PROJECT EVALUATION
Group-12

Goal of the project

- Automatically detect bird and frog species
- Real-time information
- => Automated eco-acoustic monitoring systems

Data

- 57.32 GB of audio and metadata
- audio: 1 minute long FLAC files at 48 kHz
- 4,727 samples
- 24 species

Data

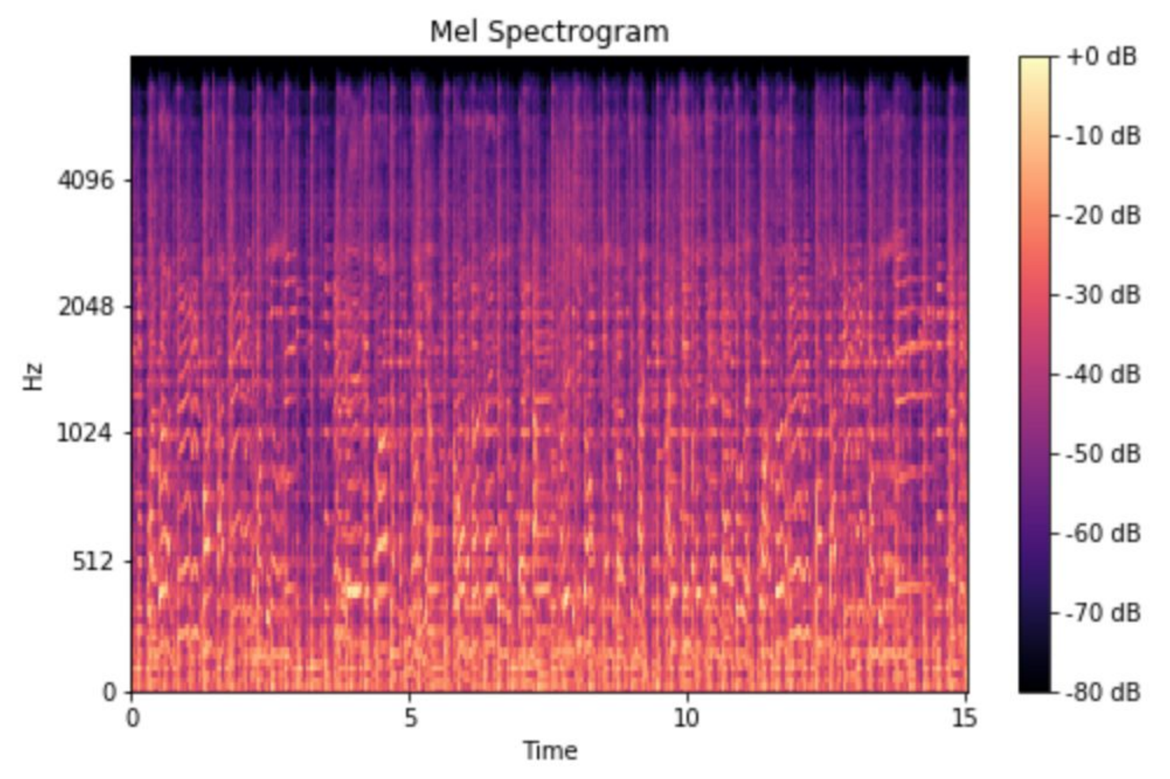
	recording_id	species_id	songtype_id	t_min	f_min	t_max	f_max
1	003bec244	14	1	44.544	2531.25	45.1307	5531.25
2	006ab765f	23	1	39.9615	7235.16	46.0452	11283.4
3	007f87ba2	12	1	39.136	562.5	42.272	3281.25
4	0099c367b	17	4	51.4206	1464.26	55.1996	4565.04
5	009b760e6	10	1	50.0854	947.461	52.5293	10852.7
6	00b404881	8	1	0.0747	3750.0	4.1973	5531.25
7	00d442df7	0	1	19.3653	5906.25	20.16	8250.0
8	011f25080	18	1	5.6853	3187.5	6.3787	5062.5
9	015113cad	15	1	50.0533	93.75	53.3973	1125.0
10	0151b7d20	1	1	46.032	3843.75	46.928	5625.0
11	01b41f92b	6	1	44.24	562.5	46.384	4406.25
12	0201197ec	10	1	43.3575	947.461	45.8014	10852.7
13	0209f7ab2	7	1	50.7573	4687.5	53.8987	11437.5
14	0268057eb	0	1	25.4987	5906.25	26.2933	8250.0
15	0275e127d	11	1	11.9989	1808.79	13.1367	5684.77
16	0295e3234	11	1	5.1606	1808.79	6.2984	5684.77
17	0297d886e	13	1	57.808	93.75	58.432	843.75
18	[illegible]	[illegible]	1	[illegible]	562.5	[illegible]	3281.25

