

# Species Identification in Noisy Soundscapes

Reproducing State-of-the-Art Solutions in  
Bioacoustic Classification

Team Name:

## Nova Matrix

Team Members:



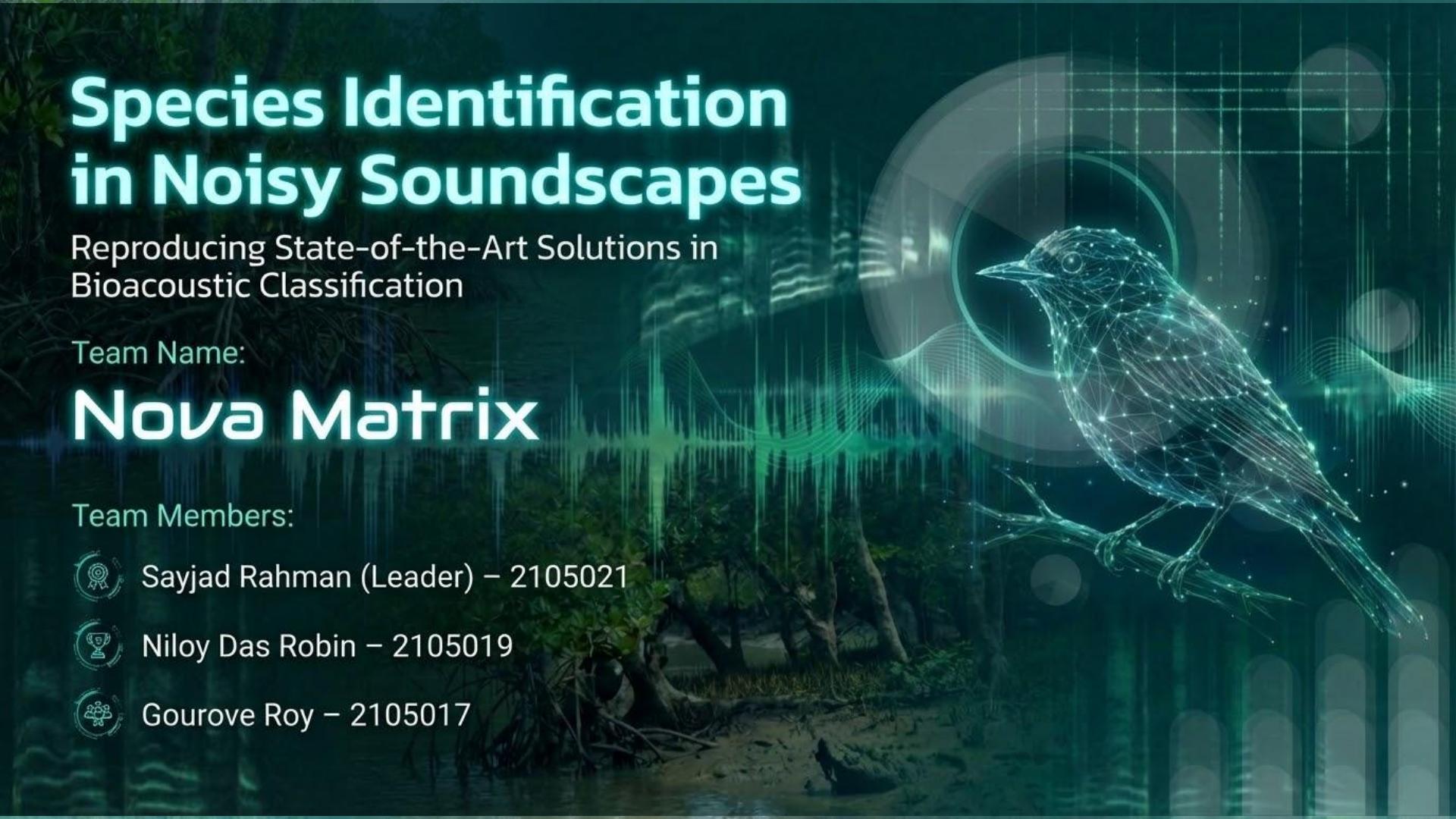
Sayjad Rahman (Leader) – 2105021



Niloy Das Robin – 2105019



Gourove Roy – 2105017



# Introduction



- **Context:** Biodiversity monitoring via Passive Acoustic Monitoring (PAM).
- **Objective:** Automate the identification of birds, amphibians, mammals, and insects from continuous audio recordings.
- **Significance:** Scalable processing of audio data to support ecological restoration and conservation efforts.



