Bioacoustics: Assessing Ecosystem Biodiversity using Transformers

UNIVERSITY OF HOUSTON

Dylan Berens, Dominic McDonald, Shruthi Yenamagandla

Natural Science & Mathematics

Abstract

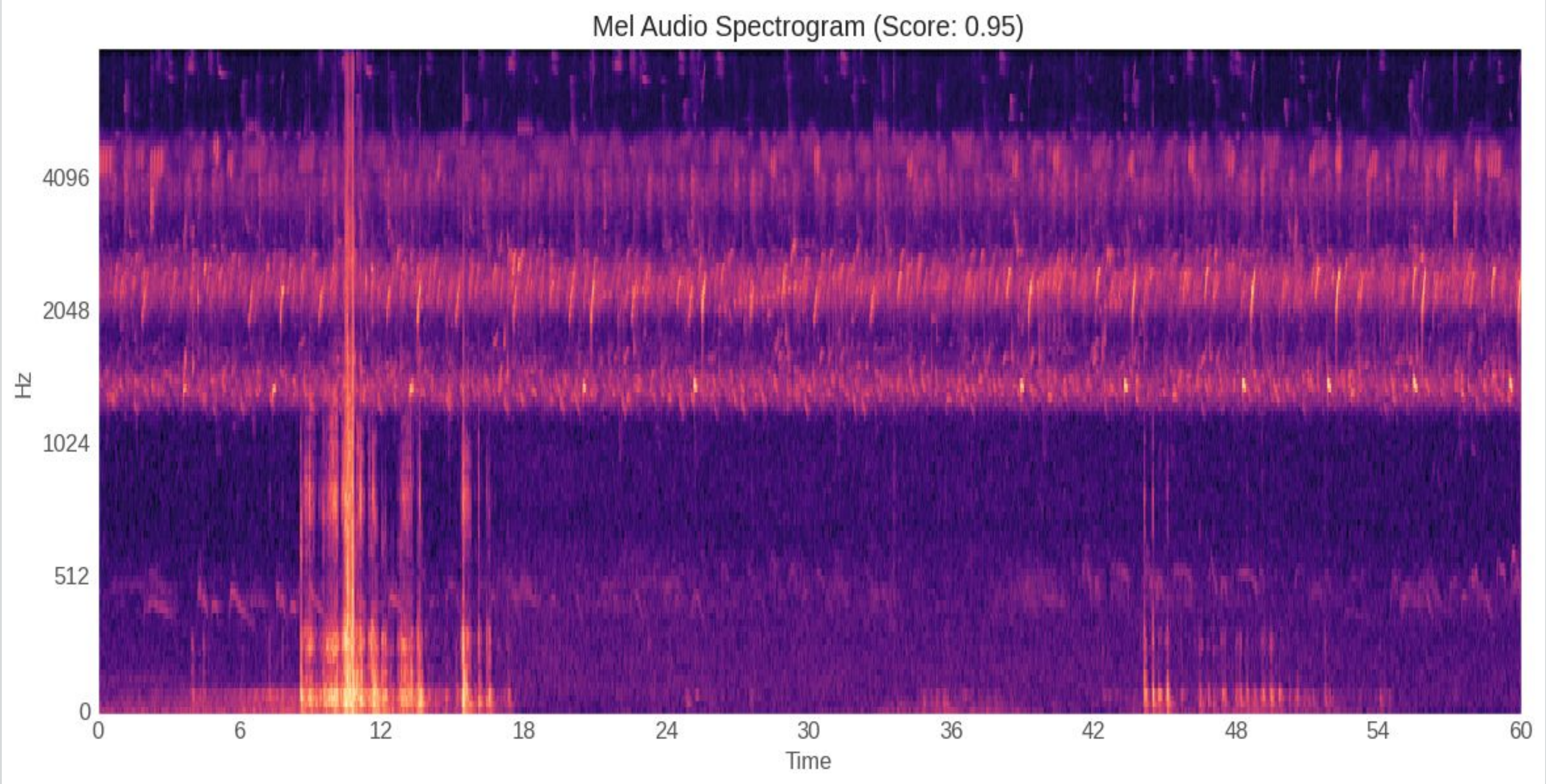
Background & Methods: Our project explores the use of Transfer Learning, Transformers and bioacoustics to assess the health of ecosystem soundscapes. Manual identification of species in field recordings is often slow, inconsistent, and prone to error. Because of this, we evaluated two modern architectures, Audio Spectrogram Transformers (ASTs) and Convolutional Neural Networks (CNNs). These models were applied to spectrogram-based audio representations. Our pipeline converted raw audio into log-mel spectrograms, applied normalization and data augmentation, and trained both models using early stopping, learning-rate scheduling, and Mean Absolute Error (MAE) as the primary loss function. We created a target variable of Acoustic Diversity Index (ADI)

