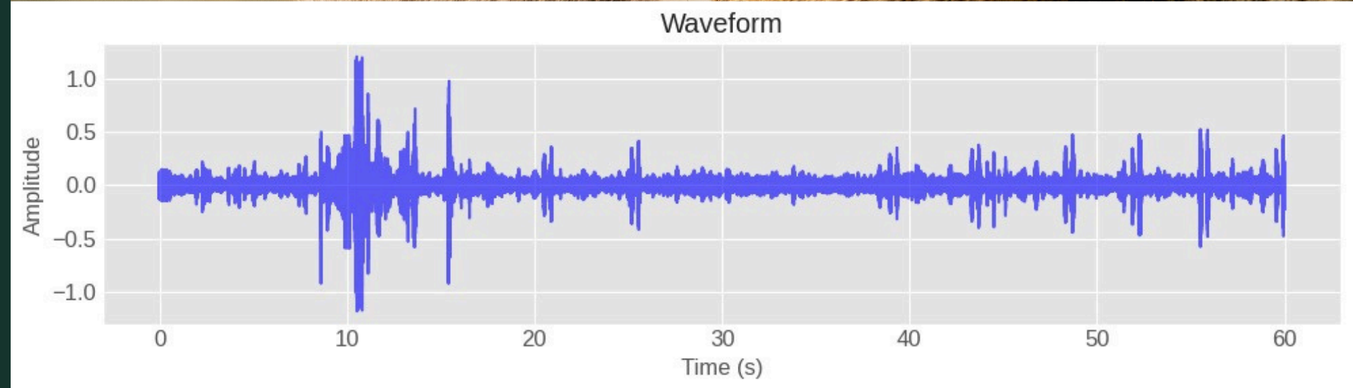
Bioacoustics: Assessing Ecosystem Biodiversity

Dylan Berens Dominic McDonald Shruthi Yenamagandla

Key Insights

- Converts raw rainforest audio recordings into Mel-spectrograms (2D representations of time/frequency patterns)
- Fine-tune Audio Spectrogram Transformer AST
- Predicted Acoustic Diversity Index (ADI) instead of species labels.
- Differential Learning Rates:
  - 1e-5 – AST backbone
  - 1e-4 – AST custom regression head



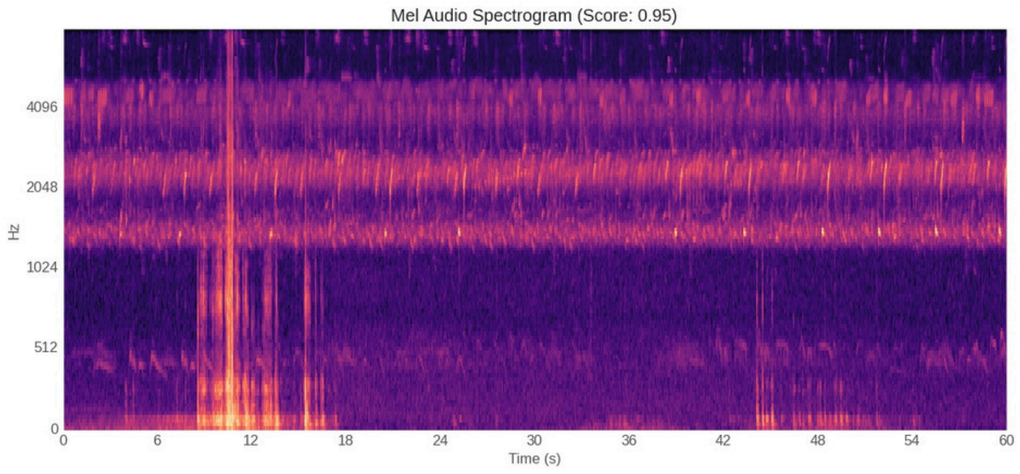
Dataset Overview

Rainforest Connection Species Audio Dataset (Kaggle):

