

Bird Vocalization Classifier

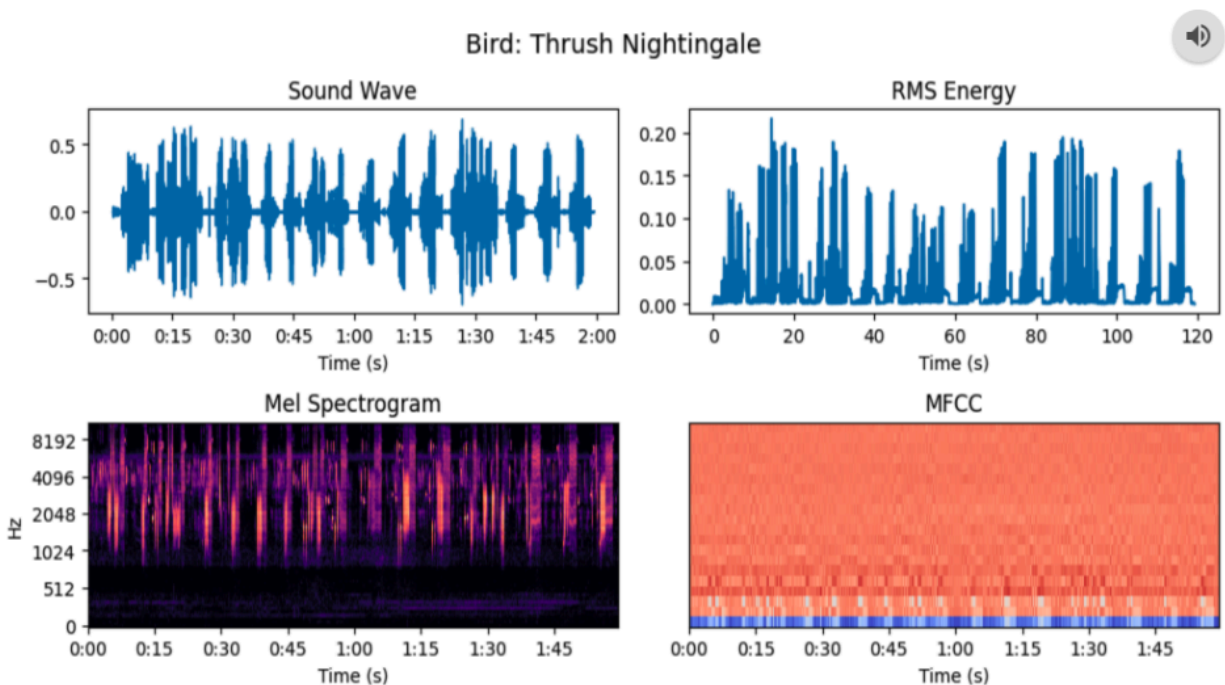
Objective

The project's primary goal was to develop a bird species classifier capable of accurately identifying species from audio recordings. The challenge was to navigate the complexities of processing and analyzing short, individual bird call recordings against a backdrop of significant class imbalance and varying audio quality.

Dataset and Methodology

The BirdCLEF 2023 dataset consisted of 16,941 audio recordings from [xenocanto.org](https://www.xenocanto.org/), encompassing 264 different bird species. The dataset featured primary attributes such as short recordings and identified bird species, along with secondary features including call type, location, and quality rating. In an effort to focus the project and address computing limitations, we refined our initial dataset of 10 bird species to three: Western Yellow Wagtail, Common Sandpiper, and Barn Swallow. These species were randomly selected from three different families to ensure diversity. To mitigate class imbalance, we down-sampled over-represented species at random, ensuring each species had approximately the same total duration of recordings, about 171-172 minutes. For our train/validation split, we adopted a strategy of random selection until the total training duration constituted 70% of the total duration for each species, with the remainder allocated to validation. This approach ensured a balanced representation of each species in both training and validation sets.