Results: The AST model demonstrated stronger predictive performance than the CNN, achieving an R² of 0.958 and a test MAE of 0.022. In comparison, the CNN reached an R² of 0.700 with a test MAE of 0.063. Our findings emphasize the importance of the model checkpoint selection.

Conclusion: Our findings demonstrate that successful Sequential Transfer Learning can vastly outperform conventional baseline methods, enabling it to recognize complex acoustic patterns and generalize well beyond the training environment. This strong generalization ability suggests that Transformer models can scale to diverse habitats, species communities, and recording conditions, broadening population-level monitoring and supporting more comprehensive biodiversity assessments.

Background

Bioacoustics has become a vital tool for wildlife monitoring, particularly because many species communicate through sound. Traditional field identification is slow, reliant on expert expertise, and difficult to scale, while autonomous recording units now generate massive amounts of audio that demand automated analysis. Spectrograms, which visually represent frequency patterns over time, enable modern deep learning models to process sound as image data, making architectures like ASTs and CNNs well-suited for extracting complex ecological information. Building on this capability, our project evaluates the performance of ASTs and CNNs in modeling biodiversity-related patterns from audio soundscapes.



Methods

Dataset: Our dataset consisted of 6,719 soundscape audio files from the Amazon Rainforest found on Kaggle, containing labeled data of 24 species of birds and amphibians, and countless unlabeled species (e.g., insects, howler monkeys, cougars)

Target Variable (ADI): Our custom target variable emphasizes the importance of species' presence across many different frequency bands, and employs 5 methods: 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 evenness of energy spread across buckets) 5.) Soft Fallback (if nothing >13.5 dB, fractional score based on energy sum)