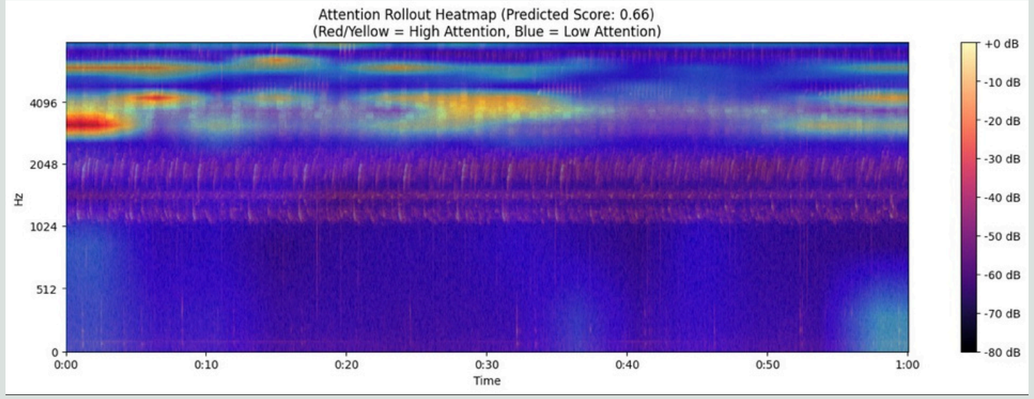
- This dataset provides tropical rainforest audio recordings of birds and frogs, enabling us to train low sample & high accuracy models that detect species support data-driven conservation decisions.
- **Type:** Audio recordings (tropical rainforest soundscapes)
- **Content:** Bird and frog species sounds + background noise (insects, howler monkeys, rain, etc.)

6,719 soundscape audio files (.flac)
+ 8 anchor files (.flac)

Mel Audio Spectrogram



Attention Rollout Heatmap



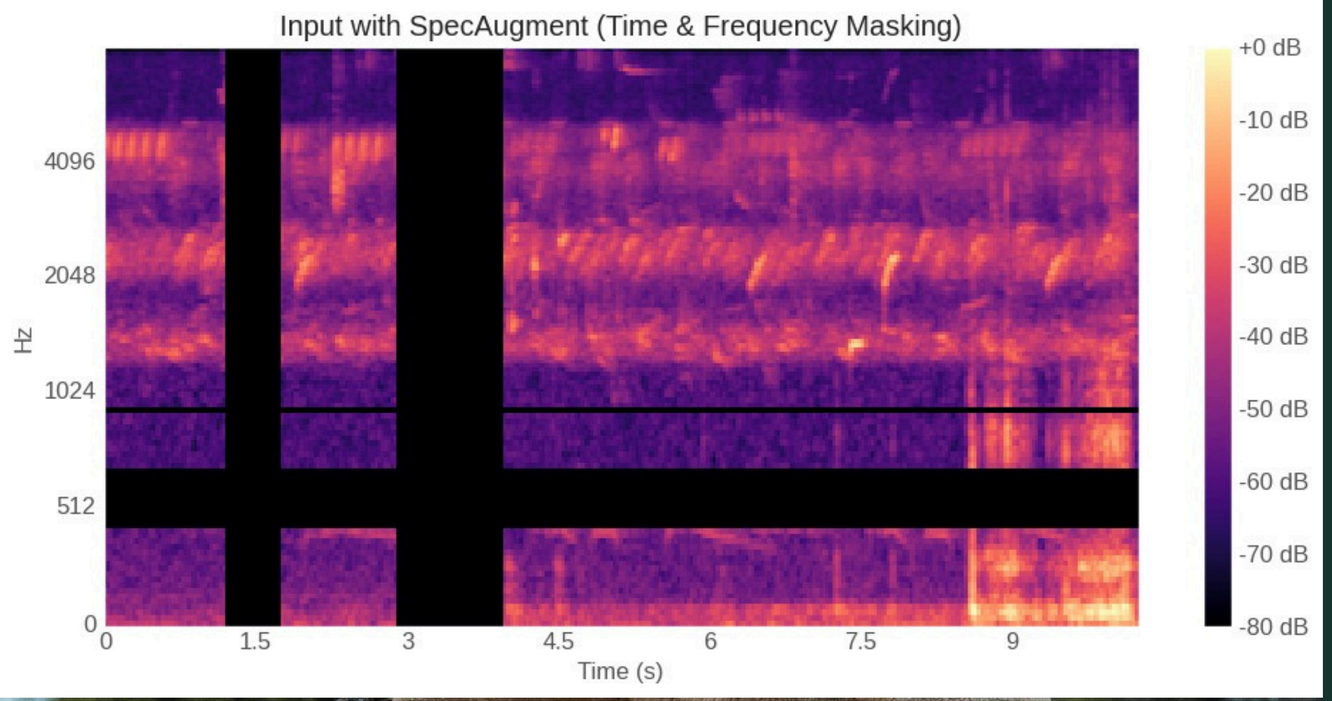
Data Visualization + ADI

Target Variable: Acoustic Diversity Index (ADI)

1. **Background Subtraction** (subtract median energy)
2. **Adaptive Thresholding** (count if 13.5 dB louder than baseline)
3. **Frequency Banding** (split into 200Hz buckets)
4. **Shannon Entropy** (check energy spread across buckets)
5. **Soft Fallback** (if nothing >13.5 dB, energy sum fractional score)

Explainable AI (XAI):

- We used Attention Rollout Heatmaps to extract the gradient from the ViT's attention head, to visualize the regions that highly influenced the model's prediction



Data Augmentation:

- **SpecAugment:**
  - Frequency Masking (horizontal Hz bars)
  - Time Masking (vertical time bars)

- **Random Tempral Cropping**
  - AST model takes input of 10.24 seconds
  - Our input data preprocessed to 60 seconds
  - Each epoch, use a different 10.24s crop

Testing/Evaluation?