Exploratory Data Analysis



This visualization presents four different audio analysis plots for a Thrush Nightingale's song, including the sound wave, Root Mean Square (RMS) energy, Mel Spectrogram, and

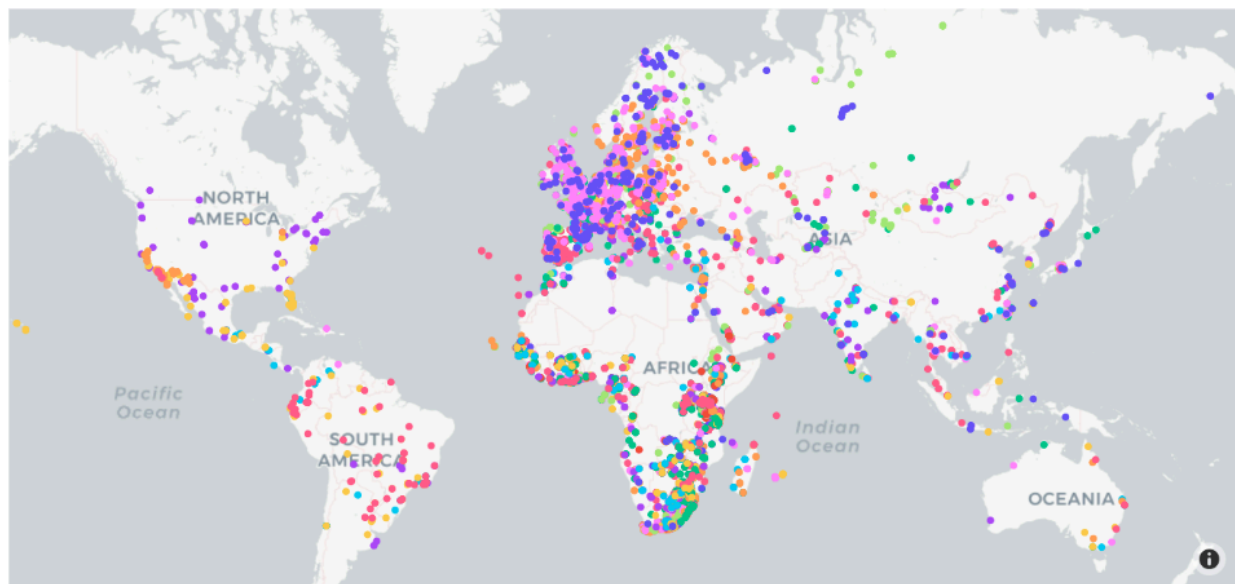
Mel-Frequency Cepstral Coefficients (MFCC), each providing a unique perspective on the audio characteristics.

Top 10 Bird Species - Taxonomy



The hierarchical plot showcases the "Top 10 Bird Species - Taxonomy," displaying vivid photographs of each bird along with their scientific classifications, offering a colorful and educational snapshot of avian diversity.

Bird Species World Map



The above map depicts a global distribution of various bird species, indicated by colored dots across continents, highlighting the geographical spread and habitats of the birds included in the study.

Feature extraction

The project utilized advanced audio processing techniques to extract meaningful features from the recordings:

Melspectrogram & MFCC: These visualizations capture the distribution of audio frequencies, transformed into the mel scale to align with human sound perception.

Chroma: This feature summarizes the 12 different pitch classes, offering insights into the harmonic content of bird calls.

RMS Energy: A measure of the signal's magnitude or "loudness" over time.

Spectral Centroid: Represents the "center of mass" of the spectrum, indicating the brightness of the sound. For testing, the models processed audio clips of 8 seconds in length, with a 4-second overlap to ensure continuity and maximize the use of the audio data. The same preprocessing techniques applied during training were used to prepare approximately 3,000 test samples, representing 35% of the total samples. Normalization of features for testing was based on scalers derived from the training data to maintain consistency and accuracy in model evaluation.

Modeling Pipeline

A range of machine learning models was evaluated:

Baseline Models: Random guessing provided a baseline accuracy of 33%.