Data

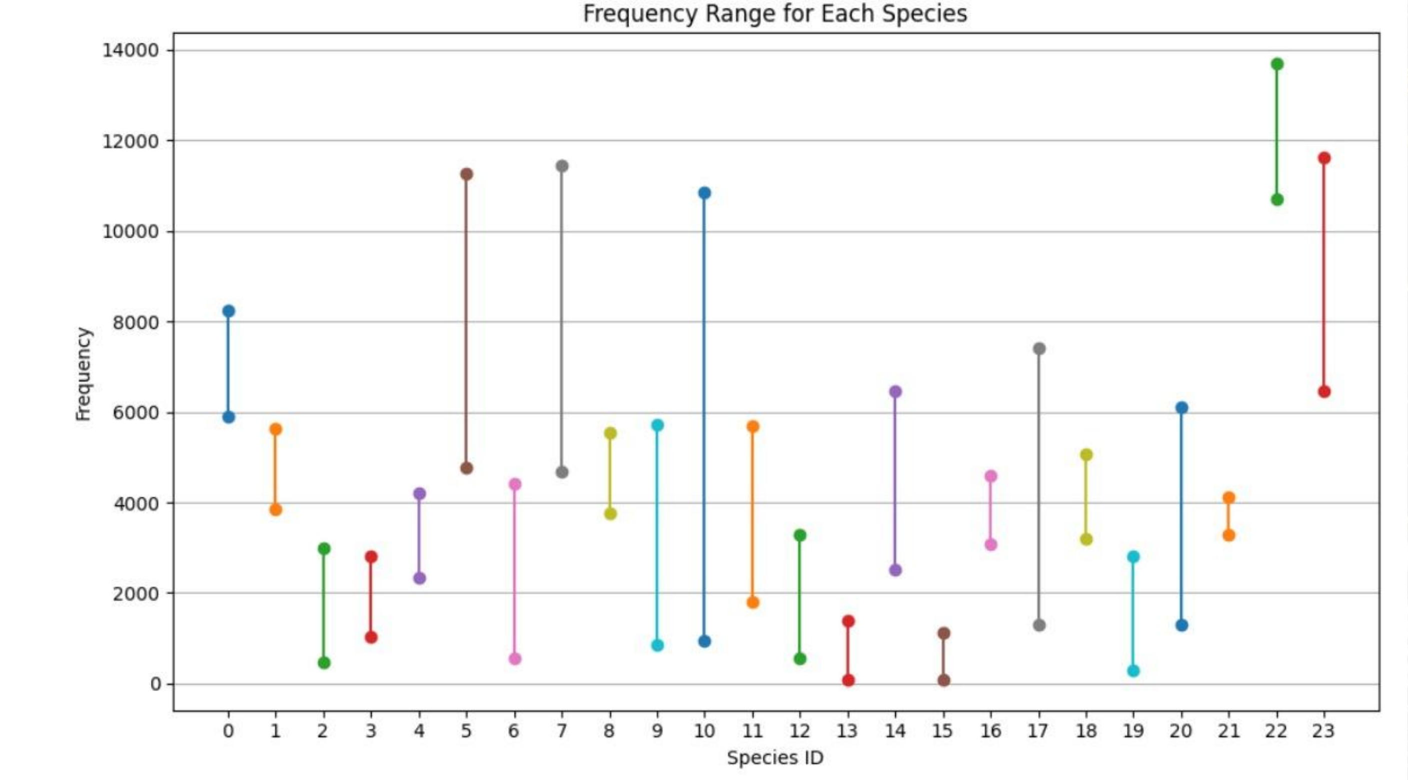
- includes a bounding box for each event
  - start/end time
  - low/high frequency
- We have strong labels in our data
- We are predicting weak labels

Mel spectrograms

- converts audio to 2D “image”
- x-axis: time
- y-axis: frequency



EDA bandwidths



Evaluation metric: LRAP

- label-ranking average precision
- 0 < LRAP <= 1 (bigger is better)
- predictions are at the audio file level, so multiple labels
- measures: “for each ground truth label, what fraction of higher-ranked labels were true labels?”