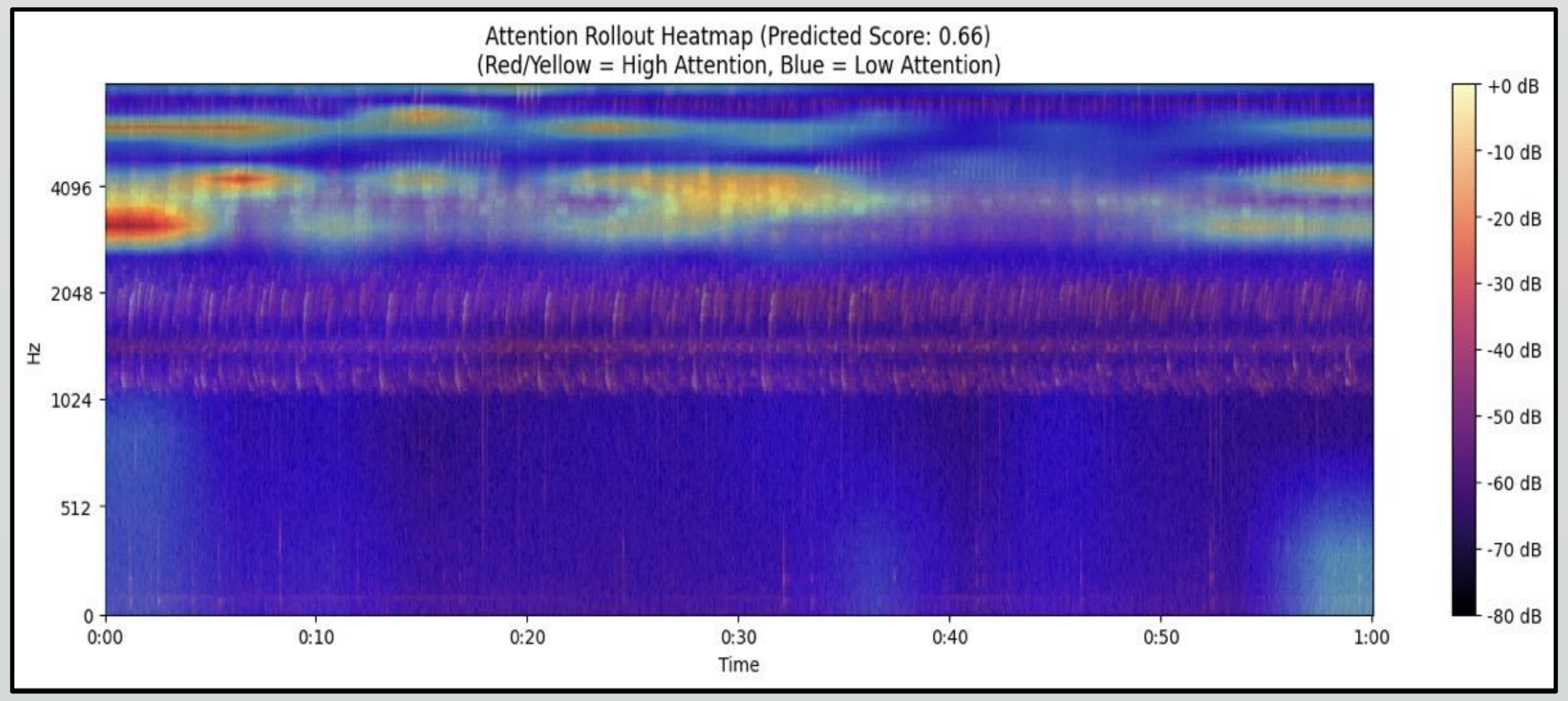
Data Preprocessing: Raw audio files (.flac) were projected onto 2D space as log-Mel Spectrograms using Librosa to visualize frequency over time, with amplitude (dB)-based coloring. We applied trimming, 16 kHz resampling, normalization, data augmentation with SpecAugment (temporal/frequency masking) & Random Temporal Cropping

Vision Transformer (AST): We developed a custom PyTorch model by stacking a 3-layer regression head on a pretrained AST that was fine tuned from a pretrained Google Vision Transformer (ViT-base), resulting in a total of 101 layers and ~86.8 million learnable parameters. We trained our AST with differential learning rates: 1e-4 for the custom regression head and 1e-5 for the backbone. We employed learning rate scheduling (ReduceLRonPlateau) and EarlyStopping as callbacks. This approach achieved an R² score of 0.958 with a test MAE of 0.022.

Convolutional Neural Network (CNN): Our baseline model was a custom 14-layer CNN built in TensorFlow. The CNN achieved an R² of 0.700 + MAE of 0.063, outperforming the original ViT due to domain shift.