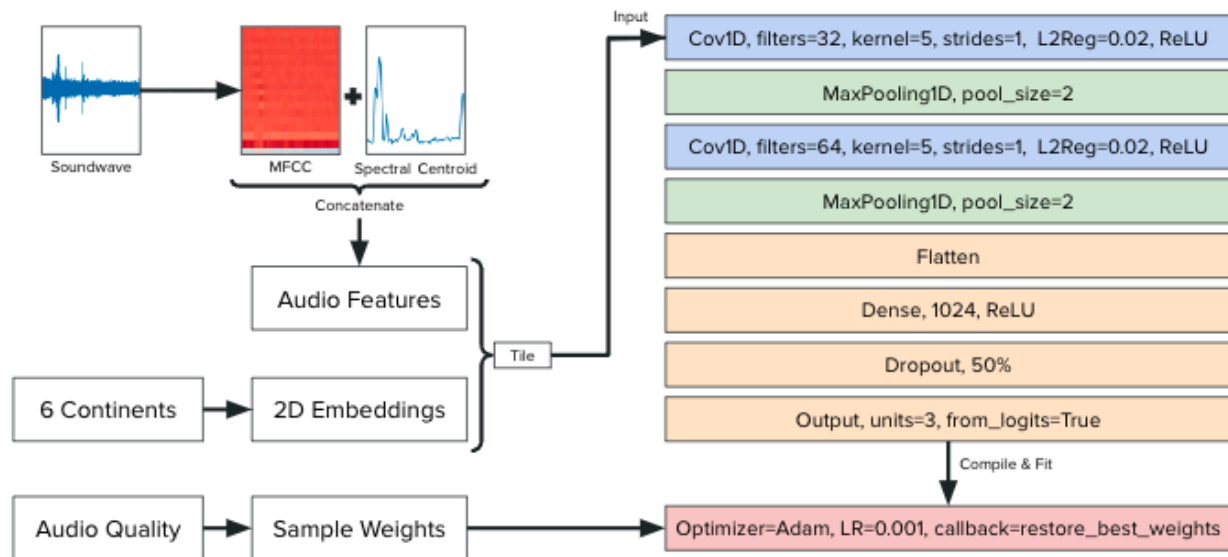
Traditional Algorithms: Included Random Forest, XGBoost, SVM, and Logistic Regression.

Deep Learning Models: Encompassed 1D CNN, 2D CNN, LSTM, GRU RNN, and Vision Transformers, with architectures tailored for audio classification.

Architectural Details

For deep learning models, architectures were specifically designed to process and learn from the complex features of audio data. For example, the 1D CNN model utilized concatenated audio features and embeddings, while the Vision Transformer applied augmented STFT spectrograms for classification.

Architecture - 1D CNN



The design of the 1D CNN model employed a functional API architecture, facilitating a seamless flow of data through the network layers. The architecture included:

Input Layer: Combined normalized, transposed audio features with 2D learned embeddings of continents, tiled along the time axis to match the audio features' shape.

Convolutional Layers: Two convolutional layers (Cov1D) were utilized, each followed by a MaxPooling layer to reduce dimensionality and capture the most salient features.

Regularization: L2 regularization was applied to convolutional layers to mitigate overfitting, complemented by a 50% dropout rate before the final output layer to further enhance generalization.