	A	B	C	D	E	F	G	H
1	labels	raw	prediction	ground truth	|L_{ij}|	rank_{ij}	ratio	
2	aardvark	8	0.6323	1	1	1	1.00	
3	baboon	7	0.2326	1	2	2	1.00	
4	cat	6	0.0856	0	0	3	0.00	
5	dog	5	0.0315	1	3	4	0.75	
6	elephant	4	0.0116	0	0	5	0.00	
7	fox	3	0.0043	0	0	6	0.00	
8	giraffe	2	0.0016	0	0	7	0.00	
9	hare	1	0.0006	0	0	8	0.00	
10			# labels	3		LRAP	0.92	
11								
12								
13								

Data challenges we faced

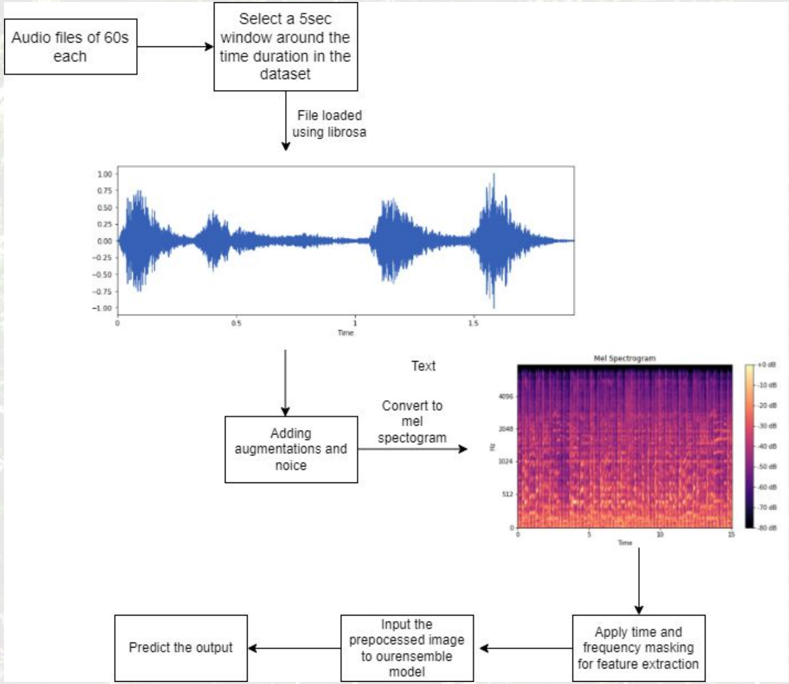
- training data is underspecified, only partially labeled and for each sample we had background noises which made it virtually impossible to distinguish.

Few preprocessing techniques

- augmentation
  - mixup
  - Gaussian noise
  - SpecAugment (time warping, freq. masks, time masks)

Algorithm and Model pipeline:

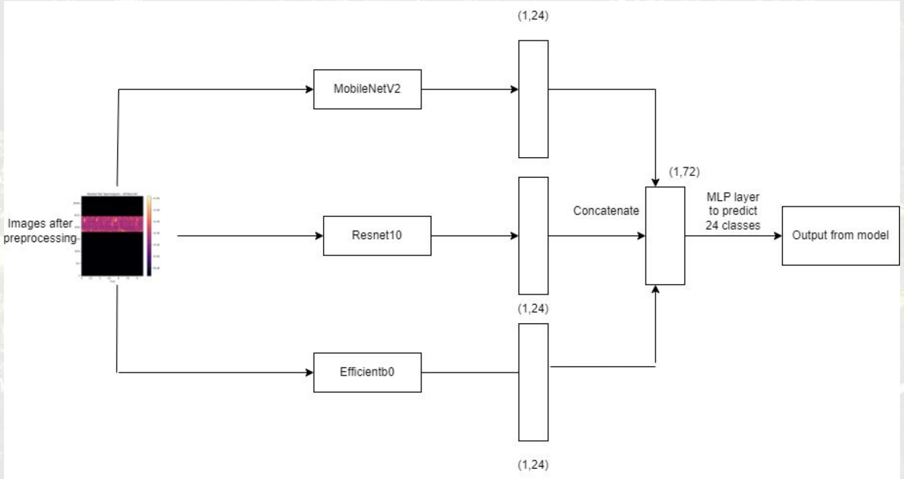
- Data preprocessing pipeline using Librosa: Load audio files and slice based on species presence, segmenting at 5-second intervals.
- Augmentation techniques: Apply augmentation with 50% probability, including adding segments from different recordings, artificial beat insertion for rhythm diversity, and Gaussian/pink noise addition for data diversity.
- Time-based preprocessing: Separate audio signals at specific time intervals when a species is heard, enhancing feature extraction and model robustness.
- Frequency and time masking: Enrich feature extraction by applying frequency and time masking techniques, contributing to a more resilient deep learning model for audio classification.
- The Preprocessed image is input to our ensemble model.
- Model predicts the output



Models Used (Continued)

- MobileNet
- EfficientNet
- ResNet
- Proposed Ensemble Model

  Fusion of EfficientNet, MobileNetV2,and ResNet50



Proposed Algorithm (Preprocessing)

1. Load audio files using Librosa.

2. Slice the audio files based on the time at which a species is heard.

3. Segment audio files into 5-second intervals.

4. Apply essential augmentation techniques with a 50% probability:

- Add audio segments from various recordings.
- Insert artificial beats to diversify rhythm patterns.
- Add Gaussian and pink noise.
- Perform time shifting and volume control modulations.
- Convert the audio data to mel spectogram
- Apply frequency and time masking to enrich feature extraction.

Proposed Algorithm (Model Implementation)

- Pass preprocessed data to the input layer of the ensemble model.
- Load pre-trained EfficientNet, MobileNetV2, and ResNet50 models.
- Modify the classification heads of each model to output the specified number of classes.
- Define an ensemble model architecture that combines the predictions of the three models.
- Create an EnsembleModel class with three backbone networks and an output layer.
- Implement the forward method to concatenate the outputs of the backbone networks and pass them through the output layer.

Conclusion

[Github Repository]

- Multiple models were tested to classify tropical rainforest species using audio signals, with the ensemble model showing the best results.
- Ensembled model designed for almost real-time predictions, doesn't require high-end GPUs, making it easier for production deployment.
- Achieved good performance with 0.925 LWLRAP score and 0.162 multiclass log losses.
- Potential for improvement with popular algorithms like Attention Mechanism, especially with more data collection for additional species.
- Model deployed on https://streamlit.io/ for accessibility and ease of use.

Experimentation and Results

- Utilized three models: ResNet, EfficientNet, and MobileNet, with their respective accuracies recorded.
- Ensembled model achieved the best results after training for 10 epochs, showcasing strong generalization.
- Key reason for employing K-fold cross-validation (k=5) was to mitigate overfitting, ensuring the model doesn't memorize training data excessively.
- K-fold cross-validation helps in generalizing better to new, unseen data by training on diverse subsets and validating on separate folds.
- Particularly beneficial for limited datasets, maximizing data use for training and validation, leading to more robust performance estimates.
- Can be combined with hyperparameter tuning methods like grid search or random search for optimal model configuration.
- Adoption of K-fold cross-validation instrumental in improving model robustness, reducing overfitting, and providing reliable performance estimates.
- Ensemble model, evaluated with K-fold cross-validation (k=5), demonstrated strong performance with a validation loss of 0.162 and a training loss of 0.154, showcasing robust generalization across unseen data.

Model	Lwlrap score
Ensembled ResNet—EfficientNet—MobileNet	0.925
ResNet	0.911
EfficientNet	0.907
MobileNet	0.891

Table 1: Comparison of Model Accuracies

Thank you