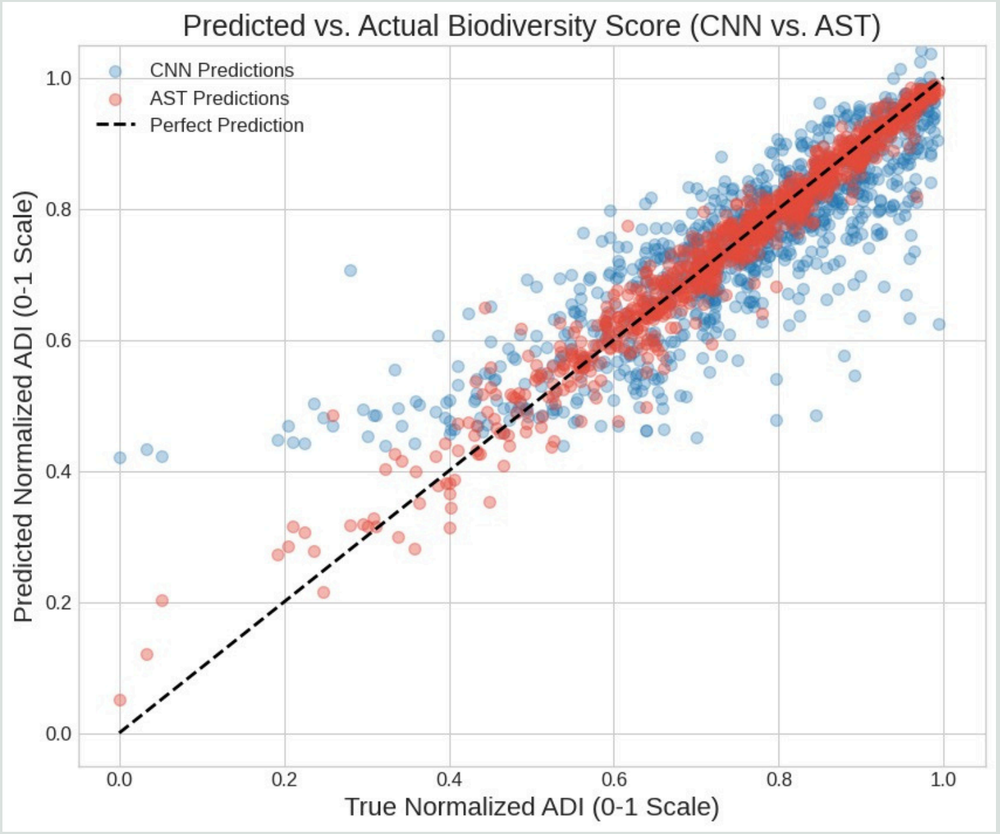
1. During testing, do this for all 6 x 10.24s slices
2. Select Top 3/6
3. Average Top 3 for score

Model Architectures

Model Definition

Audio Spectrogram Transformer (AST) [~86.6M]
Convolutional Neural Network (CNN) [~322K]
Vision Transformer (ViT-base) [~86M]

Loss: Mean Absolute Error (MAE)



CHALLENGES

- Target Variable (Species Count/ACI/NDSI/ADI)
- Keras Functional API
- Attention Rollout Heatmap blank overlay
- Reproducibility

Final Takeaways

Findings

Audio Spectrogram Transformer (AST)
- R Squared – **0.958**
- Test Loss (MAE) – **0.022**

Convolutional Neural Network (CNN)
- R Squared – **0.700**
- Test Loss (MAE) – **0.063**

Vision Transformer (ViT)*
- R Squared – **0.669**
- Test Loss (MAE) – **0.059**

Interpretation

- Transfer Learning domain shift was successfully addressed by switching from ViT-base to AST

- Selection of the appropriate pretrained model or checkpoint is crucial for predictive capability

Improvements

- AST vs. ViT ✅
- SpecAugment ✅
- ADI Parameter Tuning ✅
- Docker containerization ✅
- ~~Differential Layer Unfreezing~~
- Expand Dataset & Anchors

*ADI calculation has been adjusted since this earlier train

Thank You!