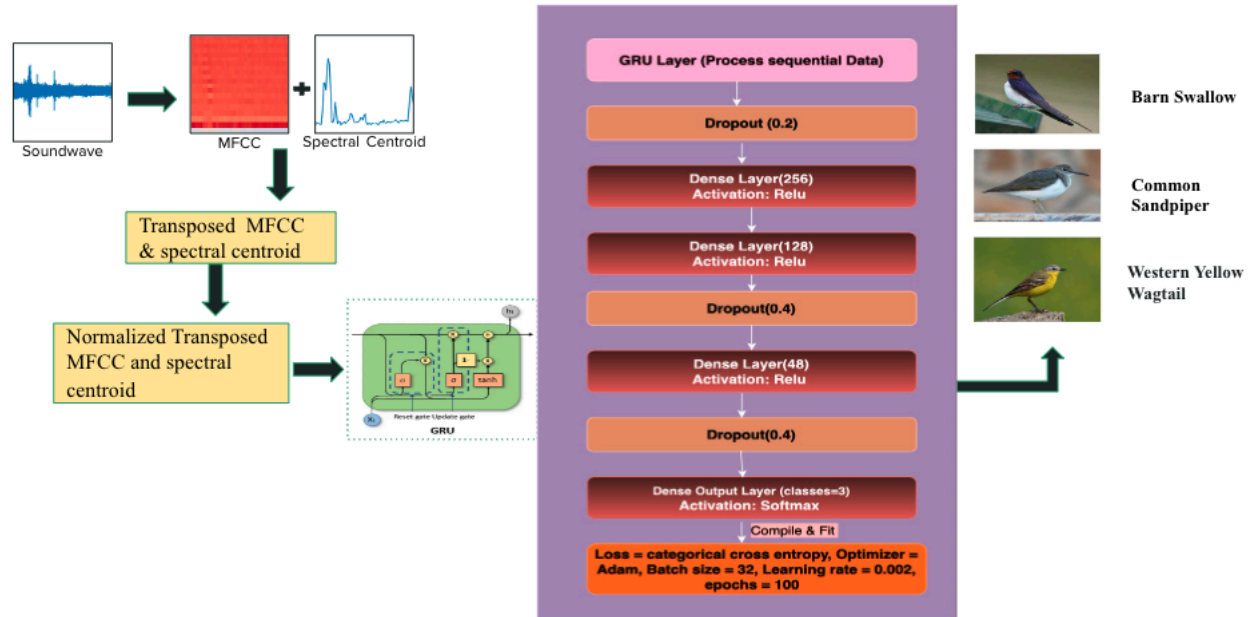
Optimization and Training: The Adam optimizer was selected for its efficiency in handling sparse gradients and adaptive learning rate capabilities. A callback mechanism was implemented to restore the model's best weights achieved at the 30th epoch, optimizing the training process.

GRU RNN:

Employing a functional API architecture allowed for a flexible and modular design approach. The architecture centered around a GRU layer with 256 units, capitalizing on the GRU's capacity for capturing temporal dependencies and sequence dynamics efficiently. To combat overfitting, the model integrated three dropout layers at critical junctures, alongside dense layers activated by ReLU, with unit counts of 128, 64, and 48. This structure was aimed at progressively refining the feature representation for accurate classification. For the final output, a softmax activation

function was utilized to facilitate multi-class classification, ensuring the model's output could be interpreted as probabilities across the different bird species.

Architecture - GRU RNN



Vision transformer

Input Spectrogram: The input to the model is an augmented STFT spectrogram, which is a visual representation of the spectrum of frequencies in the audio signal as they vary with time. This spectrogram is divided into patches (each patch could be a small time-frequency area), which are then flattened and linearly projected into a sequence of vectors (tokens). Positional embeddings are added to maintain the order of the sequence.

Transformer Blocks: The core of the ViT architecture consists of multiple Transformer blocks that process the sequence of tokens. Each block includes:

Layer Normalization (Norm): Stabilizes the learning process by normalizing the input layer.

Multi-Head Self-Attention: Allows the model to weigh the importance of different parts of the input data differently. With multiple 'heads', the model can focus on different positions of the input sequence, capturing various aspects of the data.