# The Challenge of Bioacoustic Analysis



## Data Scale & Automation

Traditional surveys are **manual & expensive**.

Passive Acoustic Monitoring (PAM) generates **petabytes of data**, requiring automated processing.



## Soundscape Complexity

Competition focus: '**Soundscapes**' analysis. Identifying species in noisy, overlapping, real-world recordings is significantly harder than clean calls.

# Motivation & Inspiration



## Ecological Motivation:

- Species are indicators of ecosystem health.
- Need to track "mobile and habitat-diverse species" to measure restoration success.



## Technical Challenge:

- **Domain Shift:** Training data (clean, focal recordings) vs. Test data (noisy, ambient soundscapes).
- **Class Imbalance:** Rare vs. common species distribution.
- **Weak Labels:** Unlabeled background species in training files.



## Inspiration:

- Recent advancements in "Noisy Student" training and Pseudo-labeling.
- The opportunity to apply Computer Vision techniques (Spectrograms) to Audio.

# Why This Project Matters



## Tackling the 'Domain Shift'

We chose this project because the core challenge is a common real-world problem: 'Domain Shift.' Models trained on clean, curated data often perform poorly when deployed in noisy, complex environments. This competition provides a rigorous testbed for advanced Semi-Supervised Learning (SSL) techniques, such as Pseudo-labeling, which have broad applicability across various machine learning domains beyond acoustic analysis.



## Real-World Conservation Impact

The successful models developed here will directly support researchers in Colombia, significantly aiding conservation efforts by automating the laborious process of filtering and analyzing thousands of hours of field recordings, and will also help to protect the wildlife of the Sundarbans, Bangladesh.

# Problem Definition



## Input & Task:

**Input:** Continuous audio files (ogg/mp3 format).

**Task:** Multi-label classification (detecting presence/absence of species).



## Taxonomy:

Birds, Amphibians, Mammals, Insects (Primary and Secondary labels).



## Key Constraints:

- Noisy Environment:** Rain, wind, anthropogenic noise, and overlapping calls.
- Human Speech:** Incidental human voices in training data that must be filtered.



## Evaluation Metric:

Macro-Averaged ROC-AUC (assessing ranking quality across all classes).

# The Core Challenge: Weakly Supervised Learning



## Focal Recordings vs. Soundscapes:

- We use 'focal' recordings where the target bird is known. The test set consists of 5-second segments from long, complex soundscapes.



## The “Noise” Problem:

- Other birds and background noise in focal recordings are often unlabeled. If the model incorrectly learns these as “noise,” it will fail.



## Robustness Requirement:

- The model must be robust enough to ignore background noise while remaining sensitive to faint calls.

# Related Works



## Efficient Bioacoustic Classification under Resource Constraints (Miyaguchi et al.)

- **Problem:** Addressed strict CPU inference limits (90-minute) for large-scale recognition.
- **Method:** Proposed optimized transfer learning (e.g., EfficientNet, Perch with TFLite) and spectrogram tokenization combined with skip-gram embeddings.
- **Outcome:** Significantly reduced inference time while maintaining competitive accuracy, highlighting the trade-off between efficiency and representational power.



## birdclef2025 Multi-Taxonomic Soundscape Recognition in Real-World Environments (Cañas et al.)

- **Scope:** Comprehensive benchmark for simultaneous acoustic identification of birds, mammals, insects, and amphibians in complex tropical soundscapes.
- **Approaches:** Summarizes dominant approaches: CNN-based sound event detection, ensemble learning, pseudo-labeling, and self-training.
- **Challenges:** Emphasizes challenges related to class imbalance, overlapping vocalizations, and generalization under field conditions.

