0.pth, sed1.pth, sed2.pth). Each model outputs a 12×207 matrix of class probabilities for a soundscape.

- Averaging Predictions:

Final predictions are computed by averaging model logits and applying a sigmoid function.

- Power Correction:

Using apply_power_to_low_ranked_cols(), we penalize classes with low max scores to sharpen predictions. Columns ranked below top-30 are squared (exponent=2).

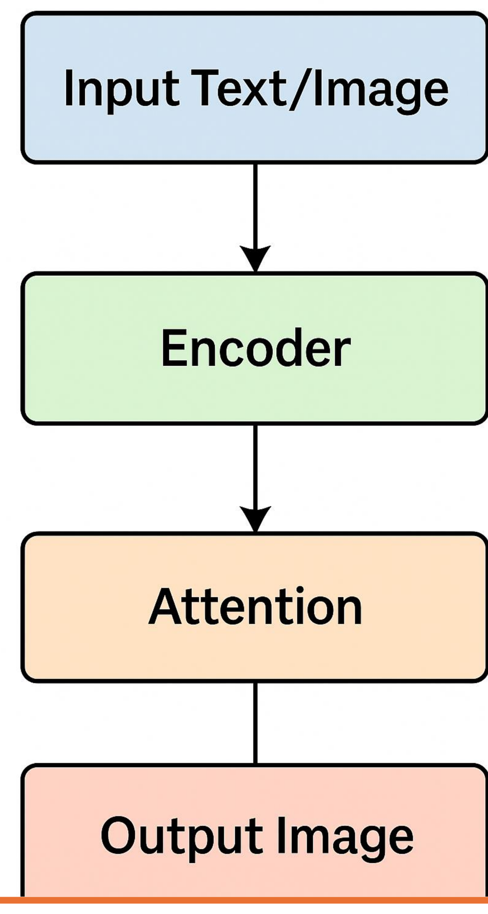
- Temporal Smoothing:

Final predictions are smoothed across 5-second windows using a rolling average:
  - Middle chunks: 0.2 × prev + 0.6 × current + 0.2 × next
  - Edge chunks: 0.8 × self + 0.2 × neighbor

Model Architecture

- Backbone Model: eca_nfnet_l0 (from timm library)
- Model Name: TimmSED
- Input: 1-channel mel spectrogram → expanded to 3 channels
- Feature Extractor: CNN encoder (final layer removed)
- Attention Head:
  - Custom attention block (AttBlockV2) Outputs soft-attended class predictions
  - Activation: sigmoid for multi-label confidence scores
- Initialization:

Xavier uniform (for weights), zeros (for bias)



Results

- Prediction Granularity:

One prediction row per 5-second chunk, with 207 columns (bird classes)

- Performance Insights:
  - Ensemble averaging and power correction improved confidence calibration
  - Temporal smoothing reduced flickering in predictions between chunks
- Output Format:

CSV with row_id + 207 columns of species scores (0–1 range)

0.872

Conclusion

- What Worked Well:
  - Mel spectrogram + CNN backbone + attention worked effectively
  - Ensemble models and score correction boosted stability
  - Temporal smoothing provided robustness in noisy environments
- Challenges:
  - Some species were hard to distinguish due to overlapping calls
  - Audio quality variability introduced inconsistency

dortmund university

Data Mining Cup
Summer Term 2025

Supervisors:
Simon Klüttermann, Steffen Maletz, Tim Katzke, Lars Grönberg