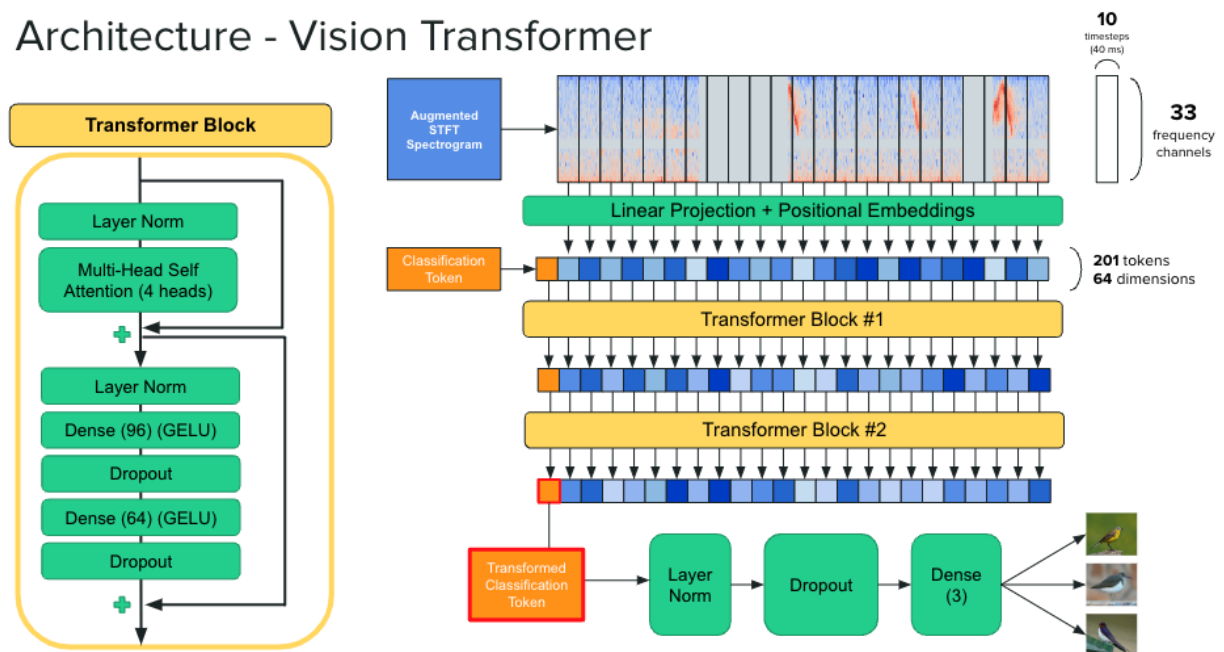
Dense Layers with GELU Activation: These fully connected layers further process the information, with the GELU (Gaussian Error Linear Unit) providing a non-linear activation function.

Dropout: Used between dense layers to prevent overfitting by randomly setting a portion of the input units to 0 during training.

Classifier Head: At the end of the Transformer blocks, the sequence of processed tokens is passed through additional layer normalization and dropout. The final 'classification token' (sometimes initialized and learned separately from the patch embeddings) is used to aggregate the information from the entire sequence and make a classification decision.

Output Layer: The last dense layer has as many units as there are classes to predict (in this case, 3 for the bird species). It uses a softmax activation function to output a probability distribution over the classes.

Architecture - Vision Transformer






Results and Evaluation

The project made significant progress from the baseline, with deep learning models showing superior performance:

- 1D CNN demonstrated an overall accuracy of 89%, with F1-scores ranging from 0.88 to 0.91 for specific bird species.

Inference results - 1D CNN

	Precision	Recall	F1-score	Support			
					  		
Barn Swallow	0.93	0.88	0.91	1288	1199	90	59
Common Sandpiper	0.84	0.96	0.89	1063	33	1019	11
Western Yellow Wagtail	0.93	0.84	0.88	1054	56	111	887
Overall Accuracy	0.89				Predicted Labels		

True Labels

- GRU RNN achieved an overall accuracy of 87%, highlighting the effectiveness of RNN architectures in handling audio data.

Inference results - GRU RNN

	Precision	Recall	F1-score	Support
Barn Swallow	0.89	0.83	0.86	1288
Common Sandpiper	0.80	0.93	0.86	1063
Western Yellow Wagtail	0.95	0.86	0.90	1054
Overall accuracy	0.87			



Confusion matrix for GRU RNN model:

	Barn Swallow	Common Sandpiper	Western Yellow Wagtail
Barn Swallow	1079	176	33
Common Sandpiper	54	996	13
Western Yellow Wagtail	77	67	910

Predicted Labels

True Labels

- Vision Transformer stood out with an overall accuracy of 94%, showcasing the advanced capability of transformer models in audio classification.

Inference results - Vision Transformer

	Precision	Recall	F1-score	Support
Barn Swallow	0.94	0.95	0.94	1288
Common Sandpiper	0.95	0.94	0.95	1063
Western Yellow Wagtail	0.93	0.94	0.93	1054
Overall Accuracy	0.94			



Confusion matrix for Vision Transformer model:

	Barn Swallow	Common Sandpiper	Western Yellow Wagtail
Barn Swallow	1220	27	41
Common Sandpiper	29	1003	31
Western Yellow Wagtail	46	22	986

Predicted Labels

True Labels