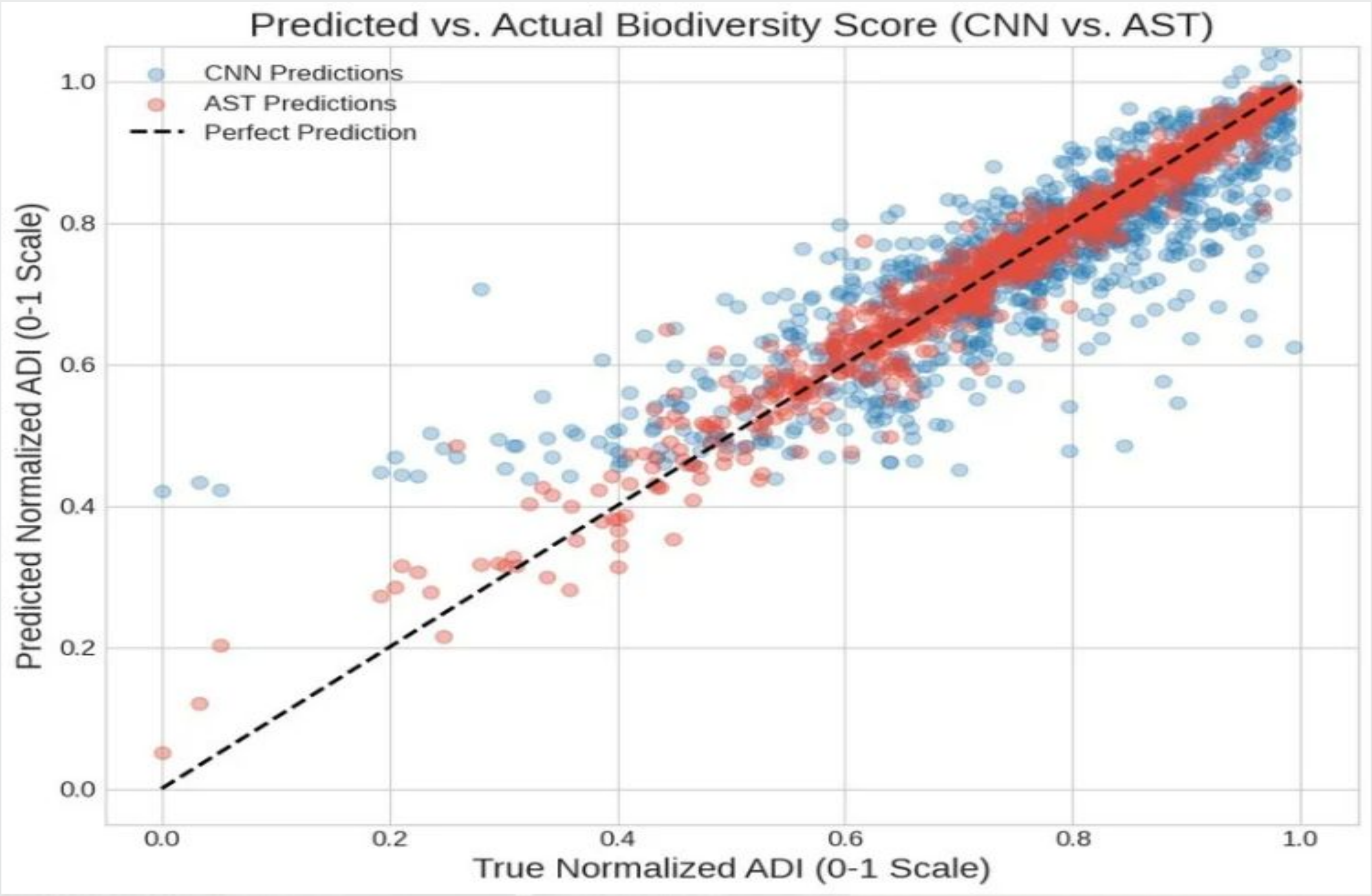
Model Evaluation: Both models were evaluated using R² and Mean Absolute Error (MAE) to measure predictive accuracy and error magnitude. Attention Rollout Heatmaps were used for Explainable AI (XAI) for the AST tuning to understand why the model was predicting.

Deployment (Docker): A lightweight version of our backend was successfully containerized using Docker, and we also made a website



Results

Vision Transformer (AST):	Convolutional Neural Network (CNN):
R²: 0.958	R²: 0.700
Test Loss: 0.022	Test Loss: 0.063



Conclusion

Effective Transfer Learning depends on mitigating domain shift, and our results highlight how crucial this consideration is. While the original ViT model achieved an R² of 0.67, our revised approach leveraging the AST model elevated performance drastically. The updated model reached an R² of 0.958, representing a substantial improvement over both the initial ViT and the baseline CNN. These findings show that selecting pretrained weights suited to the target domain greatly improves generalization and model performance in ecological audio analysis.

Future Direction

There are 4 main improvements planned for our Bioacoustics project:

1.) Expand dataset and anchors (to include Savannah, East Asia, etc)
2.) Real Time Edge Deployment: deploying the containerized backend on a Raspberry Pi in remote ecosystems in the Arctic/Costa Rica
3.) Multi-Modal Integration of temperature/humidity sensor data
4.) Seasonal/Temporal: training on data across different seasons to account for migratory patterns and fluctuations across time

Acknowledgments

Project heavily inspired by Drew Purves' appearance on the Google DeepMind Podcast with Hannah Fry, describing Perch/coral reef bioacoustics https://podcasts.apple.com/us/podcast/google-deepmind-the-podcast/id1476316441?i=1000713631651

Nouhad Rizk - Mentor | Kaggle dataset: https://www.kaggle.com/competitions/rfcx-species-audio-detection/data