# Dataset Descriptions



## Context & Objective

- **Goal:** Identify **206 species** (Birds, Amphibians, Mammals, Insects) in complex soundscapes.
- **Location:** El Silencio Natural Reserve, Colombia (Middle Magdalena Valley).
- **Challenge:** Multi-label classification with a hidden test set.



## Training Data (~7.4 GB)

- **train\_audio:** **28,564** short, labeled clips of individual species (Source: Xeno-canto, iNaturalist, CSA).
- **train\_soundscapes:** **9,726** unlabeled, 1-minute recordings from the target location (useful for background noise characterization).
- **Metadata:** Includes geolocation (lat/long), recording quality (1-5 rating), and time of day.



## Inference & Testing:

- **test\_soundscapes:** Hidden dataset of **~700** recordings (1-minute length) populated at runtime.
- **Format:** All audio is standardized to **OGG format @ 32 kHz**.
- **Submission:** Predict species probability for every 5-second segment (row\_id).

# Contribution / Goal



**Primary Goal:** Develop a robust bioacoustic classification system capable of generalizing from clean 'focal' recordings to noisy, real-world soundscapes.

## Specific Contributions:



### Domain Adaptation Strategy

We aim to implement the "**Noisy Student**" framework (Xie et al.) to bridge the gap between clean training data and noisy test environments.



### Noise Mitigation Study

We will investigate if removing anthropogenic noise (specifically human speech via **Voice Activity Detection**) reduces false positives in wilderness recordings.



### Cross-Domain Transfer Learning

We will evaluate the efficacy of using **ImageNet-pretrained encoders** (EfficientNet) on audio spectrograms compared to standard audio-only baselines.

# The Challenge and Our Solution



## The Problem: Noisy Real-World Soundscapes

Existing models struggle. Clean training data does not match **noisy test data** (wind, rain, overlapping insect sounds).



## Our Hypothesis: Semi-Supervised Learning

Standard augmentation (SpecAugment) is insufficient. We use unlabeled soundscape data in a **Semi-Supervised Learning** approach to explicitly train the model to understand and filter out 'background noise'.



## Novel Contribution: VAD for Biodiversity

We tackle the engineering challenge of applying **Voice Activity Detection (VAD)** to biodiversity monitoring to prevent incorrect identification of human speech as animal calls.



# Project Plan



## Phase 1: Foundation (25%)



- **Data Pipeline:** Build efficient audio-to-spectrogram conversion (using librosa or torchaudio).



- **EDA:** Analyze class imbalance and distribution of "background" species labels.



- **Baseline:** Train a standard Supervised CNN (EfficientNet-B0) on raw focal recordings to establish a performance floor.



## Phase 2: Robustness & Cleaning (50%)



- **Data Cleaning:** Integrate Silero VAD to filter human speech segments from the training set.



- **Augmentation:** Implement SpecAugment (Time/Frequency masking) and Mixup to simulate overlapping calls.

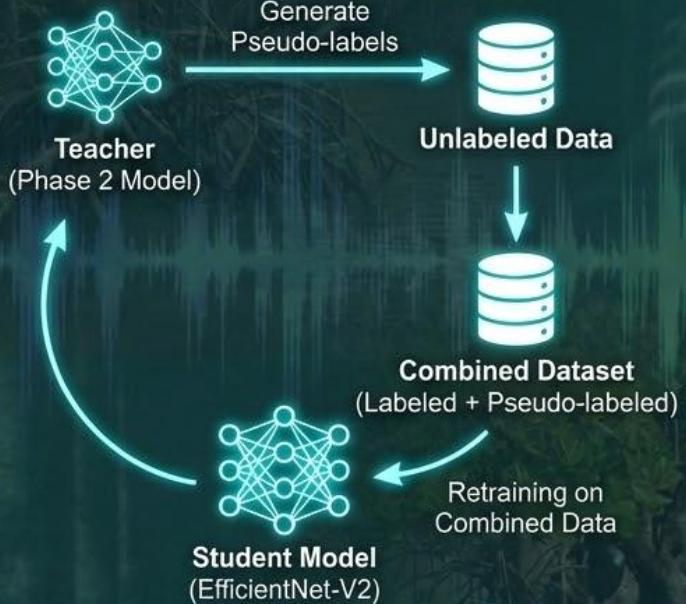


- **Evaluation:** Compare "Cleaned" vs. "Baseline" model performance on a held-out validation set.

