**BirdCLEF+ 2025**

Species identification from audio, focused on birds, amphibians, mammals and insects from the Middle Magdalena Valley of Colombia.

## Overview

# Goal of the competition

Mobile and habitat-diverse species serve as valuable indicators of biodiversity change, as shifts in their assemblages and population dynamics can signal the success or failure of ecological restoration efforts. However, conducting traditional observer-based biodiversity surveys across large areas is both costly and logistically demanding. In contrast, passive acoustic monitoring (PAM), combined with modern machine learning techniques, enables conservationists to sample across broader spatial scales with greater temporal resolution, providing deeper insights into the relationship between restoration interventions and biodiversity.

For this competition, you'll apply your machine-learning expertise to identify under-studied species based on their acoustic signatures. Specifically, you'll develop computational methods to process continuous audio data and recognize species from different taxonomic groups by their sounds. The most effective solutions will demonstrate the ability to train reliable classifiers with limited labeled data. If successful, your work will contribute to ongoing efforts to enhance biodiversity monitoring, including research initiatives in the lowlands of the Magdalena Valley of Colombia.

# Context

Humid tropical rainforests, Earth's most biodiverse and ancient ecosystems, are vital for climate regulation and water resource protection. However, rainforests face severe threats. In Colombia, a megadiverse country, the lowlands of the Magdalena Valley are a biodiversity hotspot and home to many endangered species. Over 70% of the Magdalena Valley lowland rainforests are replaced by vast pastures for cattle ranching, and illegal logging is common in forest fragment remnants. The protection of the last forest remnants and wetlands is an urgent need.

Fundación Biodiversa Colombia (FBC) collaborates with local communities, landowners, and organizations to conserve, restore, and connect fragments of forests and wetlands. Established in 2012, [El Silencio Natural Reserve](https://www.fundacionbiodiversa.org/fundacion2024/information/) protects 5,407 acres of tropical lowland forests and wetlands. Home to diverse wildlife, including 295 birds, 34 amphibians, 69 mammals, 50 reptiles, and nearly 500 plant species, El Silencio is a model for regional conservation and sustainability.

A significant part of the reserve, previously used for extensive livestock farming, is under an ecological restoration project. Through the Kaggle competition, we aim to automate detecting and classifying different taxonomic groups of soundscapes from El Silencio Natural Reserve, intending to provide a better understanding of the ecological process of the restoration projects.

The broader goals for this Kaggle competition include:

(1) Identify species of different taxonomic groups in the Middle Magdalena Valley of Colombia/El Silencio Natural Reserve in soundscape data.

(2) Train machine learning models with very limited amounts of training samples for rare and endangered species.

(3) Enhance machine learning models with unlabeled data for improving detection/classification.

Thanks to your innovations, it will be easier for researchers and conservation practitioners to understand restoration activities' effect trends accurately. As a result, they'll be able to evaluate threats and adjust their conservation actions regularly and more effectively.

This competition is collaboratively organized by (alphabetic order) the Chemnitz University of Technology, Fundación Biodiversa Colombia, Google Research, iNaturalist, Instituto Humboldt, K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology, LifeCLEF, Red Ecoacústica Colombiana, University College London, and Xeno-canto.

## Dataset Description

Your challenge in this competition is to identify which species (birds, amphibians, mammals, insects) are calling in recordings made in [El Silencio Natural Reserve](https://www.fundacionbiodiversa.org/fundacion2024/information/), Colombia. This is an important task for scientists who monitor animal populations for conservation purposes. More accurate solutions could enable more comprehensive monitoring.

This competition uses a hidden test set. When your submitted notebook is scored, the actual test data will be made available to your notebook.

## Files

**train\_audio/** The training data consists of short recordings of individual bird, amphibian, mammal and insects sounds generously uploaded by users of [xeno-canto.org](https://www.xeno-canto.org/), [iNaturalist](https://www.inaturalist.org/" \t "_blank) and the Colombian Sound Archive (CSA) of the [Humboldt Institute for Biological Resources Research in Colombia](https://www.humboldt.org.co/). These files have been resampled to 32 kHz where applicable to match the test set audio and converted to the ogg format. Filenames consist of [collection][file\_id\_in\_collection].ogg. The training data should have nearly all relevant files; we expect there is no benefit to looking for more on [xeno-canto.org](https://www.xeno-canto.org/) or [iNaturalist](https://www.inaturalist.org/" \t "_blank) and appreciate your cooperation in limiting the burden on their servers. If you do, please make sure to adhere to the scraping rules of these data portals.