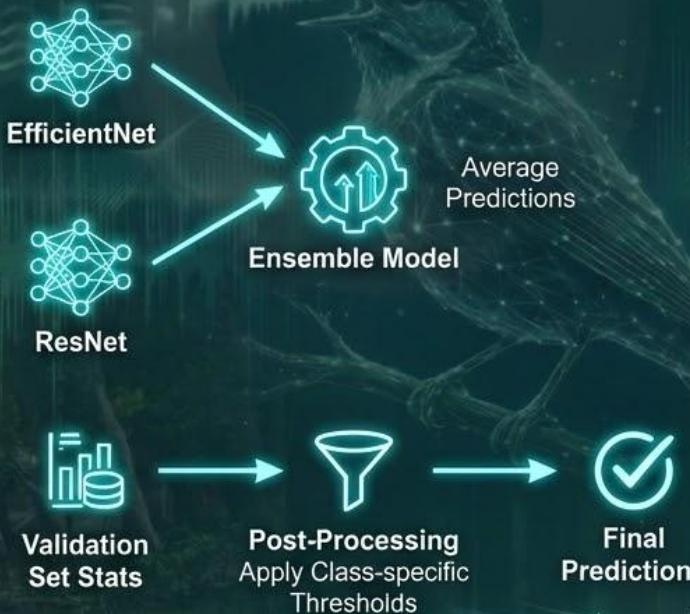




### Phase 3: Semi-Supervised Learning (75%)



### Phase 4: Optimization (100%)



# Why this order?



## Why this order? (25% Milestone)

We start simple to ensure our data pipeline works.



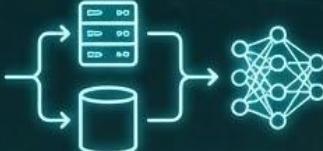
## The 50% Milestone: Speech Removal Hypothesis

This is where we test the 'Speech Removal' hypothesis. If the VAD cleaning works, our validation score should jump significantly because the model stops learning irrelevant features.



## The 75% Milestone: Semi-Supervised Learning

This is the most technically complex part. We move from Supervised Learning to Semi-Supervised Learning, which allows us to utilize the thousands of hours of unlabeled audio provided in the competition.



# Initial Progress

**Status:** Literature Review & Theoretical Framework Design

## Accomplishments:



### Literature Survey

Completed review of core methodologies including PANNs (Kong et al.) and Noisy Student training (Xie et al.).



### Dataset Analysis

Mapped out the class imbalance; identified 'Long-tailed' species that require external data augmentation.



### Architecture Selection

Decided on Mel-Spectrograms as the input feature (Log-scale, 128 mels) to leverage Computer Vision backbones.



### Feasibility Check

Verified that torchaudio can process the competition's .ogg files efficiently on our available hardware.

## Immediate Next Steps: ➤➤➤



Initialize the Git repository.

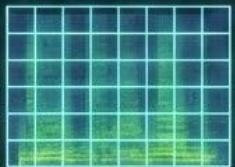


Set up the local environment with PyTorch and CUDA support.

# Theoretical Framework & Resource Planning



## Theoretical Framework



Linear Spectrogram



Mel-scale Spectrogram



Moved beyond simple “code-searching” to grasp the underlying physics of the problem.



Adopted Mel-scale spectrograms instead of linear spectrograms, as they better mimic human and animal auditory perception, which is crucial for bird call analysis.



## Resource Planning



Dataset



Low-resolution Copies



The dataset is approximately 12GB.



Developing strategies for storage and data loading, such as utilizing low-resolution copies for rapid prototyping, to accelerate iteration speed.