**test\_soundscapes/** When you submit a notebook, the **test\_soundscapes** directory will be populated with approximately 700 recordings to be used for scoring. They are 1 minute long and in ogg audio format, resampled to 32 kHz. The file names are randomized, but have the general form of soundscape\_xxxxxx.ogg. It should take your submission notebook approximately five minutes to load all the test soundscapes. Not all species from the train data actually occur in the test data.

**train\_soundscapes/** Unlabeled audio data from the same recording locations as the test soundscapes. Filenames consist of [site]\_[date]\_[local\_time].ogg; although recorded at the same location, precise recording sites of unlabeled soundscapes do NOT overlap with recording sites of the hidden test data.

**train.csv** A wide range of metadata is provided for the training data. The most directly relevant fields are:

* primary\_label: A code for the species (eBird code for birds, iNaturalist taxon ID for non-birds). You can review detailed information about the species by appending codes to eBird and iNaturalis taxon URL, such as https://ebird.org/species/gretin1 for the Great Tinamou or https://www.inaturalist.org/taxa/24322 for the Red Snouted Tree Frog. Not all species have their own pages; some links might fail.
* secondary\_labels: List of species labels that have been marked by recordists to also occur in the recording. Can be incomplete.
* latitude & longitude: Coordinates for where the recording was taken. Some bird species may have local call 'dialects,' so you may want to seek geographic diversity in your training data.
* author: The user who provided the recording. Unknown if no name was provided.
* filename: The name of the associated audio file.
* rating: Values in 1..5 (1 - low quality, 5 - high quality) provided by users of Xeno-canto; 0 implies no rating is available; iNaturalist and the CSA do not provide quality ratings.
* collection: Either XC, iNat or CSA, indicating which collection the recording was taken from. Filenames also reference the collection and the ID within that collection.

### Acknowledgements

The building of the El Silencio Dataset was supported by [Earth Species Project](https://www.earthspecies.org/), Experiment.com and Footprint Coalition under a Science Engine grant “AI for Interspecies Communication". Compiling these extensive datasets was a major undertaking, and we are very thankful to the many domain experts who helped to collect and manually annotate the data for this competition. Specifically, we would like to thank (institutions and individual contributors in alphabetic order):

[**Chemnitz University of Technology**](https://www.tu-chemnitz.de/index.html.en)  
Stefan Kahl, Mario Lasseck, and Maximilian Eibl

[**Colección de Sonidos Ambientales Mauricio Álvarez Rebolledo (IAvH-CSA)**](https://colecciones.humboldt.org.co/sonidos/)  
Maria Paula Toro-Gómez

[**Fundación Biodiversa Colombia**](https://www.fundacionbiodiversa.org/)  
Paula Caycedo, Pedro Jose Cardona, Noel Torres, Hugo Torres, Ramiro Torres, Felipe Aragón, Luis Osorio, and Santiago Rosado

[**Google Research**](https://research.google/)  
Tom Denton

[**iNaturalist**](https://www.inaturalist.org/)  
Grant van Horn

[**Instituto Humboldt**](https://www.humboldt.org.co/)  
Juan Sebastián Ulloa and Susana Rodríguez Buriticá

[**K. Lisa Yang Center for Conservation Bioacoustics**](https://www.birds.cornell.edu/ccb/)  
Stefan Kahl and Holger Klinck

[**LifeCLEF**](https://www.imageclef.org/LifeCLEF2025)  
Alexis Joly and Henning Müller

[**Universidad de los Andes**](https://www.uniandes.edu.co/)  
Jose Luis Benavides-Lopez

[**University College London**](https://www.ucl.ac.uk/biosciences/gee/people-and-nature-lab/)  
Juan Sebastián Cañas and Kate Jones

[**University of Edinburgh**](https://www.ed.ac.uk/)  
Oisin Mac Aodha

[**Xeno-canto**](https://xeno-canto.org/)  
Willem-Pier Vellinga, Bob Planqué

**Photo Credits**

Banner picture of a Olivaceous Piculet by Rosado Hidalgo  
Inset picture of a Yellow-striped poison frog by Rosado Hidalgo

### Citation

Holger Klinck, Juan Sebastián Cañas, Maggie Demkin, Sohier Dane, Stefan Kahl, and Tom Denton. BirdCLEF+ 2025. https://kaggle.com/competitions/birdclef-2025, 2025. Kaggle.

Bibtex:

@misc{birdclef-2025,

author = {Holger Klinck and Juan Sebastián Cañas and Maggie Demkin and Sohier Dane and Stefan Kahl and Tom Denton},

title = {BirdCLEF+ 2025},

year = {2025},

howpublished = {\url{https://kaggle.com/competitions/birdclef-2025}},

note = {Kaggle}

}